E-Learning on Semantic Web

Chapter · September 2023

DOI: 10.1007/978-981-99-2742-5_62

Ayesha Ameen
Deccan College of Engineering and Technology
24 PUBLICATIONS **194** CITATIONS

Ayesha Banu Mohd
Vaagdevi College of Engineering
30 PUBLICATIONS **51** CITATIONS

E-Learning on Semantic Web

Dr. Ayesha Ameen[1] [0000-0001-9424-1769] and Dr.Ayesha Banu [2][0000-0002-1646-0832]

[1] [1]Department of IT, Deccan College of Engineering and Technology, Hyderabad, 500001, Telangana, India
[2] 2Department of CSE (Data Science), Vaagdevi College of Engineering, Warangal, 506005, Telangana, India
ameenayesha@gmail.com, ayeshabanuvce@gmail.com

Abstract. E-Learning is the current trend that is gaining more significance post-pandemic, E-Learning has several distinctive features which has attracted many users across the globe and there is a huge demand for E-Learning content. E-Learning is the delivery of learning material by any electronic gadget, the communication channel used for the delivery of learning material is the World Wide Web. Currently, the content displayed on the World Wide Web is only understood by humans but not by machines, in order to make machines understand the content and make intelligent decisions the World Wide Web can be replaced by the Semantic Web. In the current paper, SemELearn a Semantic Web-based E-Learning system is proposed that utilizes the underlying characteristic feature of the Semantic Web.

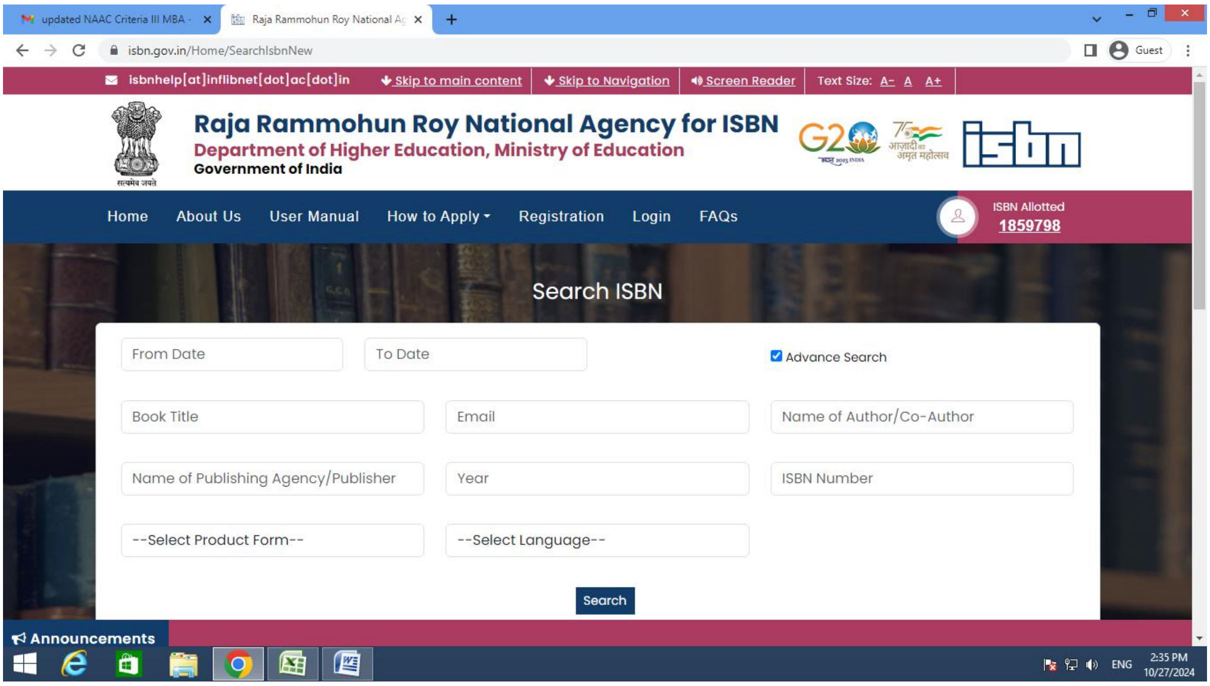
Keywords: E-Learning, Semantic Web, Ontologies, Inference.

1 Introduction

E-learning is an internet-enabled learning, it is the delivery of learning and training through digital resources like computers, tablets or phones connected to internet. E-learning use communication technologies to teach/train students /employee through various electronic media such as audio, video, multimedia, visualization technologies etc. [1]. E-learning is otherwise known as Virtual learning or Web-based learning. Instructors and students need not be available at the same place and time in E-leaning which makes it easy for learners to learn anytime, anywhere. Hence learning and teaching becomes easier, simpler and more effective by using various digital resources.

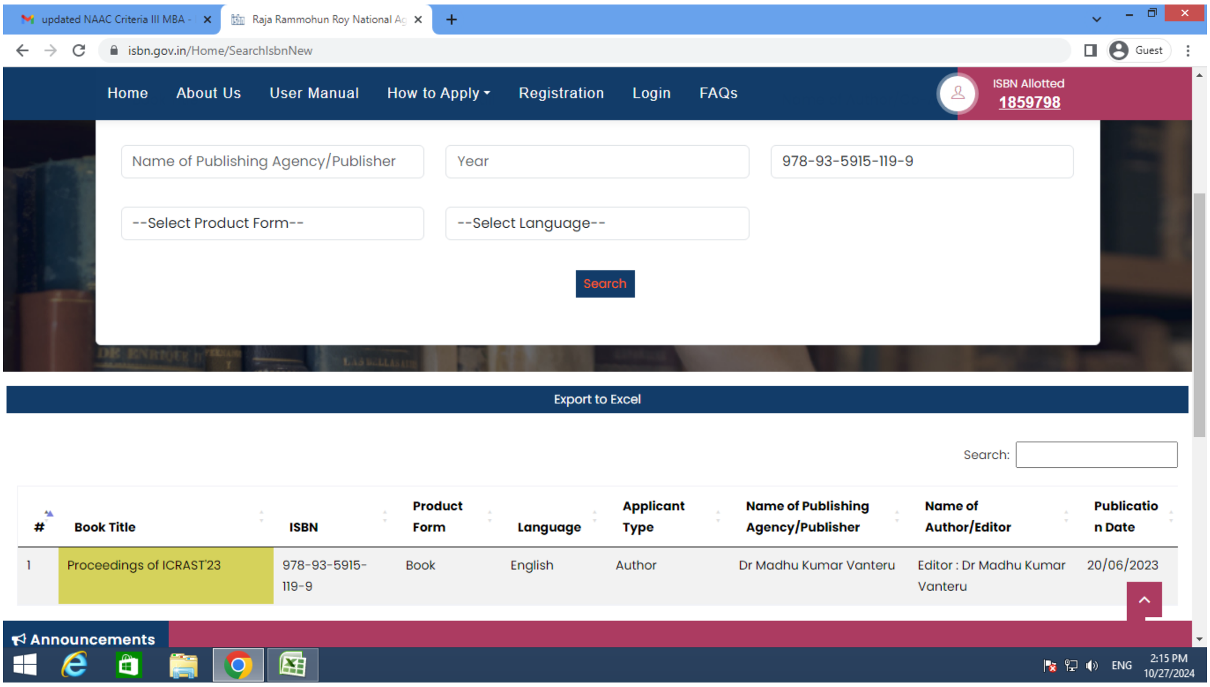
1.1 Growth of E-Learning

E-learning is growing rapidly because for any queries people first search on the internet rather than finding books or asking someone. According to Forbes, the size of global E-Learning market is expected to touch $325 Billion by 2025. In today’s context what people understand for eLearning is to get trained on any digital device. To gain



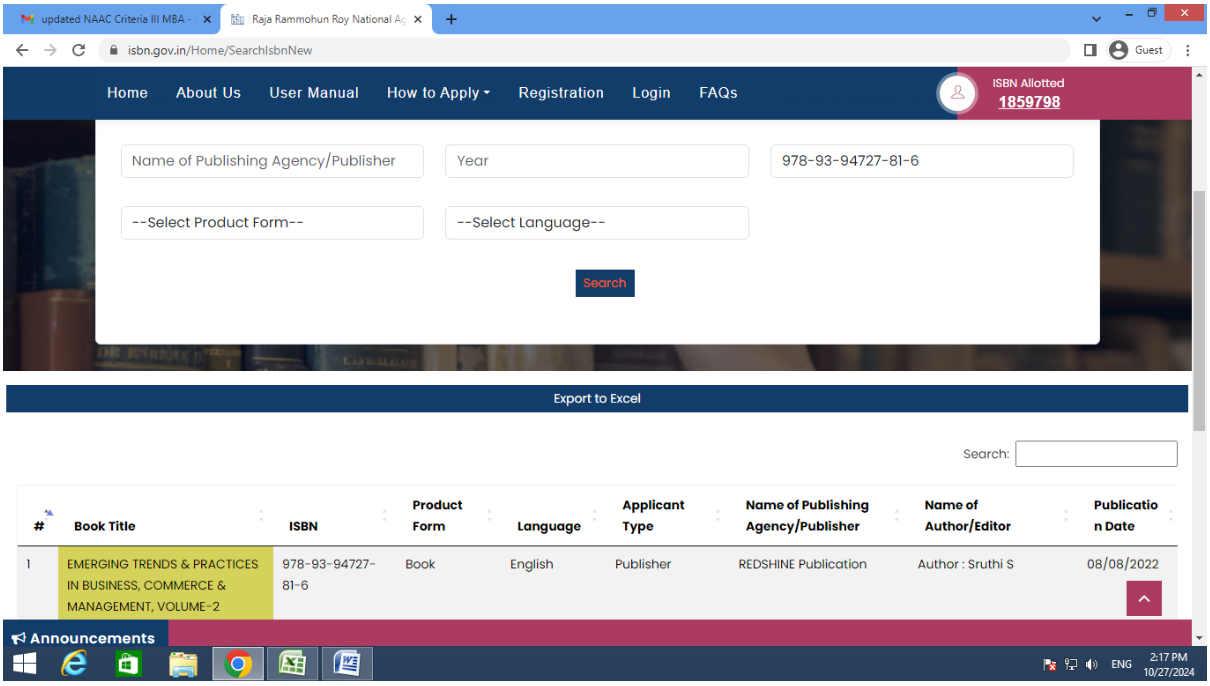
☑ Advance Search

Book Title	Email	Name of Author/Co-Author
Name of Publishing Agency/Publisher	Year	ISBN Number
--Select Product Form--	--Select Language--	



Name of Publishing Agency/Publisher	Year	978-93-5915-119-9
--Select Product Form--	--Select Language--	

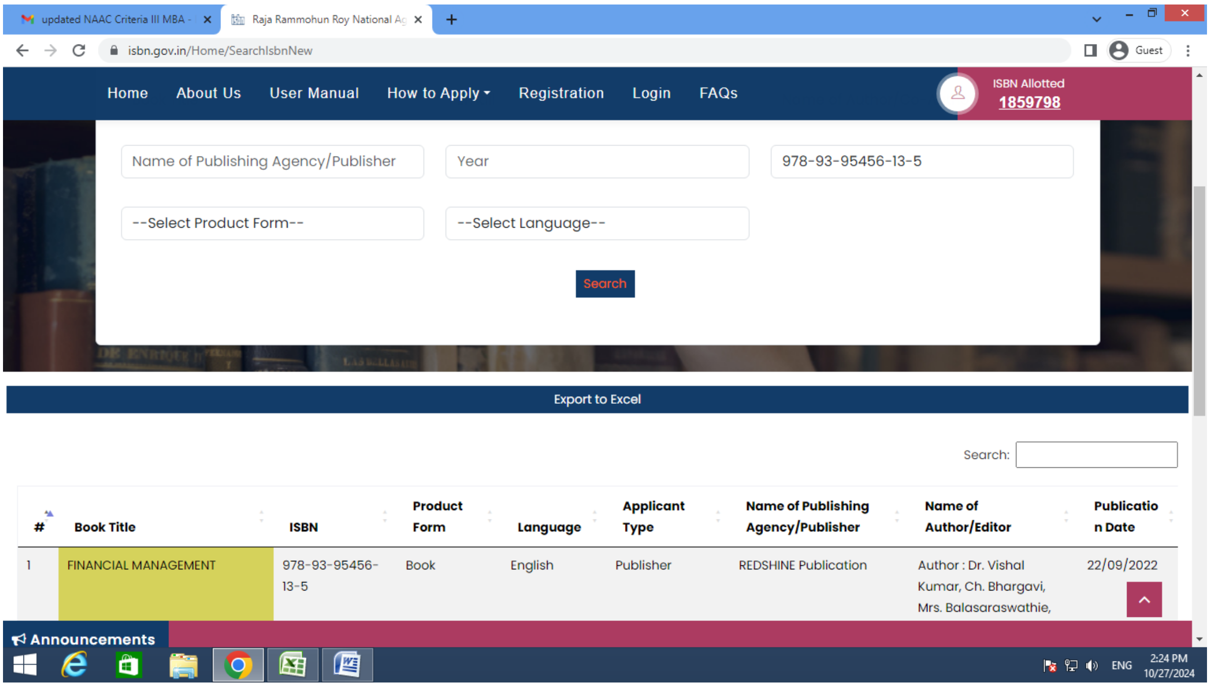
#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	Proceedings of ICRAST'23	978-93-5915-119-9	Book	English	Author	Dr Madhu Kumar Vanteru	Editor : Dr Madhu Kumar Vanteru	20/06/2023



Name of Publishing Agency/Publisher	Year	978-93-94727-81-6
--Select Product Form--	--Select Language--	

Search:

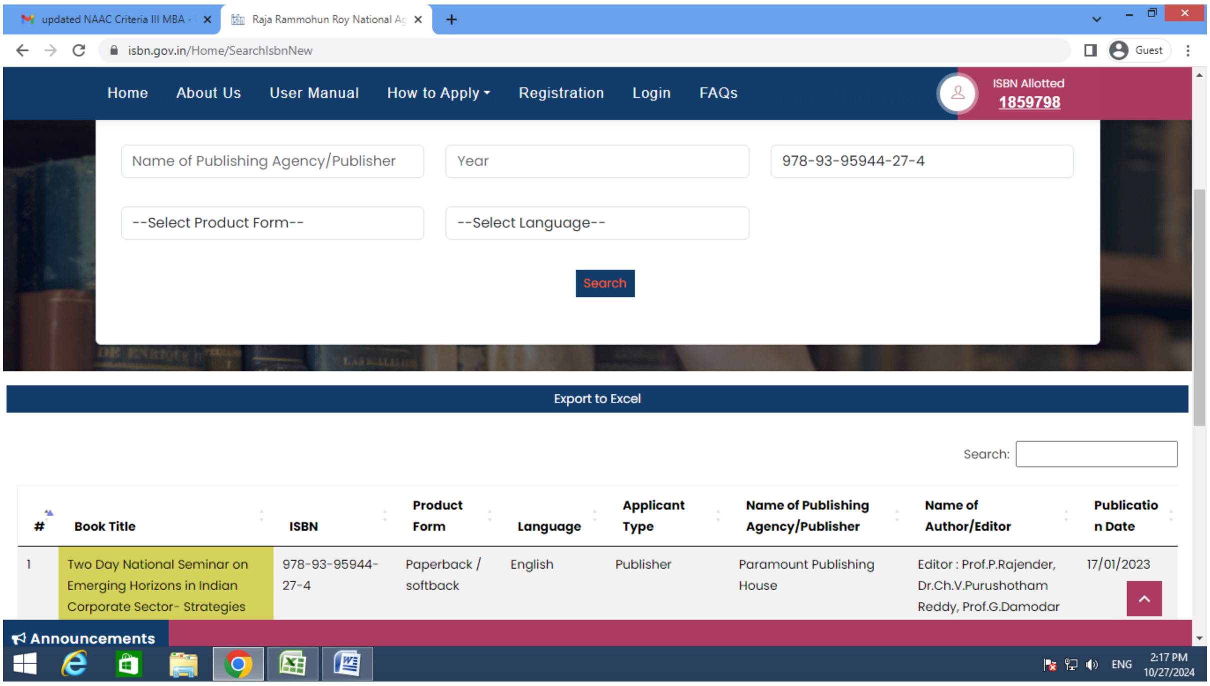
#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	EMERGING TRENDS & PRACTICES IN BUSINESS, COMMERCE & MANAGEMENT, VOLUME-2	978-93-94727-81-6	Book	English	Publisher	REDSHINE Publication	Author : Sruthi S	08/08/2022



Name of Publishing Agency/Publisher	Year	978-93-95456-13-5
--Select Product Form--	--Select Language--	

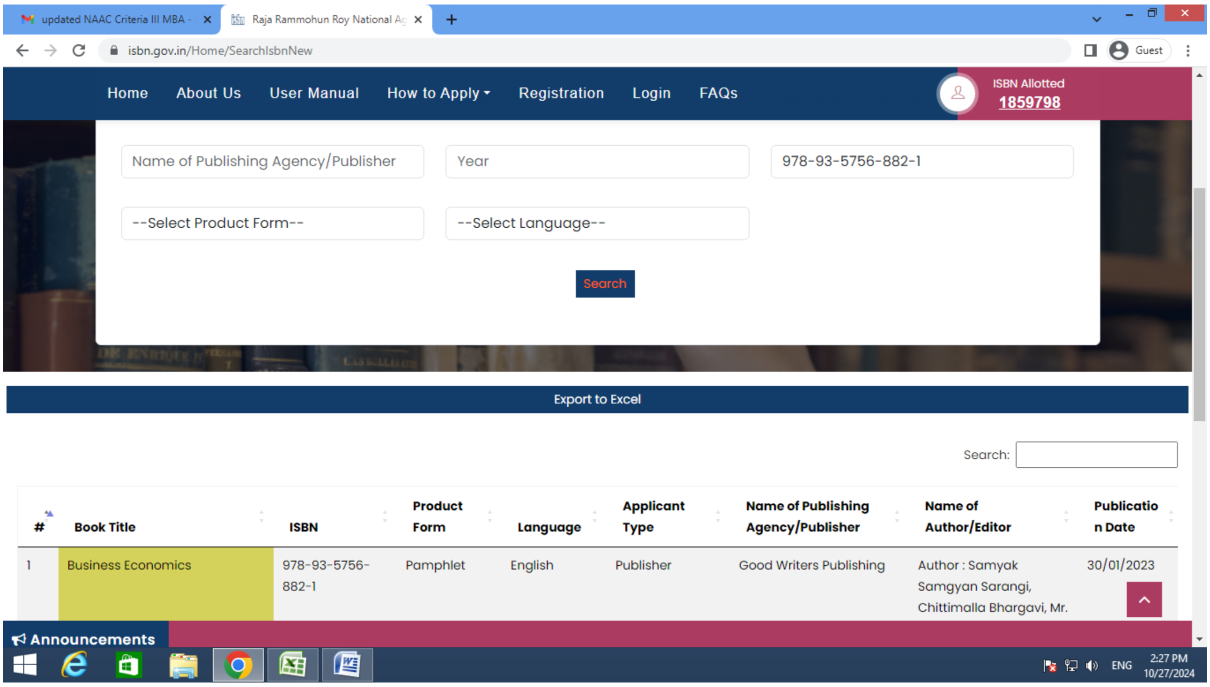
Search:

#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	FINANCIAL MANAGEMENT	978-93-95456-13-5	Book	English	Publisher	REDSHINE Publication	Author : Dr. Vishal Kumar, Ch. Bhargavi, Mrs. Balasaraswathie,	22/09/2022



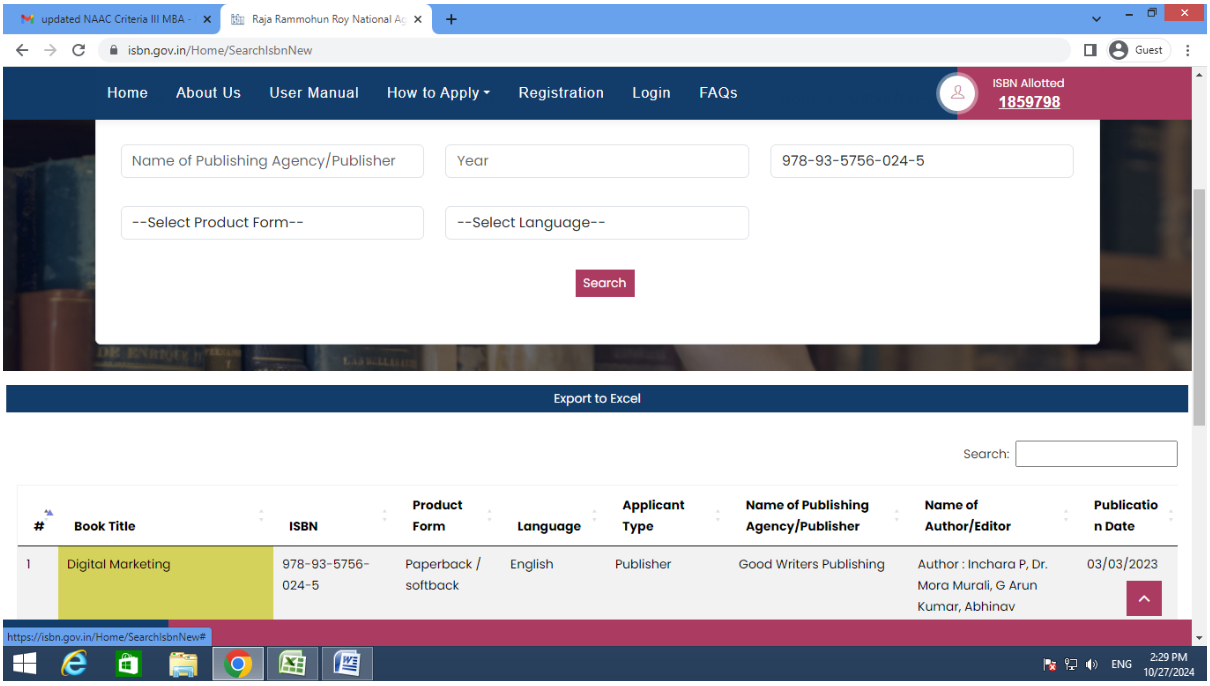
978-93-95944-27-4

#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	Two Day National Seminar on Emerging Horizons in Indian Corporate Sector- Strategies	978-93-95944-27-4	Paperback / softback	English	Publisher	Paramount Publishing House	Editor : Prof.P.Rajender, Dr.Ch.V.Purushotham Reddy, Prof.G.Damodar	17/01/2023



978-93-5756-882-1

#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	Business Economics	978-93-5756-882-1	Pamphlet	English	Publisher	Good Writers Publishing	Author : Samyak Samgyan Sarangi, Chittimalla Bhargavi, Mr.	30/01/2023



#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	Digital Marketing	978-93-5756-024-5	Paperback / softback	English	Publisher	Good Writers Publishing	Author : Inchara P, Dr. Mora Murali, G Arun Kumar, Abhinav	03/03/2023

An Impact of COVID 19- A Well-Being Perspective for a New World

Dr. Thanveer Jahan
Associate Professor, Vaagdevi College of Engineering,
Warangal, Telangana, India

Abstract:

One of the data science issue is COVID-19, which had created a major hit massively is on public health. It resulted on a major health issues and thereby resulted to deaths. The structure of the society and community is affected to concentrate on important issues such as affordability of a healthcare, availability of medicines, rights for worker's and freedom to move. Some parts of the population in the world were exposed to the complications of anxious, depression and symptoms of post-traumatic as these are related to stress The situation had become crucial for a data scientist, as there were many questions started to rise and trust the data, curves that was plotted by social media websites. The situation made them to scare or think worst that could even happen in society. It made society more panic and empowered to handle this situation. This paper concentrates on a sur vey on the issues and problems on society, children, students, teenagers such as Physical fitness, psychological and social evaluation effects of pandemic. It also focuses on a new perspective on the usage of digital devices that effected mental health.

Keywords: Data science, Psychological, Covid 19, mental health.

Introduction

Traditional health surveillance systems are well known for major time lags. The current situation clearly indicates that the, systems are critically needed locally and are robust [1]. In this situation the collection of data is very difficult for such an infectious disease. In real time data analysis of such a high resolution data has become a diffi-

cult task for a data scientist. They work in domains such as public health and also learn from various domains. Corona virus is a disease which is contagious, where mild infection are treated in home quarantine or thereby rely on hospitals or a practitioner to estimate the spread that can mislead the early stages of disease progression. The people in the society are lesser, who have actually made their presence at health facilities to test or care. The report of it can lead to focus on morbidity and mortality. The fact is that many countries don't show actual count of people having virus. The increase in the number of test will increase the count. Countries like Iceland have done systematic sampling including the people having asymptomatic symptoms [2]. The prevalence of the infection is indicated along with the containment areas. Keeping apart the conspiracy theory of the government is that, the test for caronavirus is expensive. The count collected from a country is dependent on the widespread of the virus and the financial status of a local health care facility for testing. The problem for data sampling has become concern for a data scientist in many cases. The concern of pandemic had also affected the society wherein they are separated from their loved ones, less freedom as well as uncertainty of the spread of the disease. The concern in the general public is increased working in health care centre's as to understand spread of number of cases. The families in the society were panic in storing long shelf of food items. It became more fear for them to shop in super markets or public places. Lockdown was stressful period which made society living style have completely changed. The affect of covid 19 also affected children in the society with obesity problems [3]. The lockdown in many countries have also imposed children by stopping the physical activity. Children who stay in urban or in small houses or apartments were having limited space for any physical activities. These were a one of the wave of covid 19 that affected children very badly [4].
Data collection had become essential part in this situation, which it rely on accuracy and limitations.

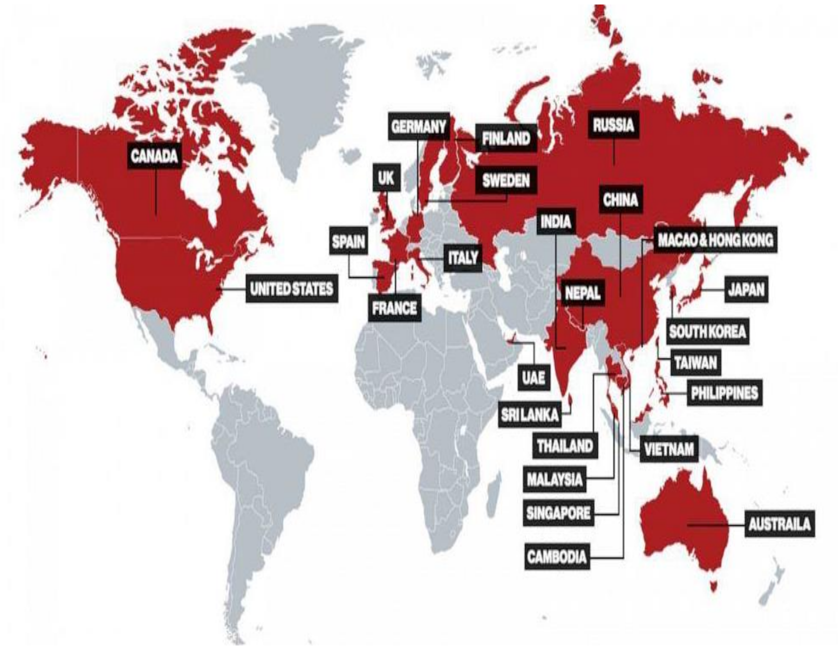
Data Modeling and Prediction

The WHO gave on information from the country china where the symptoms of the virus may vary from first day to 14 days after the

exposure [5]. It is been issued from the centre of disease control and prevention that the highest alerts from Italy.
The countries that restricted for travel in the countries such as Iran, China and South Korea. The virus outbreaks are shown in the Figure 1 below. The impact of virus had made data scientist to model the data and predict its impact on the society. The Corona virus disease made conditions worst with an outbreak of sense of fear, stress, anxiety and mental disorders. The overwhelming of the disease has caused to develop more emotions among adults and children. Overcoming with stress led the society, people made stronger [6].

To protect the mental health of the people in the society, WHO updated these measures.

1. The sense of fear is created by the digital media, reading and watching news.
2. To protect their dear ones by seeking relevant information.
3. The trigger of fear and anxiety id from the sources such as social media.



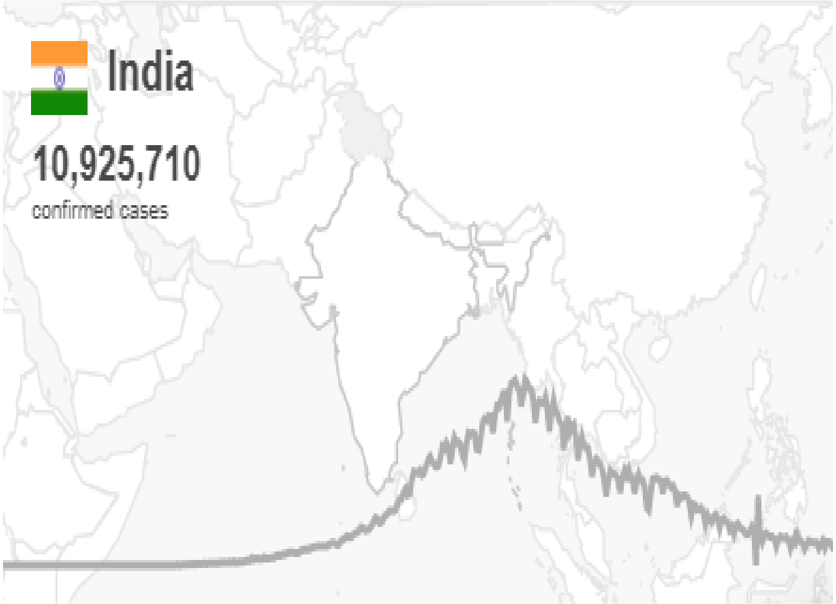


FIGURE 1: Covid 19 affected countries and Number of Cases in INDIA

Covid -19 Impact on Mental Health.

The most attractive online platforms for a young generation was used even before pandemic [8]. It has before a cup of coffee after pandemic. The schools were closed during pandemic. The communication among young children increased with this platform by playing online games and access social media platforms also. The virtual learning is been made compulsion on many school children. The physical activity was decreased as there prolonged sedentary periods, as they were using screens for long time. The screen time is been increased day by day while using these platforms[9,10]. The increased weights in the children have been increased day by day. This in turn had led to sedentary habits which have further increased in the risk for complications such as fear, anxiety, depression etc. It was then predicted by many data scientists that the prolonged closure of schools will proportional rise the obesity rate in children. The data scientists have also predicted and recorded that in December 2020, there will be an rapid increase on new obesity cases. The statistics are inferred by data scientists as shown below in the Figure 2.

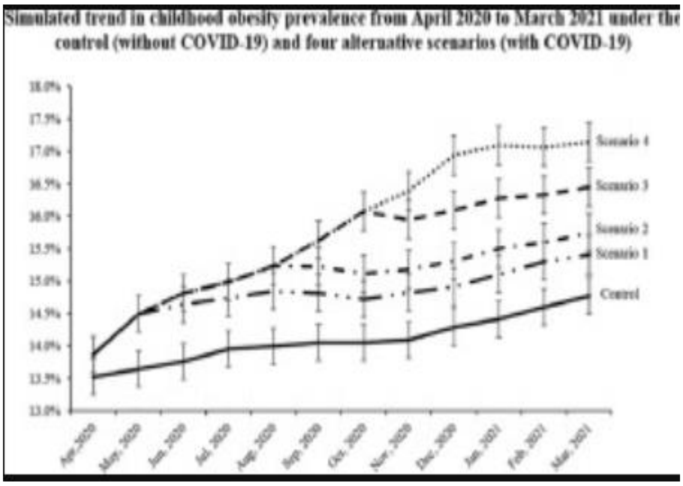


FIGURE 2: The Four Alternative Scenario Of Children Obesity During Covid 19.

Impact Of Covid19 On Unchangeable Environmental

The life style during covid 19 pandemic should have a family environment, which can change the behavior in child [11]. The foetal environment plays an important role in life course of children. As many women have a problem of obesity that is linked childhood, which can even lead to diabetes and cardiovascular diseases [23]. Pregnant women were also made to lower hospital visits as they are more vulnerable for the spread of covid 19[13,14]. The containment zones were avoided more by taking measures. The government made a mandate to stay at home. The routine checkups are cancelled temporarily. This had in turn led a pregnant woman to extra pressure and stress related issues. There was a remote antenatal care available in many remote areas [17, 18]. The stop of maternity checkups and unavailability of resources during pregnancy due to prioritize the disease have made data scientists to predict increase in the risk of death, maternal morbidity and mortality [15, 16].

An Impending Recommendation for Post covid 19

The Number of new challenges in the upcoming year 2021 are been faced many countries in the world. The year 2020 has ended up with a new scholastic year approach. The social economic conditions, emotional conditions such as stress had fallen down abruptly. The Child education is been hanged with a termination of physical learning. Children are re-introduced to schools and can facilitate the physical activity also [23]. This will not provide benefit to child behavioral health but also solves the problems related to childhood obesity. The threshold and safety measures indicators are crucial to avoid the spread of corona transmission within the school, colleges and educational institutions [7].

The wide spread of the curbing disease corona and to protect the health have become a highest priority till a good effective vaccine for covid 19 is made available. Various adhesive problems also exists in the society such as stress, mental and obesity issues[21,22]. The above problems are uncontrolled if there is a long-term extreme health and economic results [12]. There must be more support and manage system that can be dealt in the problems in obesity in children. The issues related to availability and choose of food on low budget can be handled wisely by educating parents. A necessary need of physical activity and maintaining physical distance is a need of every teenager, child and parents in the society. A very special attention should be given in the form of counseling to the pregnant ladies, who need lot of care [19,20]. They should be educated with the problems of obesity and the preventive measures taken to avoid obesity before the child is unborn. The issues should be considered as priority based by an every individual in the society, community during the pandemic.

The pandemic can be ended if there is a large share in the world that needs to get immune to the disease covid 19. Vaccines are the only one technology that is dependent in the past to lesser the death rate. The challenging task is to make these vaccine available o all people in the countries. In this connection data scientists are making their efforts in constructing the datasets for an international vaccination.

REFERENCES

[1] Bansal et al., Journal of Infectious Diseases *214*, S375–S379

[2]https://www.government.is/news/article/2020/03/15/Large-scale-testing-of-general-population-in-Iceland-underway/

[3] World Obesity Federation. Global Atlas on Childhood Obesity [Internet]. London; 2019. Available from: https://www. worldoty.org/nlsegmentation/global-atlas-on-childhoodobesity. Accessed 3 Sept 2020.

[4] González-Muniesa P, Mártinez-González M-A, Hu FB, Després JP, Matsuzawa Y, Loos RJF, et al. Obesity. Nat Rev Dis Prim [Internet]. Nature Publishing Group; 2017 [cited 2020 May 27];3: 17034. Available from: http://www.nature.com/articles/ nrdp201734. Accessed 3 Sept 2020.

[5] World Health Organization (WHO). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. 2020 [cited 2020 Apr 15]. Available from: https://www.who.int/dg/speeches/detail/who-director-general-sopening-remarks-at-the-media-briefing-on-covid-19%2D%2D-11- march-2020. Accessed 3 Sept 2020.

[6] Cuschieri S. COVID-19 panic, solidarity and equity—the Malta exemplary experience. J Public Health (Bangkok) [Internet]. Springer; 2020 [cited 2020 Jun 3];1–6. Available from: http://link. springer.com/10.1007/s10389-020-01308-w. Accessed 3 Sept 2020.

[7] Centers for Disease Control and Prevention. COVID-19 - School Reopening: Indicators to Inform Decision Making | CDC [Internet]. Centers Dis. Control Prev. 2020 [cited 2020 Sep 19]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/community/ schools-childcare/indicators.html. Accessed 3 Sept 2020.

[8] Rundle, AG., Park, Y., Herbstman, JB., Kinsey, EW., Wang Y. COVID-19-Related School Closings and Risk of Weight Gain Among Children. Obesity (Silver Spring) [Internet]. Obesity (Silver Spring); 2020 [cited 2020 May 28];28. Available from: https://pubmed.ncbi.nlm.nih.gov/32227671/. Accessed 3 Sept 2020.

[9] Pietrobelli A, Pecoraro L, Ferruzzi A, Heo M, Faith M, Zoller T, et al. Effects of COVID-19 Lockdown on Lifestyle Behaviors in Children with Obesity Living in Verona, Italy: A Longitudinal Study. Obesity [Internet]. John Wiley & Sons, Ltd; 2020 [cited 2020 May 28];oby.22861. Available from: https://onlinelibrary. wiley.com/doi/abs/10.1002/oby.22861. Accessed 3 Sept 2020.

[10]Ribeiro KD da S, Garcia LRS, Dametto JF dos S, Assunção DGF, Maciel BLL. COVID-19 and Nutrition: The Need for Initiatives to Promote Healthy Eating and Prevent Obesity in Childhood. Child Obes [Internet]. Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2020 [cited 2020 Sep 19];16:235–7. Available from:

https://www.liebertpub.com/ doi/10.1089/chi.2020.0121. Accessed 3 Sept 2020.

[11] Asigbee FM, Davis JN, Markowitz AK, Landry MJ, Vandyousefi S, Ghaddar R, et al. The Association Between Child Cooking J Diabetes Metab Disord Involvement in Food Preparation and Fruit and Vegetable Intake in a Hispanic Youth Population. Curr Dev Nutr [Internet]. Curr Dev Nutr; 2020 [cited 2020 Sep 19];4:nzaa028. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/32258989. Accessed 3 Sept 2020.

[12] Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. Int J Surg [Internet]. Elsevier; 2020 [cited 2020 may 28];78:185–93. Available from: http://www.ncbi.nlm. nih.gov/pubmed/32305533. Accessed 3 Sept 2020.

[13] Guan H, Okely AD, Aguilar-Farias N, Del Pozo Cruz B, Draper CE, El Hamdouchi A, et al. Promoting healthy movement behaviours among children during the COVID-19 pandemic. Lancet Child Adolesc Heal [Internet]. Elsevier; 2020 [cited 2020 May 28];4:416–8. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/32458805. Accessed 3 Sept 2020.

[14] Sport New Zealand IHI Aotearoa. Guidance for physical activity at COVID-19 alert level 3 | Sport New Zealand - IHI Aotearoa [Internet]. Sport New Zeal. IHI Aotearoa. 2020 [cited 2020 Sep 22]. Available from: https://sportnz.org.nz/about/news-and-media/ media-centre/guidance-for-physical-activity-at-covid-19-alertlevel-3/. Accessed 3 Sept 2020.

[15] Inchley, J, Currie, D, Budisavljevic, S, Torsheim, T, Jåstad A, Cosma A et al. Spotlight on adolescent health and well-being. Findings from the 2017/2018 Health Behaviour in School-aged Children (HBSC) survey in Europe and Canada. International report. Volume 1. Key findings. Copenhagen; 2020.

[16] Nagata JM, Abdel Magid HS, Pettee Gabriel K. Screen Time for Children and Adolescents During the Coronavirus Disease 2019 Pandemic. Obesity [Internet]. John Wiley & Sons, Ltd; 2020 [cited 2020 Sep 19];28:1582–3. Available from: https://onlinelibrary. wiley.com/doi/abs/10.1002/oby.22917. Accessed 3 Sept 2020.

[17] Tripathi M, Mishra SK. Screen time and adiposity among children and adolescents: a systematic review. J Public Health (Bangkok) [Internet]. Springer; 2020 [cited 2020 May 28];28:227–44. Available from: http://link.springer.com/10.1007/s10389-019- 01043-x. Accessed 3 Sept 2020.

[18] Marsh S, Ni Mhurchu C, Maddison R. The non-advertising effects of screen-based sedentary activities on acute eating behaviours in children, adolescents, and young adults. A systematic review. Appetite [Internet]. Appetite; 2013 [cited 2020 May 28];71:259– 73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/ 24001394. Accessed 3 Sept 2020.

[17] Franckle R, Adler R, Davison K. Accelerated weight gain among children during summer versus school year and related racial/ethnic disparities: A systematic review. Prev Chronic Dis [Internet]. Prev Chronic Dis; 2014 [cited 2020 May 28];11. Available from: https:// pubmed.ncbi.nlm.nih.gov/24921899/. Accessed 3 Sept 2020.

[18] von Hippel PT, Workman J. From Kindergarten Through Second Grade, U.S. Children's Obesity Prevalence Grows Only During Summer Vacations. Obesity (Silver Spring) [Internet]. Obesity (Silver Spring); 2016 [cited 2020 may 28];24:2296–300. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27804271. Accessed 3 Sept 2020.

[19] An R. Projecting the impact of COVID-19 pandemic on childhood obesity in the U.S.: A microsimulation model. J Sport Heal Sci [Internet]. Elsevier; 2020 [cited 2020 May 28]; Available from: https://www.sciencedirect.com/science/article/pii/ S209525462030065X. Accessed 3 Sept 2020.

[20] Jogdand SS, Naik J. Study of family factors in association with behavior problems amongst children of 6–18 years age group. Int J Appl basic Med Res [Internet]. Wolters Kluwer – Medknow Publications; 2014 [cited 2020 Sep 22];4:86–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25143882. Accessed 3 Sept 2020.

[21] Leddy MA, Power ML, Schulkin J. The impact of maternal obesity on maternal and fetal health. Rev Obstet Gynecol [Internet]. MedReviews, LLC; 2008 [cited 2020 May 28];1:170–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19173021. Accessed 3 Sept 2020.

[22] Centers for Disease Control and Prevention. Others At Risk for COVID-19 | CDC [Internet]. Centers Dis. Control Prev. 2020 [cited 2020 Sep 22]. Available from: https://www.cdc.gov/coronavirus/ 2019-ncov/need-extra-precautions/other-at-risk-populations.html. Accessed 3 Sept 2020.

[23] Esegbona-Adeigbe S. Impact of COVID-19 on antenatal care provision. Eur J Midwifery [Internet]. E.U. European Publishing; 2020 [cited 2020 Sep 22]; Available from: http://www.journalssystem. com/ejm/Impact-of-COVID-19-on-antenatal-careprovision,121096,0,2.html. Accessed 3 Sept 2020. 24. Unicef. Framework for reopening schools [Internet]. 2020. Available from: https://www.unicef.org/sites/default/files/2020-06/Framework-for-reopening-schools-2020.pdf. Accessed 3 Sept 2020.

Chapter

Multiplicative Data Perturbation Using Random Rotation Method

Thanveer Jahan

Abstract

Today's applications rely on large volumes of personal data being collected and processed regularly. Many unauthorized users try to access this private data. Data perturbation methods are one among many Privacy Preserving Data Mining (PPDM) techniques. They play a key role in perturbing confidential data. The research work focuses on developing an efficient data perturbation method using multivariate dataset which can preserve privacy in a centralized environment and allow publishing data. To carry out the data perturbation on a multivariate dataset, a Multiplicative Data Perturbation (MDP) using Random Rotation method is proposed. The results revealed an efficient multiplicative data perturbation using multivariate datasets which is resilient to attacks or threats and preserves the privacy in centralized environment.

Keywords: privacy, multiplicative data perturbation, random rotation method

1. Introduction

This chapter proposes a Multiplicative Data Perturbation method. It considers multivariate datasets to perturb using a geometric data perturbation method. Then, the perturbed data will use Discrete Cosine Transformation between a pair of data values to determine Euclidean distance. This proposal is clearly elaborated in the following section.

1.1 Background

Hybrid transformations are used to maintain statistical properties of data as well as mining utilities [1–3]. The statistical properties of data are mean and variance or standard deviation without any loss of data. A feasible solution [4] is provided to optimize the data transformations by maximizing privacy of sensitive attributes. A combined technique using randomization and geometric transformation is used to protect sensitive data. A randomized technique is represented as $D = X + R$, where R is additive noise, X is original data and D is perturbed data. A geometric transformation is used as a 2D rotation data matrix represented as $D' = R(\theta) \times D$, where D is the column vector containing original co-ordinates and D' is a column vector whose co-ordinates are rotated clockwise. The above method considered only single attributes as sensitive and rest of them as non-sensitive attributes. Data perturbation method using fuzzy logic and random rotation is proposed [5, 6].

The original data is perturbed using fuzzy based approach (M) and then random rotation perturbation is used by selecting confidential numerical attributes to get the transformed data P = M*R, where M is the dataset transformed using fuzzy based approach and R is the random dataset generated. The distorted data P is released for clustering analysis and obtained accuracy. The approach compromises in balancing privacy and accuracy. A hybrid method using SVD and Shearing based data perturbation [7] is proposed to obtain perturbed data. The approach removes the identified attributes from the dataset. These attributes are normalized using Z-score normalization to standardize to the same. Then, the dataset is perturbed using SVD transformation. Each record of the perturbed dataset is further distorted using a Shear based data Perturbation method represented as $D' = D + (Sh_D * D)$, where Sh_D is the random noise and D is the perturbed dataset obtained after SVD transformation.

The results show higher privacy is attained on hybrid methods when compared to single data perturbation methods. A hybrid technique [7, 8] based on Walsh-Hadamard Transformation (WHT) and Rotation is proposed. The Euclidean distance preserving transformation using Walsh-Hadamard (Hn) given below to generate orthogonal matrix to preserve statistical properties of the original dataset.

$$H_n = \otimes_{i=n}^{D} \quad H_2 = \frac{H_2 \otimes H_2 \cdots \otimes H_2}{n} \tag{1}$$

where H_2 is $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is a matrix and denotes the tensor or Kronecker product. Then, Rotation transformation is applied to preserve the distance between the data points. The perturbed data preserves distance between data records and maintain accuracy using classifiers. The method is limited to numerical attributes and can be extended to categorical attributes. A hybrid approach for data transformation is proposed by Manikandan et al. [9] to sanitize data and normalize the data using min-max normalization [10]. The approach transforms original data maintaining inter-relative distance among the data. Clustering analysis shows that the numbers of clusters in original data are similar to modified data. Another approach is used to modify the original data to preserve privacy with the help of inter-relative distance on categorical data is proposed [2].

The categorical data is converted into binary data and is transformed using geometric transformation. Then the clustering algorithm is used for analysis and the results for better data utilization as well as privacy preservation. The multiplicative noise is generated using random numbers with mean as 1 and is multiplied by the original data value. A random number with a short Gaussian distribution is calculated with mean as 0 and a small variance. Geetha Mary A et al. [11] proposed a non-additive method of perturbation by randomization and data is generated based on intervals on the level of privacy specified by a user. A random number is generated that is either added or multiplied with the data to generate a random modified data. The perturbed data is classified and measures using metrics.

The condensation approach is presented by Agrawal and Yui [12] for a multidimensional perturbation technique to provide privacy for multiple columns using covariance matrix. The approach was weak in protecting data privacy. Rotation perturbation was used for privacy preserving data classification [13]. Rotation perturbations are task specific and aim to have better balance between loss of information and loss of privacy. Multiplicative data perturbations include three types of perturbation techniques such as: Rotation Perturbation, Projection Perturbation and Geometric Data Perturbation.

A Rotation perturbation framework was adopted in privacy preserving data classification [14]. It is defined as $G(X) = RX$ where R is randomly generated rotation matrix and X is the original data. The benefit and weakness of this method is distance preservation and is prone to distance inference attacks. These attacks are addressed [15–17]. Chen et al. [14] proposed an improved version on resilience towards attacks. Oliveria et al. [17] proposed a scaling transformation along with random rotation in privacy preserving clustering.

A Random Projection perturbation is proposed [13, 18] to project a set of points from the original multidimensional space to another randomly chosen space. This resulted with an approximate model quality. A random projection matrix is used in privacy preserving data mining to enable an individual to choose their privacy levels.

An ideal data perturbation [19] aim with a balance tradeoff of minimizing information loss and privacy loss. However these are not balanced in the existing algorithms. Compared with the existing approaches in privacy preserving data mining, Geometric data Perturbation have significantly reduced these overcome [20].

A Geometric Data Perturbation is a sequence of random geometric transformation including multiplicative transformation (R), Translation Transformation (T) and Distance Perturbation (DP) [21, 22].

$$(X) = R(X) + T + DT \tag{2}$$

The approach has two unique characteristics. The first characteristic is to perturb the original data with geometric rotation, translation and identify rotation invariant classifiers as given in above. The second characteristic is to build privacy model by evaluating the privacy quality of perturbation method. The privacy model generated is used to analyze the attacks, such as, Naives and ICA-based reconstruction. The quality of data perturbation approach is determined by the quality of privacy preserved. It is the difficulty level in estimating the original data from perturbed ones such estimations are named as inference attacks. The attacks are categorized into three categories such as: Naives Inference, Reconstruction based inference and distance based inference. A statistical method based inference to estimate original data from perturbed named as Naives inference attack was proposed [23]. It is represented as O=P, where O is the observed data and P is the perturbed data. Reconstructing the data with perturbed and released information from data is presented. Reconstruction based attacks also called as Independent Component Analysis (ICA) [24, 25]. It is represented as, $O = E^{-1} P$, where E^{-1} is the estimation of released information of data and P is the perturbed data. Identifying the images and some relevant information of data using outliers to discover the perturbation is distance based attacks. It is represented as $O = E^{-1}P$, where E^{-1} is the mapping to estimate and P is the perturbed data. The higher the inference the more the original data is protected and preserved such that attacker cannot break the perturbation. The above attacks are analyzed with a privacy model with privacy guarantee [26]. It had failed to avoid outlier attack. The existing data perturbation techniques have contradiction between data privacy metric and mining utility [27, 28]. The multiplicative data perturbations will maximize the two levels i.e. data privacy and mining utility. The multiplicative data perturbation shows challenging features to improve data privacy during mining process as well as to preserve the model specific information.

In this chapter a survey is presented on privacy preserving data mining to protect confidential data. The drawbacks of the above existing data perturbation methods have made us to resolve the issues with balanced factors, such as, data privacy and data utility. The challenges in preserving privacy using multiplicative data perturbation have been given a new direction in this research study.

2. Proposed method

The proposed Multiplicative Data Perturbation (MDP) is shown at **Figure 1** as a block diagram.

The above block diagram considers the original dataset and deals with it in two stages. In the first stage, the original dataset is perturbed using geometric data perturbation. The geometric data perturbation generates a distorted dataset. This distorted dataset is further perturbed using Discrete Cosine Transformation in the second stage to finally generate a distorted dataset. The process of generating a distorted dataset using a geometric data perturbation comprises three steps. At the first step a random dataset is created using random values as in the original dataset. This random dataset is rotated counter clockwise and then multiplied with the original dataset. The resultant dataset obtained the above step is transposed in the second step, that is, Translation Transformation. This Transposed dataset is added with an additive noise in the third step to obtain a distorted dataset. This proposal is an algorithm for multiplicative data perturbation in the next section.

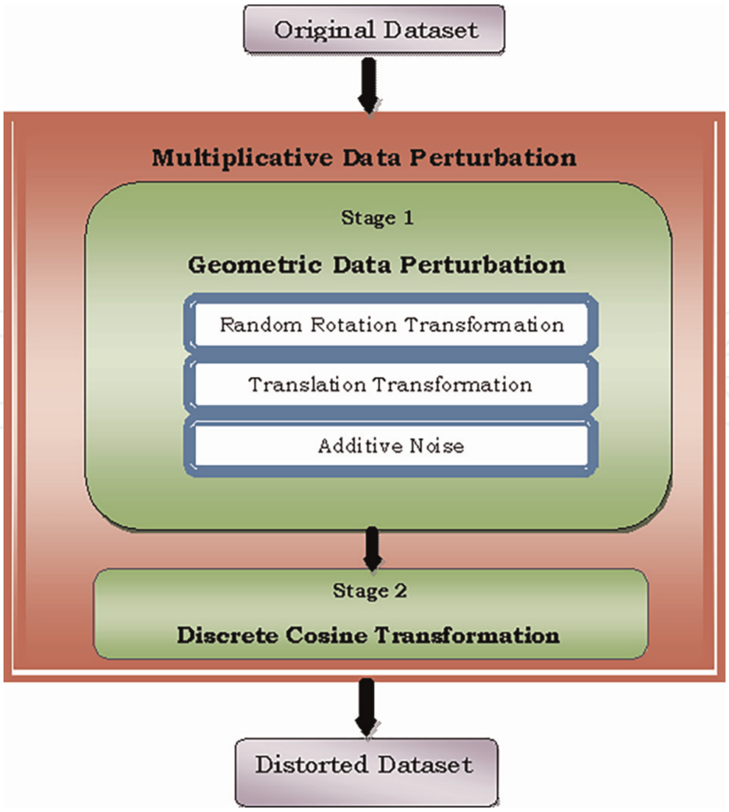

Original Dataset
Multiplicative Data Perturbation
Stage 1
Geometric Data Perturbation
Random Rotation Transformation
Translation Transformation
Additive Noise
Stage 2
Discrete Cosine Transformation
Distorted Dataset

Figure 1.
Block diagram for multiplicative data perturbation using random rotation method.

3. Proposed multiplicative data perturbation using random rotation algorithm

A proposal for multiplicative data perturbation is given in this section. The pseudo code of the proposed algorithm is listed below.

```
Algorithm:
Input: A Data Matrix $D_{p \times q}$.
Output: A Distorted Data matrices D4, D5.
Begin.
Step 1: Create a Random data matrix R with p rows and q column and Rotate the
        random data matrix as $R_{q \times p}$//counter clock wise Rotation by $90^{\circ}$.
Step 2: Construct the data matrix $X_{p \times q}$ using $R_{q \times p}$ and $D_{p \times q}$ data matrices as:
        $X_{q\times q} = R_{q\times p} * D_{p\times q}$//Multiplicative Transformation.
Step 3: Create another random data matrix $X1_{p \times q}$ with p rows, q columns with
        mean as 0 and standard deviation as 1.
Step 4: Construct the distorted data matrix $D4_{p \times q}$ using $X_{p \times q}$, Transpose of R
        and $X1_{p \times q}$ data matrices as:
        $D4 = X + R^{T} + X1$//Geometric data Perturbation.
Step 5: Call function DCT ($D4_{p \times q}$:$D5_{p \times q}$)//Discrete cosine transformation.
Step 6: The resultant distorted data matrix $D5_{p \times q}$ is output,
End.

Function DCT ($D4_{p \times q}$:$D5_{p \times q}$)//Function for Discrete Cosine Transformation.
Input: A data matrix $D4_{p \times q}$ Output: A data matrix $D5_{p \times q}$
Begin.
Step 1: Copy the data matrix D4 to a data matrix D5//alias
Step 2: For i = 1 to q.
        For k = 1 to q.
            If k = 1 then
                $D4[i] = \left(1/\sqrt{i} * X_2(i) * (\cos(3.14 * (2+1)/2i))\right)$
            Else
                $D5[i] = \left(\sqrt{2}/i * X_2(i) * (\cos(3.14 * (2+1)/2i))\right)$
            End if
        End For
Construct D5 data matrix and return as parameter.
End
```

The algorithm accepts the data matrix $D_{p \times q}$ with p rows and q columns as input. It creates a random data matrix R with p rows and q columns having random values as elements. This random data matrix R is rotated counter clockwise by 90° and then multiplied with data matrix $D_{p \times q.}$ The data matrix that results is named as data matrix $X_{p \times q.}$ Create another random data matrix X1 with p rows, q columns such that its mean is 0 and standard deviation is 1. Now, construct the distorted data matrix D4 adding the data matrices X, R^{T} and X1. This data matrix D4 is passed as a parameter to the called function DCT(). The predefined conditions are checked and data matrix D5 is updated. This data matrix D5 after completely updated is an output of the algorithm. The time complexity of the proposed MDP algorithm is found to be O(n), where n is the dimension of the dataset.

The process of updating D5 is explained with the help of an example stated below:

Example 1.1: Consider a data matrix $D_{p \times q} = \begin{bmatrix} 4 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$ where p = 2 and q = 3.

At Step 1, create a random data matrix $R_{2 \times 3}$ as given below:
$R = \begin{bmatrix} -0.3034 & -0.7873 & -1.1471 \\ 0.2939 & 0.8884 & -1.0689 \end{bmatrix}$ and rotate R counter clockwise by 90° as given below:

$$R_{3\times2} = \begin{bmatrix} -1.1471 & -1.0689 \\ -0.7873 & 0.8884 \\ -0.3034 & 0.2939 \end{bmatrix}$$

At step 2, construct the data matrix $X = D_{2 \times 3} * R_{3 \times 2}$ is given as below:

$$X = \begin{bmatrix} -6.4664 & -2.2049 \\ -2.2378 & 0.1134 \end{bmatrix}$$

At step 3, create another random data matrix X1 with 2 rows and 3 columns such that the mean is 0 and the standard deviation is 1.

$$X1 = \begin{bmatrix} -6.4664 & 1.4384 & -0.7549 \\ -2.9443 & 0.3252 & 1.3703 \end{bmatrix}$$

At step 4, construct the distorted data matrix $D4 = X\ {}^{+}\ R^{T} + X1$ as given as: $D4 = \begin{bmatrix} -3.1036 & -0.1362 & -1.3618 \\ -5.0820 & 2.1020 & 1.980. \end{bmatrix}$

At step 5, the function call DCT (D4:D5) where

$$\mathrm{DCT}(k) = f(k)\sum_{k=1}^{q} D4(q) \cos\left[(2k+1)i\pi/2q\right] \quad k = 1,2\cdots q;\ \ i = 1\cdots p \qquad (3)$$

where

$$f(k) = \begin{cases} \dfrac{1}{\sqrt{q}} & k = 1 \\ \dfrac{\sqrt{2}}{q} & 2 \le k \le q \end{cases}$$

Let k = 1, q = 1, $f(k) = \frac{1}{\sqrt{q}}$., then f(1) = 1, substituting the values in the Eq. (3)
$\mathrm{Dct}(1) = 1* - 3.1036* \cos[3*3.14/2] = -5.7881$

Let k = 2, q = 1, $f(2) = \sqrt{\frac{2}{q}}$, then f(2) = 1, substituting the above values in Eq. (3)
$\mathrm{DCT}(2) = 1* - 0.1362* \cos[(2*2)*3.14/2*2] = 1.3900$

Similarly, the remaining data values of D4 are calculated to form a D5 data matrix as given below:

$$D5 = \begin{bmatrix} -5.7881 & 1.3900 & 0.4371 \\ 1.3989 & -1.5826 & -2.3630 \end{bmatrix}$$

The constructed data matrix D5 is the output.

4. Implementation

The proposed algorithm that was discussed in the previous section is implemented in MatLab. Its source code is included. The details of implementation are furnished in this section.

The implementation utilizes the built in functions available in MatLab such as load (), size(), randn(), rot90(), dct() and normrnd(). First, a load() built-in function is used to read a data into a data matrix D. The size() function is employed to retrieve the number of rows and columns. The function randn() is used to generate a random matrix R where the size is similar to data matrix D. The data matrix R is rotated using built in function available, namely rot90(). Then, to form a data matrix X, the data matrice R is multiplied by D data matrix. Next, normrnd() is called to generate a data matrix X1 having the mean as 0, the standard deviation as 1 and the size as similar to data matrix D. The distorted data matrix D4 is constructed by adding three data matrices, X1, R^T and X2. Finally, the function DCT() is employed on distorted data matrix D4 to obtain the resultant distorted data matrix D5.

5. Experimentation

The Experimentation was conducted using desktop computer system loaded with windows XP Operating system, MatLab and Tanagra data mining tool. The experimental details are elaborated in this section. The experimentation begins with the original dataset D is given as input to the proposed MDP algorithm to obtain the distorted dataset D4 and D5. Then, the original dataset D and distorted datasets D4 and D5 are uploaded into Tanagra data mining tool after appending a class attribute. These uploaded datasets are classified using classification utility available within Tanagra data mining tool. The results of classification are analyzed thereafter.

Similarly the datasets are clustered using clustering utilities available in them. The results of clustering are also analyzed and furnished at Section 6.6 under Results and Analysis. Unified column privacy metric to analyze possibility of attacks is also discussed in this section. But, their calculation is shown in section Results and Analysis. The datasets of Credit Approval, Haber-Man, Tic-Tac-toe and Diabetes are used in this experimentation. The details of Credit Approval dataset used in this experiment is furnished here and the rest of the datasets are furnished.

A Real Time Multivariate dataset, namely, Credit Approval, is downloaded from website UCI Machine Learning Repository. The details are shown at **Table 1**. Therefore the original dataset used in the experimentation is a Credit Approval dataset. It

Dataset	Size	Description
Credit Approval	690 rows & 15 columns	It consists of information of customers details concerned with credit card applications

Table 1.
Details of credit approval dataset.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213
0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1
0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1
0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1
1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159

Table 2.
A credit approval original dataset D.

comprises 690 rows/tuples and 15 columns/attributes including one target/class attribute.

A sample list of the original dataset D with 5 rows and 14 attributes is shown at **Table 2**.

The process in the experiment is explained as below:

First, a dataset named creditapproval.txt is loaded into X data matrix with the help of load() method. Next, the size() method on X data matrix determines the number of rows p as 690 and the number of columns q as 14. The data matrix is now named $D_{p \times q}$. Then, a built- in function randn(p, q) is used to create a random data matrix R. The random data matrix R is rotated with the help of built-in function rot90(). The data matrix X is constructed using data matrix R multiplied by data matrix D. The built in function normrnd(0,1, p, q) is used to create another random data matrix X1 with p rows, q columns, such that its mean is 0 and standard deviation is 1. Construct the distorted data matrix D4 by adding three data matrices X, R^T(transpose of R), X1. The distorted data matrix D4 is given as parameter to function DCT(D4) and it returns the final distorted data matrix D5 as output. When the above process is executed in experimentation it outputs a distorted datasets D4 and D5.

6. Results and analysis

The distorted datasets D4 and D5 together with the original dataset D, respectively are appended with a class attribute, YES or NO. The original dataset D after appending with a class attribute is shown at **Table 3**.

Similarly the distorted datasets D4 and D5 are also appended with a class attribute and furnished at section 6.6 as part of Results and Analysis. The above mentioned datasets D, D4 and D5 are uploaded into Tanagra data mining tool. First, classification utility is used on the dataset D and distorted datasets D4, D5. It divides the attributes into two categories, non-class attributes and class attribute. These two categories can be two inputs to the classifier chosen from the available ones.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Class
1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213	NO
0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1	NO
0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1	NO
0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1	YES
1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159	YES

Table 3.
A credit approval original dataset with class attribute.

Suppose we select SVM (Support Vector Machine) as classifier, then, it classifies the datasets D, D4 and D5 based on class attribute into either credit card either approved or rejected. Such results are furnished at Section 6.6 under Results and Analysis. Similarly, the experimentation is repeated with Iterative Dichotomizer 3 (ID3), (Successor of ID3) C4.5, KNN (k-Nearest Neighbor) and MLP (Multi Layer Perceptron) classifiers.

The results of those experiments are furnished at Section 6.6. A Clustering utility available in Tanagra data mining tool is used to cluster the original dataset D and distorted datasets D4 and D5. Non- class attributes are considered and given as input to k-mean clustering method. As a result, categories of clusters are formed.

A unified column metric, Root Mean Square Error (RSME) is used to evaluate inference attacks. It is calculated using Eq. (3) as given below:

$$\mathrm{RSME(r)} = \sqrt{\frac{1}{q}\sum\nolimits_{i=1}^{q}(D-P)^2} \tag{4}$$

where $D = d_1, d_2 \cdots d_q$ are the original dataset values, $P = p_1, p_2 \cdots p_q$ are the perturbed dataset values and q is number of columns.

Then, privacy (D, P) $= \frac{4\sigma}{2r} = \frac{r}{2}$ (if standard deviation σ = 1). The attacks used are:

Naives inference is calculated as given in Eq. (4), where D is the original data and P = E (E is estimated or Random dataset).

Reconstruction inference is calculated as given in Eq. (4), where D is the original dataset and the Perturbed dataset

$$P = E^{-1} * P. \tag{5}$$

Distance based inference is calculated as given in Eq. (5), where D is the original dataset and P = P′ (P′ is mapped set of points of Perturbed dataset P).

The calculations of these metrics are furnished at Section 7 under Results and Analysis.

7. Results and analysis

The results obtained in the above experiment are presented in this section. The original dataset D is given as input to the proposed MDP and output distorted dataset D4 and D5 are presented below at **Table 4** and **Table 5**, respectively.

When SVM classifier is used on D, D4 and D5 datasets, the following observations are made and the same are presented at **Table 6**.

In the above **Table 4**, the first column presents the original dataset D and the distorted datasets D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples classified for credit card approved as YES. The number of support vectors available is furnished in the fourth column. Fifth column reveals the error rate of SVM classifier. The computation time is tabulated at last column.

Similarly, when ID3 and C4.5 classifiers are used on D, D4 and D5 datasets the results are tabulated at **Tables 7** and **8**.

In the above **Tables 7** and **8**, the first column presents the dataset D and distorted dataset D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples belonging to credit card approved as YES. A tree having number of nodes and leaves is furnished in the fourth column. Fifth column reveals the error rate of the ID3 and C4.5 classifiers, respectively. The computation time is tabulated at last column.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
7.2	15.69	14.65	5.3	18.7	9.8	2.57	−11.5	1.97	21.4	10.18	8.44	94.11	1.22
2.7	31.43	4.313	0.7	14.2	4.0	4.99	1.4	−6.69	4.77	−8.25	9.66	157.6	3.98
1.3	30.06	−13.5	17.8	3.13	22.0	−6.14	−4.23	5.28	−6.39	2.14	6.61	259.3	−2.37
9.7	26.01	9.224	−4.4	7.82	−10.3	−7.82	16.13	−10.2	1.78	13.12	−2.16	7.82	7.38
2.1	2.033	−9.22	−0.5	22.2	6.95	−0.10	2.54	−6.28	12.6	0.05	5.964	57.78	159.9

Table 4.
A credit approval distorted dataset D4.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
26.9	822.	111.6	43.7	19.6	108.	46.0	30.18	7.99	73.31	24.07	57.16	4.85	2.67
8.8	−1.6	10.90	−6.3	11.2	−4.44	6.46	6.85	−0.08	−0.08	−7.90	−5.33	−67.91	594.2
−4.8	−12.	19.15	−12.	12.4	−13.1	6.02	3.03	−10.18	−10.1	−2.99	−12.4	−229.7	−1.56
−2.3	2.86	5.801	4.62	3.55	5.80	5.89	−1.11	−11.90	−11.9	2.75	15.73	−12.28	1.87
22.9	−8.8	−11.61	−1.2	1.79	4.63	−8.72	8.00	−3.50	−3.50	2.75	−8.24	49.84	−1.27

Table 5.
A credit approval distorted dataset D5.

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Approved (YES)	Number of Support Vectors	Error Rate	Computation Time (ms)
Original (D)	690	589	392	0.14	1562 ms
Distorted (D4)	690	582	621	0.446	2172 ms
Distorted (D5)	690	587	632	0.315	1969 ms

Table 6.
A credit approval dataset classified using SVM.

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Approved(YES)	Tree having number of nodes and leaves	Error Rate	Computation Time(ms)
Original (D)	690	584	7 node,4 leaves	0.1464	16 ms
Distorted(D4)	690	476	3 node, 2 leaves	0.3101	31 ms
Distorted(D5)	690	580	1 node, 1 leaf	0.4464	16 ms

Table 7.
A credit approval dataset classified using ID3.

Dataset	Total Number of Tuples	Number of Training Tuples classified as Approved(YES)	Tree having number of nodes and leaves	Error Rate	Computation Time(ms)
Original (D)	690	644	67 node, 34 leaves	0.066	47 ms
Distorted(D4)	690	621	137 nodes, 69 leaves	0.101	172 ms
Distorted(D5)	690	634	157 nodes, 79 leaves	0.1246	234 ms

Table 8.
A credit approval dataset classified using C4.5.

Dataset	Total number of tuples	Number of Training Tuples Classified as Approved (YES)	Neighbors	Error Rate	Computation Time(ms)
Original (D)	690	537	5	0.2217	313 ms
Distorted(D4)	690	485	5	0.297	422 ms
Distorted(D5)	690	673	5	0.3145	391 ms

Table 9.
A credit approval dataset classified using KNN.

When KNN classifier is used on D, D4 and D5 datasets the following observations are made and presented at **Table 9**.

In the above **Table 9**, the first column presents the original dataset D and distorted datasets D4 and D5. The number of tuples in the

datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples classified as credit card approved as YES for KNN classifier. The fourth column displays the number of neighbors. The fifth column reveals the error rate of KNN classifier. The computation time is tabulated in the last column.

Similarly, the results are tabulated at **Table 10** when MLP classifier is used on D, D4 and D5 datasets.

In the above **Table 10**, the first column presents the original dataset D and the distorted datasets D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of tuples classified for credit card approved as YES. The maximum number of

Dataset	Total Number of Tuples	Number of tuples Classified as Approved (YES)	Max Iteration	Train Error Rate	Computation Time(ms)
Original (D)	690	620	100	0.0924	578 ms
Distorted(D1)	690	552	100	0.168	562 ms
Distorted(D2)	690	589	100	0.347	625 ms

Table 10.
A credit approval dataset classified using MLP.

iteration for MLP classifier is furnished in the fourth column. The fifth column reveals the training error rate of KNN classifier. The computation time is tabulated in the last column. Based on the results presented above the accuracy of classification of datasets is presented at **Table 11**. The accuracy is the percentage of tuples that were correctly classified by a classifier.

The above **Table 11** presents the accuracy of the classifiers for Credit Approval, Haber Man, Tic-Tac-Toe and Diabetes datasets. The first column presents the dataset D, the distorted datasets D4 and D5. The second column presents the accuracy of classification obtained on Credit Approval dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The third column presents the accuracy of classification obtained on Haber Man dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The fourth column presents the accuracy of classification obtained on Tic-Tac-Toe dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The fifth column presents the accuracy of classification obtained on Diabetes dataset using SVM, ID3, C4.5, KNN and MLP classifiers.

It is observed that accuracy of C4.5, KNN and MLP classifiers are better than the accuracy of the other classifiers for distorted dataset D5 compared to distorted dataset D4.

The above **Table 12** presents the comparison of accuracy. The first column presents the distorted dataset D4 and D5. The second column presents the accuracy obtained on the proposed MDP using Credit approval, Tic-Tac-Toe and diabetes datasets for SVM and KNN classifiers. The third column presents the accuracy for the existing geometric data perturbation methods using Credit approval, Tic-Tac- Toe and Diabetes datasets for SVM and KNN classifiers. It is observed that the accuracy on the datasets using our proposed MDP was found better than the accuracy of the Existing Geometric data perturbation. Moreover, their accuracy was found only on SVM and KNN classifiers for Credit Approval, Tic-Tac-Toe, and Diabetes datasets only.

The proposed MDP has given good accuracy for distorted dataset D5 compared to distorted dataset D4, whereas the literature does not show any accuracy for distorted data D5.

The results of k-means clustering are shown below at **Table 13**, when k = 2 (form two clusters).

In the above **Table 13**, the first column presents the dataset D, D4, and D5. The number of objects in the dataset considered for the experiment can be seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster 2. The computational time is presented in the last column. Based on the results presented above the misclassification error rate of datasets is presented at **Table 14**.

Dataset	Credit Approval					Haber Man					Tic-Tac-Toe					Diabetes				
	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP
Original (D)	85	84	93	98	89	75	86	88	89	90	97	89	87	98	97	89	86	82	89	90
Distorted (D4)	86	70	90	88	80	68	78	79	75	76	98	73	77	96	89	80	83	79	79	85
Distorted (D5)	88	84	91	97	87	72	85	85	89	90	99	88	87	98	97	83	85	81	89	90

Table 11.
Accuracy of classifiers (%).

Dataset	Proposed Multiplicative Data Perturbation (MDP)						Existing Geometric Data Perturbation Method					
	Credit Approval		Tic-Tac-Toe		Diabetes		Credit Approval		Tic-Tac-Toe		Diabetes	
	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN
Distorted (D4)	86	88	98.7	98	80	79	86.5	82.9	98	99.5	77	73.5
Distorted (D5)	88	97	99	98.5	83.4	89	—		—		—	

Table 12.
Comparison of accuracy.

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Computation time (ms)
Original (D)	690	259	431	94 ms
Distorted (D4)	690	391	299	109 ms
Distorted (D5)	690	336	354	125 ms

Table 13.
Clustering on credit approval dataset for k = 2.

Dataset	PROPOSED MULTIPLICATIVE DATA PERTURBATION (MDP)			
	Credit Approval	Haber Man	Tic-Tac- Toe	Diabetes
Distorted (D4)	0.389	0.189	0.035	0.03
Distorted (D5)	0.22	0.100	0.031	0.02

Table 14.
Comparison of misclassification error-rate.

The above **Table 14** presents the misclassification error rate. The first column presents the distorted dataset D4 and D5. The second column presents the error rate obtained on the proposed MDP using Credit Approval, Haber Man, Tic-Tac-Toe and Diabetes datasets.

In the privacy metric mentioned in Section 1.5 in Eq. 1.2, the detailed calculation of privacy quality to analyze attacks is shown below:

Consider the data matrix $D = \begin{bmatrix} 4 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$ the corresponding distorted data matrix using the proposed MDP is given below:

$P = \begin{bmatrix} -5.7881 & 1.3900 & 0.4371 \\ 1.3989 & -1.5826 & -2.3630 \end{bmatrix}$, E is the estimated values (Random) as given below:

$E = \begin{bmatrix} -3.5441 & 1.3900 & 0.3211 \\ 0.9321 & 2.4567 & -6.7860 \end{bmatrix}$ and calculating $D'' = R^{-1}*P$ is given below

$D'' = \begin{bmatrix} -2.1461 & 0.2800 & 0.3211 \\ 1.8421 & 4.6767 & 4.6130 \end{bmatrix}$ and calculating P' is given below

Attacks	**Proposed MDP Method**				**Existing Geometric Data Perturbation Method**		
	Credit Approval	**Haber Man**	**Tic-Tac-Toe**	**Diabetes**	**Credit Approval**	**Tic-Tac-Toe**	**Diabetes**
Naives	1.743	1.129	1.564	1.512	1.345	1.234	1.456
Reconstruction	1.467	1.841	1.489	1.893	1.287	1.450	1.921
Distance	1.527	1.980	1.901	1.452	1.556	1.784	1.356

Table 15.
Analysis on attacks.

$$P' = \begin{bmatrix} -1.9261 & 0.6800 & 1.3211 \\ 3.6821 & 1.6821 & -4.5920 \end{bmatrix}$$

Then, substitute the above data matrices in eq. 1.2 to analyze the following attacks:

Naives-based Inference Attack: The RMSE is calculated by substituting the data matrices D and E. The result for RMSE r, obtained is as given below:

$$r = \sqrt{\frac{1}{3}\sum_{i=1}^{2}((D) - (E))^2} = 1.9221, \text{ Privacy } (D, P) = r/2 = 0.6796$$

Reconstruction -based Inference Attack: The RSME r is calculated by substituting the data matrices D and D″. The result r obtained is as given below:

$$r = \sqrt{\frac{1}{3}\sum_{i=1}^{2}((D) - (D''))^2} = 1.6794, \text{ Privacy } \left(D, D''\right) = r/2 = 0.839$$

Distance -based Attack: The RSME r is calculated by substituting the data matrices D and P′. The result r obtained is as given below:

$$r = \sqrt{\frac{1}{3}\sum_{i=1}^{2}((D) - (P'))^2} = 1.70261, \text{ Privacy } (D, P') = 0.851$$

Similarly the RMSE r is calculated for the original D and distorted datasets D4 and D5 and the results are furnished at **Table 15** as shown below.

In the above **Table 15**, the first column presents the Naives based, Reconstruction based and Distance -based attacks. The second column displays RMSE (Root Mean Square Error) r is calculated for the proposed MDP method on Credit Approval, Haber Man, Tic-Tac- Toe and Diabetes datasets. The third column reveals the RMSE calculated for existing hybrid methods on Credit Approval and Diabetes datasets. It is observed that the RMSE r for proposed MDP method on distance -based attack is high compared to RMSE for the existing geometric data perturbation methods. The metric for the proposed MDP shows better quality in preserving the confidential data and provides high uncertainity to reconstruct the original data.

8. Conclusion

A Multiplicative Data Perturbation algorithm by combining a Geometric Data Perturbation method and Discrete Cosine Transformation is proposed in this chapter.

The proposed MDP is successfully implemented using different multivariate datasets mentioned above.

The experiments on those datasets resulted to classify accurately and create accurate number of clusters. Based on the result analysis, it is resolved that our proposed MDP algorithm is efficient to preserve confidential data during perturbation and ensures privacy while being resilient against possible of attacks the proposed methods considered a univariate datasets ex: Terrorist. A multivariate dataset is considered and a multiplicative data perturbation (MDP) was explored to effectively perturb the data in a centralized environment. This method has resulted in perturbing the data effectively and be resilient towards attacks or threats while preserving the privacy.

The research studies can explore the privacy issues on a Big Data as a future scope of research work in the following directions:

Improving Data Analytic techniques –Gather all data, filter them out with certain constraints and use to take confident decision.

Algorithms for Data Visualization- In order to visualize the required information from a pool of random data, powerful algorithms are crucial for accurate results.

In future scope includes, research can include many various methods explore many methods. These latest methods can show various results.

Author details

Thanveer Jahan
Vaagdevi College of Engineering, India

*Address all correspondence to: tanveer_j@vaadevi.edu.in

IntechOpen

References

[1] Li L, Zhang Q. A privacy preserving clustering technique using hybrid data transformation method. In: 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009). Vol. 2010. Nanjing: IEEE; 2009. pp. 1502-1506. DOI: 10.1109/GSIS.2009.5408151

[2] Natarajan AM, Rajalaxmi RR, Uma N, Kirubhkar G. A hybrid transformation approach for privacy preserving clustering of categorical data. In: Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. Dordrecht: Springer. 2007. pp. 403-408. DOI: 10.1007/978-1-4020-6268-1_72

[3] Selva Rathnam S, Karthikeyan T. A survey on recent algorithms for privacy preserving data mining. International Journal of Computer Science and Information Technologies. 2015;**6**(2): 1835-1840

[4] Patel A, Patel K. A hybrid approach in privacy preserving data mining. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). Vol. 2. Ahmedabad, Gujarat, India: IEEE; 2016. p. 3

[5] M. Naga Lakshmi and K. Sandhya Rani, "A privacy preserving clustering method based on fuzzy approach and random rotation perturbation", Publications of Problems & Application in Engineering Research-Paper, Vol. 04, Issue No. 1, pp. 174-177, 2013.

[6] Mary AG. Fuzzy–based random perturbation for real world medical datasets. International Journal of Telemedicine and clinical Practices. 2015;**1**(2):111-124. DOI: 10.1504/IJTMCP.2015.069749

[7] M. Naga Lakshmi, K Sandhya Rani," Privacy preserving hybrid data transformation based on SVD"," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, 2013, 2278-1021

[8] Jalla HR, Girija PN. An efficient algorithm for privacy preserving data mining using hybrid transformation. International Journal of Data Mining & Knowledge Management Process. 2014; **4**(4):45-53. DOI: 10.5121/ijdkp.2014.4404

[9] Manikandan G, Sairam N, Saranya C, Jayashree S. A hybrid privacy preserving approach in data mining. Middle- East Journal of Scientific Research. 2013; **15**(4):581-585. DOI: 10.5829/idosi.mejsr.2013.15.4.1.991

[10] Saranya C, Manikandan G. Study on normalization techniques for privacy preserving data mining. International Journal of Engineering and Technology (IJET). 2013;**5**(3):2701-2704

[11] Geetha Mary AN, Iyenger NSC. Non-additive random data perturbation for real world data. Procedia Technology. 2012;**4**:350-354. DOI: 10.1016/j.protcy.2012.05.053

[12] Aggarwal CC, Yu PS. A condensation approach to privacy preserving data mining. In: Proceedings of International Conference on Extending Database Technology (EDBT). Vol. 2992. Heraklion, Crete, Greece: Springer; 2004. pp. 183-199

[13] Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering (TKDE). 2006;**18**(1):92-106

[14] Chen K, Liu L. “A Random Rotation Perturbation Based Approach to Privacy Preserving Data Classification”, CC-Technical Report GIT-CC-05-12. USA: Georgia Institute of Technology; 2005

[15] Lui K, Giannella C, Kargupta H. An Attacker’s view of distance preserving maps for privacy preserving data mining. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases(Pkdd’06). Berlin, Heidelberg: Springer-Verlag; 2006

[16] Xu H, Guo S, Chen K. Building confidential and efficient query services in the clod with RASP data perturbation. IEEE Transactions on Knowledge and Data Engineering. 2014;**26**(2):322-335

[17] Oliveira SR, Zaiane OR. Privacy preserving clustering by data transformation. Journal of Information and Data Management (JIDM). 2010; **1**(1):37–51

[18] Guo S, Wu X. Deriving private information from arbitrarily projected data. In: Proceedings of the 11th European conference on principles and practice of knowledge Discovery in databases (PKDD07). Warsaw, Poland. 2007

[19] Balasubramaniam S, Kavitha V. A survey on data retrieval techniques in cloud computing. Journal of Convergence Information Technology. 2013;**8**(16):15-24

[20] Liu J, Yifeng XU. Privacy preserving clustering by random response method of geometric transformation. Harbin, Heilong Jiang, China: IEEE. 2010: 181-188. DOI: 10.1109/ICICSE.2009.31

[21] Balasubramaniam S, Kavitha V. Geometric data perturbation-based personal health record transactions in cloud computing. The Scientific World Journal. 2015;**2015**:927867, 1-927869. DOI: 10.1155/2015/927867

[22] Chen K, Lui L. Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining. London: Springer-Verlag Limited; 2010

[23] Hyvarinen AK, Oja E. Independent Component Analysis. New York/Chichester/Weinheim/Brisbane/Singapore/Toronto: Wiley-Interscience; 2001

[24] Brankovic L, Estivill-Castro V. Privacy issues in knowledge discovery and data mining. In: Proceedings of Australian Institute of Computer Ehic Conference (AICEC99). Melbourne, Victoria, Australian: Lecture Notes in Computer Science. 1999;**4213**:297-308. DOI:10.1007/11871637_30

[25] Liu K, Giannella C, Kargupta H. An Attacker’s view of distance preserving maps for privacy preserving data mining. In: European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). Berlin, Germany; 2006

[26] Li L, Zhang Q. A privacy preserving clustering technique using hybrid data transformation method. In: Grey Systems and Intelligent Services, 2009 GSIS 2009, IEEE International Conference. Nanjing, China: IEEE; 2010. DOI: 10.1109/GSIS.2009.5408151, 08

[27] Rajesh N, Sujatha K, Kumar AALS. Survey on privacy preserving data mining techniques using recent algorithms. International Journal of Computer Applications Foundation of Computer Science (FCS). 2016;**133**(7):30-33

[28] Patel L, Gupta R. A survey of perturbation technique for privacy-preserving of data. International Journal of Emerging Technology and Advanced Engineering Website. 2013;**3**(6):162-166

DATA PERTURBATION METHOD TO PRESERVE PRIVACY

Blue Ink
Publishing House

Dr. Thanveer Jahan

A Book

On

DATA PERTURBATION METHOD TO PRESERVE PRIVACY

By

Dr. THANVEER JAHAN

Professor and Head CSE (AI & ML)

Vaagdevi College of Engineering

Warangal

Blue Ink
Publishing House

PREFACE

Today's applications rely on large volumes of personal data being collected and processed regularly. Many unauthorized users try to access this private data. Data perturbation methods are one among many Privacy Preserving Data Mining (PPDM) techniques. They play a key role in perturbing confidential data. The research work focuses on developing an efficient data perturbation method using multivariate dataset which can preserve privacy in a centralized environment and allow to publish data.

The contributions are made in this book are used to accomplish the above mentioned proposal. Initially a univariate dataset is considered and matrix decomposition method is modified to perturb the data. To overcome the information loss observed, a fuzzy logic based data perturbation is proposed. Even though the accuracy of classification as well as clusters were found satisfactorily in the above methods they used only univariate dataset.

ACKNOWLEGMENT

It gives me great pleasure to acknowledge with gratitude the help and guidance extended to me by a host of people for the successful completion of my research work.

My special thanks to my beloved husband **Dr Md. Anwar Miya**, my children, **Tanisha Aanam** and **Ashiqa Aanam**, for their love, understanding and cooperation during my research work.

Words are insufficient to express my thanks for my friends and well-wishers supporters who directly or indirectly have always been there by my side to encourage and overcome all the problems that I have came across through.

I always see God through my parent's blessings. With the grace blessings, kindness and mercy of Allah, reached my destination.

I prayerfully dedicate my book to my beloved parents:

Md. Mahaboob Ali and Shaheda Begum

Date: 29-03-2023 **Dr. Thanveer Jahan**

Title of the Book: **“DATA PERTURBATION METHOD TO PRESERVE PRIVACY”**

Edition: First- 2023

Disclaimer

The author is solely responsible for the contents published in this book. The publishers or editors don’t take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the editors or publishers to avoid discrepancies in future.

ISBN: 978-93-92473-55-5

MRP: ₹199/-

PUBLISHER & PRINTER: Blue Ink publishing house

1st floor opposite to Shree lekha Theatre Chikkagamalur-

577101, Karnataka, India.

Tel.: +91- 8050855786

E-mail: blueinkpublisherofficial@gmail.com

IMPRINT: Blue Ink print

About Author

Dr.Thanveer Jahan is a professor and head of the department CSE(AI & ML). She is presently working in Vaagdevi College of Engineering, Warangal, Telangana, India Her research area is Data mining, Machine Learning, Privacy and Security. She has published in SCI/ Scopus/ UGC journals and many publications in International Conference also. She has been awarded as women researcher, Outstanding Researcher Award, Best Faculty in Machine Learning, Best Researcher and Best Paper Presentation in 2020, 2021, 2022. She also published a patent titled "HIV Smart : A Health care system that monitors Peediatric Antiretroviral Therapy(ART) Drug Dosages" , Application Id: 201941010348 in 2019 and a patent titled "Health Analyzing Office Chair Seat" , Application Id: 346758-001 in 2021.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Acronyms and Abbreviated Terms

PPDM	Privacy Preserving Data Mining
SMC	Secure Multiparty Computation
SVD	Singular Value Decomposition
SSVD	Sparsified Singular Value Decomposition
GDP	Geometric Data Perturbation
ETS	Experiment Threshold Strategy
WWW	World Wide Web
METS	Modified Experiment Threshold Strategy
FMF	Fuzzy Membership Function
MMDM	Modified Matrix Decomposition Method
VD	Value Difference
ME	Misclassification Error
FDP	Fuzzy Data Perturb
ICA	Independent Component Analysis
SVM	Support Vector Machines
ID3	Iterative Dichotomizer 3
C4.5	Successor of ID3
BSTD	Bias of Standard Deviation
MF	Membership Function

CHAPTER 1

INTRODUCTION

Today's applications rely on large volumes of personal data being collected regularly. A number of data mining techniques are used to analyze them. During this process many unauthorized users try to access private data. Therefore it needs protection from unauthorized users with the help of privacy preserving data mining (PPDM) techniques. The PPDM techniques are used to mask or hide private information and preserve an individual's personal data. In the field of research, PPDM plays an important role during data publishing.

Data publishing is dependent on the environments. In privacy preserving data mining, the data is published either in a centralized or in a distributed environment. In the distributed environment, the data is located or distributed on various different sites. Secure multiparty computations are used in distributed environment. In this environment privacy preserving techniques will protect the individual's data by integrating the data from multiple sites. In centralized environment an authorized user owns the data and has central authority to protect this private data.

The PPDM techniques can be classified as Reconstruction methods, Heuristic methods and Cryptographic methods, which are Secure Multi-party Computations. Among them heuristic methods cannot be used for reconstructing the data. Therefore it fails in preserving privacy. The cryptographic methods concentrate on the distributed environments to securitize the multi-party computations. The reconstruction methods are used to build the original data from the transformed one. It also preserves privacy in a better way. The methods used in reconstruction comprise data perturbation methods, aggregation, swapping and randomization.

Data perturbation is one of the popular methods in PPDM. It refers to transformation of data. These methods include some constraints, such as, privacy protection metric, accuracy of mining results, computation and applicability. However, the challenging issues in data perturbation are balance of privacy protection and data utility.

These data perturbation methods can be classified into two approaches. The first approach is based on probability distribution. It replaces the data with another sample distribution or by the distribution itself. The second approach is value distortion approach.

In this approach the data values are perturbed directly by adding either additive noise or multiplicative noise. The advantage of random noise additive method is that the distribution reconstruction algorithm will recover the perturbed data and distribution. A Large volume of data can be easily perturbed in the value distortion approaches using additive noise and multiplicative noise. These include matrix decomposition methods, fuzzy-based approaches, hybrid transformations and geometric data perturbation methods.

The matrix decomposition method was proposed to satisfy the distortion of data using singular value decomposition (SVD) and Sparsified Singular Value Decomposition (SSVD). It works well in preserving privacy as well as maintaining the utility of datasets. The result concluded that feature selection should be performed prior to distortion of data. It will affect and compromise the accuracy of data published. The matrix decomposition methods suffer from loss of information due to data distortion. To avoid this loss of information in the original data and to distort only confidential information present in the dataset, a fuzzy logic-based perturbation was explored. The confidential numerical attributes in original dataset are transformed into fuzzy data using Fuzzy Membership Function (FMF).

The resultant clusters maintain privacy and at the same time preserve the original values. The drawback of the above method is high processing time which can be reduced by applying different fuzzy membership functions. In a way to extend the fuzzy logic-based data perturbation, different fuzzy membership functions were used to distort data, such as, Z-FMF, T-FMF, and Gaussian-FMF. From all those Z-FMFs data resulted in a low misclassification error rate. The results show a higher data utility during analysis of clusters. The fuzzy-based transformations minimized the loss of information in the original data. Various data perturbation methods as stated above have motivated us to investigate the problem that to develop an efficient Data Perturbation method has to be developed so as to preserve privacy in a centralized environment and allow publication of the data and avoid attacks or threats on it while mining the data.

The chapters in this book focus on the Two phases to arrive at a solution. The Two phases are illustrated as follows.

Phase I:

The Matrix Decomposition method considered in earlier research works to perturb the data resulted in a compromised accuracy. In order to improve this accuracy and develop an efficient data perturbation method, a Modified Matrix Decomposition method is proposed. The modification suggested here is to drop the features selected in the original dataset below a threshold. Then decompose the matrix in contrast to the existing matrix decomposition method. This proposal is clearly explained in chapter 3

The Modified Matrix Decomposition is implemented in Java with the help of utility packages such as JamaMatrix, Scanner and inputMismatchException. First the original data is read in the form of matrix using tobeMatriced() built-in method. The formed data matrix is further decomposed using methods, such as, svd(), getSingularValues() and rank(). The get() and set() method drops the data values to less than threshold to obtain final distorted data. The details of implementation are elaborated in chapter 3.

An exhaustive experimentation on final distorted data resulted better accuracy. But it is observed that there is a loss of information in the process. The experiments and results on this implementation are described in chapter 3.

Phase II:

To avoid this loss of information due to the modification proposed in the modified matrix decomposition, the Fuzzy Logic is considered in the process of perturbing the data. However, fuzzy logic requires large processing time due to its membership functions to create fuzzy datasets. It is considered to propose fuzzy membership function only to those identified confidential values of the original data and create fuzzy datasets, which are the perturbed data. This proposal is clearly explained in chapter 4.

This proposal is implemented in MatLab. It creates a MEX file and uses built-in functions, such as, load(), size(), min(), max() and avg(). The original data is loaded and further used smf(), zmf() and tmf() fuzzy membership functions to create fuzzy data. The details on implementation are presented in chapter 4.

The exhaustive experiments revealed that accuracy is further improved considerably with minimal loss of information. However this proposal suffered from a poor privacy. The experiments and results on this implementation are presented in chapter 4.

All the above proposed methods are considered univariate datasets when they are experimented with an objective to maximize the data privacy and efficiency. The above mentioned proposals are to be realized on a Data Analysis system is shown at Fig1.1.

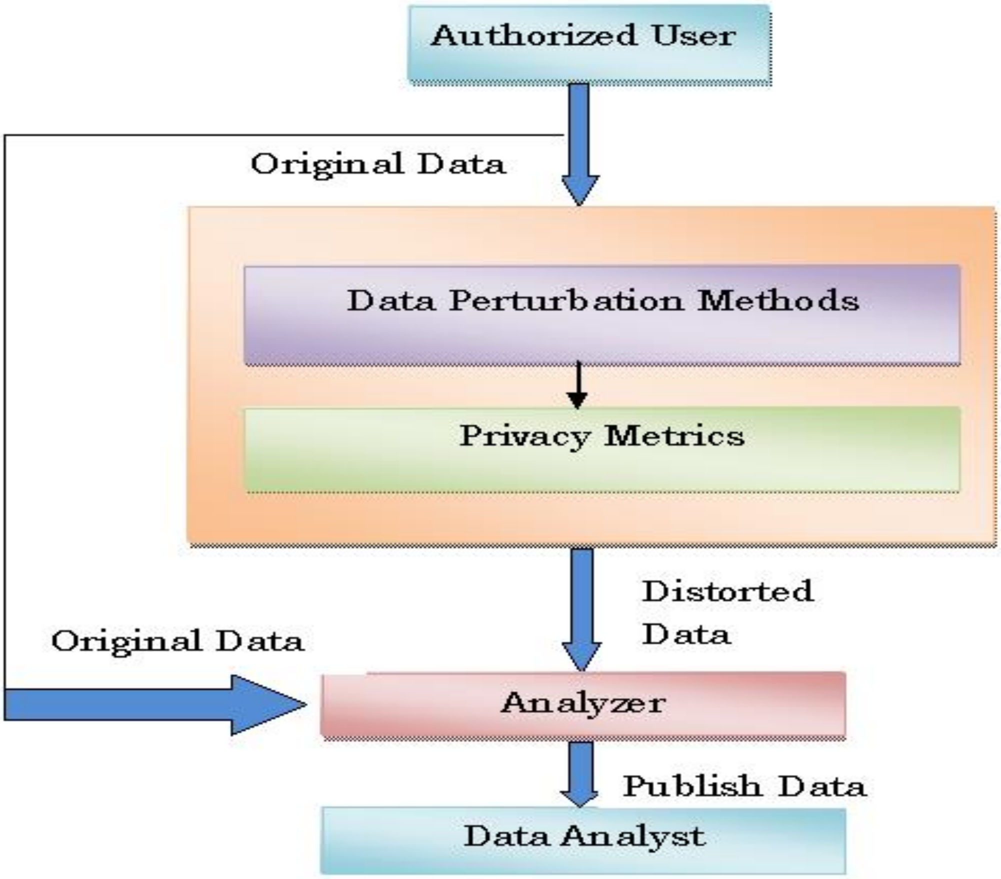


Figure 1.1: A Data Analysis System

The above Fig1.1 indicates the Data Analysis System. An authorized user possesses the original data. This original data is to be perturbed by the data perturbation method. This perturbed data is evaluated using different privacy metrics. The original and perturbed data are analyzed with mining utilities at Analyzer. The Analyzer also checks the possibility of threats or attacks. Thus the well-protected data is published to the Data Analyst.

An efficient data perturbation method on a multivariate dataset is successfully implemented. It is experimentally proved that it is efficient and the first of its kind. The system preserves privacy very effectively. Further, it can be explored for an unstructured multivariate datasets across the internet (WWW).

Chapter 2 presents a survey of the related research work carried out on the subject under investigation. The Modified Matrix Decomposition Method which is used to perturb the data, its implementation and the different experiments conducted on it together with the results and their analysis are discussed in **Chapter 3**. The **Chapter 4** proposes a Fuzzy Logic based Data Perturbation Method and its implementation, experimentation, results and analysis are also presented in this chapter. The experiments conducted on it together with results and analysis is furnished followed by **Bibliography**.

CHAPTER 2

SURVEY

Data mining is a well-established technique for extracting knowledge or information from large amount of data [1, 2]. Recent advances rely mostly on applications that constitute quick retrieval while protecting sensitive information of an individual. Therefore privacy preserving data mining is an upcoming research direction with in data mining [3, 4]. Recent developments in the area of data storage, data processing and rapid development in the internet has given rise to privacy preserving data mining as a new challenge to the research community [5]. In order to build secure systems for public use we should trim the private sensitive data of an individual while blocking all the inferring channels [6].

A number of effective methods for privacy preserving data mining have been proposed [15-100]. But most of these methods resulted either in information loss or privacy loss or both. In other words the secure systems suffered from side effects, such as, either the reduced data utility or downgraded efficiency of data mining [7,8]. Therefore the essential problem under context is a tradeoff between the data utility and disclosure risk. This chapter analyzes the representative methods for privacy preserving data mining from among existing techniques and points out their merits and demerits.

Privacy preserving data mining [9, 10, 11] techniques are categorized into three such as, Heuristic, Secure Multiparty Computation and Reconstruction methods.

Heuristic methods are also called as Anonymization methods [12]. In these methods personal data is integrated and protected by removing key identifiers such as, id, name and age from individual records. However the combination of other record attributes (quasi-identifiers) can be used to exactly identify individual records. For example: attributes such as birth, gender, pin code are available in public records of a voters list. When these attributes are available in a given dataset such as medical data they can be used to infer the identity of the corresponding individual with high probability by linking operation as shown at Fig 2.1.

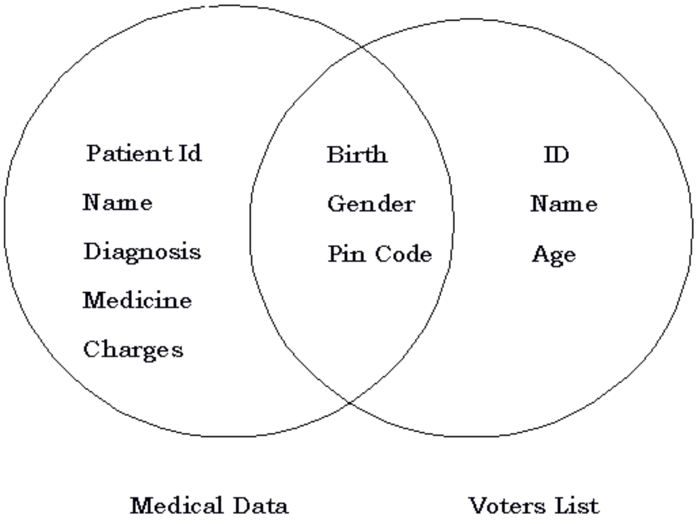


Figure 2.1: A Sample Record Linkage

The privacy in the above data is preserved using k- anonymity. An individual is indistinguishable from at least k-1 other ones in the anonymized dataset. For example: Table 2.1 is anonymous to Table 2.2 as show below.

Table 2.1: Original Data

Name	Birth	Gender	Pin Code	Disease
Ram	12/3/1987	M	6001	Flu
Sham	15/7/1999	M	6002	Cancer
Gita	4/9/1988	F	6009	Fever

Table 2.2: Anonymous Data

Birth	Gender	Pin Code
1987	M	600*
1999	M	600*
1988	F	600*

This k-anonymity is achieved using generalization and suppression. Generalization replaces a value with a less specific but semantically consistent one. For example date of birth is generalized to a range such as year, so as to reduce the risk of identification.

Suppression does not release a value at all. The Anonymization methods [13,14] reduce the risk identification with the use of public records while reducing the accuracy of application of the transformed data. An optimal k-Anonymization [15] for a given dataset proposes to perturb the input dataset as little as to achieve k-anonymity and it is quantified by a given cost metric.

A full domain generalization [16] map the entire domain of each quasi identifier attributes in original data to a more general in its domain general hierarchy. A Top Down Specialization (TDS)[17] is used to generalize table to satisfy the anonymity requirement while preserving privacy using classification. TDS generalizes the table by specializing the information in a top down manner iteratively starting from the most general state into a specific state

A theoretical algorithm using generalization and specialization named as Minimal Generalization (MinGen) [18] used to provide k-anonymity with minimal distortion. The algorithm MinGen comprises of two steps. In the first step, the original data should satisfy the k-anonymity requirement. These attacks do not guarantee privacy. To solve the severe privacy problems a novel and powerful definition is proposed namely, *l*-diversity [19]. The disadvantage of heuristics methods are based on generalization and suppression technique which suffer from high information loss and low usability. A new kind of algorithms using clustering technique was proposed [20-24] which have reduced information loss to some extent. These methods are also prone to attacks such as background knowledge and homogeneity attacks. The major weakness of these methods is that they cannot reconstruct the data and hence fail in preserving privacy.

The upcoming distributed data mining [25] opportunities, where people conduct jointly, mining tasks based on the private inputs they supply. These mining tasks could occur between mutual untrusted parties or even between competitors, therefore protecting privacy becomes a primary concern in distributed data mining.

Both the methods specified above are based on the special encryption protocol known as Secure Multiparty Computation (SMC) technology [26]. The SMC defines two basic adversarial models: Semi-Honest model and Malicious Model [27]. The Semi-Honest adversaries follow the protocol faithfully but try to infer private data of other parties during execution; malicious adversaries may do anything to infer secret information. They abort the protocol at any time, send spurious messages, spoof messages and collude with others.

Two kinds of secure sub protocols are d discussed here on horizontally partitioned and vertically partitioned data setting [28]. Secure sum: This protocol securely calculates the sum of values obtained from multiple sites. Let each site i has a value v and all sites want to compute securely compute $S=v_1+v_2+\dots\dots v_n$, where v_i is known in the range [0…m]. For example, in horizontally partitioned association rule mining [29] setting a user can securely calculate global support count of an itemset using secure sum protocol. The above methods can ensure that the transformed data is exact and secure with a lower efficiency. However, it is important to develop efficient mining protocols that remain secure and private even if some of the parties involved behaved maliciously [30-32].

In order, to efficient protocol for a specific problem of Secure Multiparty Computation Lindel.et.al [33] considered distributed ID3 for the task of decision tree learning. The protocol considered the computation seen by the players obtain random shares v_1 & v_2, such that, their sum equals an appropriate intermediate values. Efficiency is achieved by having the parties do most of the computation independently.

Kantarcioglu and Clifton [34] addresses secure computation of association rules over horizontally partitioned data. The association rules are computed without disclosing individual transactions is straight forward. We can compute global support and confidence of an association rule AB => C, knowing only local supports of AB and ABC and the size of database.

$$\text{Support}_{AB \Rightarrow C} = \frac{\sum_{i=1}^{sites} \text{support\_countABC(i)}}{\sum_{i=1}^{sites} \text{database\_size(i)}}$$

$$\text{Support}_{AB} = \frac{\sum_{i=1}^{sites} \text{support\_countAB(i)}}{\sum_{i=1}^{sites} \text{database\_size(i)}}$$

$$\text{Confidence}_{AB \Rightarrow C} = \frac{\text{SupportAB} \Rightarrow C}{\text{Support AB}}$$

A-priori algorithm (Association rule mining) is extended to the distributed case using the following lemma: if a rule has support>k%, it must have support>k% on at least one of the individual's sites. The method of computing secured association rules consists two phases. The two phases are discovering candidate itemsets and determine which candidate itemsets meets global support/confidence thresholds. The first phase uses encryption where each site encrypts frequent itemsets and then passed to other parties, until all sites encrypt their itemsets. Then, these itemsets are passed to a common party to eliminate duplicates and to begin decryption, where each site decrypts each itemset and passes it to other sites. The final result comprises of common itemsets. In the second phase, each of the locally supported itemsets is tested to check if it is supported globally then each computes their local support. Each site chooses a random value R and adds to R the amount by which its support for ABC exceeds the minimum support threshold. This value is passed to other site (adds excess support). The resulting value is tested using secure comparison to check if it exceeds the Random value. If so itemset is supported globally. Secure cryptographic based methods are used to minimize information shared while adding little overhead to the mining task.

Generating association rules on vertically partitioned data is proposed [35], where the transactions are distributed across sources. It considers a vertical partitioned data between two parties A, B and generates Boolean association rules. The absence or presence of an attribute is represented as 0 or 1. Transactions are strings of 0 or1. To find the particular itemset is frequent, then count the number of records where values for all attributes in the itemset is 1. Let X and Y be the columns of the database, x_i =1 if row i has value 1 for attribute

X. The dot product of two cardinality n vectors X and Y is defined as in Equation 2.1:

$$X.Y=\sum_{i=1}^{n} xi * yi \qquad \text{Eq 2.1}$$

Determine if the two-itemsets <XY> are frequent thus reduces to test if <X.Y> >=k. This dot product protocol then securely computes without either side disclosing its vector. It showed good individual security with communication cost comparable to that required to build a centralized data warehouse. It is limited to only Boolean association rule mining. Non-categorical attributes and quantitative association rules are significantly more complex.

Du and Zhan [36] proposed to build a decision tree classifier on a private data over vertically partitioned data. Masooda Modak et.al [37] proposed a secure association rule over horizontally and vertically distributed data. Let the number of sites >=3, then each site has a private transaction database (DB). The goal is to discover the association rules that satisfy the given support and confidence, such that disclosure is limited and no site should able to learn contents of a transaction at any other site

The methods in cryptography-based techniques [38-40] are based on computation and communication costs. The communication cost requires a protocol based on the number of messages shared between multiple sites [41]. The computation cost depends on the algorithm to perform encryption and decryption methods. The use of cryptographic techniques does not assure data privacy at each site. Thus, the data quality at each site is compromised.

The reconstruction method is used to reconstruct the transformed data into the original data. It preserves privacy in a better way by satisfying the two criteria, namely, the distorted data must not be discovered by the opponent (attacker), and the statistical properties of the distorted data are to be maintained like original data [42]. The methods in reconstruction are Data Perturbation, Aggregation and Swapping. The methods of aggregation and swapping [43] do not support large data, whereas data perturbation is efficient for large data [44]. It shows most challenging issues in preserving privacy [45, 46].

Data Perturbation has a major role among all the methods [47]. It efficiently transforms data and recovers knowledgeable pattern from the transformed one. It is a study used in statistical databases community. The approaches in data perturbation are categorized as value distortion approach and probability distribution approach [48] as shown below in Fig 2.2.

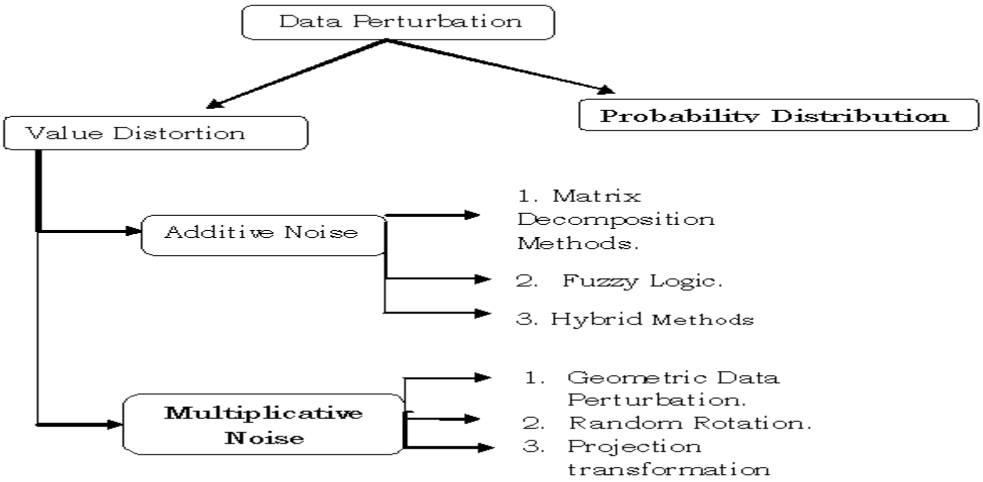


Figure 2.2: Classification of Data Perturbation Methods

The first approach is based on probability distribution. The probability distribution approach replaces the data with another sample distribution or by the distribution itself [49]. Whereas, the value distortion [50] the data values are transformed or changed. These transformations can be done by adding either Additive noise or Multiplicative noise.

Kargupta et. al. [51] proposed a random value distortion to mask the data for preserving privacy. The method modifies the data values using Value distortion approach. A Randomization method is developed to mine association rules from a categorical data to preserve privacy [52]. A new formula is incorporated to support and variance prediction to a data mining algorithm. The approach is restricted with class of privacy breaches for a given level of privacy. A Randomized Response scheme [53] is proposed based on the statistical community to collect sensitive information from the individuals in such a way that the survey interviewers and the ones who process the data do not know which of two alternative questions the respondent have answered. The question of which specific technique will prove superior is a matter of investigation.

Huang et.al [54] proposed two data reconstruction methods based on data correlations. One method use Principal Component Analysis (PCA) to control redundancy. In this scheme a data reconstruction problem is an estimated problem, given a disguised data Y, X is found such that the posterior probability P(X/Y) is maximized. Then, another method, Bayes estimate technique is used to solve X as the final reconstructed data. The method shows accurate results compared to other data reconstruction methods that exploit data correlation. The drawback is that the partial knowledge of a disguised dataset can compromise privacy.

Rizvi and Haritsa [55] proposed Mining Associations with Secrecy Konstraints (MASK) to mine secret constraints using association rules. It is based on probabilistic perturbation of data. The perturbed data is supplied to the data miner along with the description of distortion procedure. The method is further enhanced with generalization and quantitative association rules. Privacy metric is used to evaluate the distortion approach. The methods are simple and useful to hide individual data in privacy preserving data mining.

Perturbation based PPDM [56] are proposed to perturb individual values to preserve privacy before data is published. In this survey different approaches are discussed for data publishing scenarios along with the challenges. The typical concerns include the degradation of data/service quality, loss of valuable information, increased costs and increased complexity.

Data is perturbed by applying adding noise, data transpose matrix, decomposition methods. A matrix decomposition method satisfies the distortion data using a Singular Value Decomposition (SVD) [57, 58]. The threshold ε property is used to drop the elements in U_k and V_k. A Sparsified SVD [59] is represented as $D" = U'_k S_k V'^T_k$. It works well in preserving privacy as well as maintaining the utility of datasets. The data distortion and SVD computation are considered to be data preprocessing and preparation procedure. A reasonable amount of computational cost in this phase may be tolerable because of the limited number of attributes utilized in this method.

A structural partition and Sparsified Singular value decomposition [60] is proposed. The proposed strategies perform a dimension or rank reduction and conduct sparsification operation on selected parts of the original dataset. Three different matrixes structural partition strategies, such as object- based partition, feature based partition and hybrid partition are used to partition the original data into several sub matrices. Sparsified SVD (SSVD) is then applied on the selected submatrix to perturb the partial information in the subset. The distorted submatrix is then combined with the original unperturbed part of the matrix to form new perturbed dataset. Data mining utilities are used on the perturbed data matrix. The performance of the feature based partition is feasible and efficient for privacy preserving. To reduce the computational costs of sparsified SVD (SSVD) method, substantially, alternative methods are explored.

For this purpose a CLUST-SVD [61] was proposed. In this method the original data is quantified an amount of information is preserved after perturbation. It is mainly focused to protect the individual's private data and providing accurate data for clustering analysis. It distorts only sensitive continuous attributes using SVD method to ensure privacy requirements while preserving general features for k-means clustering analysis. This technique desires complete accuracy for clustering analysis and complete privacy to perform well. To enhance performance of SSVD, filter-based feature selection is used for data distortion reduced feature space [62]. Data is perturbed after selecting features as the discarded data has less perturbation values. Then the data is perturbed using a modified Experiment Threshold strategy (ETS) for a matrix sparsification called METS and takes negative carry values into consideration. The result concluded that feature selection should be performed before distortion of data and compromised accuracy. The perturbed data published can have little effects on correct predict rate, but can significantly result with better feature selections [63, 64]. The data perturbation using matrix decomposition methods ensures privacy but has a disadvantage of dimensionality reduction. The sparsification strategies used in decomposition methods suffer from loss of information [65].

B. Karthikeyan et.al, to overcome the above drawback and to maintain similar dimensions on distorted datasets, proposed fuzzy based perturbation [66]. Fuzzy logic is applied to protect individual private data using S-based fuzzy membership function. The, k-means clustering algorithm is applied on transformed data. It maintains privacy relatively between original and transformed data. The nature of the Fuzzy Membership Function (FMF) used affects the processing time of the algorithm. Further, it can be improve using different fuzzy membership functions.

V. Vallikumari et.al [67] proposed a holistic method to achieve maximum privacy with no loss of information and minimum overheads. M. Nagalakshmi et.al proposed [68] a fuzzy based data transformation such as Z-based FMF, Triangular FMF and Gaussian FMF. The approach have satisfied privacy requirements as well as retained clustering quality. In second case a hybrid is proposed with a combination of fuzzy data transformation and Random Rotation (RRP).

Mary et.al proposed [69] extension of the fuzzy logic-based perturbation, such as T-FMF and Gaussian FMF using random data that is added or multiplied with the data and obtained a random modified data.

Manikandan et al proposed [70] a survey on fuzzy membership functions and their efficiency to preserve private data. Fuzzy logic can be used to preserve privacy using perturbation. The membership functions used are T-FMF, Bell-shaped FMF, Gaussian FMF, S-based FMF. Among all these S-FMF is an efficient procedure that can be used to obtain sanitized data.

Hybrid transformations are used to maintain statistical properties of data as well as mining utilities [71-73]. The statistical properties of data are mean and variance or standard deviation without any loss of data.

A feasible solution [74] is provided to optimize the data transformations by maximizing privacy of sensitive attributes. A combined technique using randomization and geometric transformation is used to protect sensitive data. A randomized technique is represented as D = X +R, where R is additive noise, X is original data and D is perturbed data. A geometric transformation is used as a 2D rotation data matrix represented as D’= R(θ)×D, where D is the column vector containing original co-ordinates and D’ is a column vector whose co-ordinates are rotated clockwise. The above method considered only single attributes as sensitive and rest of them as non sensitive attributes.

Data perturbation method using fuzzy logic and random rotation is proposed [75].

CHAPTER 3

DATA PERTURBATION METHOD USING MODIFIED MATRIX DECOMPOSITION

3.1 Introduction

This chapter proposes a data perturbation method by modifying matrix decomposition method. Since the existing matrix decomposition method perturbs the data, the present study modifies the feature selection and decomposes the matrix of select features. This proposal is illustrated in the following section.

3.2 Proposed Modified Matrix Decomposition Method (MMDM)

The Modified Matrix Decomposition Method is proposed as shown in the block diagram at Fig 3.1 given below.

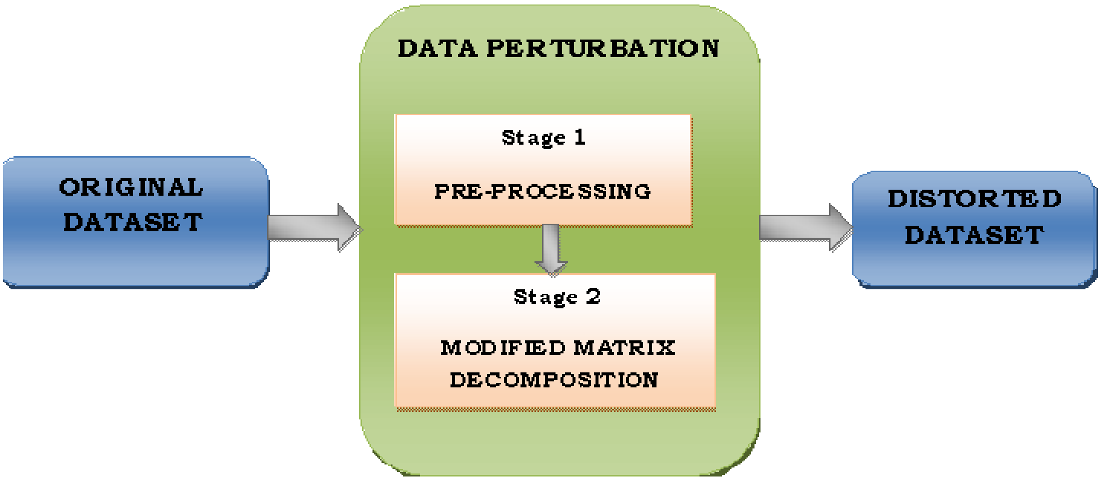


Figure 3.1: Block Diagram for Modified Matrix Decomposition Method

The above block diagram considers the original dataset and deals with it in two stages. In first stage, it is proposed to preprocess the original dataset to extract the best features as shown at Fig 3.2. In the next stage this featured dataset is decomposed using the proposed MMDM to perturb the data.

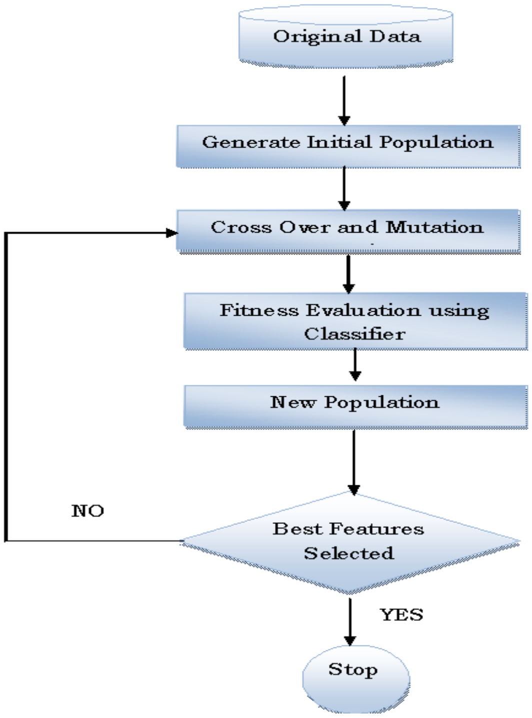


Figure 3.2: Flow Chart for Feature Extraction

The above Fig 3.2 demonstrates the flow for extraction of the best features from the original dataset. It is done in a preprocess, with the help of data mining tool, which employs a genetic algorithm and allows preprocessing of the original dataset. First, an original dataset is considered and an initial population is generated comprising parent strings. From this generated population good parent strings are selected and made to perform cross over as well as mutation operator. The cross over will combine a parent string with an offspring. In these combinations the mutation will result in a better string. The resulting population is evaluated using a classifier for its fitness. Accordingly, a new population is generated. Check whether the best features are selected. If the best features are not selected, repeat the cross over and mutation operation step. Otherwise, preprocess is complete. Then, the second stage is a proposal for modified matrix decomposition algorithm. This is explained in the next section.

3.3 Proposed Modified Matrix Decomposition Algorithm

A proposal for modifying matrix decomposition algorithm is provided in this section. The pseudo code of the proposed algorithm is listed below.

Algorithm Modified Matrix Decomposition Method (MMDM)

```
Input: A Data Matrix $D_{p\times q}$.
Output: A Distorted Data Matrix D1 & D2.
Begin
Step 1:  Decompose $D_{p\times q}$ using Singular Value Decomposition Method
         into three data matrices U, S, $V^T$ where
         U is p×p orthonormal matrix of $D_{p\times q}$.
```

```
            V is p×q diagonal matrix of $D_{p\times q}$.
            $V^T$ is q×q orthonormal matrix of $D_{p\times q}$.
Step 2: Compute rank r, of $D_{p\times q}$. where r is min (p, q).
Step 3: Reconstructing Data matrices U, S, $V^T$ as $U_r$, $S_r$, $V_r^T$ depending
        upon on rank r where
        $U_r$ consists of first r columns of U.
        $S_r$ consists of first r non-zero diagonal values of S.
        $V_r^T$ consists of first r rows of $V^T$.
Step 4: Construct the distorted data matrix D1= $U_r$ $S_r$ $V_r^T$.
Step 5: Select Threshold € range of values and use.
Step 6: In the Data matrix $U_r$
            For i=1 to p do
            For j=1 to r do
               If |u [i, j]|<€   then
                 u [i, j] = 0  //dropping the values in $U_r$
                 Construct U1   //After dropped elements in $U_r$
               End If
            End For
            End For
Step 7:  In the Data matrix $V_r^T$
            For i=1 to r do
            For j=1 to q do
               If |v [i, j]|< € then
                 v[i, j] =0   //dropping the values in $V_r^T$
                 Construct V1 //After dropped elements in $V_r^T$
              End if
           End For
           End For
Step 8: Construct a distorted data matrix D2= U1 $S_r$ $V1^T$.
End
```

The algorithm accepts data matrix D as input with p rows and q columns. The data matrix $D_{p\times q}$ is decomposed into three data matrices using singular value decomposition method. The decomposed data matrices are U, S, V^T, where U is a p×p orthonormal matrix, S is a diagonal p×q and V^T is a q×q orthonormal matrix. The Rank r for the data matrix $D_{p\times q}$ is calculated as the minimum of p and q. Depending upon the rank r, the data matrices U, S, V^T are reconstructed as U_r, S_r, V_r^T. The data matrix U_r consists of first r columns of U, S_r comprises first r non-zero diagonal values of S, V_r^T comprises first r rows of V data matrix.

Construct the distorted data matrix as D1= U_r, S_r, V_r^T. Select threshold € from a given range of values and drop the values in the data matrices U_r based on the threshold value and construct U1. Similarly, drop the values in V_r^T and construct $V1^T$. The output is the distorted data matrix D2 constructed using U1 S_r $V1^T$. The time complexity of the proposed MMDM algorithm is found to be O(n), where n is the dimension of the dataset.

Example 3.1: Consider a data matrix $D_{3\times 2}$ ($D_{p\times q}$ where p = 3 and q = 2) as below.

$$D_{3\times 2} = \begin{pmatrix} 1 & 10 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}_{3\times 2}$$

After Step 1, the $D_{3\times 2}$ is decomposed using a singular value decomposition method that is available, and three data matrices U, S & V^T are formed as a result. These decomposed data matrices are listed below.

$$U = \begin{bmatrix} 0.7018 & -0.4173 & -0.5774 \\ 0.0104 & 0.1864 & -0.5774 \\ 0.7123 & 0.3992 & 0.5774 \end{bmatrix} \qquad S = \begin{bmatrix} 14.3015 & 0 \\ 0 & 1.2111 \\ 0 & 0 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 0.1494 & 0.9888 \\ 0.9888 & -0.1494 \end{bmatrix}$$

At step 2, select a rank r as min (p, q) = 2.

At step3, reconstruct the data matrices U, S, V as U_r, S_r, V_r^T depending upon rank r listed as:

$$U_r = \begin{bmatrix} 0.7018 & -0.4173 \\ 0.0104 & 0.8164 \\ 0.7123 & 0.3992 \end{bmatrix}$$ (First 2 columns of data matrix U)

$$S_r = \begin{bmatrix} 14.3015 & 0 \\ 0 & 1.2111 \end{bmatrix}$$ (First 2 non-zero diagonal elements of data matrix S)

$$V^T = \begin{bmatrix} 0.1494 & 0.9888 \\ 0.9888 & -0.1494 \end{bmatrix}$$ (First 2 rows of data matrix V^T)

In Step 4, Construct distorted data matrix D1 = U_r S_r V_r^T is given below: $D1 = \begin{bmatrix} 0.9998 & 9.999 \\ 0.9999 & -0.0006 \\ 2.000 & 10.0006 \end{bmatrix}$

At Step 5 select the threshold € in the range of values and considered threshold as 0.001.

In Step 6 & 7, the data values less than threshold in the data matrices U_r and V^T_r are set to zero and constructed data matrices U1 & $V1^T$ is given below:

$$U1 = \begin{bmatrix} 0.7018 & 0 \\ 0.0104 & 0.8164 \\ 0.7123 & 0.3992 \end{bmatrix} \qquad V1^T = \begin{bmatrix} 0.1494 & 0.9888 \\ 0.9888 & 0 \end{bmatrix}$$

At Step 8 construct the resultant distorted data matrix D2 = U1 S_r $V1^T$ given as: $D2 = \begin{bmatrix} 1.4995 & 9.9244 \\ 0.9999 & 0.1471 \\ 2.000 & 10.0729 \end{bmatrix}$

This implementation detail of this proposal is explained in the next section.

3.4 Implementation

The proposed algorithm that was discussed in the previous section is implemented using Java. Its source code is included at APPENDIX I (Page No.130). The details of implementation are furnished in this section.

The implementation imports java packages and utility packages, such as, JamaMatrix, SingularValueDecomposition, Scanner and inputMismatchException. The source code first utilizes a built-in method called Scanner() to read the data from a specified file. Then, tobeMatriced() method from JamaMatrix package is employed to arrange those data values into data matrix $D_{p\times q}$. A svd() method of SingularValueDecomposition utility package is used to decompose the data matrix $D_{p\times q}$ into three data matrices U, S, V^T. Now, the source code with the help of a built-in method called rank() is utilized to compute the rank of the data matrix $D_{p\times q.}$ Next, a getSingularValues() method decomposes the three data matrices, based on the value of rank, into U_r S_r V_r^T. Then construct the distorted data matrix D1 using U_r S_r V_r^T.

At this juncture a display function is used in the source code which facilities an assignment of threshold ε. In order to drop the data values in U_r, V_r^T which are less than a specified threshold, a get() method is employed. The updation of U1 and $V1^T$ data matrices is carried out using a set() method. Finally, construct distorted data matrix D2 with U1 S_r $V1^T$ data matrices.

3.5 Experimentation

The Experimentation was conducted using a desktop computer system loaded with windows XP Operating System, Java development kit 1.5 and Tanagra data mining tool. The experimentation details are elaborated in this section.

The experimentation begins with preprocessing a dataset to extract the best features using data mining tool. These best features form an original dataset D. This original dataset D is given as input to the proposed MMDM algorithm to obtain distorted datasets D1, D2. Then, the original dataset D and distorted datasets D1, D2 are uploaded into Tanagra data mining tool after appending a class attribute. These uploaded datasets are classified using classification utility available within Tanagra data mining tool. The results of classification are analyzed thereafter. Similarly, the datasets are clustered using clustering utility available in it. The results of clustering are also analyzed. Now, privacy metrics are calculated on those datasets. These privacy metrics are discussed in this section. But, their calculation is shown at section 3.6 under Results and Analysis. The details of dataset employed in this experimentation are furnished below.

A Real Time univariate dataset namely, Terrorist is downloaded from the website of DARPA. The details are shown at Table 3.1.
Therefore, the original dataset used in the experimentation is a terrorist dataset. It comprises of 1000 rows/tuples and 42 columns/attributes including one target/class attribute.

Table 3.1: Details of Terrorist Dataset

Dataset	Size	Description
Terrorist	1000 rows & 42 column	It consists of information of terrorists and their details

The dataset shown at Table 3.1 is considered for preprocessing. Initially a population size of 41 attributes (other than class attributes) is considered for the genetic algorithm used to preprocess. A cross over rate of 0.4 and mutation rate of 0.02 are obtained. It employs a fitness evaluator to generate a new population. The extracted features in this new population are checked. If the selected best features are the required confidential attributes, then, stop. Otherwise, repeat the process from cross over and mutation. A sample list of required confidential attributes is given below at Table 3.2, together with a sample original dataset D with 5 rows and 10 attributes.

Table 3.2: A Terrorist Original Dataset D

ID	Age	Nationality	Siblings	Pilot Training	Temporary Address	Wedding Address	Time	Date	Place
1	3	1	0	0	0	0	10	0	10
2	3	1	0	0	0	0	10	0	10
2	3	0	0	0	0	0	0	0	8
4	3	0	0	0	0	0	7	0	10
4	3	0	0	0	0	1	0	1	8

The process in the experiment is explained as follows:

First, a sample original dataset is read into D matrix with the help of tobeMatrixed() method. This method scans row by row all the elements from a text file named Terrorist.txt. The singular value decomposition (svd) method is applied on the data matrix D and assigned to s. A get() method is used on s to decompose U matrix from D. Similarly matrices S and V are also decomposed from D using the same get() method on s. The rank() method is applied on s and obtained rank of this singular value decomposition s as r. Based on this rank r as 25 the getSingularValues() method obtains the singular values as svalues. These svalues form a D1 distorted data matrix. The threshold value is selected from a specified range of values (0.001 to 0.9).

The selected threshold value is read and assigned to a variable thr as 0.001. The threshold thr value is used to form data matrices U1 and V1 from D1 distorted data matrix. First data matrix D1 is decomposed using singular value decomposition svd() method and assigned to s1. A get() method is used on s1 to decompose U1 and V1 data matrices. Now the decomposed data matrices U1, V1 are searched for values less than threshold thr using get() method. All the values which are less than threshold values are set to zero using set() method. The updated data matrices U1 and V1 are utilized to construct a D2 distorted data matrix with the help of U1, S and $V1^T$ by multiplying all of them.

When the above process is executed in experimentation, it outputs a distorted dataset D1 and then it finally constructs another distorted dataset D2. They are furnished in section 3.6 under Results and Analysis section. Each of these distorted datasets D1 & D2 together with the original dataset D, respectively are appended to a class attribute (CA), such as, YES or NO as shown at Table 3.2. However, it is reproduced at table 3.3.

Table 3.3: A Terrorist Original Dataset with Class Attribute

ID	Age	Nationality	Sibling	Pilot Training	Temporary Address	Wedding Address	Time	Date	Place	CA
1	3	1	0	0	0	0	10	0	10	Yes
2	3	1	0	0	0	0	10	0	10	No
2	3	0	0	0	0	0	0	0	8	Yes
4	3	0	0	0	0	0	7	0	10	Yes
4	3	0	0	0	0	1	0	1	8	No

Similarly the distorted datasets D1, D2, are also appended to a class attribute and furnished in section 3.6 as part of Results and Analysis. The above mentioned datasets D, D1, D2, are uploaded into Tanagra data mining tool. First classification utility is used on the dataset D and distorted datasets D1, D2. It divides the attributes into two categories namely non-class attributes and class attribute. These two categories can be two inputs to the classifier chosen among available ones. Suppose we select SVM (Support Vector Machines) as classifier, then, it classifies the datasets D, D1, D2, based on class attribute into either a terrorist or non terrorist datasets. Such results are furnished at section 3.6 under Results and Analysis.

Similarly, the experimentation is repeated with Iterative Dichotomizer 3 (ID3), and C4.5 classifiers. The results of those experiments are furnished at section 3.6. A Clustering utility available in Tanagra data mining tool is used to cluster the original dataset D and distorted datasets D1and D2. Non-class attributes are considered and given as input to k-mean clustering method. As a result, two categories of clusters are formed, namely terrorist clusters and non terrorist clusters. The results of clustering are furnished at section 3.6.

In order to ensure privacy certain privacy metrics considered are Value Difference (VD) and Position difference. These privacy metrics are discussed below:

Value difference (VD): The value difference (VD) in the datasets is the ratio of the difference between original dataset D and distorted dataset |D’| to Frobenius norm (F) as given below at Eq 3.1.

$$VD = \frac{\left| |D| - |D'| \right| F}{|D| F} \qquad \text{Eq 3.1}$$

Position difference: The position difference is the privacy metric based on four parameters, such as RP, RK,CP and CK.

RP denotes attribute order change given at Eq 3.2 stated below:

$$RP = \left(\left(\sum_{i=1}^{p} \sum_{j=1}^{q} |\mathrm{Ord\,ij} - \mathrm{ord'ij}| \right) \Big| (p * q) \right) \qquad \text{Eq 3.2}$$

RK denotes percentage attribute order change given at Equation 3.2

$$RK = \left(\sum_{i=1}^{p} \sum_{j=1}^{q} \mathrm{RK\,ij} \right) / (p * q) \qquad \text{Eq 3.3}$$

The above equation 3.3 can be abstracted as:

$$RK^{i}_{j} = \begin{cases} 1, & \text{if} \quad Ord^{i}_{j} = Ord'^{i}_{j} \\ 0, & \text{otherwise} \end{cases}$$

Where p is the data objects and q is the attributes in the dataset. Ord_{ij} is the attribute order change in original dataset D, while Ord'_{ij} is the attribute order change in distorted dataset.

CP denotes the data objects change for each attribute given at Eq 3.4.

$$CP = (\textstyle\sum_{i=1}^{p} |\,ordDvi - ord'Dvi|)|p \qquad \text{Eq 3.4}$$

The above equation Eq 3.4 can be abstracted as below:

$$CK = \begin{cases} 1, \text{ if } OrdDV_{i} = OrdDV'_{i} \ , \\ 0, \text{ otherwise} \end{cases}$$

The calc [illegible] ırnished at the section 3.6 under Results and Analysis.

3.6 Results and Analysis

The results obtained in the above experiment are presented in this section. The original dataset D furnished at Table 3.2 (Page No: 46) is given as input to the proposed MMDM and output, a distorted dataset D1, is presented below at Table 3.4.

Table 3.4: A Terrorist Distorted Dataset D1

ID	Age	Nationa-lity	Siblings	Pilot Training	Temp Address	Wedding Address	Time	Date	Place
1	3	1	-1.5E-4	0	0	8.60E-16	10	8.60E-1	10
2	3	1	-7.39E-	0	0	4.16E-16	10	4.16E-1	10
2	3	4.00E-15	2.90E-4	0	0	-1.67E-15	-4.4E	-1.67E-	8
4	3	4.44E-15	3.28E-4	0	0	-1.78E-15	7	-1.78E-	10
4	3	-5.74E-1	-1.83E-	0	0	1	-3.8E	1	8

Similarly, a distorted dataset D2 is presented below at Table 3.5.

Table 3.5: A Terrorist Distorted Dataset D2

ID	Age	Nationality	Siblings	Pilot Training	Temp Address	Wedding Address	Time	Date	Place
1.8	0.79	0.152523	1.4654	0.09	7.807	8.641117	1.586	0.03284	0.05
2.7	0.02	-0.00861	2.8131	0.69	6.883	6.910741	1.166	0.05605	0.03
2.9	0.74	0.061803	1.4680	0.00	5.193	5.45031	0.972	0.10272	0.03
4.2	0.39	0	0.8734	0	3.271	3.73363	0.680	0.00649	0.00
4.8	0.37	0.0766	-1.412	-0.43	4.565	5.050918	0.839	0.05608	-0.4

When SVM classifiers are used on D, D1, D2 the following observations are made and they are represented at Table 3.6.

Table 3.6: A Terrorist Dataset Classified using SVM

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Terrorist(YES)	Number of Support Vectors	Error Rate	Computation Time (ms)
Original (D)	1000	728	647	0.300	2359 ms
Distorted (D1)	1000	710	647	0.310	2437 ms
Distorted(D2)	1000	731	660	0.300	2331 ms

In the above Table 3.6, the first column presents the original dataset D and distorted datasets D1 and D2. The number of tuples in the datasets considered for experimenting is seen in the second column. The third column displays the number of training tuples belonging to terrorist category as YES. The number of support vectors available is furnished in the fourth column. The Fifth column reveals the error rate of SVM classifier. The computation time is tabulated in the last column. Similar results are tabulated at Table 3.7 and 3.8, when the ID3 and C4.5 classifiers are used on D, D1 and D2.

Table 3.7: A Terrorist Dataset Classified using ID3

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Terrorist (YES)	Tree having Number of Nodes and Leaves	Error Rate	Computation Time(ms)
Original (D)	1000	729	1 node,1 leaf	0.300	4355 ms
Distorted(D1)	1000	712	1 node, 1 leaf	0.300	4345 ms
Distorted(D2)	1000	752	1 node, 1 leaf	0.300	4321 ms

Table 3.8: A Terrorist Dataset Classified using C4.5

Dataset	Number of Tuples	Number of Training Tuples Classified as Terrorist (YES)	Tree having Number of Nodes and Leaves	Error Rate	Computation time (ms)
Original (D)	1000	739	23nodes, 12 leaves	0.292	47 ms
Distorted (D1)	1000	742	23 nodes, 12 leaves	0.299	31 ms
Distorted (D2)	1000	769	41 nodes, 21 leaves	0.291	30 ms

In the above Tables 3.7 and 3.8, the first column presents the dataset D and distorted datasets D1 and D2. The numbers of tuples in the datasets considered for experimenting are seen in the second column. The third column displays the number of training tuples belonging to terrorist category as YES. A tree having the number of nodes and leaves is furnished in the fourth column. The fifth column reveals the error rate of ID3 and C4.5 classifier respectively. The computation time is tabulated in the last column.

Based on the results presented above, the accuracy of classification of datasets is presented at Table 3.9. The accuracy is the percentage of tuples that are correctly classified by a classifier and is given by the following equation 3.5.

$$\text{Accuracy} = \frac{\text{Number of Training Tuples Classified}}{\text{Total Number of Tuples}} \qquad \text{Eq 3.5}$$

Table 3.9: Accuracy of Classifiers

Dataset	Modified Matrix Decomposition Method (MMDM)				Existing Matrix Decomposition Methods			
	TERRORIST				WBC		SONAR	
	Features	SVM	ID3	C4.5	Features	SVM	Features	SVM
ORIGINAL (D)	10	72.8	72.9	73.9	32	97.1	60	75.9
DISTORTED (D1)	10	71.0	71.2	74.2	12	92.2	19	76.4
DISTORTED (D2)	10	73.1	75.2	76.9	12	90	13	75.0

The above Table 3.9 presents the accuracy of the classifiers. The first column presents the dataset D and distorted datasets D1, D2. The second column presents the accuracy of classification obtained on the proposed MMDM. The third column presents the accuracy of classification for the existing matrix decomposition methods. The accuracy of classifications that are listed in the second column can be consistent with C4.5 compared to SVM and ID3 on the proposed MMDM. However, the existing methods employed only an SVM classifier and the datasets were WBC as well as SONAR.

The results of k-means clustering are shown below at Table 3.10, when k=2 (form two clusters).

Table 3.10: Clustering on Terrorist Dataset for k = 2

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Computation time (ms)
Original (D)	1000	830	170	37 ms
Distorted (D1)	1000	830	170	31 ms
Distorted (D2)	1000	810	190	30 ms

In the above Table 3.10, the first column presents the dataset D, D1 and D2. The number of objects in the dataset considered for the experiment can be seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster2. The computational time is presented in the last column.

The results of k-means clustering are shown below at Table 3.11 for k=3.

Table 3.11: Clustering on Terrorist Dataset for k = 3

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Number of Objects in Cluster 3	Computation time (ms)
Original (D)	1000	750	120	130	43 ms
Distorted (D1)	1000	780	120	100	41 ms
Distorted (D2)	1000	650	160	190	40 ms

In the above Table 3.11, the first column presents the dataset D, D1, and D2. The number of objects in the dataset considered for the experiment is seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster2. The fifth column tabulates the number of objects in cluster3. The computational time is presented in the last column. Similarly the results of k-means clustering are shown below at Table 3.12 for k=4.

Table 3.12: Clustering on Terrorist Dataset for k=4

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Number of Objects in Cluster 3	Number of Objects in Cluster 4	Computation time (ms)
Original (D)	1000	450	125	315	110	75 ms
Distorted (D1)	1000	480	120	300	100	71 ms
Distorted (D2)	1000	500	130	240	130	70 ms

In the above Table 3.12, the first column presents the dataset D, D1, and D2. The numbers of objects in the dataset considered for the experiment are seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster2. The fifth column tabulates the number of objects belonging to cluster3. The sixth column displays number of objects belonging to cluster4. The computational time is presented in the last column.

Based on results presented above the misclassification error rate of datasets is presented at Table 3.13. The misclassification error rate is given by the following Equation 3.6:

$$\mathrm{ME} = \frac{1}{p} * \sum_{i=1}^{k}(|\mathrm{Clusteri}(D)| - |\mathrm{Clusteri}(D')|) \qquad \text{Eq 3.6}$$

Where clusteriD is the i^{th} cluster in the original dataset D, clusteriD' is the i^{th} cluster in the distorted dataset, p is the number of objects and k is the number of clusters.

Table 3.13: Misclassification Error rate on Datasets

Data Objects	Original Dataset(D)			Distorted Dataset(D1)			Distorted Dataset(D2)		
	K=2	K=3	K=4	K=2	K=3	K=4	K=2	K=3	K=4
100	0.00	0.01	0.002	0.01	0.06	0.003	0.00	0.02	0.002
500	0.00	0.02	0.04	0.00	0.05	0.04	0.00	0.02	0.01

In the above Table 3.13, the first column presents the number of data objects from a dataset D, D1 and D2. The second, third and fourth column tabulate, the misclassification error rate for the original dataset D, distorted dataset D1 and D2 and the number of clusters using k=2, k=3 and k=4. It is observed that the error rate is less for k=2 on original dataset D and distorted dataset D2 than D1.

The privacy metrics mentioned above at section 3.5 from Eq 3.1 to 3.4 (Page No.51), the detailed calculation of privacy metrics are shown below:

Example 3.2: Consider the data matrix $D = \begin{bmatrix} 2 & 1.5 & 3 & 0.2 \\ 1 & 1.4 & 4 & 1.2 \\ 4 & 1.3 & 5 & 0.6 \\ 3 & 2.3 & 6 & 1.1 \end{bmatrix}$

The corresponding distorted data matrix is shown below as D'

$D' = \begin{bmatrix} 0.2 & 0.2 & -3.8 & -0.1 \\ 1.7 & 0.3 & -4.3 & -0.7 \\ 3.6 & 0.5 & -5.3 & 0.4 \\ 1.9 & 0.8 & -6.5 & -0.1 \end{bmatrix}$ According to equation 3.1 the value difference $VD = \frac{19.83}{11.41}$ then VD= 1.7383

Similarly the position difference holding four metrics is also RK, RP, CK and CP is also calculated.

RP from equation 3.2, in the data matrix D, the order for the first attribute is represented as $\mathrm{ordD1} = [2 \quad 1 \quad 4 \quad 3]^T$.

Similarly the order for the first attribute in distorted data matrix D' is $\mathrm{ordD'1} = [1 \quad 2 \quad 4 \quad 3]^T$. Therefore the total order change for the first attribute is 2. Similarly, calculating the order change for all attributes, RP=1.34. RK from Eq 3.3 is calculated as 0.34.

CP from equation 3.4, in the data matrix D, the order change of data values for all attributes is $D = [2 \quad 1 \quad 3 \quad 4]^T$ and similarly for the distorted data matrix$D' = [3 \quad 4 \quad 1 \quad 2]^T$. Therefore the total change is obtained as 8, CP=2 and CK as 0.34

The above metrics are calculated for the original dataset D and distorted dataset D1, D2. The results are furnished below at Table 3.14.

Table 3.14: Privacy Metrics for Distorted Datasets

Dataset	Modified Matrix Decomposition Method (MMDM)					Existing Matrix Decomposition Methods				
	VD	RP	RK	CP	CK	VD	RP	RK	CP	CK
Distorted (D1)	0.062	651.2	0.02	32.2	0.12	0.36	666.9	0.07	21.28	0.39
Distorted (D2)	0.72	677.5	0.00	37.1	0.00	0.74	666.4	0.05	36.42	0.00

In the above Table 3.14, the first column presents the distorted datasets D1 and D2. The second column displays the privacy metrics calculated for the proposed MMDM, such as, VD, RP, RK CP and CK. The third column reveals the privacy metrics calculated for the Existing Matrix Decomposition Method. It is observed that the metrics RP and CP are high for the proposed MMDM of distorted D2 for the Existing Matrix Decomposition Method. The VD, RK and CK are less for proposed MMDM compared to the existing matrix decomposition method. The privacy metrics for the proposed MMDM shows better quality in preserving confidential data.

3.7 Conclusion

This chapter focuses on modifying the existing matrix decomposition methods to perturb the data. This proposal is implemented successfully. Exhaustive experiments using different classifiers and k-means clustering resulted in accurate classification and clusters. Even though the data perturbation is efficient and privacy is preserved, this implementation suffered some information loss during the perturbation process.

CHAPTER 4

DATA PERTURBATION USING FUZZY LOGIC

4.1 Introduction

This chapter proposes a fuzzy logic based data perturbation method. It considers different fuzzy membership functions to perturb the data and overcomes the drawback of losing information in the process. The proposal is explained as below.

4.2 Proposed Fuzzy Data Perturbation (FDP)

The Fuzzy Data Perturbation (FDP) is proposed as shown below in the block diagram at Fig 4.1. This block diagram is explained as follows.

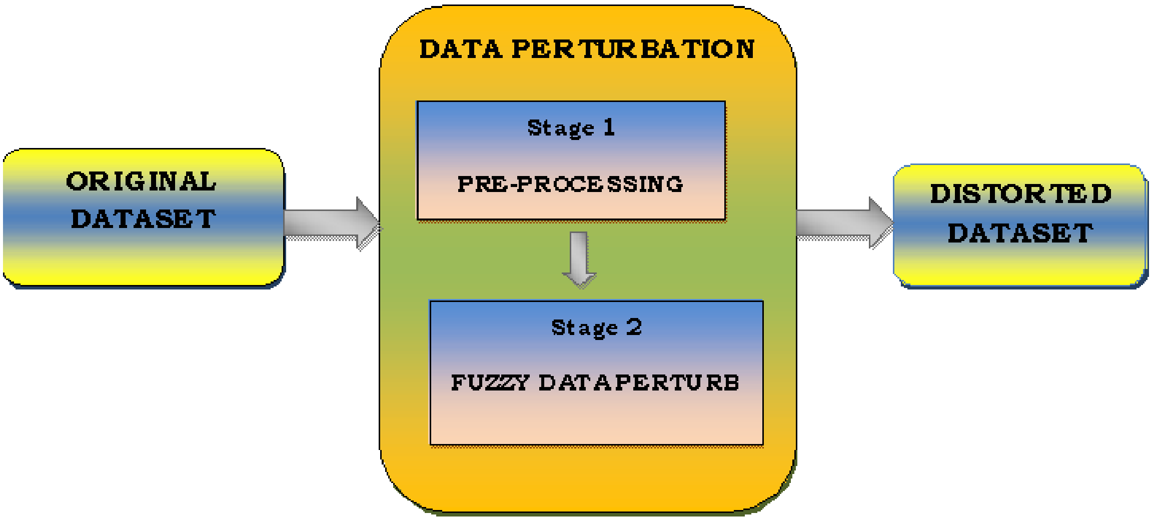


Figure 4.1: Block Diagram for Fuzzy Data Perturbation

The above block diagram at Fig 4.1 considers the original dataset and deals with it in two stages. In the first stage the best features of the original dataset are extracted. This extraction is known as preprocessing. It is similar to the explanation furnished in chapter 3. second stage considers the best featured dataset is distorted using the proposed The fuzzy data perturbation. This is explained in the next section.

4.3 Proposed Fuzzy Data Perturb Algorithm

A proposal for fuzzy data perturbation is provided in this section. The pseudo code of the proposed algorithm is listed below:

Algorithm Fuzzy Data Perturb (FDP)

```
Input:  A Data matrix $D_{p×q}$
Output: Distorted Dataset Ds/ Dz /Dt.
Begin
Step 1: Consider a Data Matrix $D_{p×q}$, and Create A,B & C as below:
          A= min (D) // Minimum matrix of 1×q decomposed from D
                        with minimum elements of each column.
          B= max (D) // Maximum matrix of 1×q decomposed from D
                        with maximum elements of each column.
```

```
               C= Avg (D) // Average matrix of 1×q decomposed from D with
                                  average elements of each column.
Step 2:  Create Fuzzy Membership function of D by
               Call MF (D, A, B, C: Ds/ Dz/ Dt); //semicolon (:) indicates
               end of sequence of data matrices and beginning of return
               parameter (dataset).
Step 3: The Distorted data matrices Ds/Dz/Dt are the output.
 End
```

The algorithm accepts the data matrix D having p rows and q columns as input. This data matrix $D_{p\times q}$ is used to create A, B and C data matrices. The data matrix A is created using minimum values of each column in D as row and q column. Similarly, B and C are created using maximum and average values of each column in D. Now call the membership functions to create the fuzzy datasets. Once the fuzzy datasets are formed, the fuzzy data matrices Ds, Dz and Dt are constructed.

Membership Function: MF (D, A, B, C : Ds/Dz/ Dt)

Input: A Data Matrix $D_{p\times q}$, A, B & C (row vectors)

Output: A Fuzzy dataset $Ds_{p\times q}$ /$Dz_{p\times q}$ / $Dt_{p\times q}$.

```
Begin
Step 1: Copy D Data Matrix to Ds, Dz & Dt Data Matrices //Aliasing D
Step 2: In the Data Matrix $D_{p\times q}$ do
          For i=1 to p do
          For j=1 to q do
          If ( (D[i, j] <= A[j]) && (D[i, j] <= B[j] ) ) then //Nested if.
                Dt[i, j] = ( D[i, j] - A[j] )/ ( B[j] -A[j] ) ;
                  If (D[i,j] <= ( A[j] + B[j] ) /2)  then
                     Ds [i, j] = (2 * ( D[i, j] - A[j] )/ ( B[j] - A[j] ) )$^2$);
                     Dz [i, j] = (1 - (2 * ( D[i, j] - A[j] ) / (B[j] - A[j])$^2$ );
                 Else
                    Ds[i ,j] = (1 – ( 2 * (D[i, j] - A[j] ) / ( B[j] - A[j] )$^2$ );
                     Dz[i, j] = (2 * (D[i, j] - A[j] ) / ( B[j] - A[j] )$^2$);
                End if
            Else                                    //Nested If condition.
            if ( ( D[i, j] >= B[j] ) && ( D[i, j] <= C[j] ) ) then
              Ds[i, j] =1; Dz[i, j] = 0; Dt[i, j] = ( C[j] – D[i, j] ) / ( C[j] - B[j] );
            Else
               Dt[i, j] = 0; Ds[i, j] =0;
            End if
            End if
           Construct Ds, Dz and Dt Data Matrices with updated values.
          End For
         End For
Step 3:  Return of dataset Ds/Dz / Dt as parameter.
End
```

This membership function is called using D, A, B, C data matrices as parameters in the above mentioned algorithm FDP. First, D data matrix is copied to Ds, Dz and Dt data matrices. Then, predefined conditions are checked to update Ds, Dz and Dt. This process of updating values in this membership function is explained using an example. This membership function returns the updated datasets Ds, Dz and Dt. The time complexity of the proposed FDP algorithm is found to be O(n), where n is the dimension of the dataset.

Example 4.1: Consider a data matrix $D_{3\times2}$ ($D_{p\times q}$ where p=3 and q=2) given as $D_{3\times2}= \begin{bmatrix} 1 & 5 \\ 0.9962 & 5 \\ 0.8956 & 3 \end{bmatrix}$

The data matrices A, B and C are created from the above mentioned data matrix $D_{3\times2}$. These data matrices are listed below:

$A = \begin{bmatrix} 0.8956 & 3 \end{bmatrix}$ //Minimum element of $D_{3\times2}$ data matrix//

$B = \begin{bmatrix} 1 & 5 \end{bmatrix}$ //Maximum element of $D_{3\times2}$ data matrix//

$C = \begin{bmatrix} 0.9639 & 4.3333 \end{bmatrix}$ //Average element of $D_{1\times2}$ data matrix//

When membership function is called in the algorithm FDP, the data matrices D, A, B and C are passed as parameters.

Now, to form a fuzzy data matrix $Ds_{3\times2}$ from these data matrices, an S-based fuzzy membership function is given at Eq 4.1.

$$f(x;a,b)=\begin{cases} 0, & x\le a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a\le x\le \frac{a+b}{2} \\ 1\text{-}2\left(\frac{x-b}{b\text{-}a}\right)^2, & \frac{a+b}{2}\le x\le b \\ 1, & x\ge b \end{cases} \qquad \text{Eq 4.1}$$

Where x is a data value of $D_{3\times2}$, a is the data value of A data matrix and b is the data value of B data matrix.

Let the data value x=1, a=0.8956,b=1. Substitute these values in Eq 4.1. According to the conditions stipulated, the $f(x;a,b)$ satisfies the third value. The other conditions fail to satisfy the value of $f(x;a,b)$

Therefore $f(x;a,b) = 1 - 2\left(\frac{x-b}{b-a}\right) = 1$

In the similar manner by considering the values of the data matrix, $D_{3\times2}$ are used to calculate the remaining elements in data matrix $Ds_{3\times2}$.The updated $Ds_{3\times2}$ is given as $DS = \begin{bmatrix} 1 & 1 \\ 0.9974 & 1 \\ 0 & 0 \end{bmatrix}$

Similarly, to form a fuzzy data matrix $Dz_{3\times2}$ from the data matrices D, A, B, a Z-based fuzzy membership function is given in Eq 4.2.

$$f(x;a,b)=\begin{cases} 1, & x\le a \\ 1-2\left(\frac{x-a}{b-a}\right)^2, & a\le x\le \frac{a+b}{2} \\ 2\left(\frac{x-b}{b\text{-}a}\right)^2, & \frac{a+b}{2}\le x\le b \\ 0, & x\ge b \end{cases} \qquad \text{Eq4.2}$$

Where x is a data value of $D_{3\times2}$, a is the data value of A data matrix and b is the data value of B data matrix.

Let the data value x=1, a=0.89, b=1. Substitute these values in Eq 4.2. According to the conditions stipulated, the $f(x; a, b)$ satisfies the third value. The other conditions fail to satisfy the value of $f(x; a, b)$

Therefore $f(x; a, b) = 2 * (\frac{x-b}{b-a})^2 = 0$

In the similar manner by considering the values of the data matrix $D_{3\times2}$ are used to calculate the remaining elements in data matrix $Dz_{3\times2}$.The updated $Dz_{3\times2}$ is given as $Dz = \begin{bmatrix} 0.9951 & 0 \\ 0.9954 & 0 \\ 1 & 0 \end{bmatrix}$

Finally, to form a fuzzy data matrix $Dt_{3\times2}$ from the data matrices D, A, B, C and T-based fuzzy membership function are given in Eq 4.3.

$$f(x; a, b, c) = \begin{Bmatrix} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & c \leq x \end{Bmatrix} \qquad \text{Eq 4.3}$$

Where x is a data value of $D_{3\times2}$, a is a data value of A data matrix and b is the data value of B data matrix while c is the data value of C data matrix.

Let the data value x=1, a=0.89, b=1. Substitute these values in Eq 4.2. According to the conditions stipulated, the $f(x; a, b, c)$ satisfies the third value. The other conditions fail to satisfy the value of $f(x; a, b, c)$.

Therefore, $f(x; a, b, c) = \left(\frac{c-x}{c-b}\right) = 1$

In the similar manner by considering the values of the data matrix $D_{3\times2}$ are used to calculate the remaining elements in data matrix $Dt_{3\times2}$. The updated $Dt_{3\times2}$ is given as $Dt = \begin{bmatrix} 1 & 1 \\ 1 & 0.9951 \\ 0 & 0.9954 \end{bmatrix}$

The fuzzy data matrices Ds, Dz and Dt are the parameters returned. These constructed fuzzy data matrices Ds, Dz and Dt are the outputs. The implementation detail of this proposal is explained in the next section.

4.4 Implementation

The proposed algorithm that was discussed in the previous section is implemented using MatLab. Its source code is included in APPENDIX II (Page No.132). The details of implementation are furnished in this section.

The Implementation utilizes built in functions available in MatLab such as, load(), size(), min(), max() and avg(). These are used to create a MEX file. First, a load() built in function, is called to read the data into a data matrix. Then, size() function is employed to read the size of a data matrix. Then min(), max() and avg() built in function available are used to decompose a data matrix D into data matrices A, B and C.

Next, smf(), tmf() and zmf() built in fuzzy membership functions are utilized to create the data matrices Ds, Dz and Dt.

4.5 Experimentation

The experimentation was conducted using desktop computer system loaded with windows XP Operating system, MatLab and Tanagra data mining tool. The experimental details are elaborated in this section.

The experimentation begins by preprocessing a dataset to extract the best features using data mining tool. These best features form an original dataset D. This original dataset D is given as input to the proposed FDP algorithm to obtain the distorted datasets Ds, Dz and Dt. Then, the original dataset D and distorted datasets Ds, Dz and Dt are uploaded into Tanagra data mining tool after appending a class attribute. These uploaded datasets are classified using classification utility available within Tanagra data mining tool. The results of classification are analyzed thereafter. Similarly the datasets are clustered using clustering utilities available in them. The results of clustering are also analyzed and furnished in section 4.6 under Results and Analysis. The datasets, such as, Terrorist, Hepatitis, Fertility and Cancer, are used in this experimentation downloaded from UCI machine Learning Repository. The details of Terrorist dataset used in this experimentation are furnished.

A Real Time univariate dataset, namely, Terrorist, is downloaded from website DARPA [101]. Their details of the same are shown at Table 3.1. The Terrorist dataset is considered for preprocessing with a population size of 41 attributes, other than class attribute. The details of preprocessing, together with a sample Terrorist original dataset D with 5 rows and 10 attributes are shown at Table 3.2.

The process in the experiment is elaborated as below: First, a dataset named terrorist.txt is loaded into X data matrix with the help of load method. Next, the size() method on X data matrix determines the number of rows p as 1000 and the number of columns q as 10. The data matrix X is now named data matrix $D_{p\times q.}$ Then, the built in functions, min(D) is used to decompose data matrix D and obtain data matrix A. Similarly, max (D) and avg(D) built-in functions are used to decompose data matrix D and obtain B, C. The decomposed data matrices A, B and C comprises of one row and q columns. Now, the fuzzy membership functions are called smf(), zmf() and trimf() respectively. The above fuzzy membership function uses the data matrices D, A, B and C as input parameters.

The data matrix $D_{3\times 2}$ is identified as Ds, Dz and Dt data matrices. Then, S-based fuzzy membership function smf(D,[A,B]) is used. The function checks the pre-defined conditions on the data values of D, A, B data matrices and updates its corresponding value into Ds data matrix. Similarly, Z-based fuzzy membership function zmf(D,[A,B]) is used. The function checks the pre-defined conditions on the data values of D, A, B data matrices and updates its corresponding value into Dz data matrix. Finally, T-based fuzzy membership function tmf(D,[A,B]) is used. The function checks the pre-defined conditions on the data values of D, A, B and C data matrices and updates its corresponding value into Dt data matrix. The resultant data matrix Ds or Dz or Dt are the distorted or Fuzzy data matrix as output. When the above process is executed in experimentation. It outputs a distorted datasets Ds, Dz and Dt. They are furnished in section 4.6 under Results and Analysis section. Each of these distorted datasets Ds, Dz and Dt together with original dataset D, respectively are appended with a class attribute, YES or NO, as mentioned at Table 4.1.

Table 4.1: A Terrorist Original Dataset D with Class Attribute

ID	Age	Nationality	Sibling	Pilot Training	Temporary Address	Wedding Address	Time	Date	Place	CA
1	3	1	0	0	0	0	10	0	10	Yes
2	3	1	0	0	0	0	10	0	10	No
2	3	0	0	0	0	0	0	0	8	Yes
4	3	0	0	0	0	0	7	0	10	Yes
4	3	0	0	0	0	1	0	1	8	No

The above mentioned datasets D, Ds, Dz and Dt are uploaded into Tanagra data mining tool. First, classification utility is used on dataset D and distorted datasets Ds, Dz, Dt. It divides the attributes into two categories namely non-class attributes and class attributes. These two categories can be two inputs to the classifier chosen from among available ones.

Suppose we select SVM (Support Vector Machines), as classifier, then, it classifies the datasets D, Ds, Dz and Dt based on class attribute and produces the results, respectively. The details of classification results are furnished at section 4.6 under Results and Analysis. Similarly, the experimentation is repeated with Iterative Dichotomizer 3 (ID3), and C4.5 classifiers. The results of these experiments are furnished at section 4.6.

A Clustering utility available in Tanagra data mining tool is used to cluster the original dataset D and distorted datasets Ds, Dz and Dt. Non class attributes are considered and given as input to k-mean clustering method. As a result, number of clusters is formed. The results of clustering are furnished at section 4.6.

4.6 Results and Analysis

The results obtained in the above experiment are presented in this section. The Terrorist original dataset D furnished at Table 3.2 (Page No.46) is given as input to the proposed FDP. It outputs three distorted datasets Ds, Dz and Dt. The terrorist distorted dataset Ds is presented at Table 4.2.

Table 4.2: A Terrorist Distorted Dataset Ds

ID	Age	Nationality	Siblings	Pilot Training	Temp Address	Wedding Address	Time	Date	Place
0	1.0	0	0	0	0	0	1	0	1

0.5	1.0	0	0	0	0	0	1	0	1
0.5	1.0	0	0	0	0	0	0	0	1
1.0	1.0	0	0	0	0	0	1	0	1
1.0	1.0	0	0	0	0	0	0	0	1

Similarly, a terrorist distorted datasets Dz and Dt are shown at Table 4.3 and Table 4.4

Table 4.3: A Terrorist Distorted dataset Dz

ID	Age	Nationality	Siblings	Pilot Training	Temp Address	Wedding Address	Time	Date	Place
1	0	1	1	1	1	1	0	1	0
0.5	0	1	1	1	1	1	0	1	0
0.5	0	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	0	1	0
0	0	1	1	1	1	1	1	1	0

Table 4.4: A Terrorist Distorted dataset Dt

ID	Age	Nationality	Siblings	Pilot Training	Temp Address	Wedding Address	Time	Date	Place
0	1.0	0	0	0	0	0	1	0	0
1.0	1.0	0	0	0	0	1	0	0	0
1.0	1.0	0	0	0	0	0	0	0	0
1	1.0	0	0	0	0	1	1	0	0
1	1.0	0	0	0	0	0	1	0	0

When SVM classifier is used on Terrorist original dataset D and distorted datasets Ds, Dz and Dt, the following observations are made and the same are presented at Table 4.5.

Table 4.5: A Terrorist Dataset Classified using SVM

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Terrorist (YES)	Number of Support Vectors	Error Rate	Computation Time (ms)
Original (D)	1000	728	647	0.30	2359 ms

Distorted (Ds)	1000	770	740	0.30	2250 ms
Distorted(Dz)	1000	770	720	0.30	2266 ms
Distorted (Dt)	1000	750	696	0.50	2266 ms

In the above Table 4.5, the first column presents the original dataset D and distorted datasets Ds, Dz and Dt. The numbers of tuples in the datasets considered for experimenting are seen in the second column. The third column displays the number of training tuples belonging to terrorist category as YES. The number of support vectors available is furnished in the fourth column. The fifth column reveals the error rate of SVM classifier. The computation time is tabulated in the last column. Similar results are tabulated at Table 4.6 and 4.7 when the ID3 and C4.5 classifiers are used on D, Ds, Dz and Dt.

Table 4.6: A Terrorist Dataset Classified using ID3

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Terrorist(YES)	Tree Having Number of Nodes and Leaves	Error Rate	Computation Time(ms)
Original (D)	1000	729	1 node,1 leaf	0.30	4355 ms
Distorted(Ds)	1000	770	1 node,1 leaf	0.30	16 ms
Distorted(Dz)	1000	770	1 node,1 leaf	0.30	16 ms
Distorted (Dt)	1000	690	1 node,1 leaf	0.45	16 ms

In Tables 4.6 and 4.7 below, the first column presents the dataset D and distorted datasets Ds, Dz and Dt. The number of tuples in the datasets considered for experimenting is seen in the second column. The third column displays the number of training tuples belonging to terrorist category as YES. A tree having the number of nodes and leaves are furnished in the fourth column. The fifth column reveals the error rate of ID3 and C4.5 classifiers, respectively. The computation time is tabulated in the last column.

Table 4.7: A Terrorist Dataset Classified using C4.5

Dataset	Number of Tuples	Number of Training Tuples Classified as Terrorist (YES)	Tree having Number of Nodes and Leaves	Error Rate	Computation time (ms)
Original (D)	1000	739	23 nodes, 12 leaves	0.29	47 ms
Distorted (Ds)	1000	779	21 node, 10 leaves	0.29	16 ms
Distorted (Dz)	1000	778	21 nodes, 11 leaves	0.29	31 ms

Distorted (Dt)	1000	720	21 nodes,11 leaves	0.30	31 ms

In the similar manner the results obtained on Hepatitis, Fertility and Cancer datasets are furnished in APPENDIX V (Page No.136). Based on the results presented above, the accuracy of classification of datasets is presented at Table 4.8.

The accuracy is the percentage of tuples correctly classified by a classifier and is given by the Equation 3.5 (Page No. 53).

Table 4.8: Accuracy of Classifiers (%)

Dataset	Terrorist			Hepatitis			Fertility			Cancer		
	SVM	ID3	C4.5	SVM	ID3	C4.5	SVM	ID3	C4.5	SVM	ID3	C4.5
Original (D)	72.8	72.9	73.9	86	87	88	95	95	96	79	81	82
Distorted (Ds)	77	77.2	77.9	86.4	87.2	88.6	93	95	96	79.7	82	82.9
Distorted (Dz)	77	77.2	77.9	86	86.4	86.8	91	94	95	79.4	80.6	82.9
Distorted (Dt)	75	69.0	72.8	84.2	84	85.8	90	91	91	77.3	80.4	82.2

The Table 4.8 below presents the accuracy of the classifiers for Terrorist, Hepatitis, Fertility and Cancer datasets. The first column presents the dataset D, distorted datasets Ds, Dz and Dt. The second column presents the accuracy of classification obtained on Terrorist dataset using SVM, ID3 and C4.5 classifiers. The third column presents the accuracy of classification obtained on Hepatitis dataset using SVM, ID3 and C4.5 classifiers. The fourth column presents the accuracy of classification obtained on Fertility dataset using SVM, ID3 and C4.5 classifiers. The fifth column presents the accuracy of classification obtained on Cancer dataset using SVM, ID3 and C4.5 classifiers. It is observed that accuracy of C4.5 classifier is better than all the classifiers on distorted dataset Ds and Dz. The results of k-means clustering are shown below in Table 4.9, when k=2 (form two clusters).

Table 4.9: Clustering on Terrorist Dataset for k = 2

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Computation time (ms)
Original (D)	1000	170	830	37 ms
Distorted (Ds)	1000	20	980	61 ms
Distorted (Dz)	1000	21	979	62 ms

Distorted (Dt)	1000	30	970	62 ms

In the above Table 4.9, the first column presents the dataset D, Ds, Dz and Dt. The number of objects in the dataset considered for the experiment is seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster2. The computational time is presented in the last column.

In the similar way the results obtained on Hepatitis, Fertility and Cancer datasets are furnished. Based on results presented above the misclassification error rate of datasets is presented at Table 4.10. The misclassification error rate is given by the following Equation 3.6.

Table 4.10: Comparison of Misclassification Error Rate

Dataset	Proposed FDP Method				Existing Fuzzy Based Approaches		
	Terrorist	Hepatitis	Fertility	Cancer	Iris	Wine	Credit-G
Distorted(Ds)	0.048	0.08	0.06	0.02	-	-	-
Distorted(Dz)	0.055	0.07	0.05	0.028	0.097	0.089	0.012
Distorted(Dt)	0.061	0.04	0.075	0.03	0.15	0.17	0.222

The Table 4.10 below presents the misclassification error rate. The first column presents the distorted dataset Ds, Dz, Dt. The second column presents the error rate obtained on proposed FDP using Terrorist, Hepatitis, Fertility and Cancer datasets. The third column presents the error rate seen in the literature for the existing fuzzy based approaches. The error rate listed in the second column is less on our proposed FDP. However, the existing fuzzy-based approaches were employed only on distorted Dz and Dt. Moreover, their results were only shown on IRIS, WINE and CREDIT-G datasets. There are no traces of other classifiers employed in the Existing Fuzzy methods.

4.6 Conclusion

A Fuzzy logic-based data perturbation is proposed in this chapter. It is successfully implemented using different fuzzy membership functions. Exhaustive experiments using different datasets resulted in better classification and clusters. This method of data perturbation overcame the information loss but at the cost of privacy.

BIBIOLOGRAPHY

[1] Han Jiawei, M. Kamber, Data Mining: Concepts and Techniques, Beijing: China Machine Press, pp:1-40, 2006.

[2] Kamber, Pei ,Han, “Data Mining: Concepts and Techniques”, 3rd Edition. San Francisco: Morgan Kaufmann, July 2011.

[3] R.Agrawal and R. Srikant, “Privacy–Preserving data mining”, In Proceedings of the 2000 ACM SIGMOD International Conference on management of data, San Diego, CA, pp: 86-97, 2003.

[4] Elisa B`1ertino, Dan Lin and Wei Jiang, “A Survey of Quantification of Privacy Preserving Data Mining Algorithms”, Advances in Databases Systems, Springer US, Vol. 3, PP:183-205, 2008.

[5] S.Canetti, M.Fungini, G.Martella and P.Samarati “Database Security”, Addison Weley, 1995.

[6] Shipra Agrawal and Jawant Haritsa, “A Framework for High – Accuracy Privacy Preserving Mining”, In proceedings of the 21st International Conference on Data Engineering, PP:19-204, 5-8 April 2005.

[7] V. S. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin and Y.Theodoridis, “State-of-the-art in privacy preserving data mining”, SIGMOD Record, 33(1), PP:50-57, 2004.

[8] Asmaa Hatem Rashid and Norizan Binti Mohd Yasin , “Privacy preserving data publishing: Review”, In International Journal of Physical Sciences, Vol. 10(7), DOI: 10.5897/IJPS11.1795, ISSN 1992 – 1950, PP: 239-247, 16 April 2015.

[9] Yousra Abdul Alsahib S.Aldeen, Mazleena Salleh, Mohommad Abdur Razzaque, “A Comprehensive review on privacy preserving data mining”, In Springer Plus 4:694, DOI**:** 10.1186/s40064-015-1481, Oct 2015.

[10] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, 2014, “A Review on Privacy Preserving Data Mining: Technique and Research Challenges, International Journal of Computer Science and Information Technologies”, Vol. 5 (2), pp: 2310-2315, 2014.

[11] Malik, M.B., Ghazi, M.A., Ali, R., “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, Third International Conference on Computer and Communication Technology (ICCCT), pp: 26 – 32, 2012.

[12] L. Sweeney, “Replacing Personally Identifiable Information in Medical Records”, In Proceedings of Journal of the American Medical Informatics Association, 1996.

[13] Chaum DL, “ Untraceable electronic mail, return addresses, and digital pseudonyms Communication”, ACM 24(2), PP:84-90, 1981.

[14] Loukides G, G koulalas Divanis A, Shao J, “ An Efficient and Flexible Anonymization of Transaction data”, Knowledge Information Systems, 36(1), PP:153–210, 2012.

[15] R. Bayardo, R. Agrawal, “Data Privacy Through Optimal k-Anonymization”, In Proceedings the 21st International Conference on Data Engineering, pp:217-228, 2005.

[16] K. Lefevre, J. Dewittd, R. Ramakrishnan, “Incognito: Efficient Full-Domain k-Anonymity”, In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp:49-60, 2005.

[17] B. Fung, K. Wang, P. Yu, “Top-down Specialization for Information and Privacy Preservation”, In Proceedings of the

21st IEEE International Conference on Data Engineering, pp:205-216, 2005.

[18] L.Sweeney, “Achieving k-Anonymity Privacy Protection using Generalization and Suppression”, In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol.10, no.5, pp: 571-588, 2002.

[19] A. Machanavajjhala, J. Gehrke, D. Kifer, “l-Diversity: Privacy Beyond k-Anonymity”, ACM Transactions on Knowledge Discovery from Data, pp:24-35, 2007.

[20] N.H. Li, T.C. Li, “t-Closeness: Privacy beyond k-Anonymity and l-Diversity”, In Proceedings of the 23rd International Conference on Data Engineering, pp: 106-115, 2007.

[21] G. Loukides, J.H. Shao, “An Efficient Clustering Algorithm for k-Anonymisation”, International Journal of Computer Science And Technology,Vol.23, No.2, pp.188-202, 2008.

[22] J.L. Lin, M.C. Wei, “Genetic Algorithm-Based Clustering Approach for k-Anonymization”, International Journal of Expert Systems with Applications, Vol.36, No.6, pp.9784-9792, 2009.

[23] L.J. Lu, X.J. Ye, “An Improved Weighted-Feature Clustering Algorithm for k-Anonymity”, In Proceedings of the 5th International Conference on Information Assurance and Security, pp: 415-419, 2009.

[24] Z.H. Wang, J. Xu, W. Wang, B.L. Shi, “Clustering-Based Approach for Data Anonymization”, Journal of Software, Vol.21, No.4, pp.680-693, 2010.

[25] N.R Adam, J.C. Worthmann. “Security-control methods for statistical databases: a comparative study”, In proceedings of ACM Computing Surveys,21(4), PP: 515-556,1989.

[26] Raju R, Komalavalli R, Kesavakumar V, “ Privacy maintenance collaborative data mining: a practical approach”, In: 2nd international conference on emerging trends in engineering and technology (ICETET), PP: 307–311, 2009.

[27] Pingshui Wang, “Survey on Privacy Preserving Data Mining”, In International Journal of Digital Content Technology and its Application”, Vol 4,No. 9, pp:1-7, 2010.

[28] I. Ioannidis, A. Grama, M.J. Atallah, “A Secure Protocol for Computing Dot-Products in Clustered and Distributed Environments”, In Proceedings of the 31st International Conference on Parallel Processing, pp.379-384, 2002.

[29] Nguyen XC, Le HB, Cao TA, “ An Enhanced Scheme for Privacy Preserving association rules mining on Horizontally Distributed databases”, In IEEE RIVF International conference on computing and communication technologies”, Research Innovation and Vision for the Future, http://doi.org/10.1109/rivf.2012.6169821, pp: 1-4, 2012.

[30] Mi Wen, RongXing Lu, Jingshen Lei, Xiaohui Lang, “ECQ: An Efficient Conjunctive Query Scheme over Encrypted Multidimensional Data in smart grid”, Global Communications Conference (GLOBECOM) IEEE, pp:796-801, 2013.

[31] Li Yaping Chen Minghau, Li Qiwei, Zhang, Wei, "Enabling Multilevel Trust in Privacy Preserving Data Mining", In Knowledge and Engineering IEEE Transactions, Vol 24, No. 9, pp:1598-1612, 2012.

[32] Kokkinos Y,Margaritis K, "Distributed privacy preserving P2P data mining via probabilistic neural network committee machines, Fourth International Conference on Information, Intelligence, Systems and Applications (IISA) pp.1-4,2013.

[33] Lindell, Yehuda, Pinkas, "Privacy preserving data mining", In Proceedings of the Advances in Cryptology–CRYPTO, pp:36–54, 2000.

[34] M. Kantarcioglu, C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.9, pp: 1026-1037, 2004.

[35] J. Vaidya, C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.639-644, 2002.

[36] W.L.Du, Z.J.Zhan, "Building Decision Tree Classifier for Vertically Partitioned Data", In Proceedings of the IEEE International Conference on Data Mining Workshop on Privacy, pp: 1-8, 2002.

[37] Masooda Modak and Rizwana Shaikh, "Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy", In 7th International Conference on Communication, Computing and Virtualization, Procedia Computer Science 79, Published by Elsevier B.V, 1877-0509,doi:10.1016/j.procs.2016.03.126, 2016.

[38] Inan A, Saygin Y , "Privacy preserving spatio-temporal clustering on horizontally partitioned data", In Lecture Notes in Computer Science including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatic, 6202 LNAI, http://doi.org/10.1007/978-3-642-16392-0_11,pp:187–198, 2010.

[39] Om Kumar CU, Tejaswi K, Bhargavi P, "A distributed cloud—prevents attacks and preserves user privacy". In 15th international conference on advanced computing technologies, ICACT. http://doi.org/10.1109/ICACT.2013.6710509, 2013.

[40] R.Canetti, " Security and Composition of multiparty of multiparty Cryptography Protocols" , In Proceedings of Journal of Cryptology, PP:143-202,2000.

[41] Kamakshi P, Babu AV, "Preserving privacy and sharing the data in distributed environment using cryptographic technique on perturbed data" , 2(4), 2010.

[42] Liew CK, Choi UJ, Liew CJ," A Data Distortion by Probability ", ACM Transaction Database System 10, pp: 395-411, 1985.

[43] Fienberg S.E. and McIntyre J ,“Data Swapping: Variations on a theme by Dalenius and Reiss” , In Journal of Official Statistics, Vol .21, PP: 309- 323, 2005.

[44] S.Kavitha and P.Raja Vadhana P, “Data Privacy Preservation using Various Perturbation Techniques”, In International Jounal of Innovative Research in Computer and Communication Engineering, Vol.3, Issue 2, pp:1039-1042, Febrauary 2015.

[45] D. Agrawal and C .C. Aggarwal, “On the Design and Quantification of Privacy Preserving Data Mining Algorithms”, In Proceedings of the 20th ACM SIGACT-SIGMOD SIGART Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 2000.

[46] Matwin S, “Privacy-preserving data mining techniques: survey and challenges. In: Discrimination and privacy in the information society, Springer, Berlin, Heidelberg, PP: 209–221,2013.

[47] Pathak F.A.N, Pandey S.B.S, “An Efficient Method for Privacy Preserving Data Mining in Secure Multiparty Computation, Nirma University, In International Conference on Engineering (NUICONE), pp:1-3, 2013.

[48] Liew CK, Choi UJ, Liew CJ,” A Data Distortion by Probability “, ACM Transaction Database System 10, pp: 395-411, 1985.

[49] R. Agrawal, A. Evfimieski and R.Srikanth, “Information sharing across private databases” In proceedings of the 2003 ACM SIGMOD International Conference on management of data, pp:86-97, SanDiego, CA, 2003.

[50] Bo Peng, Xingyu Geng., JunZhang., “Combined Data Distortion Strategies for Privacy-Preserving Data Mining”, In proceedings of 3rd International Conference on Advanced Computer Theory and Engineering (1CACTE) ,2010.

[51] H. Kargupta, S. Dutta, Q.Wang,K. Sivakumar, “On the privacy Preserving Properties of Random Data Perturbation Techniques”, In Proceedings of the 3rd International conference on Data Mining,pp.99-106,2003.

[52] A.Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, “Privacy Preserving Mining of Association Rules”, In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pp:217-228, 2002.

[53] S. L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias”, J. Am. Stat. Assoc., vol.60, no.309, pp.63-69,1965

[54] Z. Huang, W. Du, B. Chen, “Deriving Private Information from Randomized Data”, In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland,USA, pp: 37-48, 2005.

[55] S. Rizvi, J. Haritsa, “Maintaining Data Privacy in Association Rule Mining”, In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[56] Fung BCM, Wang K, Chen R, Yu PS, “Privacy-Preserving Data Publishing”, A survey on recent developments and Computing. 5(4), pp:1-53, 2010

[57] S. Xu, J. Zhang, D. Han and J. Wang, “Data distortion for Privacy Protection in a Terrorist Analysis System”, In Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics, Atlanta GA, USA, PP: 459–464, May 19-20, 2005.

[58] S. Xu, J. Zhang, D. Han, J. Wang, “Singular value decomposition based data distortion strategy for privacy protection”, Knowledge and Information Systems, 2006, 10(3): 383–397.

[59] J.Gao, J. Zhang. “Sparsification strategies in latent semantic indexing”, In Proceedings of the 2003 Text Mining Workshop, M.W. Berry and W.M. Pottenger, (ed.), pp. 93-103, San Francisco, CA, May 3, 2003.

[60] J. Wang, W. Zhong, S. Xu and J. Zhang, “Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation”, In Proceedings of the 2006 International Conference on Information & knowledge Engineering, Las Vegas, PP:114-120, 2006.

[61] N. Maheswari and K. Duraiswamy, “CLUST-SVD: Privacy Preserving Clustering in Singular Value Decomposition”, World Journal of Modelling and Simulation, England, UK, ISSN: 1746-7233, Vol. 4, Issue No. 4, ISSN: 1746-7233, PP: 250-256, 2008.

[62] Pengpeng Lin, Jun Zhang, Ingrid St. Omer, Huanjing Wang, and Jie Wang, “A Comparative Study on Data Perturbation with Feature Selection”, In Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, Vol. I, ISBN: 978-988, ISSN: 2078-0958,PP: 1-766, March 16-18, 2011

[63] L. Rokach, O. Maimon, “Theory and application of feature decomposition”, In Proceedings of the First IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, pp:473–480, 2001.

[64] O. Maimon, L. Rokach, “Improving supervised learning by feature decomposition”, In Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems, Lecture Notes in Computer Science, Springer-Verlag,

[65] Sumit Ghosh, Qutaiba Razouqi, H.Jerry Schumachader and Aivars Celmins, “Survey of Recent Advances in Fuzzy Logic in Telecommunications Networks and New Challenges”, IEEE Transactions on Fuzzy Systems, Vol. 6, No. 3, August 1998. pp:178–196, 2002.

[66] B. Karthikeyan, G. Manikandan and V. Vaithiyanathan, “A Fuzzy Based Approach for Privacy Preserving Clustering”, Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, PP: 118–122, October31, 2011.

[67] V. Vallikumari, S. Srinivasa Rao, K.V.S.V.N Raju, K.V. Ramana and B.V.S. Avadhani, "Fuzzy Based Approach for Privacy Preserving Publication of Data", IJCSNS, Vol. 8, Issue No.1, PP: 115-122, January 2008.

[68] M.Naga Lakshmi and K.Sandhya Rani, "Privacy Preserving Clustering Based on Fuzzy Data Transformation Methods", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277128X, Vol. 3, Issue No. 8, PP: 1027-1033, August 2013.

[69] Geetha Mary A,N. Ch.S.N Iyenger, " Non-Additive Random Data Perturbation for Real World Data" , Procedia Technology 4(2012), Published by Elsevier LTD, doi:10.1016/j.protcy.2012.05.053, pp:350-354, 2012.

[70] Manikandan G, Sairam N, Harish V, and Nooka Saikumar, "Survey on the use of Fuzzy Membership Functions to Ensure Data Privacy", in Research Journal of Pharmaceutical, Biological and Chemical Sciences, ISSN: 0975-8585, May–June 2016, RJPBCS7(3)Page No.344-348.

[71] Liming Li and Qishan Zhang , "A privacy preserving clustering technique using hybrid data transformation method", Grey Systems and Intelligent Services ,2009 GSIS 2009,IEEE international conference , DOI 10.1109/GSIS.2009.5408151, 08, February 2010.

[72] Dr A.M. Natarajan, R. R. Rajalaxmi, N. Uma and G. Kirubhkar," A Hybrid Transformation Approach for Privacy Preserving Clustering of Categorical data", Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, Online ISBN: 978-1-4020-6268-1, DOI 10.1007/978-1-4020-6268-1_72, PP:403-408,

[73] S.Selva Rathnam, Dr T.Karthikeyan, " A Survey on Recent Algorithms for Privacy Preserving data Mining", In International Journal of Computer Science and Information Technologies, Vol .6(2), pp.1835-1840,2015.

[74] Anjana Patel & Prof Khyati Patel, "A Hybrid Approach in Privacy Preserving Data Mining", JARIIE-ISSN(O)-2395-4396, Vol-2, Issue-3, 2016.

[75] M. Naga Lakshmi and K. Sandhya Rani, "A Privacy Preserving Clustering Method Based on Fuzzy Approach And Random Rotation Perturbation", Publications of Problems & Application in Engineering Research-Paper, Vol. 04, Issue No. 1, ISSN: 2230-8547, E-ISSN: 2230-8555, PP: 174-177, 2013.

Chapter 13

Big Data Analytics – Tools and Techniques – Application in the Insurance Sector

Ayesha Banu

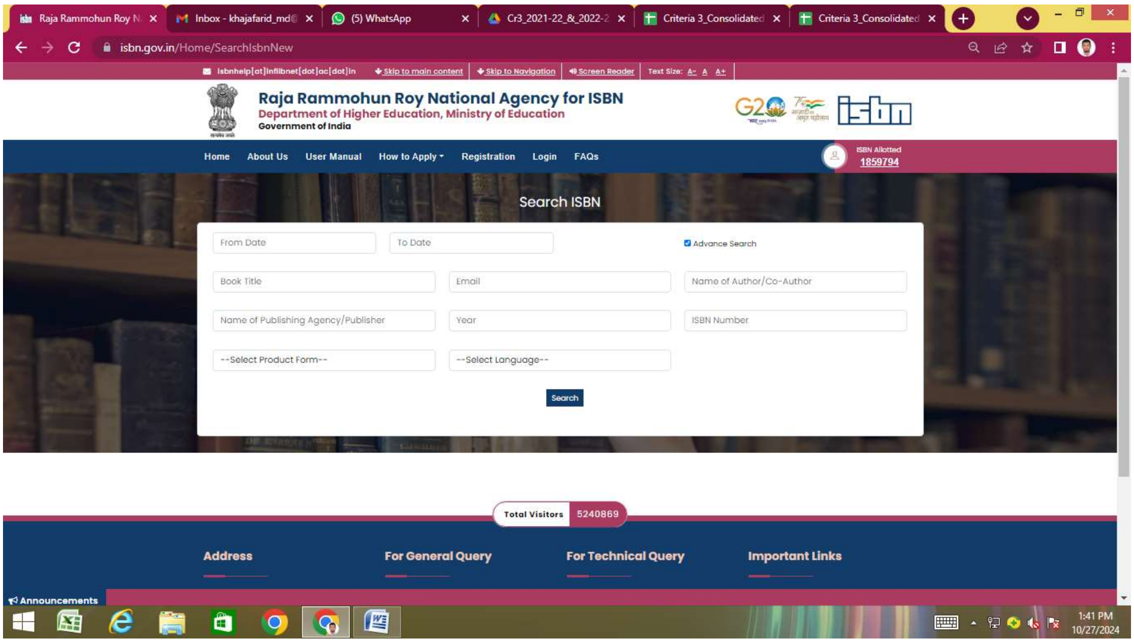
Abstract

Introduction: The Internet has tremendously transformed the computer and networking world. Information reaches our fingertips and adds data to our repository within a second. Big data was initially defined as three Vs, where data come with greater variety, increasing volumes and extra velocity. Big data is a collection of structured, unstructured and semi-structured data gathered from different sources and applications. It has become the most powerful buzzword in almost all the business sectors. The real success of any industry can be counted based on how the big data is analysed, potential knowledge is discovered and productive business decisions are made. New technologies such as artificial intelligence and machine learning have added more efficiency to storing and analysing data. This big data analytics (BDA) becomes more valuable to those companies, focusing on getting insight into customer behaviour, trends and patterns. This popularity of big data has inspired insurance companies to utilise big data at their core systems and advance the financial operations, improve customer service, construct a personalised environment and take all possible measures to increase revenue and profits.
Purpose: This study aims to recognise what big data stands for in the insurance sector and how the application of BDA has opened the door for new and innovative changes in the insurance industry.
Methodology: This study describes the field of BDA in the insurance sector, discusses the benefits, outlines tools, architectural framework, the method, describes applications in general and specific and briefly discusses the opportunities and challenges.
Findings: The study concludes that BDA in insurance is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however, there remain challenges to overcome.

Big Data: A Game Changer for Insurance Industry, 191–212

doi:10.1108/978-1-80262-605-620221013

Civil engineering



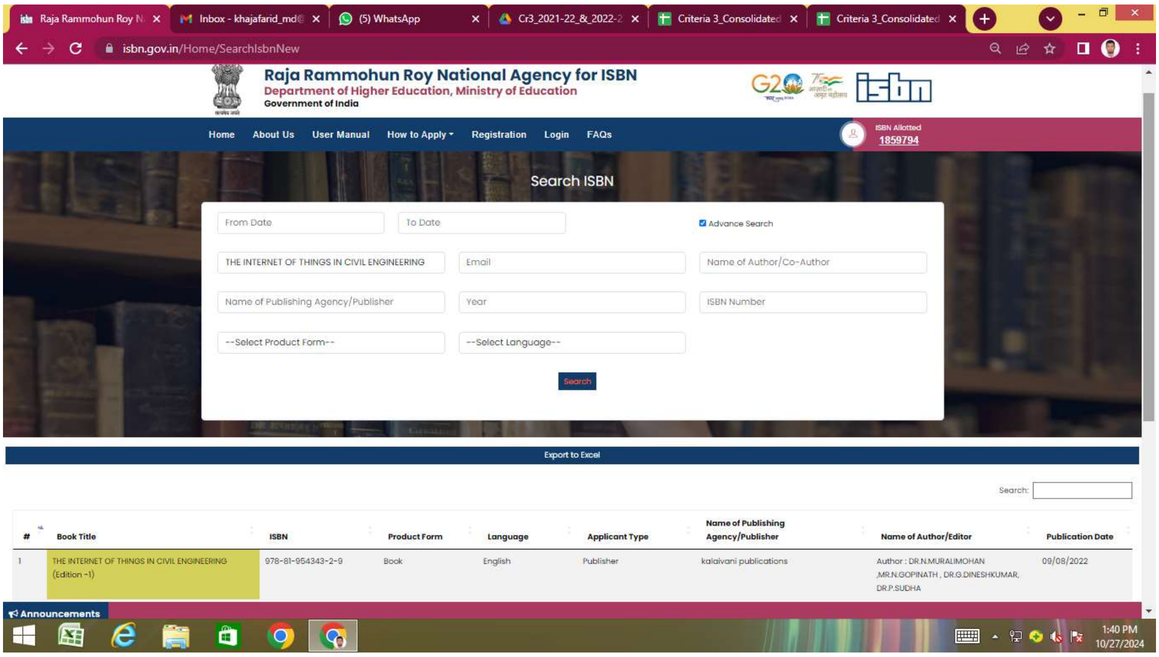
From Date | To Date | ☑ Advance Search

Book Title | Email | Name of Author/Co-Author

Name of Publishing Agency/Publisher | Year | ISBN Number

--Select Product Form-- | --Select Language--

2022-2023



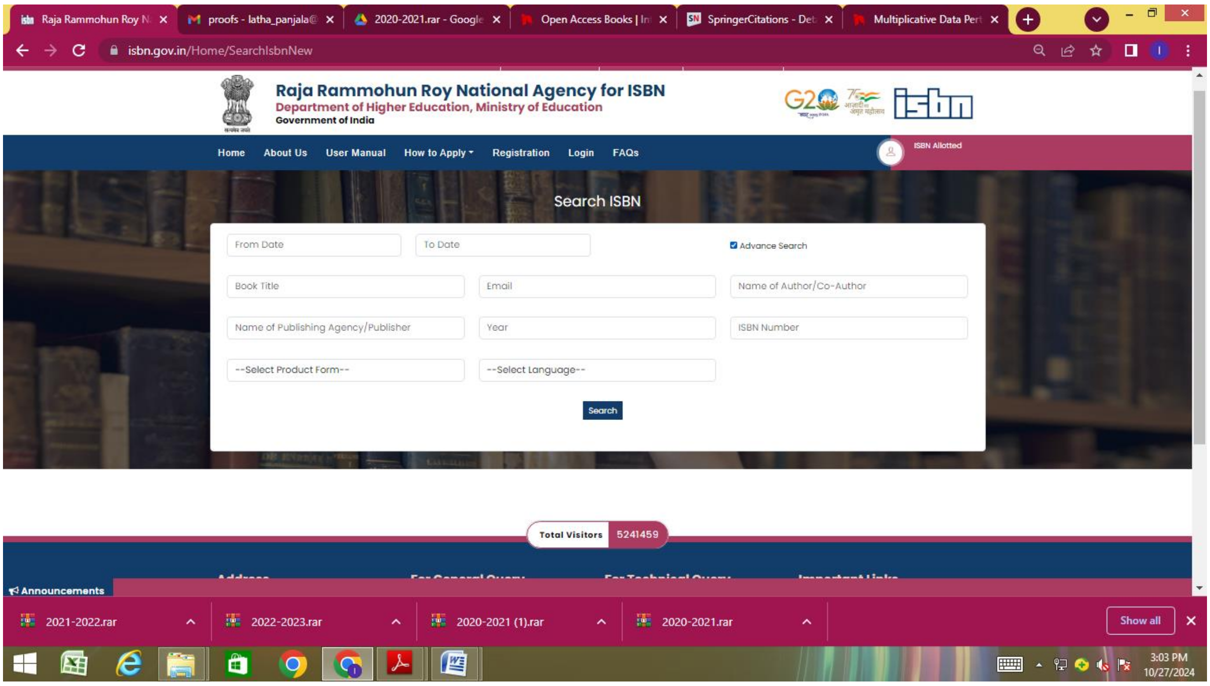
From Date | To Date | ☑ Advance Search

THE INTERNET OF THINGS IN CIVIL ENGINEERING | Email | Name of Author/Co-Author

Name of Publishing Agency/Publisher | Year | ISBN Number

--Select Product Form-- | --Select Language--

#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	THE INTERNET OF THINGS IN CIVIL ENGINEERING (Edition -1)	978-81-954343-2-9	Book	English	Publisher	kalaivani publications	Author : DR.N.MURALIMOHAN ,MR.N.GOPINATH , DR.G.DINESHKUMAR, DR.P.SUDHA	09/08/2022



Raja Rammohun Roy National Agency for ISBN
Department of Higher Education, Ministry of Education
Government of India

Search ISBN

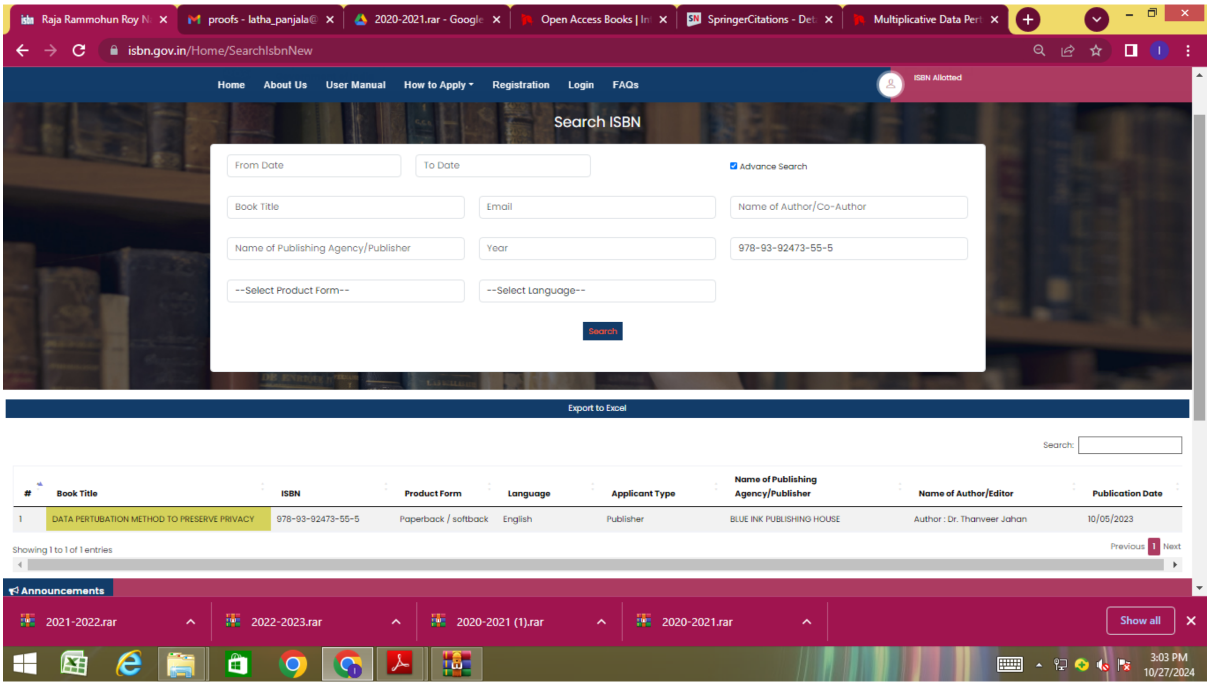
From Date | To Date | ☑ Advance Search

Book Title | Email | Name of Author/Co-Author

Name of Publishing Agency/Publisher | Year | ISBN Number

--Select Product Form-- | --Select Language--

Total Visitors 5241459



Search ISBN

From Date | To Date | ☑ Advance Search

Book Title | Email | Name of Author/Co-Author

Name of Publishing Agency/Publisher | Year | 978-93-92473-55-5

--Select Product Form-- | --Select Language--

Search:

#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	DATA PERTUBATION METHOD TO PRESERVE PRIVACY	978-93-92473-55-5	Paperback / softback	English	Publisher	BLUE INK PUBLISHING HOUSE	Author : Dr. Thanveer Jahan	10/05/2023

Showing 1 to 1 of 1 entries

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Subscribe Subscribe Cart Create Account Persc Sign

Browse My Settings Help Institutional Sign In

Institutional Sign In

All

ADVANCED SEARCH

Conferences > 2022 13th International Confe... ?

Performance Analysis of Windowing algorithms during Filtering of an Additive White Gaussian Noisy EMG Signal

Publisher: IEEE Cite This PDF

Hemant Kumar Gupta ; Neetu Gupta ; M. Shashidhar ; M.A.Himayath Shamshi All Authors •••

86
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Downl
PDF

Abstract:EMG signal has extensive application in analyzing muscular disorder. The bio-potential across the human muscles can be measured through Surface Electromyography. Real tim... **View more**

Metadata

Abstract:

EMG signal has extensive application in analyzing muscular disorder. The bio-potential across the human muscles can be measured through Surface Electromyography. Real time sEMG signal includes some artifacts like ECG, EEG signals and 50Hz power supply noise. To remove these artifacts and noise components filtering is needed. In this paper, different windowing functions like Hamming, Hanning and Rectangular window of FIR filter are applied on raw sEMG signals. The performances of these window techniques are analyzed based on power to average power ratio, signal to noise ratio, average power and EMG rejection ratio. An EMG data of a healthy person from MATLAB data base has been used as standard data base for processing. Since EMG data is random in nature so AWGN is mixed with the clean EMG data. The noisy signal is processed and comparison of parameters is performed. It is observed that rectangular window is most suitable windowing technique then hamming and hanning window.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 03-05 October 2022

DOI: 10.1109/ICCCNT54827.2022.9984628

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

▸ **ISBN Information:**

Conference Location: Kharagpur, India

Contents

I. Introduction

In current years the Electromyography signal processing has shown tremendous advantages to find the abnormalities in muscles or nerves [1]. EMG signals are analyzed to detect medical abnormalities. It is used to find the muscle disorder, nerve disorder and also helpful to find the disorders which affects the interconnection between muscles and nerves [2]. On the basis of methodology adopted to acquire, the EMG signals are broadly classified as intramuscular and surface EMG signals [3][4]. During Intramuscular Electromyography signals needles or wires are inserted inside the muscle for signal acquisition whereas in surface Electromyography (sEMG) sensors are placed over the muscle of the skin during muscle movement [5]. Surface EMG is a strategy used to acquire physiological understanding into muscle compression attributes [6]. sEMG signal is a portrayal of the mind boggling voltage changes that happen as activity possibilities instigate muscle compression [7]. Action potential signals can be estimated with a non-intrusive sEMG cathode connected to the muscle skin [1][6].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

More Like This

Modified 2D median filter for impulse noise suppression in a real-time system
IEEE Transactions on Consumer Electronics
Published: 1995

Adjustable flow control filters and reflective memories as support for distributed real-time systems
Second Workshop on Parallel and Distributed Real-Time Systems
Published: 1994

Show More

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE Account

- » Change Username/Password
- » Update Address

Purchase Details

- » Payment Options
- » Order History
- » View Purchased Documents

Profile Information

- » Communications Preferences
- » Profession and Education

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

Performance Analysis of Windowing algorithms during Filtering of an Additive White Gaussian Noisy EMG Signal

Publisher: IEEE

Cite This PDF

Hemant Kumar Gupta ; Neetu Gupta ; M. Shashidhar ; M.A.Himayath Shamshi All Authors •••

86
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Downl
PDF

Abstract:EMG signal has extensive application in analyzing muscular disorder. The bio-potential across the human muscles can be measured through Surface Electromyography. Real tim... **View more**

Metadata

Abstract:

EMG signal has extensive application in analyzing muscular disorder. The bio-potential across the human muscles can be measured through Surface Electromyography. Real time sEMG signal includes some artifacts like ECG, EEG signals and 50Hz power supply noise. To remove these artifacts and noise components filtering is needed. In this paper, different windowing functions like Hamming, Hanning and Rectangular window of FIR filter are applied on raw sEMG signals. The performances of these window techniques are analyzed based on power to average power ratio, signal to noise ratio, average power and EMG rejection ratio. An EMG data of a healthy person from MATLAB data base has been used as standard data base for processing. Since EMG data is random in nature so AWGN is mixed with the clean EMG data. The noisy signal is processed and comparison of parameters is performed. It is observed that rectangular window is most suitable windowing technique then hamming and hanning window.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 03-05 October 2022

DOI: 10.1109/ICCCNT54827.2022.9984628

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

▸ **ISBN Information:**

Conference Location: Kharagpur, India

Contents

I. Introduction

In current years the Electromyography signal processing has shown tremendous advantages to find the abnormalities in muscles or nerves [1]. EMG signals are analyzed to detect medical abnormalities. It is used to find the muscle disorder, nerve disorder and also helpful to find the disorders which affects the interconnection between muscles and nerves [2]. On the basis of methodology adopted to acquire, the EMG signals are broadly classified as intramuscular and surface EMG signals [3][4]. During Intramuscular Electromyography signals needles or wires are inserted inside the muscle for signal acquisition whereas in surface Electromyography (sEMG) sensors are placed over the muscle of the skin during muscle movement [5]. Surface EMG is a strategy used to acquire physiological understanding into muscle compression attributes [6]. sEMG signal is a portrayal of the mind boggling voltage changes that happen as activity possibilities instigate muscle compression [7]. Action potential signals can be estimated with a non-intrusive sEMG cathode connected to the muscle skin [1][6].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

More Like This

Modified 2D median filter for impulse noise suppression in a real-time system
IEEE Transactions on Consumer Electronics
Published: 1995

Adjustable flow control filters and reflective memories as support for distributed real-time systems
Second Workshop on Parallel and Distributed Real-Time Systems
Published: 1994

Show More

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

» Profession and Education

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Subscribe Subscribe Cart Create Account Persc Sign

Browse My Settings Help Institutional Sign In

Institutional Sign In

All

ADVANCED SEARCH

Conferences > 2022 13th International Confe...

Performance Analysis of Windowing algorithms during Filtering of an Additive White Gaussian Noisy EMG Signal

Publisher: IEEE Cite This PDF

Hemant Kumar Gupta ; Neetu Gupta ; M. Shashidhar ; M.A.Himayath Shamshi **All Authors** •••

86
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Abstract

Document Sections

Authors

Figures

References

Keywords

Metrics

More Like This

Downl
PDF

Abstract:EMG signal has extensive application in analyzing muscular disorder. The bio-potential across the human muscles can be measured through Surface Electromyography. Real tim... **View more**

Metadata

Abstract:

EMG signal has extensive application in analyzing muscular disorder. The bio-potential across the human muscles can be measured through Surface Electromyography. Real time sEMG signal includes some artifacts like ECG, EEG signals and 50Hz power supply noise. To remove these artifacts and noise components filtering is needed. In this paper, different windowing functions like Hamming, Hanning and Rectangular window of FIR filter are applied on raw sEMG signals. The performances of these window techniques are analyzed based on power to average power ratio, signal to noise ratio, average power and EMG rejection ratio. An EMG data of a healthy person from MATLAB data base has been used as standard data base for processing. Since EMG data is random in nature so AWGN is mixed with the clean EMG data. The noisy signal is processed and comparison of parameters is performed. It is observed that rectangular window is most suitable windowing technique then hamming and hanning window.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 03-05 October 2022

DOI: 10.1109/ICCCNT54827.2022.9984628

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

▸ **ISBN Information:**

Conference Location: Kharagpur, India

Contents

I. Introduction

In current years the Electromyography signal processing has shown tremendous advantages to find the abnormalities in muscles or nerves [1]. EMG signals are analyzed to detect medical abnormalities. It is used to find the muscle disorder, nerve disorder and also helpful to find the disorders which affects the interconnection between muscles and nerves [2]. On the basis of methodology adopted to acquire, the EMG signals are broadly classified as intramuscular and surface EMG signals [3][4]. During Intramuscular Electromyography signals needles or wires are inserted inside the muscle for signal acquisition whereas in surface Electromyography (sEMG) sensors are placed over the muscle of the skin during muscle movement [5]. Surface EMG is a strategy used to acquire physiological understanding into muscle compression attributes [6]. sEMG signal is a portrayal of the mind boggling voltage changes that happen as activity possibilities instigate muscle compression [7]. Action potential signals can be estimated with a non-intrusive sEMG cathode connected to the muscle skin [1][6].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

More Like This

Modified 2D median filter for impulse noise suppression in a real-time system
IEEE Transactions on Consumer Electronics
Published: 1995

Adjustable flow control filters and reflective memories as support for distributed real-time systems
Second Workshop on Parallel and Distributed Real-Time Systems
Published: 1994

Show More

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE Account

- » Change Username/Password
- » Update Address

Purchase Details

- » Payment Options
- » Order History
- » View Purchased Documents

Profile Information

- » Communications Preferences
- » Profession and Education

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Subscribe Subscribe Cart Create Account Perso Sign

Browse My Settings Help Institutional Sign In

Institutional Sign In

All

ADVANCED SEARCH

Conferences > 2022 13th International Confe...

Parametric Evaluation for Muscle Interfacing Devices during Finger Movement of Paralytic Patients

Publisher: IEEE

Cite This PDF

Hemant Kumar Gupta ; Neetu Gupta ; M. Shashidhar ; M. A. Himayath Shamshi **All Authors**

14
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Downl
PDF

Abstract: Electromyography (EMG) is a non-invasive and relatively inexpensive technique that can provide considerable quantitative information about muscle activation patterns. Ele... **View more**

Metadata

Abstract:

Electromyography (EMG) is a non-invasive and relatively inexpensive technique that can provide considerable quantitative information about muscle activation patterns. Electromyography signal has extensive application in analyzing paralysis or muscle twitching and muscular disorder. To record the EMG signal from paralytic patient small activities should be identified like single finger movements as well as combined finger movements. For muscle computer interface devices, which are an extensive technology of human computer interface (HCI), certain parameters i.e. average power of muscle and average power of muscle tissues must be calculated. For finding the muscle disorder EMG signal must have low level of noise and should be processed by filtration techniques. Filtering of EMG signal by various means may cause the filtration of EMG components with noise. EMG rejection ratio is a parameter which shows the amount of EMG signal filtration with noise components. The real time EMG signal has been acquired from sixty paralytic patients and processed by dc offset, rectification and filtering algorithms to form envelope of EMG signal. The computer based system presented in this thesis is capable to find the certain parameters of EMG signal which have extensive use in designing the muscle interfacing devices for paralytic patients.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 03-05 October 2022

DOI: 10.1109/ICCCNT54827.2022.9984252

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

▸ **ISBN Information:**

Conference Location: Kharagpur, India

Contents

I. Introduction

An electric voltage is generated during the contraction of muscles. These electric voltages can be measured by the EMG signals [1]. The contraction and relaxation of muscles is controlled by the neurological system [2][3]. Hence, the EMG signals are managed by the neurological system and depend on the physiological and anatomical properties of muscles. In neurological EMG, the response of artificial muscle is analyzed under static conditions, when the muscle is electrically stimulated [4][5][6]. In kinesiological EMG, the neuromuscular activation of muscles is studied when the muscle is under postural task, work condition, treatment and training regimes and functional movement [3].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

Loading [MathJax]/extensions/MathZoom.js

» Profession and Education

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

Loading [MathJax]/extensions/MathZoom.js

Parametric Evaluation for Muscle Interfacing Devices during Finger Movement of Paralytic Patients

Publisher: IEEE Cite This PDF

Hemant Kumar Gupta ; Neetu Gupta ; M. Shashidhar ; M. A. Himayath Shamshi **All Authors** •••

14
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Downl
PDF

Abstract: Electromyography (EMG) is a non-invasive and relatively inexpensive technique that can provide considerable quantitative information about muscle activation patterns. Ele... **View more**

Metadata

Abstract:

Electromyography (EMG) is a non-invasive and relatively inexpensive technique that can provide considerable quantitative information about muscle activation patterns. Electromyography signal has extensive application in analyzing paralysis or muscle twitching and muscular disorder. To record the EMG signal from paralytic patient small activities should be identified like single finger movements as well as combined finger movements. For muscle computer interface devices, which are an extensive technology of human computer interface (HCI), certain parameters i.e. average power of muscle and average power of muscle tissues must be calculated. For finding the muscle disorder EMG signal must have low level of noise and should be processed by filtration techniques. Filtering of EMG signal by various means may cause the filtration of EMG components with noise. EMG rejection ratio is a parameter which shows the amount of EMG signal filtration with noise components. The real time EMG signal has been acquired from sixty paralytic patients and processed by dc offset, rectification and filtering algorithms to form envelope of EMG signal. The computer based system presented in this thesis is capable to find the certain parameters of EMG signal which have extensive use in designing the muscle interfacing devices for paralytic patients.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

ng [MathJax]/extensions/MathZoom.js

Date of Conference: 03-05 October 2022

DOI: 10.1109/ICCCNT54827.2022.9984252

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

▶ **ISBN Information:**

Conference Location: Kharagpur, India

Contents

I. Introduction

An electric voltage is generated during the contraction of muscles. These electric voltages can be measured by the EMG signals [1]. The contraction and relaxation of muscles is controlled by the neurological system [2][3]. Hence, the EMG signals are managed by the neurological system and depend on the physiological and anatomical properties of muscles. In neurological EMG, the response of artificial muscle is analyzed under static conditions, when the muscle is electrically stimulated [4][5][6]. In kinesiological EMG, the neuromuscular activation of muscles is studied when the muscle is under postural task, work condition, treatment and training regimes and functional movement [3].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

More Like This

Human-computer interaction: The usability test methods and design principles in the human-computer interface design
2009 2nd IEEE International Conference on Computer Science and Information Technology
Published: 2009

Coarse-to-fine particle filters for multi-object human computer interaction
2009 IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
Published: 2009

Show More

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

Loading [MathJax]/extensions/MathZoom.js

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Subscribe Subscribe Cart Create Account Perso Sign

Browse My Settings Help Institutional Sign In

Institutional Sign In

All

ADVANCED SEARCH

Conferences > 2022 13th International Confe...

Parametric Evaluation for Muscle Interfacing Devices during Finger Movement of Paralytic Patients

Publisher: IEEE Cite This PDF

Hemant Kumar Gupta ; Neetu Gupta ; M. Shashidhar ; M. A. Himayath Shamshi **All Authors** •••

14
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Abstract

Document Sections

I. Introduction

II. Literature Review

III. Fundamental knowledge and proposed algorithms

IV. Results and Discussions

V. Conclusion and Future Scope:

Authors

Figures

References

Keywords

Metrics

More Like This

Downl
PDF

Abstract: Electromyography (EMG) is a non-invasive and relatively inexpensive technique that can provide considerable quantitative information about muscle activation patterns. Ele... **View more**

Metadata

Abstract:

Electromyography (EMG) is a non-invasive and relatively inexpensive technique that can provide considerable quantitative information about muscle activation patterns. Electromyography signal has extensive application in analyzing paralysis or muscle twitching and muscular disorder. To record the EMG signal from paralytic patient small activities should be identified like single finger movements as well as combined finger movements. For muscle computer interface devices, which are an extensive technology of human computer interface (HCI), certain parameters i.e. average power of muscle and average power of muscle tissues must be calculated. For finding the muscle disorder EMG signal must have low level of noise and should be processed by filtration techniques. Filtering of EMG signal by various means may cause the filtration of EMG components with noise. EMG rejection ratio is a parameter which shows the amount of EMG signal filtration with noise components. The real time EMG signal has been acquired from sixty paralytic patients and processed by dc offset, rectification and filtering algorithms to form envelope of EMG signal. The computer based system presented in this thesis is capable to find the certain parameters of EMG signal which have extensive use in designing the muscle interfacing devices for paralytic patients.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 03-05 October 2022

DOI: 10.1109/ICCCNT54827.2022.9984252

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

ISBN Information:

Conference Location: Kharagpur, India

Contents

I. Introduction

An electric voltage is generated during the contraction of muscles. These electric voltages can be measured by the EMG signals [1]. The contraction and relaxation of muscles is controlled by the neurological system [2][3]. Hence, the EMG signals are managed by the neurological system and depend on the physiological and anatomical properties of muscles. In neurological EMG, the response of artificial muscle is analyzed under static conditions, when the muscle is electrically stimulated [4][5][6]. In kinesiological EMG, the neuromuscular activation of muscles is studied when the muscle is under postural task, work condition, treatment and training regimes and functional movement [3].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

IEEE Account

- » Change Username/Password
- » Update Address

Purchase Details

- » Payment Options
- » Order History
- » View Purchased Documents

Profile Information

- » Communications Preferences
- » Profession and Education

Loading [MathJax]/extensions/MathZoom.js

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

Loading [MathJax]/extensions/MathZoom.js

Conferences > 2022 13th International Confe...

Implementation of Deep Learning Based Compression Technique and Comparative Analysis With Conventional Methodologies

Publisher: IEEE Cite This PDF

Neetu Gupta ; Hemant Kumar Gupta ; Vibhakar Pathak ; Prakash Pareek **All Authors** •••

3 Cites in Papers

61 Full Text Views

Alerts

Manage Content Alerts

Add to Citation Alerts

Abstract

Document Sections

Authors

Figures

References

Citations

Keywords

Metrics

More Like This

Downl PDF

Abstract: Recently, the movement in information science and advancement completely impacts the human ordinary everyday practice. The size of mechanized data is filling rapidly to a... **View more**

Metadata

Abstract:

Recently, the movement in information science and advancement completely impacts the human ordinary everyday practice. The size of mechanized data is filling rapidly to achieve images of unrivalled grade. The image compression works with to send large size images with insignificant bytes and to restore the image with incredible quality on social event. The focuses are to make an extra little data size; better nature of data during entertainment and connect transmission of data over confined information move limit with security. In this paper, conventional compression techniques i.e. Luminous DCT, Biorthogonal DWT, Quadtree Fractal and Huffman compression, are implemented and analyzed based on compression efficiency parameters like compression ratio, PSNR, MSE and SSIM. Here a deep learning based compression technique using stack autoencoder model is also proposed and implemented to compress the image data and result are compared with conventional compression techniques based on compression efficiency parameters. Simulation results show that proposed deep learning based compression scheme provides high quality reconstructed images having satisfactory compression ratio.

Published in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 03-05 October 2022 **DOI:** 10.1109/ICCCNT54827.2022.9984363

Date Added to IEEE *Xplore*: 26 December 2022

Publisher: IEEE

ISBN Information:

Conference Location: Kharagpur, India

Contents

I. Introduction

The transmission and accessing of information over telecommunication network and internet in form of multimedia is growing rapidly [1]. Since the images cover major part of the multimedia data and they use most of the bandwidth during transmission over internet. With the advancement of high resolution digital cameras, the storage and transmission of these high quality images is an important issue [2]. Since the high quality digital images of 512x512 pixels has 2,62,144 elements so these images require large memory space to store and higher bandwidth to transmit. Due to large size of these images the downloading time is also much higher. Therefore the compression of image data is very necessary before transmission of information [3].

Sign in to Continue Reading

Authors

Figures

References

Citations

Keywords

Metrics

More Like This

Fractal Image Compression based on Discrete Wavelet Transform
2022 OITS International Conference on Information Technology (OCIT)
Published: 2022

Automatic Detection of Early Esophageal Cancer from Endoscope Image Using Fractal Dimension and Discrete Wavelet Transform
2015 12th International Conference on Information Technology - New Generations
Published: 2015

Show More

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

» Profession and Education

- » Technical Interests

Need Help?

- » **US & Canada:** +1 800 678 4333
- » **Worldwide:** +1 732 981 0060
- » Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies

IEEE.org IEEE *Xplore* IEEE SA IEEE Spectrum More Sites Subscribe

Subscribe Cart Create Account Perso Sign

Browse My Settings Help Institutional Sign In

Institutional Sign In

All

ADVANCED SEARCH

Conferences > 2022 Second International Con...

Energy-Efficient Cooperative Communications System

Publisher: IEEE

Cite This PDF

Uppula Kiran ; Krishan Kumar **All Authors**

41
Full
Text Views

Alerts

Manage Content Alerts
Add to Citation Alerts

Abstract

Document Sections

I. Introduction

II. System Model

III. Average Energy Efficiency Analysis

IV. Results and Discussions

V. Conclusion

Authors

Figures

References

Keywords

Metrics

More Like This

Downl
PDF

Abstract: In a situation where all relays can pay attention to each other, that is, hidden relays are present, proposed energy efficiency and low-cost co-operation system. The anal... **View more**

Metadata

Abstract:

In a situation where all relays can pay attention to each other, that is, hidden relays are present, proposed energy efficiency and low-cost co-operation system. The analytical and simulation results monitor whether the proposed system works significantly better than direct transfers, high quality transfer options, all transfer options, and the current high-performance interaction system. In the current case the integrated slides with location visibility and a complete time-based selection option, in which the slides can sense the transmission and see a different modern region, the proposed efficient overhead transmission system is proposed, making the channel state information (CSI) response from the destination useful. To avoid potential conflict between relay transmissions during the best transfer selection, the provided method of relays is selected to accurately monitor the monitoring intervals before the transmission is proposed. in addition, the effect of today is a wide range of new relays, the number of selected modern relays and a collection of fashionable forward space in energy efficiency(EE) is being investigated. The results of the simulation indicate that the proposed joint venture system gains a higher EE than direct communication, satisfactory transfer options, and all the options available for group transfers.

Published in: 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)

Date of Conference: 21-22 April 2022 **DOI:** 10.1109/ICAECT54875.2022.9807900

Date Added to IEEE *Xplore*: 01 July 2022

Publisher: IEEE

ISBN Information:

Conference Location: Bhilai, India

Contents

I. Introduction

In order to reap the benefits of the overall operation of the joint venture, the choice of relay plays an important role. In the middle selection, one node is selected to act as a central controller that collects all the important facts and selects one or more transmissions to facilitate verbal exchange between the supply and destination. Important companies may be prominent: single transfer option schemes the use of a single node to facilitate negotiations and a few transfer option schemes, while a few transfer relays statistics in the destination [12] [16].

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

More Like This

Energy Efficiency of The Cooperative Communication Systems With Energy Harvested at Source and Relay Using Battery Power at Relay
2022 International Electronics Symposium (IES)
Published: 2022

Energy Efficiency in D2D Cooperative Communication System UAV-Assisted for Energy Harvesting Process at Source and Relay
2022 International Electronics Symposium (IES)
Published: 2022

Show More

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

Loading [MathJax]/extensions/MathZoom.js

» Profession and Education

VARDHAMAN COLLEGE OF ENGINEERING
VARDHAMAN
ESTD. 1999

AIP
American Institute of Physics

VARDHAMAN COLLEGE OF ENGINEERING

Department of Electronics and Communication Engineering

ICSVCE 2022

CERTIFICATE OF PARTICIPATION

This is to certify that

Uppula Kiran

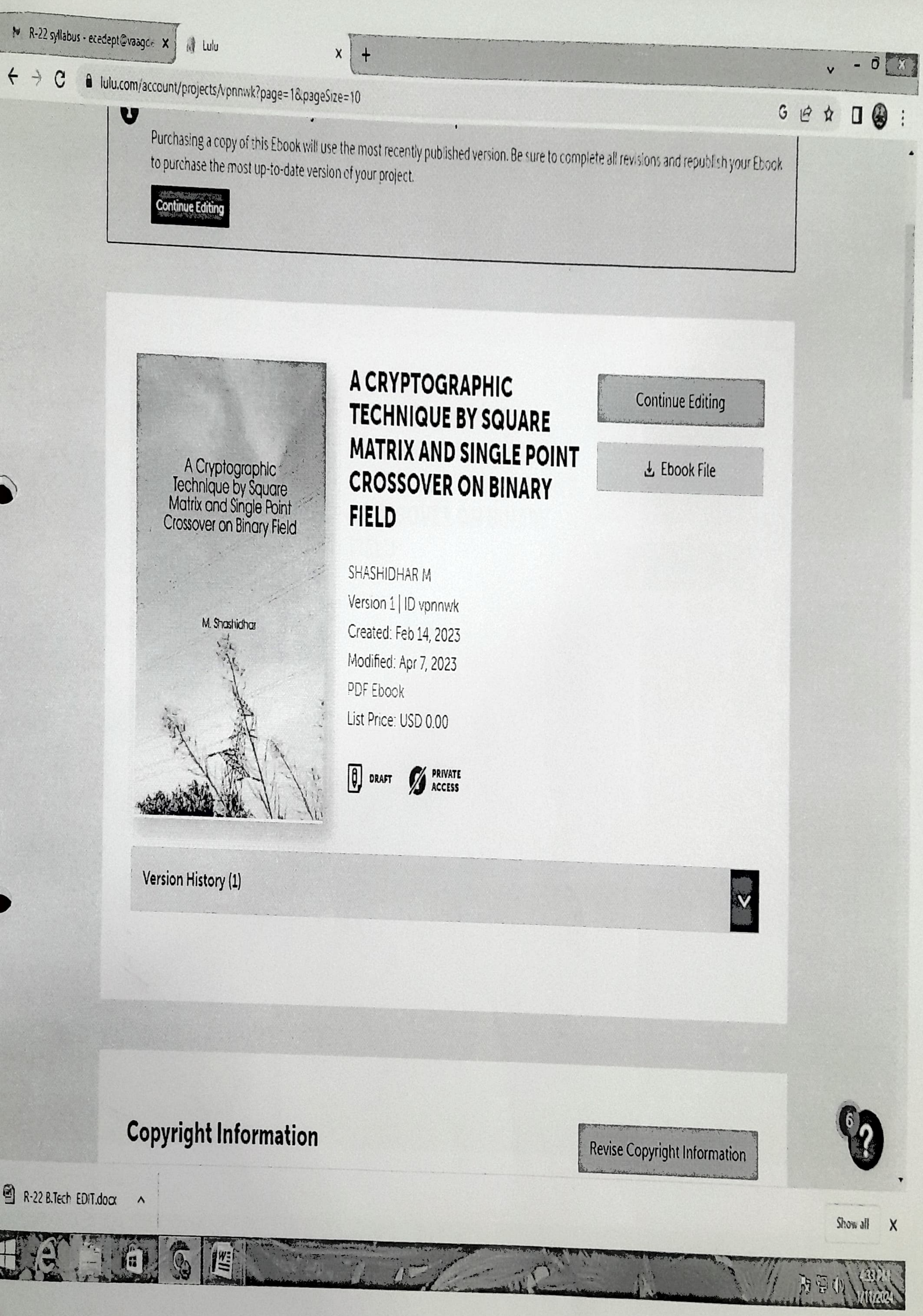
has presented paper entitled

An Optimized Relay Selection to Improve Reliability and Reduce Energy Consumption in Cooperative Networks

Paper Id: **4566** in International Conference on Advances in Signal Processing, VLSI, Communications and Embedded Systems (ICSVCE – 2022) organized by Department of Electronics and Communication Engineering, Vardhaman College of Engineering, Hyderabad during 29th – 30th July, 2022.

Dr. G. A. E. Satish Kumar
General Chair

Dr. J. V. R. Ravindra
General Chair


R-22 syllabus - ecedept@vaagc
Lulu
lulu.com/account/projects/vpnnwk?page=1&pageSize=10
Continue Editing
A Cryptographic Technique by Square Matrix and Single Point Crossover on Binary Field
M. Shashidhar
Continue Editing
Ebook File
DRAFT
PRIVATE ACCESS
Revise Copyright Information
R-22 B.Tech EDIT.docx
Show all

Purchasing a copy of this Ebook will use the most recently published version. Be sure to complete all revisions and republish your Ebook to purchase the most up-to-date version of your project.

A CRYPTOGRAPHIC TECHNIQUE BY SQUARE MATRIX AND SINGLE POINT CROSSOVER ON BINARY FIELD

SHASHIDHAR M

Version 1 | ID vpnnwk

Created: Feb 14, 2023

Modified: Apr 7, 2023

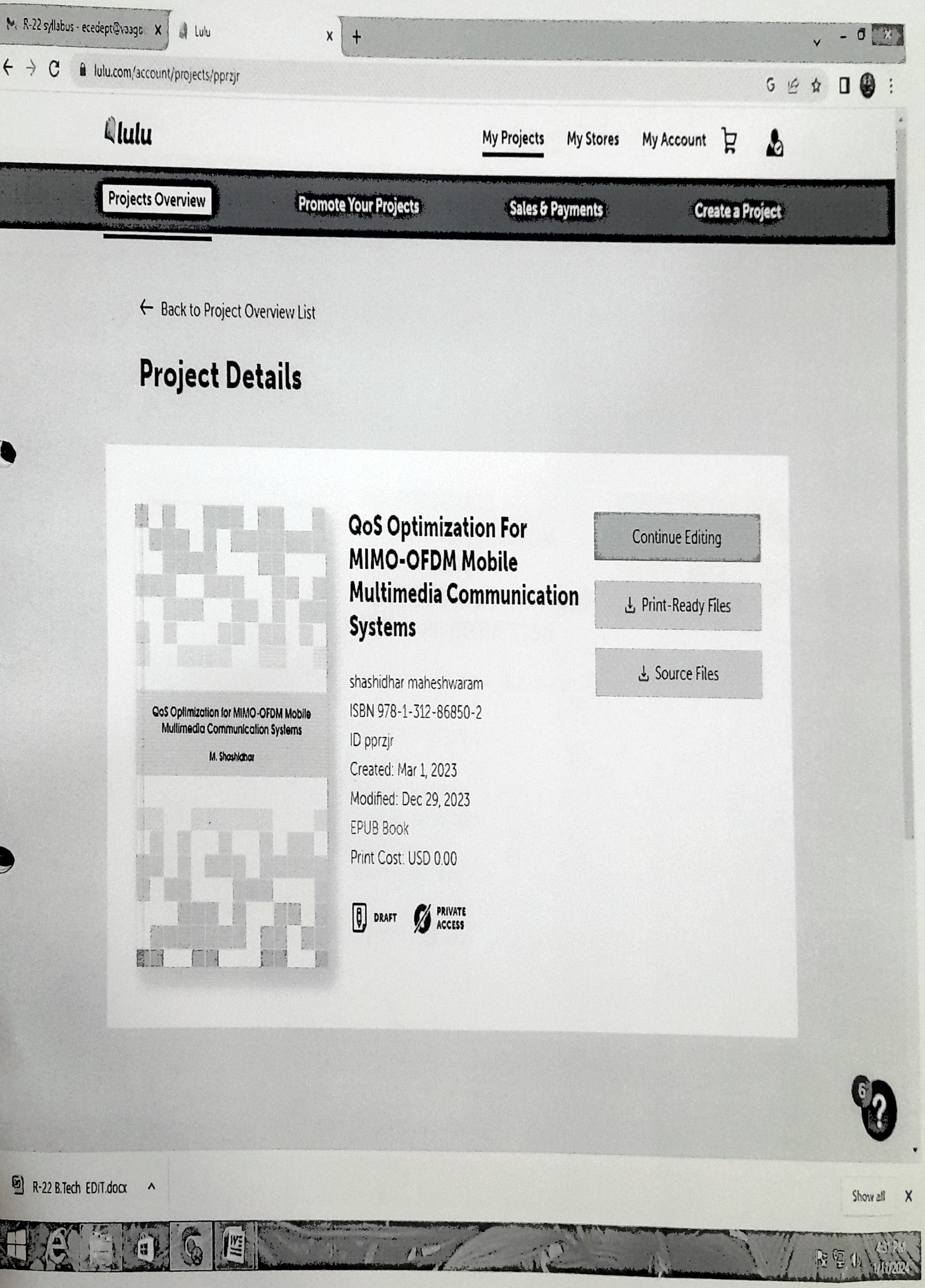
PDF Ebook

List Price: USD 0.00

DRAFT PRIVATE ACCESS

Version History (1)

Copyright Information



Project Details

QoS Optimization For MIMO-OFDM Mobile Multimedia Communication Systems

QoS Optimization for MIMO-OFDM Mobile Multimedia Communication Systems

M. Shashidhar

shashidhar maheshwaram

ISBN 978-1-312-86850-2

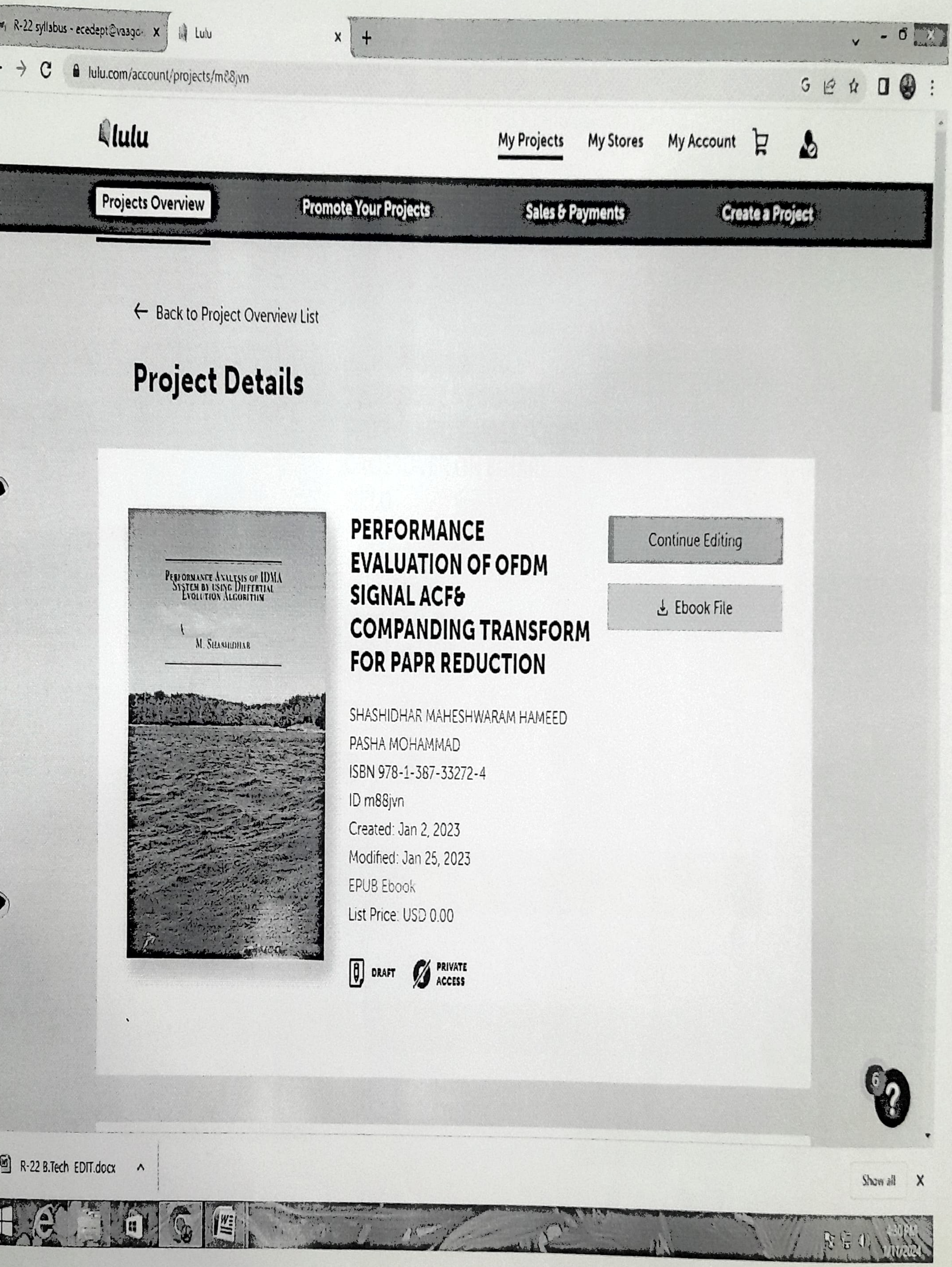
ID pprzjr

Created: Mar 1, 2023

Modified: Dec 29, 2023

EPUB Book

Print Cost: USD 0.00



Project Details

PERFORMANCE EVALUATION OF OFDM SIGNAL ACF& COMPANDING TRANSFORM FOR PAPR REDUCTION

SHASHIDHAR MAHESHWARAM HAMEED PASHA MOHAMMAD

ISBN 978-1-387-33272-4

ID m88jvn

Created: Jan 2, 2023

Modified: Jan 25, 2023

EPUB Ebook

List Price: USD 0.00

About the Authors

Dr.B.Srikanth working as an Assistant Professor in the Department of Electronics and Communication Engineering Vardhaman College of Engineering has about 12 years of teaching and 2 years of industry experience. He received his B.Tech degree in Electronics and Communication Engineering from JNTU Hyderabad and M.Tech degree in VLSI and Embedded Systems with distinction from Kakatiya University, Warangal. He received Ph.D degree in Electronics and Communication Engineering from Koneru Lakshmaiah University, Andhra Pradesh state. He has published 15 research papers in refereed international journals and 10 research papers in the proceedings of various international conferences. He has received several best paper awards for his research papers at various international conferences. His areas of research include low power VLSI design, arithmetic circuits, memory circuits, mixed signal VLSI design, FPGA. He is an active member of IEEE, IETE and ISTE.

Dr.M.Shashidhar working as an Professor in the Department of Electronics & Communication Engineering, Vaagdevi College of Engineering, has about 22 years of teaching experience. He received his B.Tech degree in Electronics & Communication Engineering with First class from JNTUH, Hyderabad and M. Tech. degree (Digital Communications) first class with distinction from Kakatiya University, Warangal. He received Ph.D. degree in Electronics & Communication Engineering from JNTUH University, Telangana state. He has published 25 research papers in refereed international journals and 16 research papers in the proceedings of various international conferences. He has received several best paper awards for his research papers at various international conferences. He received Ambitions award Hyderabad for education excellence "Best Teacher Award in ECE" in 2021. He received Dr. Sarvepalli Radha Krishna "Best Teacher Award" from Society for Learning Technologies, Vijayawada in 2021. His areas of research include wireless communication, Signal Processing, 5G Communications. He is an active member of IETE and ISTE.

Mr.D.Sreenivasulu Reddy worked as an Assistant Professor in the Department of Electrical and Electronics Engineering, Sree Vidyanikethan Engineering College, has about 12 years of teaching experience. He received his B.Tech. degree in Electrical and Electronics Engineering with distinction from JNTUH and M.Tech. degree in Power Systems Engineering with distinction from Sree Vidyanikethan Engineering College, Tirupati. He is doing as Research Scholar for Ph.D degree in CHRIST (DEEMEND TO BE UNIVERSITY), BANGALORE. He has published 10 research papers in refereed International journals and 13 research papers in the proceedings of various International conferences. His areas of research include Power Distributed Generations, Electrical Vehicles, Voltage Stability of electric power grids, Renewable energy systems and Microgrid. He is an active member of IEEE.

Dr.K.Kalaiselvan working as an Professor in the Department of Electronics and Communication Engineering, Roever Engineering College, Perambalur has about 16 years of teaching experience. He received his B.E degree in Electronics and Communication Engineering and M.E. degree in Power Electronics and Drives with distinction from Anna University, Chennai. He received Ph.D degree in Power Electronics and Drives from Anna University, Tamilnadu state. He has published 14 research papers in refereed International journals and 20 research papers in the proceedings of various international conferences. He has received several best paper awards for his research papers at various international conferences. His areas of research include Multilevel inverter, Voltage regulators, and power grids, Renewable energy systems and E vehicle. He is an active member of IEEE.

Infinite Research
Registered under MSME
Government of India
https://www.infinite-research.org
info@infiniteresearch.org
infiniteresearchorg@gmail.com
+91 7995871450, +91 8919552865

ISBN 978-81-965046-4-9
9 788196 504649

DIGITAL SYSTEM DESIGN

Digital System Design

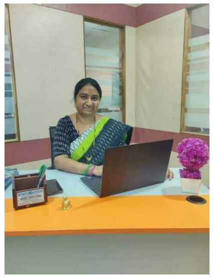
Dr. B. Srikanth | Dr. M. Shashidhar
Mr. D. Sreenivasulu Reddy | Dr. K. Kalaiselvan

Association Rule Mining (ARM) in data mining provides quality association rules based on measures such as support and confidence. These rules are interpreted by domain experts for making well-informed decisions. However, there is an issue with ARM when the dataset is subjected to changes from time to time. Discovering rules by reinventing wheel, scanning entire dataset every time in other words, consumes more memory, processing power and time. This is still an open problem due to proliferation of different data structures being used for extracting frequent item sets. An algorithm is proposed for update of mined association rules when dataset changes occur. The proposed algorithm outperforms the traditional approach as it mines association rules incrementally and dynamically updates mined association rules.

Satyavathi Nedendla
Balakrishna Eppakayala
Rama Abbidi

Mining Association Rules from Incremental Data set

Incremental Mining



Dr.N.Satyavathi working as a Head of the Department ,CSE in an Prestigious Institution. Having 17 Years of Experience in Teaching. Published various research papers in various Scopus journals and International Conferences. Continuing research in the area of datamining.

9 786204 980942

LAP LAMBERT Academic Publishing

Satyavathi Nedendla
Balakrishna Eppakayala
Rama Abbidi

Mining Association Rules from Incremental Data set

Satyavathi Nedendla
Balakrishna Eppakayala
Rama Abbidi

Mining Association Rules from Incremental Data set

Incremental Mining

LAP LAMBERT Academic Publishing

Cover image: www.ingimage.com

Publisher:
LAP LAMBERT Academic Publishing
is a trademark of
Dodo Books Indian Ocean Ltd., member of the OmniScriptum S.R.L Publishing group
str. A.Russo 15, of. 61, Chisinau-2068, Republic of Moldova Europe
Printed at: see last page
ISBN: 978-620-4-98094-2

ABSTRACT

Association Rule Mining (ARM) in data mining provides quality association rules based on measures such as support and confidence. These rules are interpreted by domain experts for making well-informed decisions. However, there is an issue with ARM when the dataset is subjected to changes from time to time. Discovering rules by reinventing wheel, scanning entire dataset every time in other words, consumes more memory, processing power and time. This is still an open problem due to proliferation of different data structures being used for extracting frequent item sets. We proposed an algorithm for update of mined association rules when dataset changes occur. The algorithm is known as FIN-INCRE which exploits the Preorder Coded Tree used by FIN algorithm for fast item set mining. The proposed algorithm outperforms the traditional approach as it mines association rules incrementally and dynamically updates mined association rules.

Following are important outcomes of this research

i) Fast mining of frequent item sets.

ii) Incremental mining of frequent item sets in case of data insertions.

iii) Incremental mining of frequent item sets in case of data deletions.

iv) Incremental mining of frequent item sets in case of support change.

INDEX

Chapter 1 : Introduction

Association rule mining is the process of finding patterns, associations and correlations among sets of items in a database. The association rules generated have an antecedent and a consequent. An association rule is a pattern of the form X & Y ⇒ Z [support, confidence], where X, Y, and Z are items in the dataset. The left hand side of the rule X & Y is called the antecedent of the rule and the right hand side Z is called the consequent of the rule. This means that given X and Y there is some association with Z. Within the dataset, confidence and support are two measures to determine the certainty or usefulness for each rule. Support is the probability that a set of items in the dataset contains both the antecedent and consequent of the rule or $P (X \cup Y \cup Z)$. Confidence is the probability that a set of items containing the antecedent also contains the consequent or $P (Z| X \cup Y)$. Typically, an association rule is called strong if it satisfies both a minimum support threshold and a minimum confidence threshold that is determined by the user.

Association rule mining has been around in data mining. It is widely used data mining technique which retrieves patterns or association rules from large databases. The discovered trends or patterns can provide business intelligence. Association rule is the rule which tells the association between data objects that gives the latent relationships between different attributes in given dataset. Association rule mining in fact enables to discover rules that can help in making well informed decisions. The discovery of rules is further made more qualitative using statistical measures such as support and confidence. High quality rules can help in making more accurate decisions. Association rule mining is done in two phases. In the first phase frequent item sets are generated.

In the second phase rules are generated based on the support and confidence requirements. Association rule mining has many applications in the real world. They include medical diagnosis, GIS, relational database, large database and distributed databases etc. An important research area with association rule mining is to improve underlying data structure to improve the efficiency of ARM process. Another important problem identified in the literature is that when new data is added, the present ARM algorithms are mining the rules from the scratch again. This is time consuming, and involves unnecessary processing. Therefore, this thesis is aimed at proposing a novel algorithm that can use more efficient data structure and also support association rule mining dynamically and incrementally avoiding unnecessary mining process.

1.1 Background

ARM has been a persistent topic in the domain of data mining for number of years. Plentiful research is found on ARM and it proved its utility. Agrawal, Imielinsky & Swami proposed the first algorithm called AIS [1] for association rule mining. Later Agrawal &Srikant proposed two new algorithms called Apriori and AprioriTID [2]. These algorithms outperformed prior algorithms. They proposed a new algorithm called AprioriHybrid by combining the features of Apriori and AprioriTID.

Based on the principle of Apriori many other algorithms like DHP [21], Partitioning [25], DIC [20], and ECLAT [31]. Drawbacks of Apriori are: 1). More number of database scans. 2). Candidate generation process. To overcome these drawbacks, Pattern-Growth algorithm called FP-Growth [13] is proposed for association rule mining. FP-Growth is an association rule mining method without candidate item set generation. Later many Pattern-Growth association rule mining methods are proposed: H-MINE [22], ParticiaMine [23], RElim [5], PPV [9], Prepost [10], NSFI [4], FIN (Zhi-hong & Sheng-long, 2014) [32] for association rule mining.

The proposed algorithms: Apriori-based and tree-based are best suitable mining association rules in case of static dataset. However, when the database is subjected to update (new tuples are inserted/old tuples are deleted/modified), the discovered association rules must be updated because some of the old rules may become uninteresting and the rules which are considered as uninteresting before database updates may become interesting after database updates. Hence there is a need of some techniques which involves incremental maintenance of association rules that does not use entire database (Original part plus updated part) but uses only updated part of the database to maintain association rules.

Hence based on the Apriori principle, many algorithms were developed for incremental association rule mining: FUP and FUP2 (Cheung, 1996) [6] Algorithms utilizing negative borders [27], UWEP [3], DELI [16], MAAP [33]. These algorithms are still like Apriori in which many number of candidate item are generated and also scanning of entire database is required to mine incremental association rules.

To overcome these drawbacks tree-based incremental association rule mining algorithms: DB-tree & PotFP-tree [12], FELINE [8], FOLD-GROWTH [30], AFPIM [15], IFP-Growth [28], AFOPT [19], CAN tree [17], EFPIM [18], CP-tree [26], FUFP-tree [14], BIT-FP-Growth [29] were developed. HUI-list-INS (2015) [34] Incremental mining algorithm to handle transaction insertions is developed. It generates frequent item sets without candidate generation method and uses a data structure called utility list.

A new Incremental Relational Association Rule Mining (IRARM) [11] approach for mining interesting relational association rules is developed. A system [24] is developed for incremental mining. Original database is represented in the form of COMVAN tree and frequent item sets are mined using COMVAN tree. Many algorithms have been proposed for incremental mining. But still these algorithms suffer from the following drawbacks:

- Required to scan original database many times.
- Works only in case of insertions.
- Works only in case of Deletions.
- Doesn't work in case of support change.
- Data structure used in mining is not efficient in terms of time complexity/space complexity.

Hence a there is need for an algorithm for ARM, which overcomes the drawbacks of existing algorithms, must be developed.

1.2 Problem Definition

Association Rule Mining (ARM) in data mining provides quality association rules based on measures such as support and confidence. These rules are interpreted by domain experts for making well-informed decisions. However, there is an issue with ARM when the dataset is subjected to changes from time to time. Discovering rules by reinventing wheel, scanning entire dataset every time in other words, consumes more memory, processing power and time. This is still an open problem due to proliferation of different data structures being used for extracting frequent item sets. Therefore, provided a dataset D with generated association rules R, the problem of updating R incrementally when D is subjected to modifications and deletions. The R also needs to be updated with support measure's threshold is changed. This is the challenging problem considered in this research.

1.3 Scope of the Research

The scope of the research encompasses the proposal of faster frequent item set mining algorithm based on POC tree and generate Association Rules. Association rules are to be updated incrementally when the original database is subjected to modifications and deletions or when support threshold for ARM is changed. It is achieved by defining FIN-INCRE algorithm that performs incremental update of association rules without scanning entire database each time.

1.4 Research Objectives

The aim of the research is to investigate the present state of the art on incremental association rule mining and propose a framework for automatic generation and update of association rules effectively. To achieve the aim of this research, the following objectives are conceived.

1. To investigate and review literature to explore the existing mechanisms available for incremental update of association rules in data mining domain.
2. To identify an algorithm that generates frequent item sets faster besides helping in the creation of association rules.
3. To propose an algorithm that not only considers dynamic generation of association rules but update them in response to change drivers such as arrival of new transactions, deletion of existing transactions and change in the support measure.
4. To realize a complete framework for automatic generation and update of association rules incrementally and evaluate the performance the proposed algorithms.

1.5 Research Contributions

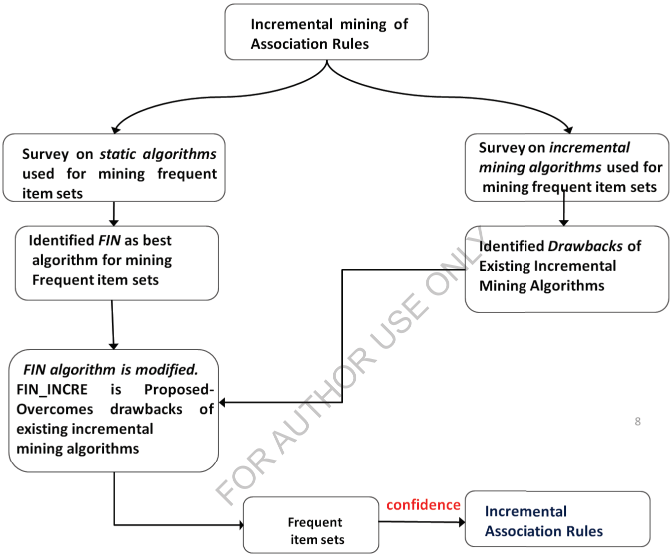
This research is aimed at proposing a framework for fast frequent item set mining and generation of incremental, high quality and association rules. It explores state of the art in this area besides proposing necessary algorithms for efficient mining of association rules incrementally. As per data dynamics, the rules are automatically updated with the underlying algorithms in the proposed framework.

The contributions of the thesis are as follows.

- Literature review has been made to ascertain useful insights into state-of-the-art approaches for incremental association rule mining. This has resulted in understanding about the existing algorithms used for dynamically generating association rules and updating them from time to time.
- A fast item set mining algorithm by name FIN is identified. This algorithm uses a data structure for holding data in tree format. The data structure is named as POC tree. This will help in faster navigation of entities and thus help in generating item sets faster. It plays crucial role when incremental association rule mining is implemented.
- An incremental association rule mining with novel approach is proposed. It is encapsulated in the proposed algorithm known as FIN-INCRE which not only generates association rules faster but also responds to change drivers such as arrival of new transactions, deletion of existing transactions and changes in the support. These factors cause the algorithm to have incremental generation and update of association rules instead of generating rules from the scratch. Thus it leverages the state of the art in terms of speed, efficiency and reduction of overhead.
- A prototype application is built to demonstrate proof of the concept. The application shows the utility of the proposed FIN-INCRE algorithms for effectiveness in generation of association rules and dynamically updating the rules when change drivers are triggered. The application provides intuitive interface to show the utility of the proposed framework.

1.6 Overview of Research work carried out

The overview of the research carried out is described here. It is based on the objectives provided in section 5. Keeping the aim of the research in mind, various aspects of incremental mining of association rules are explored considering need for efficiency. Towards this end, a methodology is proposed as shown in Figure 1.6 which outlines the research carried out in this thesis.



Incremental mining of Association Rules
Survey on static algorithms used for mining frequent item sets
Survey on incremental mining algorithms used for mining frequent item sets
Identified FIN as best algorithm for mining Frequent item sets
Identified Drawbacks of Existing Incremental Mining Algorithms
FIN algorithm is modified. FIN_INCRE is Proposed- Overcomes drawbacks of existing incremental mining algorithms
8
Frequent item sets
confidence
Incremental Association Rules

Figure 1.6: Overview of the research carried out

Association rule mining provides rules that help in making expert decisions. They reflect patterns or customer behaviour in the given data corpus. First, a fast item set mining algorithm named FIN is identified from the literature which will help in fast mining of frequent item sets. Afterwards, a novel algorithm known as FIN-INCRE is proposed to have frequent item sets for incremental dataset that reduce processing overhead and time taken. Various metrics are used to evaluate the results of the empirical study. Thus a complete framework is realized for dynamic update of association rules. The proposed framework shows improved performance over the state of the art.

1.7 Organization of the Synopsis

The research carried out in this Synopsis is organized into several chapters. The essence of these chapters is provided here.

Chapter 1 covers introduction to the research, state of the art, problem definition, motivation, scope of the research, research objectives, methodology and research contributions.

Chapter 2 reviews relevant literature pertaining to the study area. It covers review of literature on present state of the art pertaining to dynamic update of association rules, different means of improving association rule generation through various approaches employed for frequent itemset mining.

Chapter 3 presents a fast itemset mining algorithm. The methodology used includes both frequent itemset mining as well as association rule generation.

Chapter 4 presents a novel approach in incremental mining of interesting and high quality association rules. It throws light into different aspects of incremental mining of association rules such as the need for incremental mining and the drivers of such mining phenomenon. It provides the functioning of the proposed FIN-INCRE algorithm that automatically updates association rules when new transactions are added, existing transactions are deleted and when support needed for generating association rules is changed. FIN-INCRE takes care of the three change drivers to enhance efficiency in generating association rules.

Chapter 5 presents results of empirical study and evaluates the results. It throws light into the datasets used for experiments, experimental setup and evaluation methodology used for arriving at performance comparison between the proposed method and state of the art. The outcomes of the evaluation led to conclusions made.

Chapter 6 concludes the research besides providing directions for possible future scope of the research. From the research carried out on the incremental association rule mining, conclusions are made.

Chapter 2 : Literature Review

The problem of Association rule mining is to find out association rules that satisfy the predefined measures: support & Confidence. Association Rule mining process involves two steps: 1) Finding frequent item sets that satisfy predefined minimum support. Frequent item sets are defined as the set of items that frequently occur together in the given transaction. 2) Generating association rules that satisfy predefined minimum confidence. Association rule mining process is shown in the figure 2.

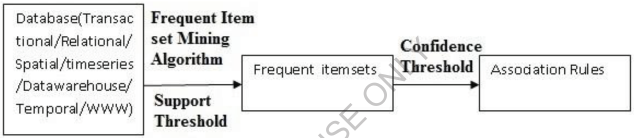

Database(Transactional/Relational/Spatial/timeseries/Datawarehouse/Temporal/WWW)
Frequent Item set Mining Algorithm
Support Threshold
Frequent itemsets
Confidence Threshold
Association Rules

Figure 2: The process of Association Rule Mining

Generating association rules from frequent item sets is a direct method. Hence, most of the researchers concentrated on developing an efficient algorithm for finding frequent item sets.

Initially algorithms that work on static databases were developed. There is an issue with ARM when the dataset is subjected to changes from time to time. Discovering rules by reinventing wheel, scanning entire dataset every time in other words, consumes more memory, processing power and time. Hence Incremental Association Rule Mining Algorithms came into existence.

2.1 Static Association Rule Mining algorithms

In 1994, Agrawal and Srikant presented an efficient algorithm called Apriori for mining frequent item sets. Later many algorithms based on

Apriori have been developed. Apriori algorithm uses candidate set generation approach to mine frequent item sets. Candidate set generation approach is costly, when database contains more number of transactions. Hence to overcome this approach a Pattern-Growth mining method called “FP-Growth” was proposed. The main idea of Pattern-Growth Algorithms is to use divide and conquer approach for mining frequent item sets, compress the given database by using some efficient data structure and then continue mining frequent item sets from the compressed structure. Later many approaches based on FP-Growth are developed for mining frequent itemsets.

Overview of each apriori-based association rule mining algorithms and Pattern-Growth algorithms is presented in subsections.

2.1.1 Apriori-based Association Rule Mining Algorithms

Apriori algorithm is found to be efficient in mining frequent item sets. Many improvements have been made to the Apriori and many other algorithms have developed with little modifications. Such algorithms are discussed below.

A. Apriori-Tid Algorithm

Apriori-TID [2] is a variant of the Apriori algorithm. It was proposed by Rakesh Agrawal and Ramakrishnan Srikant in the same article in which Apriori is proposed. A method used by AprioriTid is different when compared with Apriori but outputs produced by the two algorithms are same.

B. Apriori-Hybrid Algorithm

Apriori-Hybrid algorithm [2] is also proposed by Rakesh Agrawal and Ramakrishnan Srikant in the same article in which Apriori is proposed. In the earlier passes Apriori works better than Apriori-Tid but in later passes Apriori-Tid works better than Apriori. Hence, a hybrid algorithm called Apriori Hybrid can be designed that works as Apriori in the earlier passes and as Apriori-Tid in later passes. Apriori-Hybrid is very efficient in case of large databases.

C. Direct Hashing and Pruning Algorithm

This Direct Hashing and Pruning [21] is proposed by Jong Soo Park, Ming- Syan Chen and Philip S.Yu. The following procedure is followed in this technique: The algorithm finds frequent 1- item sets by scanning the given database. At the same time, possible frequent 2-item sets are generated and hashed them to a hash table. Now, this hash table is used to reduce the number of candidate item sets and to find the final set of frequent item sets.

D. Dynamic Item Set Counting Algorithm

Dynamic item set counting [20] is developed by Sergey Brin, Rajiv Motwani, Jeffry D.Ulman and Shalom Tsur. The basic algorithm is as follows: Given database is divided into k number of partitions. In the first partition, start counting the 1-item sets. Next, in the second partition, start counting 1-itemsets and also start counting the 2-item sets by making use of 1-itemsets obtained from the first partition. Now in the third partition, start counting the 1-item sets and also start counting 2-item sets, 3-item sets by using the results obtained from first and second partitions.

E. Partition Algorithm

Partition algorithm [25] was proposed by Ashok Savasere, Edward Omiecinski, Shamkant Navathe.The main idea behind the algorithm is it divides the given database into equal size partitions and apriori property is used for each partition and finds large item sets in each partition. To get the final set of frequent item sets for the given database performs union operation on large item sets obtained from each partition. The efficiency of the algorithm depends on database size, partition size and number of local large item sets generated.

F. Eclat (Equivalence CLAss Transformation) MaxEclat, Clique, MaxClique, TopDown and ApprClique algorithms

Zaki. M [31] proposed six efficient algorithms : Eclat (Equivalence CLAss Transformation), MaxEclat, Clique, MaxClique, TopDown and ApprClique for fast mining of frequent item sets.

The main idea of these algorithms is as follows:

1) Frequent 1- item sets and frequent 2-item sets are computed by using a vertical tid-list database format and frequent 2-item sets are computed by simply performing tid-list intersections.

2) On the set of frequent 2-item sets, sub lattices are generated by using any one of the relation: prefix-based equivalence relation or maximal-clique-based pseudo equivalence relation.

3)As a final step to generate frequent item sets, these sub lattices are processed one at a time by making use of any one of the search procedure: Bottom-up/Top-Down/Hybrid.

2.1.2 Pattern-Growth Association Rule Mining Algorithms

Pattern algorithms are developed to overcome the drawback of Apriori & Apriori-based algorithms. The main idea of Pattern-Growth Algorithms is to use divide and conquer approach for mining frequent item sets, compress the given database by using some efficient data structure and then continue mining frequent item sets from the compressed structure.

Pattern Growth algorithms are efficient and scalable when compared to Apriori-based algorithms. First developed Pattern Growth algorithm is “FP-Growth”. Later many algorithms are developed which are discussed in next section.

A. FP-Growth Algorithm

FP-Growth algorithm [13] mines frequent item sets without candidate generation approach. The main idea of FP-Growth is as follows: 1) constructs FP-tree (Frequent pattern tree) for the given database. Header Table is maintained for the FP-tree. 2) Consider items from the Header Table (from the last item), find conditional pattern base, construct conditional pattern tree for each item.3) Finally Frequent item sets are mined from the conditional pattern tree.

B. H-MINE

The main idea of HMine algorithm [22] is, for the given database HStruct structure is constructed. Mining is performed on HStruct structure to find frequent item sets. H-mine is efficient and scalable at mining very large databases.

C. Patricia mine

A compressed patricia trie [23] is used to represent the database. Patricia trie is a modification of standard trie. To find the frequent item sets, patricia trie is traversed using a technique called “item-guided traversal”.

D. RElim

This algorithm is similar to H-mine.RElim [5] finds frequent items, starting with a pre-processing step and then recursively eliminates least frequent item. RElim uses a “linked list” data structure in the process of mining frequent item sets.

E. PPV

PPV[9] is a vertical mining algorithm. It uses a tree structure called PPC-tree(Preorder Postorder Coded-tree) to represent the database. The tree is traversed using preorder and postorder techniques and each node in the tree is assigned a pre-order and post-order code, based on which, each frequent item can be represented by Node-list.

F. Prepost

PrePost [10] uses a data structure called "N-List" for representing frequent items. The difference between PPV and Prepost is,PPV uses candidate generation approach for mining frequent item sets but Prepost mines frequent item sets without candidate generation approach.

G. NSFI (N-list and subsume based algorithm for mining Frequent item sets)

It is an enhanced version of Prepost algorithm. This algorithm creates N-List using a hash table and uses an enhanced N-intersection algorithm for mining frequent item sets. NSFI algorithm [4] outperforms Prepost in terms of runtime and memory usage.

H. FIN (Fast mining of frequent item sets using Node-sets)

FIN[32] algorithm is used for fast mining of Frequent item sets.FIN approach is: Given database is pre-processed and infrequent items are removed .POC- tree (Pre order coded tree) is constructed for the pre-processed data. From which Frequent-1 item sets and frequent 2-iem sets are determined. Set Enumeration tree is constructed for each frequent 2-item set to find frequent k-item sets.

2.2 Incremental Association Rule Mining algorithms

Incremental ARM has gained significance in rendering data mining service for more efficient means of obtaining Knowledge from large databases. The aim of it is to ensure the generation of rules based on the newly added records and update the existing rules in fairly less amount of time. Though it is incremental mining of rules, it needs to consider already generated to have updated rules. Of late many algorithms came into existence to leverage ARM process in that way. They are broadly categorized into Apriori-based methods and Pattern-Growth ones.

2.2.1 Apriori-based Incremental Mining Algorithms

There are many incremental ARM algorithms based on Apriori mechanism were developed.

A. FUP Algorithm

Fast UPdate (FUP) [6] is the algorithm that exhibits Incremental ARM. It supports incremental update of mined association rules in case of new records insertion. It can update the rules based on the changes made to database incrementally. The algorithm contains much iteration and each iteration generates a candidate set, subjected to the frequent item sets already mined in the previous iteration.

B. FUP2 Algorithm

This algorithm is proposed by Cheung et al. [6], an extension to the existing FUP algorithm. As mentioned above FUP is able to generate association rules incrementally when latest transactions are added or existing ones are deleted. The FUP2, on the other hand, supports incremental ARM in case of both new record insertions and deletion of existing records.

C. Algorithm Utilizing Negative Borders

This algorithm [27] makes use of the notion of negative borders. Thus it can improve the performance of FUP-based algorithms. First of all, it generates frequent item sets related to an increment in the database D'. It goes for a full scan of the entire database D only when an item set is outside the negative border. In such cases only it needs one scan of the entire database. Its drawback is that the size of candidate dataset may be increased as it considers negative border closure.

D. *DELI (*Difference Estimation for Large Item sets)

An efficient algorithm [16] called DELI is proposed for an incremental ARM. When database update occurs, DELI uses a sampling technique to decide whether it is necessary to generate a new set of association rules or not. If an estimate is small, it treats an old set of rules are a good approximation to the new set of rules. It waits until more changes are made to the database and DELI algorithm is applied again. If the estimate is large then the FUP2 algorithm is applied to generate a new set of rules. DELI finds to be more efficient than FUP2.

E. UWEP (Update with Early Pruning)

This algorithm [3] is another kind where incremental ARM takes place. However, it uses a property known as early pruning. It has an advantage over FUP-based algorithm as it is capable of pruning the supersets of originally mined frequent item set. When D is the dataset, the UWEP algorithm prunes as and when needed and does not wait until k-the iteration. This can improve its performance significantly. The rationale behind this is that early pruning can avoid unnecessary processing of certain records while focusing only on the incremental updates.

F. MAAP and PELICAN

MAAP [33] generates frequent itemsets of large size depending on the mined itemsets that are frequent. If an item set denoted as k-itemset is frequent, its subsets are also added to frequent item sets denote as L' according to the Apriori property. It results in the reduction of computational complexity. Afterward, other frequent itemsets are identified based on the concept of level-wise item set generation. The MAAP and PELICAN algorithms are closer to FUP2 but they aim at maintaining minimum frequent itemsets as the database is subjected to changes from time to time. MAAP computes maximum frequent item sets by using Apriori property while the PELICAN does the same using lattice decomposition and vertical database format. As these two makes use of maximum frequent item sets only, they are able to reduce space and time complexity while mining association rules incrementally.

2.2.2 Pattern-Growth Incremental Mining Algorithms

Many pattern-Growth algorithms were developed which are used for incremental mining of Association Rules.

A. DB-tree Algorithm

This algorithm [12] is another kind where incremental ARM takes place. However, it uses a property known as early pruning. It has an advantage over FUP-based algorithm as it is capable of pruning the supersets of originally mined frequent item set. When D is the dataset, the UWEP algorithm prunes as and when needed and does not wait until k-the iteration. This can improve its performance significantly. The rationale behind this is that early pruning can avoid unnecessary processing of certain records while focusing only on the incremental updates.

B. PotFP-tree Algorithm

PotFP-tree [12] follows different approach when compared to that of DB-tree algorithm. It stores only potentially frequent items apart from the frequent 1-itemsets. It uses the notation of tolerance and a new parameter ***t*** denoting tolerance to determine a potentially frequent item.

As a matter of fact, FP-tree is a subset of PotFP-tree and also DB-tree for that matter. Therefore FP-tree is projected from either of them before extracting frequent item sets.

C. FELINE

CATS stand for Compressed and Arranged Transaction Sequence [8]. CATS tree has certain similarities with the features of FP-tree. Another important fact is that the DB-tree and the CATS tree are almost identical as they store all items irrespective of whether they are frequent. This seemingly significant feature makes the CATS tree right candidate to avoid re-scans of original database when incremental updates are needed. However, the way it is constructed differs from that of DB-tree and FP-tree. Ordering of global supports is employed in FP-tree while ordering local supports is employed by CATS-tree with respect to all frequent items in the path. From the original database CATS tree is built and it supports traversal in both directions in order to extract frequent item sets. But the construction of CATS tree is a complicated process and suitable for static databases in general.

D. CAN Tree (Canonical – Order Tree)

Originally CAN tree [17] is equipped with all tuples of given database. The items in the constructed tree are arranged in canonical order or alphabetical. In other words it supports an alphabetical order that has its utility in the data retrieval. Once the decision is made on the kind of ordering, that is followed by the tree while making subsequent changes in response to the changes made in the database. Then FP-growth kind of algorithm can be employed on the tree to generate frequent item sets. When database is subjected to frequent changes, CAN tree is an ideal candidate to reflect it. Neither it needs rescan of the original database fully nor CAN the reconstruction of a new tree in order to support incremental ARM. However, it needs more space in case of large CAN tree that also consumes more time for generating frequent item sets.

E. CP-tree (FP Growth algorithm*)*

It is the novel tree structure that is known as Compact Pattern Tree (CP-tree) [26]. It is convenient to capture data from database incrementally as database undergoes changes from time to time. Thus it achieves a structure known as frequency-descending structure. This tree is built in two steps. In the first step, inserts given transactions into the CP-tree which is according to the sort order that prevails with I-list. Then the frequency count is updated from time to time. The second step is for reconstruction. It is responsible to rearrange frequency-descending order of items besides updating the tree nodes to reflect new I-list. Alternative execution of these two phases makes the construction of CP tree and update of the same. CP-tree outperforms its predecessors in terms of memory consumption and execution time though construction process is complex and the cost of adjusting nodes in tree is high.

F. BIT_FP Growth algorithm

This algorithm is based on FP-tree. The BIT algorithm [29] performs merging of consecutive FP-trees in order to have final representative FP-tree that reflects the data present in the database. In case of large databases, the BIT algorithm is found to reduce execution time when it is compared with the performance of other incremental ARM algorithms. Its construction of tree takes less time though the merging process is complex and consumes more memory. Therefore, this algorithm is best used for batch processing.

G. FEEPAMT algorithm

An algorithm FEEPAMT [35] is proposed for mining incremental

Association Rules and also reduces system complexity in terms of time and space. To mine the association rules: For the given database, Canonical Ordered tree with Multiple Values Node tree or COMVAN [16] tree is constructed. Frequent item sets are generated from COMVAN tree.

Chapter 3 : Proposed System and Methodology

An incremental algorithm known as FIN_INCRE is proposed for discovering frequent item sets which eliminates drawbacks of existing algorithms.The proposed algorithm works well in the following cases:

1) If new transactions are inserted in to the original database.
2) If any transactions are deleted from original database.
3) If support change occurs.

3.1 Conceptual Overview of Proposed Methodology

The methodology used to generate association rules incrementally is presented in Figure 3.1. It is conceptual and provided with a simple dataset for comprehending the methodology.

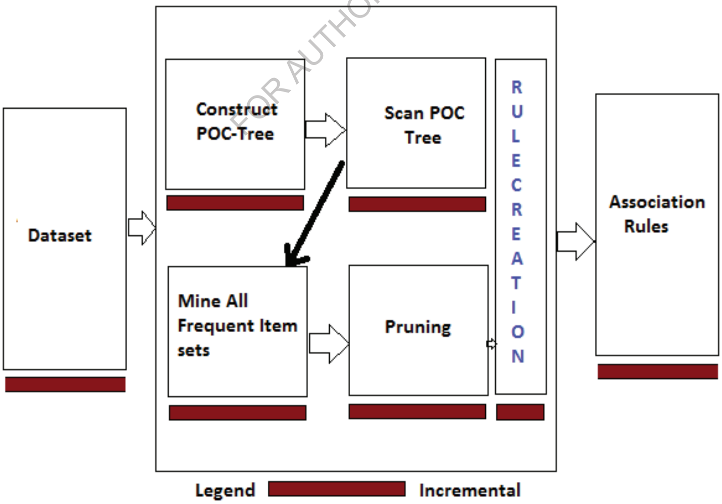


Figure 3.1: Conceptual overview of the proposed methodology

As presented in Figure 3.1, it is evident that initially POC tree for original transactional dataset is generated. Then frequent item sets are generated from the POC tree. Afterwards, the proposed algorithm named FIN_INCRE considers original POC tree and the incremental dataset and updates the POC tree and then finds frequent item sets for the incremental dataset without undergoing the entire ARM process again. Thus the proposed approach updates association rules rather than creating from the scratch. When the dataset is subjected to new records, then the whole methodology becomes incremental in nature.

3.2 Steps involved in Proposed Approach

The following steps are involved in the proposed approach:

Step 1: Scan original database, construct POC-tree and frequent 1-item sets.

Step 2: Traverse the tree using pre-order traversal technique and find frequent 2-item sets.

Step 3: Construct Set Enumeration tree for each frequent 2-item set and find frequent k-item sets.

Step 4: Using confidence measure, generate association rules for the frequent item sets in step 3.

Step 5: When database update occurs, new POC-tree is constructed by scanning only updated portion of the database:

Case 1: If old transactions are deleted from the original database:

Step a: Calculate new minimum support threshold value.

Step b: Find frequent and infrequent items in deleted records.

Step c: (i) Find items which are frequent in both original and deleted records, and find whether they are still frequent. If they are still frequent, update original POC-tree otherwise delete those items from original POC-tree.

(ii)Find items which are frequent in original and infrequent in deleted records, and find whether they are still frequent. If they are still frequent, update original POC-tree otherwise delete those items from POC-tree.

(iii) Find items which are infrequent in both original and deleted Records, and find whether they are frequent. If they are frequent in updated database, add those items at the end of the Original POC-tree.

Step d: An Updated POC-tree is generated.

Case 2: If new transactions are inserted in the original database:

Step a: Calculate new minimum support threshold value.

Step b: Find frequent and infrequent items in inserted records.

Step c: (i) Find items which are frequent in both original and infrequent in inserted records and find whether they are still frequent in updated database. If some of them are not frequent in update database, delete those items from original POC-tree otherwise add those items to “add to tree list”

(ii)Find items which are frequent in both original inserted records,and find whether they are still frequent. If they are still frequent in updated database, add those items to “add to tree” list. Otherwise delete those items from original POC-tree

(iii) Find items which are infrequent in both original and deleted records, and find whether they are frequent. If they are frequent, add those items at the end of the original POC-tree.

(iv) Find items which are infrequent in original and frequent in deleted records and find whether they are frequent in updated database. If they are frequent, add those items to “add to tree” list.

Step d: Sort the items in “add to tree” list and update original POC-tree to generate Updated POC-tree.

Case 3: If Minimum support threshold value is changed:

Case (i): If new support value > old support value, some of the infrequent items may become frequent, so identify such items and insert them into the original POC-tree to generate updated POC-tree.

Case (ii): If new support value < old support value, some of such items and delete them from the original POC-tree to generate updated POC-tree.

Step 6: Original POC-tree and updated POC-tree are compared to generate updated frequent item sets.

Step 7: Only for the new frequent item sets Association Rules are generated.

Step 8: These rules are combined with the original rules to form a set of updated rules for the updated database.

Chapter 4 : Proposed Algorithms

An incremental algorithm known as FIN_INCRE is proposed for discovering frequent item sets. In other words, it provides association rules incrementally when aforementioned changes are made to database or support threshold.

4.1 Algorithm FIN-INCRE

This is the main algorithm which is meant for incremental mining of association rules.

Input: Transactional Database, Minimum support Threshold MS.

Output: Frequent k-item sets.

Steps:

1. Calculate Support value (SV) from predefined MS.
2. Sv= (n*MS)/100.

Where n=Number of transactions in D,

MS=Minimum Support threshold specified in percentage.

Sv=Minimum support value.

3. Scan the given database and get support 'S' of each item, LoIinOg.
4. Add the items to list of Frequent items 'F1' or 'FinOg' if its S>=Sv, otherwise add to 'IFinOg'
5. Delete the items from given database, if its S<Sv.
6. Sort each transaction in descending order of their 'S'.
7. Construct POC-Tree (using **POC TREE Generation algorithm**) for the sorted database.
8. For each item in F1 find Nodesets.

   Nodeset of an item= {S, PC}

```
//where S=support of an item, PC=Preorder code of item.
  9. Scan the POC-Tree to get frequent 2-itemsets.
         a. For each item ‘i’ in POC-tree
              // Consider the items based on their Pre order code.
           i.  Find 2-itemsets registering the item ‘i’.
                    2-itemsets= {$ix_1$, $ix_2$, $ix_3$,..$ix_n$}
                  //where $x_1$, x2, x3, xn are ancestors of Item ‘i’.
          ii.  Repeat above step for all the items in POC-Tree to
               get complete set of 2-item sets, f2.
         b. Get each item from list f2 check its ‘S’, if it is >=Sv
               then add the 2-itemset to F2.
                //where F2 is a set of frequent 2-itemset.
         c. For each frequent 2-itemset say ‘xy’ in F2,
               findNodeset. To find Nodeset of ‘xy’:
           i.  Find Nodeset of ‘x’ ,Say (1,2)
          ii.  Find Nodeset of ‘y’ ,Say (3,4)
         iii.  Find whether ‘1’ is ancestor of ‘3’.If yes then
               Nodeset of ‘xy’= (3, 4).
          iv.  Repeat the step (iii) for all the items in F2 to get
               final set of Nodesets.
  10.      For each item in F2, construct set enumeration tree
  11.      For each itemset in set enumeration tree, find Nodeset
           to determine whether it is frequent or not.
  12.      If it is frequent add the itemset to list of frequent k-item
           sets.
  13.      Otherwise don’t add it to list frequent k-item sets and
           also ignore its children in next consideration.
  14.      Repeat the steps 11 to 13, for all the item sets in set
           enumeration tree.
```

```
15.     Repeat step 10 to 14 for all the items in F2 to get
        complete set of frequent k-items, $F_k$.
16.     $F_k=F_kUF_1UF_2$.
17.     If the given database is updated:
            Case 1: If new transactions are inserted then call
                UPOCinINS algorithm to  generate POC-Tree for
                the Updated database.
            Case 2: If some of the transactions are deleted then call
                UPOCinDEL algorithm to generate POC-Tree for
                the updated database.
            Case 3: If Minimum support value is changed then call
                UPOCinSup algorithm to  generate POC-Tree for
                the updated database.
18.     Once POC-Tree is generated for the updated database,
    repeat steps 9 to 16 to find complete set frequent k-item sets
    for the updated database.
```

Algorithm 4.1: Proposed FIN_INCRE algorithm for incremental ARM

As presented in Algorithm 4.1, the FIN algorithm is used for mining frequent item sets for the original database D by using steps 1 to 16 in **FIN_INCRE** Algorithm, which is described as follows: 1). Original database is represented in the form of a POC-Tree . 2). Frequent 1-items & Frequent 2-items are determined from POC-Tree. 3). For each frequent 2-itemset *set enumeration tree* is constructed. 4). Frequent k-item sets are generated using Set enumeration tree.

When new transactions are added/old transactions are deleted/user specified support changes, the item which is frequent may become infrequent and infrequent item sets may become frequent. The proposed incremental mining algorithm finds the items which became infrequent after adding new transactions, deletes them from Original

POC-Tree and also finds items which became frequent after adding new transactions and adds them to original POC-Tree. In the process of updating POC-Tree, item deletion is done before item insertion. In this way POC-Tree is updated and the performance of proposed algorithm is greatly improved.

4.2 Algorithm POC-TREE Generation

This is the algorithm used to generate POC tree which is the underlying structure for faster generation of item sets.

Input: Sorted Database D.

Output: POC-Tree.

Steps:

1. Initially root node of a POC-Tree is "null".
2. Get the first transaction from D' and create a first branch (with each node corresponds to an item in the transaction and also item's frequency is shown, which is 1 for first transaction) of POC-Tree.
3. Get the next transaction from D' and insert into the tree using any common prefix path that may appear.
4. Repeat step 3 until all transactions in D' are processed.
5. Traverse the tree (which is obtained in step 4) using preorder traversal technique and assign preorder code to the nodes in the tree while traversing.
6. As a result of step 5 final POC-Tree will obtain.

Algorithm 4.2: POC-Tree generation

As presented in Algorithm 4.2, it takes sorted database as input and constructs POC-tree that will be used for generating frequent item sets.

4.3 Algorithm UPOCinINS:

This algorithm is used to have support for incremental update of association rules by returning POC-tree for the updated database when some transactions are updated.

```
Input: Original Database D, LoIinOg, F1, IFinOg, MS, Newly added
transactions and Original POC-Tree
Output: POC-Tree for the updated database.
Steps:
    1. Scan new transactions, get the items and their support (S)
       'LoIinNew'
    2. Calculate MSI.
    3. MSI= (m*MS)/100.
         //Where MSI=Minimum support threshold for Inserted
                         transactions,
                     m=number of transactions newly inserted,
                     MS=Minimum support threshold defined for
                         original database.
    4. Find the List 'FinNew' and 'IFinNew'
         // where FinNew=List of Frequent items in new transactions
            &  IFinNew=List of Infrequent items in new transactions.
         a. Read an item from 'LoIinNew'.
         b. If its S >=MSI then add it 'FinNew' list.
         c. Otherwise add it to 'IFinNew' list.
         d. Repeat steps a to c until 'LoIinNew' becomes empty.
    5. Find the list 'LoIinUp' by merging 'LoIinOg' & 'LoIinNew'.
       //Where 'LoIinUp'=List of Items in Updated database.
```

```
6. Calculate MSU.
7. MSU=(n*MS)/100
     //Where MSU=Minimum Support threshold for Updated database,
             MS=Minimum threshold defined for Original database,
             n=Total number of transactions in Updated database.
8. Find the List FinUp&IFinUp
     //FinUp&IFinUP is the list of Frequent & Infrequent items in Updated database respectively.
     a. Read an item from 'LoIinUp'.
     b. If its S >=MSU then add it 'FinUp' list.
     c. Otherwise add it to 'IFinUp' list.
     d. Repeat steps a to c until 'LoIinUp' becomes empty.
9. To update POCtree, process the following steps:
        Case I:
             a.  $CI= FinOg \cap FinNew$
             //where CI contains list of items which are frequent in both Original and newly added transactions.
             b.  Read an item say 'x' from CI and search for 'x' in 'FinUp', if it returns true then add it to 'AddtoTree' list (which will be processed later).
             //'AddtoTree' contains list of items which are frequent in original, in newly added transactions and also in Updated database.
             c.  Otherwise delete it from Original POC-Tree.
```

```
            d.  Repeat steps b & c until the list CI becomes empty.
    Case II:
            a.  CI= FinOg ∩ IFinNew
            //where CI contains list of items which are frequent in Original database and infrequent in newly added transactions.
            b.  Read an item say ‘x’ from CI and search for ‘x’ in ‘FinUp’, if it returns true then add it to ‘AddtoTree’ list (which will be processed later).
                //’AddtoTree’ contains list of items which are Frequent in original, infrequent in newly added transactions but frequent in updated database.
            c.  Otherwise delete it from Original POC-Tree.
            d.  Repeat steps b & c until the list CI becomes empty.
    Case III:
            a.  CI= IFinOg ∩ FinNew
            //where CI contains list of items which are infrequent in Original database and frequent in newly added  transactions.
            b.  Read an item say ‘x’ from CI and search for ‘x’ in ‘FinUp’, if it returns true then add it to ‘AddtoTree’ list and also to Re_examine list (which will be processed later).
            //’AddtoTree’ contains list of items which are infrequent in original, frequent in newly added transactions and frequent in Updated database.
```

```
                c. Otherwise delete it from Original POC-Tree.
                d. Repeat steps b & c until the list CI becomes
                   empty.
                e. Transaction Number corresponding to the
                   items present in 'Re_examine' list is added
                   to 'Re_examineTrans'list (Which will be
                   processed later).
      i.   Get a transaction number from the list
           'Re_examineTrans' read the corresponding
           Transactions from 'D' and add it Original POC-Tree.
      ii.  Repeat step (ii) until all the values in 'Re_examineTrans'
           list is processed.
      iii. Read all the transactions corresponding to items
           present in 'AddtoTree' list and insert them into Updated
           POC-Tree which is obtained in step (ii).
      iv.  At the end of step (iv) Updated POC-Tree will be
           generated.
```

Algorithm 4.3: UPOCinINS algorithm

As presented in Algorithm 4.3, the **UPOCinINS** is used to update POC-Tree in case of insertions. Inputs for the algorithm are: List of items in Original Database, Frequent 1-itemsets, infrequent item sets in the original database, newly added transactions, Original POC-Tree.

Scan newly added transactions and get the items present in new transactions and also their supports (step 1). Calculate minimum support in newly added transaction (using steps 2 & 3). Find items which are frequent and also infrequent in new transactions (step 4). Find the list of items along with their support in updated database. Calculate minimum support for updated database and then find List of frequent & infrequent items in updated database (steps 3 to 6).

To update POC-Tree the following procedure is followed:

1. Items which are frequent in both original & inserted records are identified and tested whether they are still frequent in updated database. If such item exists then they are added to “addtoTree” list, otherwise the items are deleted from POC-Tree (step 7 – case (i)).
2. Items which are frequent in original database & infrequent in newly inserted records are identified and tested whether they are still frequent in updated database. If such items exist, then they are added to “addtoTree” list. Otherwise the items are deleted from POC-Tree (step 7 – case (ii)).
3. Items which are infrequent in original database & frequent in newly inserted records are identified and tested whether they are still frequent in updated database. If such items exist, then they are added to “addtoTree” list and also to “Re_examine” list. Transaction number corresponds to the items present in “Re_examine” lists are determined and are added to “Re_examineTrans” list. Read all the transactions present in “Re-examineTrans” list and “addtoTree” list, insert them into POC-Tree (step 7- case (iii)).
4. Finally updated POC-Tree will be obtained.

4.4 Algorithm UPOCinDel:

This algorithm is used to have support for incremental update of association rules by returning POC-tree for the updated database when transactions are deleted from the database.

Input: D, LoIinOg, FinOg, IFinOg, MS, d and original POC-Tree
D= Original Database,
LoIinOg=List of Items in Original Database,

```
FinOg=Frequent items in original database,
IFinOg=Infrequent items in Original database,
 MS=Minimum support threshold,
d=List containing transaction numbers that are to be deleted
Output: POC-Tree for the updated database.
Steps:
   1. Read transactions corresponding to the transaction numbers
      of the list d.
   2. Get the items and their support (S) and prepare a list
      'LoIinDel'.
   3. Calculate MSD.
   4. MSD= (m*MS)/100.
      //Where MSD=Minimum support threshold for Deleted
                   transactions,
                m=number of transactions that are to be deleted,
               MS=Minimum support threshold defined for
                   original database.
   5. Find the List 'FinDel' and  'IFinDel'
        // where FinDel=List of Frequent items in deleted
                 transactions
                 IFinDel=List of Infrequent items in deleted
                       transactions.
        a. Read an item from 'LoIinDel'.
        b. If its S >=MSD then add it 'FinDel' list.
        c. Otherwise add it to 'IFinDel' list.
        d. Repeat steps a to c until 'LoIinDel' becomes empty.
   6. 'LoIinUp' = LoIinOg - LoIinDel.
     //Where 'LoIinUp'=List of Items in Updated database.
   7. Calculate 'MSU'.
   8. MSU=((n-m)*MS)/100
```

```
  // Where MSU=Minimum Support threshold for Updated
                    database,
                  MS=Minimum threshold defined for Original
                      database,
                  n=Total number of transactions in Original
                     database.
                  m=number of transactions that are to be deleted.
 9.  Find the List FinUp & IFinUp
      //FinUp & IFinUP are the list of Frequent & Infrequent items
         in updated database respectively.
        a.  Read an item from 'LoIinUp'.
        b.  If its S >=MSU then add it 'FinUp' list.
        c.  Otherwise add it to 'IFinUp' list.
        d.  Repeat steps a to c until 'LoIinUp' becomes empty.
 10.        To update POC-Tree, process the following steps:
Case I:
                        a.  $CI= FinOg \cap FinDel$
                //where CI contains list of items which are frequent
                in both original and in newly added transactions.
                        b.  Read an item say 'x' from CI and search
                            for 'x' in 'FinUp', if it returns true then
                            add it to 'UpdateTree' list (which will be
                            processed later).
                          // 'UpdateTree' contains list of items which
                         are frequent in Original, in deleted
                         transactions and  also in Updated database.
                        c.    Otherwise delete it from Original POC-
                            Tree.
                        d.  Repeat steps b & c until the list CI
                            becomes empty.
```

```
Case II:
        a. $CI= FinOg\cap IFinDel$
        //where CI contains list of items which are frequent in original database and infrequent in deleted transactions.
        b. Read an item say 'x' from CI and search for 'x' in 'FinUp', if it returns true then add it to 'UpdateTree' list (which will be processed later).
          // 'UpdateTree' contains list of items which are frequent in original, infrequent in deleted transactions but frequent in Updated database.
        c. Otherwise delete it from Original POC-Tree.
        d. Repeat steps b & c until the list CI becomes empty.
        e. Read all the transactions corresponding to items present in 'UpdateTree' list and delete them from original POC-Tree.
  Case III:
        a. $CI= IFinOg \cap IFinDel$
       //where CI contains list of items which are infrequent in both Original database and in deleted transactions.
        b. Read an item say 'x' from CI and search for 'x' in 'FinUp', if it returns true then add it to 'AddtoTree' list (which will be processed later).
         //'AddtoTree' contains list of items which are infrequent in original, in deleted transactions
```

```
   but frequent in Updated database.
c. Repeat step b until the list CI becomes
   empty.
d. Read the transactions containing items
   present in 'AddtoTree' list and insert them
   into POC-Tree (obtained after processing
   Case II (e)).
e. At the end of step (d), Updated POC-Tree will
   be generated.
```

Algorithm 4.4: UPOCinDEL algorithm

As presented in Algorithm 4.4, **UPOCinDEL** is used to update POC-Tree in case of Deletions. Inputs for the algorithm are: List of items in Original Database, Frequent 1-itemsets, infrequent item sets in the original database, List of transactions that are to be deleted, Original POC-Tree.

Scan transactions that are to be deleted and get the items present and also their supports (step1 & 2). Calculate minimum support in Deleted transaction (using steps 3 & 4). Find items which are frequent and also infrequent in new transactions (step 5). Find the list of items and their support in updated database (step 6). Calculate minimum support for updated database and then find List of frequent & infrequent items in updated database (steps 7 to 9).

To update POC-Tree the following procedure is followed:

1. Items which are frequent in both original &deleted records are identified and tested whether they are still frequent in updated database. If such items exist then they are added to "UpdateTree" list, otherwise the items are deleted from POC-Tree (step 10 – case (i)).
2. Items which are frequent in original database & infrequent in

deleted records are identified and tested whether they are still frequent in updated database. If such items exist, then they are added to “UpdateTree” list. Otherwise the items are deleted from POC-Tree (step 10 – case (ii)).

3. Items which are infrequent in original database &infrequent in deleted records are identified and tested whether they are still frequent in updated database. If such items exist, then they are added to “addtoTree” list. Transaction number corresponds to the items present in “addtoTree” list is determined and are inserted into POC-Tree (step 10- case (iii)).
4. Finally updated POC-Tree will be obtained.

4.5 AlgorithmUPOCinSup:

This algorithm is used to update POC tree when minimum support threshold is changed by the user.

```
Input: D, LoIinOg, FinOg, IFinOg, MS, Original POC-Tree and NMS
          //where   D= Original Database,
                    LoIinOg=List of Items in Original Database,
                    FinOg=Frequent items in original database,
                    IFinOg=Infrequent items in Original database,
                    MS=Minimum support threshold,
                    NMS=New Minimum support threshold.
          Output: POC-Tree for the new minimum support threshold
                  defined.
          Steps:
             1. Calculate Sv.
             2. Sv=(n*NMS)/100;
                  //where Sv=support value,
                          n=number of transactions in D,
                NMS=new minimum support threshold
                    defined.
   3. If NMS < MS:
        a. Read an item from 'IFinOg', if its count >=Sv
           then delete it from 'IFinOg' list add it to list
           'AddtoTree' list and also add itto 'FinOg' list.
        b. Repeat step (a) for all the items in 'IFinOg' list.
        c. Read the transactions containing items present
           in 'AddtoTree' list and insert them to Original
           POC-Tree.
        d. At the end of Step (c) Original POC-Tree is
           updated.
   4. Otherwise
        a. Read an item from 'FinOg', if its count <Sv then
           delete it from 'FinOg' list add it to list
           'DelfromTree' list and also add it to 'IFinOg' list.
        b. Repeat step (a) for all the items in 'FinOg' list.
        c. Read the transactions containing items present
           in 'DelfromTree' list and Delete them from
           original POC-Tree.
        d. At the end of Step (c) original POC-Tree is
           updated.
```

Algorithm 4.5: UPOCinSup algorithm

As presented in Algorithm 4.5, **UPOCinSup** is used to update POC-Tree in case of support change. Inputs for the algorithm are: List of items in Original Database, Frequent 1-itemsets, infrequent item sets in the original database, new minimum support, Original POC-Tree.

If new minimum support specified is less than old specified minimum support then there is a chance that some of the infrequent items may become frequent according to the new specified minimum support. Identify the items which became frequent and add them to “addtoTree” list and to list of frequent item sets, delete them from list of infrequent item sets. Read the transactions containing the items in “addtoTree” list and insert them into POC-Tree (step 3).

If new minimum support specified is greater than old specified minimum support, then there is a chance that some of the frequent items may become infrequent according to the new specified minimum support. Identify the items which became infrequent and add them to “DelfromTree” list and to the list of infrequent item sets, delete them from list of frequent item sets.Read the transactions containing the items in “DelfromTree” list and delete them from POC-Tree (step 4).

4.6 Generating Association Rules

Association Rules are generated from frequent item sets by using “confidence” measure, which is defined as follows:

Confidence (A->B) =P (B/A) =Support (AUB)/Support (A) //where {A, B} is a frequent item set.

4.6.1 Algorithm for Generating Association Rules:

Association rules are generated after obtaining frequent item sets. Here is the algorithm used to achieve this.

Steps: Start

1. Let F (for example F = {A, B, C, AB, AC,}) be the set of Frequent item sets.
2. Generate all no-empty subsets, for each frequent item set.
3. For every non-empty subset (for eg: {AB}) find confidence value.
4. If its value is greater than specified threshold value then

```
        output the rule(i.e., A->B or B->A)
End
```

Algorithm 4.6.1: Association Rule Generation algorithm

As presented in Algorithm 4.6.1, the association rules are generated based on the frequent item sets that are obtained from POC tree. The algorithms proposed above realize the system which generates association rules .Incremental ARM is realized against modifications in database and change in the minimum support threshold.

Chapter 5: Experimental Results and Discussions

This section presents results of experiments and evaluates them. The results of generating high quality association rules with FIN algorithm and the results of the FIN-INCRE algorithm are provided in the following sub sections.

5.1 Results of FIN Algorithm

This section presents experimental results of the FIN algorithm. The results of the algorithm are compared with the state of the art.

5.1.1 Results based on Execution Time

This sub section provides the experimental results of FIN compared with the state of the art algorithms named Apriori and FP-Growth in terms of execution time. The time taken for generating frequent itemsets is considered and shown in Table 5.1.1.

Table 5.1.1: Shows the performance comparison in terms of execution time

Frequent Itemset Mining Algorithm	**Execution Time (seconds)**	
	100000 Instances	**200000 Instances**
Apriori	45.26	81.54
FP-Growth	32.64	59.45
FIN	21.78	38.67

As presented in Table 5.1.1, the execution time taken by the three frequent itemset mining algorithms known as Apriori, FP-Growth and FIN is provided against different dataset size.

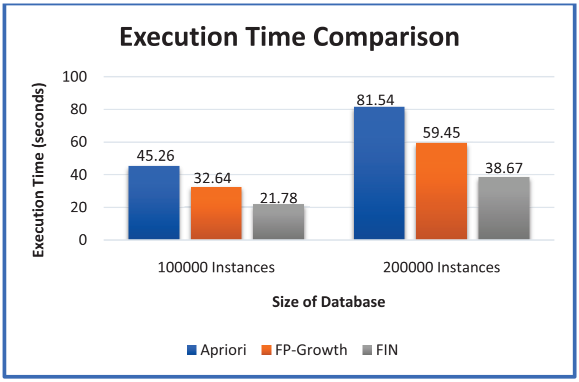


Figure 5.1.1: Execution time comparison

As presented in Figure 5.1.1, the frequent itemset mining algorithms such as Apriori, FP-Growth and FIN are compared in terms of execution time. The data sizes such as 100000 instances and 200000 instances are provided in horizontal axis. The execution time taken by the algorithms with different dataset size is provided in vertical axis. The results revealed that the size of dataset has its influence on the execution time. When data size is increased, the execution time also increased. Another important observation is that the FIN algorithm has outperformed the other two algorithms. It needed 21.78 seconds with dataset containing 100000 instances while its predecessors such as Apriori and FP-Growth needed 45.26 seconds and 32.64 seconds respectively. When the dataset size is 200000 instances, the FIN algorithm needed 38.67 seconds while its predecessors such as Apriori and FP-Growth needed 81.54 seconds and 59.45 seconds respectively. Thus the results revealed that FIN has significant performance improvement over the other algorithms.

5.1.2 Results based on Memory Usage

This sub section provides the experimental results of FIN compared with the state of the art algorithms named Apriori and FP-Growth in terms of memory usage. The memory consumed for generating frequent itemsets is considered and shown in Table 5.1.2.

Table 5.1.2: Shows the performance comparison in terms of memory usage

Frequent Itemset Mining Algorithm	Memory Usage (MB)	
	100000 Instances	200000 Instances
Apriori	34.3175	61.4523
FP-Growth	25.8211	48.7463
FIN	22.8377	41.1514

As presented in Table 5.1.2, the memory usage of the three frequent itemset mining algorithms known as Apriori, FP-Growth and FIN is provided against different dataset size.

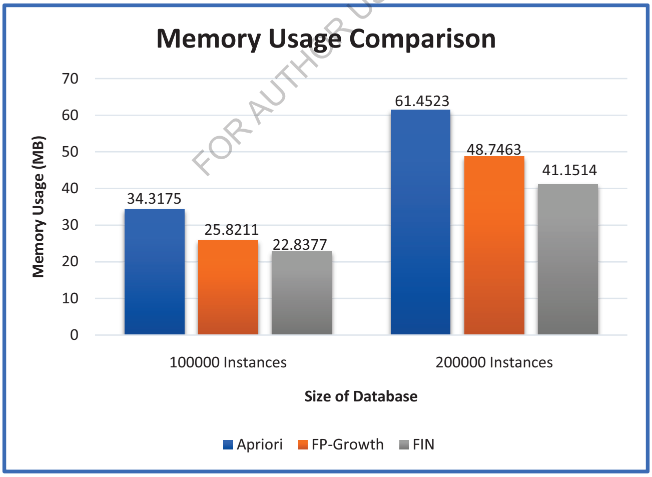


Figure 5.1.2: Memory usage comparison

As presented in Figure 5.1.2, the frequent itemset mining algorithms such as Apriori, FP-Growth and FIN are compared in terms of memory usage. The data sizes such as 100000 instances and 200000 instances are provided in horizontal axis. The memory usage taken by the algorithms with different dataset size is provided in vertical axis. The results revealed that the size of dataset has its influence on the memory usage. When data size is increased, the memory usage is also increased. Another important observation is that the FIN algorithm has outperformed the other two algorithms. It needed 22.8377MB with dataset containing 100000 instances while its predecessors such as Apriori and FP-Growth needed 34.3175MB and 25.8211MBrespectively. When the dataset size is 200000 instances, the FIN algorithm needed 41.1544MB while its predecessors such as Apriori and FP-Growth needed 61.4523MB and 48.7463MB respectively. Thus the results revealed that FIN has shown significant performance improvement over the other algorithms.

5.2 Results of FIN-INCRE Algorithm

This section provides experimental results of FIN-INCRE with benchmark datasets.

5.2.1 Datasets Used

The details of datasets used for empirical study are provided in Table 5.2.1. The incremental association rule mining is achieved with FIN-INCRE and the results are presented in section 5.2.2 & 5.2.3.

Table 5.2.1: Summary of datasets used

Database	Avg. Length	#Items	#Trans
Mushroom	23	119	8124
Connect13	43	130	67557
T25I10D100K	25	990	99822

These datasets are collected from UCI machine learning repository. We modified FIN algorithm in order to update generated association rules based on the changes made to dataset. The algorithm consumes less memory and takes very less time for execution. Our algorithm FIN_INCRE is as shown below. This algorithm is used once FIN obtains frequent item sets into the underlying data structure.

5.2.2 Results based on Execution Time

This sub section provides the experimental results of FIN_INCRE compared with the state of the art algorithm named BIT-Growth in terms of execution time. The time taken for incrementally generating association rules is considered.

Table 5.2.2: Shows the performance comparison in terms of execution time

Incremental ARM Algorithms	Execution Time (seconds)	
	Insertions	Deletions
BIT-Growth	37	34
FIN_INCRE	18	13

As presented in Table 5.2.2, the execution time taken by the two incremental ARM algorithms known as BIT-Growth and the proposed FIN_INCRE is provided in case of insertion of new transactions and deletion of existing ones.

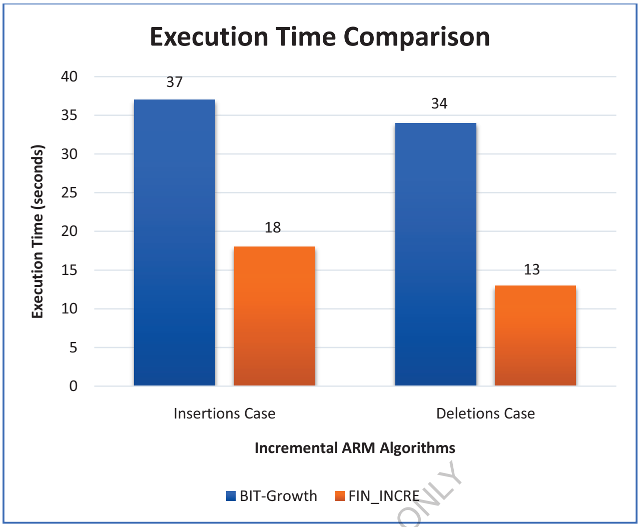


Figure 5.2.2: Execution time comparison

As presented in Figure 5.2.2, the incremental ARM algorithms BIT-Growth and FIN_INCRE are compared in terms of execution time. The two cases such as inserting new transactions into an original database from which association rules are already generated by scanning POC tree and deletions from the database are provided in horizontal axis. When new transactions are added or existing ones are deleted, the algorithms incrementally update association rules rather than reinventing the wheel again. The time taken for the incremental update of association rules is provided in the vertical axis. For insertions case, the execution time needed by the existing BIT-Growth algorithm is 37 seconds while the proposed FIN_INCRE needed only 18 seconds. For deletions case, the execution time needed by the existing BIT-Growth algorithm is 34 seconds while the proposed FIN_INCRE needed only 13 seconds. From the results, it is observed that the performance of FIN_INCRE is significantly better than the state of the art algorithm.

Since the time taken for generating association rules is an important consideration, the proposed algorithm does well in reducing time taken. The reason behind the performance enhancement lies in the usage of fast frequent item set mining approach from POC-tree. At the same time, the quality of association rules is increased with the FIN_INCRE as it dynamically filters rules using the combined metrics comprising of subjective and objective measures. The execution time is computed at runtime by finding start time just before execution of the algorithm starts and end time after completion of the algorithm. The start time is subtracted from the end time to arrive at the execution time in seconds.

5.2.3 Results based on Memory Usage

This sub section provides the experimental results of FIN_INCRE compared with the state of the art algorithm named BIT-Growth in terms of memory usage. The memory required for incrementally generating association rules is considered.

Table 5.2.3: Shows the performance comparison in terms of memory usage

Incremental ARM Algorithms	**Memory Usage (MB)**	
	Insertions	**Deletions**
BIT-Growth	14.01563	12.02637
FIN_INCRE	11.49316	9.753906

As presented in Table 5.2.3, the memory used for execution by the two incremental ARM algorithms known as BIT-Growth and the proposed FIN_INCRE is provided in case of insertion of new transactions and deletion of existing ones.

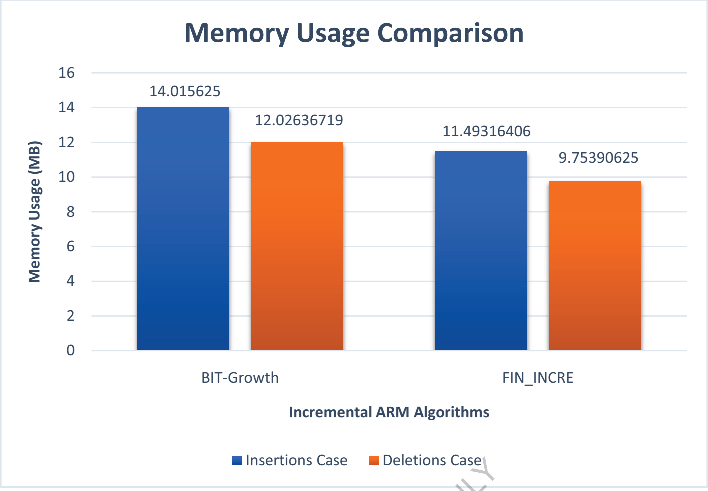


Figure 5.2.3: Memory usage comparison

As presented in Figure 5.2.3, the incremental ARM algorithms BIT-Growth and FIN_INCRE are compared in terms of memory usage. The two cases such as inserting new transactions into an original database from which association rules are already generated by scanning POC tree and deletions from the database are provided in horizontal axis. When new transactions are added or existing ones are deleted, the algorithms incrementally update association rules rather than reinventing the wheel again. The memory consumed for the incremental update of association rules is provided in the vertical axis. It is measured in Mega Bytes (MB). For insertions case, the memory needed by the existing BIT-Growth algorithm is 14.01563 MB while the proposed FIN_INCRE needed only 11.49316 MB. For deletions case, the execution time needed by the existing BIT-Growth algorithm is 12.02637 MB while the proposed FIN_INCRE needed only 9.753906 MB. From the results, it is observed that the performance of FIN_INCRE is significantly better than the state of the art algorithm in memory conservation. Since the main memory is an asset to any computing facility, the proposed algorithm does well in reducing memory usage while incrementally generating association rules. The reason behind the performance enhancement lies in the usage of fast frequent item set mining approach from POC-tree which lessens overhead of the algorithm. The memory usage is computed at runtime by finding memory in use just before execution of the algorithm starts and memory usage after completion of the algorithm. The latter is subtracted from the former to arrive at the memory usage in MB.

Chapter 6 : Conclusion and Future Scope for Research

- Extensive review of literature has provided useful insights on the need for current research carried out on the automatic incremental association rule mining which is much desired in the industries where data mining algorithms like ARM is widely used.
- An algorithm is proposed to achieve faster and high quality association rules by using POC tree as underlying data structure and combination of subjective and objective measures. This has resulted in faster generation of association rules besides obtaining the quality expected.
- An algorithm known as FIN-INCRE is proposed to achieve incremental association rule mining. Many other algorithms by name UPOCinINS, UPOCinDEL and UPOCinSup are proposed to help FIN-INCRE for realizing incremental update when data is subjected to changes, when data is deleted and when threshold for minimum support is modified.
- UPOCinINS algorithm provides incremental frequent item sets when original data in the database is modified with new transactions. When the database records are modified, the UPOCinDEL algorithm generates frequent item sets incrementally. When minimum support is changed by user of the application, then the UPOCinSup generates frequent item sets incrementally. The results of these three algorithms are used by FIN-INCRE to achieve the desired objective of generating faster and high quality association rules incrementally.
- A prototype application is built to demonstrate proof of the concept. The application shows the utility of the proposed FIN-

INCRE algorithm and its effectiveness against changes made to database and change in the minimum support threshold.

The work done in this thesis has led to many useful insights and helped the researcher to have meaningful directions for future work. The following are possible future enhancements.

1. An important direction for future work is to enhance the proposed framework to deal with big data in cloud computing environment.
2. With the emergence of cloud computing and big data eco-system, it is interesting to adapt the proposed algorithms for distributed programming frameworks like Hadoop to process large volumes of data.
3. Yet another direction for future work is to improve the proposed framework to realize a reusable service known as Incremental Association Rule Mining as a Service (IARMaaS) and deploy in cloud for rendering on-demand mining services.

REFERENCES

1. Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on very Large Databases, Santiago, Chile, pp. 487-499.
2. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C.
3. Ayan, N.F., Tansel, A.U., & Arkun, M.E. (1999). An Efficient Algorithm to Update Large Itemsets with Early Pruning. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 287-291.
4. Bay, V.O., Tuong, Le., FransCoenen., & Tzung-Pei Hong. (2013). Mining frequent itemsets using the N-list and subsume concepts. International Journal of Machine Learning and Cybernetics, DOI: 10.1007/s13042-014-0252-2.
5. Borgelt, C. (2005). Keeping Things Simple: Finding Frequent Item Sets by Recursive Elimination. Proc. Workshop Open Software for Data Mining (OSDM'05 at KDD'05, Chicago, IL), pp. 66– 70. ACM Press, New York, NY, USA.
6. Cheung, D., Han, J.M., & Wong, C.Y. (1996). Maintenance of Discovered Association Rules in Large Databases. An Incremental Updating Technique. Proceedings of the 12th International Conference on Data Engineering, pp. 106-114.
7. Cheung, D., Lee, S.D., & Kao, B. (1997). A General Incremental Technique for Updating Discovered Association Rules. Proceedings of the Fifth International Conference on Database Systems for Advanced Applications, pp. 185-194.
8. Cheung, W., & Zaiane, O.R. (2003). Incremental mining of frequent patterns without candidate generation or support constraint. Proceedings of the IDEAS, IEEE Computer Society Press, Los

Alamitos, CA, pp. 111-116.

9. Deng, Z. H., & Wang, Z. H. (2010). A new fast vertical method for mining frequent itemsets. International Journal of Computational Intelligence Systems, 3(6), 733–744.

10. Deng, Z. H., Wang, Z. H., & Jiang, J. J. (2012). A new algorithm for fast mining frequent itemsets using N-lists, pp. 55, doi: 10.1007/s11432-012-4638-z.

11. Diana-Lucia, Miholca., Gabriela, Czibula., & Liana, Crivei. (2018). A new incremental relational association rules mining approach. International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, Belgrade, Serbia

12. Ezeife C.I., & Su Y. (2002). Mining Incremental Association Rules with Generalized FP-Tree. Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol 2338. Springer, Berlin, Heidelberg.

13. Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. Proceedings of the ACM-SIGMOD International Conference on Management of Data.

14. Hong, T. P., Lin, J. W., & We, Y. L. (2008). Incrementally Fast Updated Frequent Pattern Trees. Expert Systems with Applications, vol. 34, pp. 2424-2435.

15. Koh, J. L., Shieh, S. F. (2004). An efficient approach for maintaining association rules based on adjusting FP-tree structures. Proceedings of the DASFAA, Springer-Verlag, Berlin Heidelberg, New York, pp. 417-424.

16. Lee, S.D., David, Cheung, W., & Ben Kao. (1998). Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules. Data Mining and Knowledge Discovery, Volume 2 Issue 3, pp. 233-262, Kluwer Academic Publishers Hingham, MA, USA.

17. Leung, C. K., Khan, Q.I., & Hoque, T. (2005). CanTree: A Tree

Structure for Efficient Incremental Mining of Frequent Patterns. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM’05), 2005.

18. Li, X., Deng, X., & Tang, S. (2006). A Fast Algorithm for Maintenance of Association Rules in Incremental Databases. Proceeding of International Conference on Advance Data Mining and Applications, pp.56-63.
19. Liu, G., Lu, H., Lou, W., & Xu. (2004). Efficient Mining of Frequent Patterns using Ascending Frequency Ordered Prefix-Tree. Data Mining and Knowledge Discovery, vol. 9, pp. 249-274.
20. Motwani, R., Ullman, J.D., & Brin, S. (1997). Dynamic Itemset Counting and Implication Rules For Market Basket Data. ACM SIGMOD International Conference on Management of Data, vol. 26, no. 2, pp.55–264.
21. Park, J.S., Chen, M., & Yu, P.S. (1995). An Effective Hash Based Algorithm for Mining Association Rules. ACM SIGMOD International Conference on Management of Data.
22. Pei, J., Han, J., Nishio, S., Tang, S., & Yang, D. (2001). H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Databases. Proceedings of International Conference on Data Mining.
23. Pietracaprina, A., & Zandolin, D. (2003). Mining frequent itemsets using Patricia tries. Proceedings of the FIMI.
24. Archana Gupta , Akhilesh Tiwari and Sanjeev Jain , “A system for Incremental Association rule mining without candidate generation”,International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 7, July 2019, https://sites.google.com/site/ijcsis,ISSN 1947-5500.
25. Savasere, A., Omiecinski, E., & Navathe, S. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. Proc. 21st Very Large Data Bases Conference.
26. Tanbeer, S.K., Ahmed, C.F., Jeong, B.S., Lee, YK. (2008). CP-Tree:

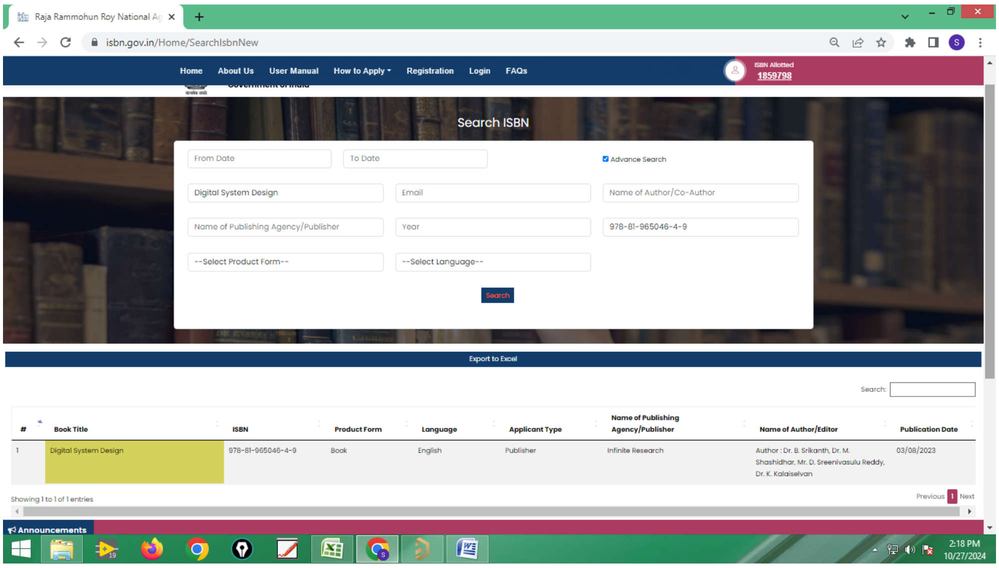
A Tree Structure for Single-Pass Frequent Pattern Mining. Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, vol 5012. Springer, Berlin, Heidelberg.

27. Thomas, S., Bodagala, S., Alsabti, K., & Ranka, S. (1997). An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases. In KDD'97, New Port Beach, California.
28. Tong, Yi., Baowen, Xu., & Fangjun, Wu. (2004). A FP-Tree Based Incremental Updating Algorithm for Mining Association Rules, Issue No.5, pp. 703-710.
29. Totad, S.G., Geeta, R.B., & Prasad Reddy, P.V.G.D. (2012). Batch incremental processing for FP-tree construction using FP-Growth algorithm. Knowledge and Information Systems, Volume 33, Issue 2, pp. 475-490, https://doi.org/10.1007/s10115-012-0514-9.
30. Woon, Y., Kwong, E.W., Wee Keong., & Lim, Ee-Peng. (2004). A Support-Ordered Trie for Fast Frequent Itemset Discovery. Knowledge and Data Engineering, IEEE Transactions on. 16. 875 - 879. 10.1109/TKDE.2004.1318569.
31. Zaki, M. (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, vol. 12, no.3, pp. 372-390.
32. Zhi Hong., Deng, H., & Sheng Long, L. V. (2014). Fast mining frequent itemsets using Nodesets. Expert Syst. Appl. 41(10): 4505-4512.
33. Zhou, Z., & Ezeife, C. I. (2001). A Low-Scan Incremental Association Rule Maintenance Method based on the Apriori property. Proceedings of the 14th Canadian Conference on Artificial Intelligence.
34. Jerry Chun-Wei Lin, Wensheng Gan, Tzung-Pei Hong, and Binbin Zhang, "An Incremental High-Utility Mining Algorithm with Transaction Insertion," The Scientific World Journal, vol. 2015, Article ID 161564, 15 pages,

2015. https://doi.org/10.1155/2015/161564.

More
Books!
yes
I want morebooks!

2022-2023



Search ISBN

From Date

To Date

☑ Advance Search

Digital System Design

Email

Name of Author/Co-Author

Name of Publishing Agency/Publisher

Year

978-81-965046-4-9

--Select Product Form--

--Select Language--

Search

Export to Excel

Search:

#	Book Title	ISBN	Product Form	Language	Applicant Type	Name of Publishing Agency/Publisher	Name of Author/Editor	Publication Date
1	Digital System Design	978-81-965046-4-9	Book	English	Publisher	Infinite Research	Author : Dr. B. Srikanth, Dr. M. Shashidhar, Mr. D. Sreenivasulu Reddy, Dr. K. Kalaiselvan	03/08/2023

Showing 1 to 1 of 1 entries