Cerebration of Privacy Preserving Data Mining Algorithms

P. Rajesh, G. Narsimha

Abstract—**Recent advances in data intensive information processing systems is becoming increasingly important to make decisions in business organizations. Precious & sensitive knowledgeable patterns may reside in the process of business analysis. Recently Privacy preserving data mining have acquired a great extent by data mining and Information security community researchers in development of techniques that integrates privacy concerns. Many researchers have projected and implemented numerous algorithms for privacy preserving in data mining that mainly focus on cluster based, classification approach, association and outsourcing of sensitive knowledgeable patterns obtaining from data mining systems for the intended persons. This paper presents and evaluates comparative based analytical results of privacy preserving data mining algorithms empirically. Various kinds of analytical measures were employed and plotted them against the different data sets on a graph, in order to find out the effectiveness and efficiency of privacy preserving data mining algorithms.**

Index Terms—**Analytical Results, Comparison, Data mining, Privacy Preserving**

I. INTRODUCTION

DATA mining has emerged as key technique in modern science and technology. It has become well accepted domain area by all most all industries like Business, Financial, Retail, Biological, Intrusion detection etc. Data mining techniques are the way of changing the data into useful information [1]. However, Data mining demonstrate some negative social potential perception insights in the direction of privacy invasion.

Business Data is the valuable resource of any organizations to extract new patterns (predicting, Behavioral, cluster, information processing) to make decisions, to increase net growth of the company and try to provide potential outstanding business services to their respective customers [2].

Business operational application services like banks, insurance companies, Credit card transactions generates an incredible quantity of sensitive data day-to-day everywhere due to communications in technology and sharing [3], [4]. There is a emergent need to defend sensitive information of organization data, employees information, production database, customers information in enterprise where such sensitive data may exist in [5]. Sensitive data is the information that is confined against unwarranted disclosure.

Corporations often restrict to gain accessing sensitive information of users, by deploying various kinds of methods that integrates privacy concern, however sensitive data is exposed to be vulnerable by partners, adversary, competitors. [6], [7], [8], [9], [10], [11]. Apart from that accumulate all the data in single repository of data warehouse causes violation of security exposure.

So there is an vital need to construct accurate models of privacy preserving data mining algorithms without access to precise information and not disclosing the confidential data. Researcher's forums are much interest in addressing wide variety of challenges that come across in privacy preserving data intensive information processing systems.

Privacy preserving data mining is expected to be a multibillion dollar industry by the year 2015. Computer Emergency Response Team (CERT-In), Information and Communication Technology (ICT) develops novel sophisticated strengthening technique up to date to protect sensitive information from cyber critters.

The rest of paper is structured as follows. Section 2 describes Brief explanation of proposed a new privacy preserving data mining techniques. Evaluation and comparative based analytical results of privacy preserving data mining were presented in section 3. Section 4 consists of conclusion and future scope.

II. PRIVACY PRESERVING DATA MINING ALGORITHMS

Fundamental data mining applications poses challenges include mine knowledge patterns from noisy data, classify knowledge in very skewed data, distributed data, Issues related to data privacy, autonomously controlled information, Business and Big data analytics. Privacy preserving data mining particularly deals with respect to analysis of sensitive data. This section describes our new approaches of privacy-preserving data mining techniques briefly which were published in the earlier papers.

A. *Fuzzy Privacy Preserving Classification Approach*

Fuzzy based privacy preserving technique in data mining transforms the original sensitive information of data into fuzzy values and classifies them into the patterns using alpha cut property. By using this technique the sensitive data will be transformed in to uncertain manner that protects

Manuscript received July 7, 2014; revised August 5, 2014.

F. P. Rajesh is a student in the Department of Computer Science and Engineering with the Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, India, 522 508; (e-mail: rajesh.pleti@gmail.com).

S. Dr. G. Narsimha is an Associate Professor in the Department of Computer Science and Engineering with Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, India, (e-mail: narsimha06@gmail.com).

the privacy and becomes difficult for the intruder to know the information.

we consider the execution time from submitting of search user query onwards to until getting the fuzzy classification results. Fuzzy classification may be done using direct (sensitive) or indirect approach (non sensitive). Direct fuzzy classification takes place when decisions are made by sensitive attributes. Indirect fuzzy classification takes place when decisions are made according to non sensitive attributes that are highly correlated with the biased sensitive attributes [12].

Experimental Evaluation of fuzzy classification results are based on indirect querying, most robust to noisy data, and uncertainty data while preserving privacy.

B. MFI Privacy Preserving Document Clustering

Text, data mining analytics has become central powerful techniques recognized broadly due to vast amounts of data sources generated by academic, economic and social activities are increasingly available in electronic form accessible to more readers.

According to survey, in every year 1.5 million articles are added to internet. Text document cluster mining offers a solution to these problems, by employing new insights methods of natural language processing, enhanced techniques from information retrieval.

The main aim of MFI privacy preserving document clustering is to find out similar kind of hierarchical documents by MFI similarity measure but not the same content in every document (Duplicate documents).Providing privacy preserving of documents is by avoiding duplicate documents. There by we can protect the privacy of individual copy rights of document authors [13].

C. CRT(Chinease Remainder Theorem) Based Approach

The conditions for processing sensitive personal data are more difficult to satisfy. Information Technology is involved in storing, sharing and securing the data, however it is in the dark concerning how data is shared and used.

People would like their personal sensitive data to work for them collected by organizations, (like tax, health records, pensions, council services and so on) potential risks, and benefits involved to society, individuals to sharing the data.

Privacy preserving data mining using CRT is one of the techniques that provides a solution to share sensitive information to group of intended persons by dynamically generated secret key [14].

III. COMPARITVE BASED ANLYTICAL RESULTS

The objective of this paper is to evaluate efficiency and analyze the cerebration based analytical results of privacy preserving data mining algorithms on different data sets.

The efficiency, privacy level, accuracy of privacy preserving data mining algorithms depends on diverse dynamic factors including kind of sensitive, non sensitive attributes involved in data set, size of database, type of search querying, type of knowledge to be extracted, privacy protection measures imposed, utility of data [15], [16], [17], [18].

By make use of our novel privacy preserving data mining algorithms on different data sets (Bank client dataset, Hospital dataset, University dataset, document dataset) [19], [20] we observe the computation time, communication cost to share the sensitive data among parties will be reduced . We developed proposed privacy preserving data mining techniques using Java, XAMPP web server 1.8.0 including features like PHP, MYSQL, Apache Tomcat.

The above three privacy preserving data mining algorithms offer the tradeoff among utility and privacy problem in taking into consideration of algorithmic requirements, privacy at the similar time.

The subsequent datasets have been employed for our study. The bank training data set are biased with 16 attributes, 4211 instances contains sensitive attributes like job, balance, housing loan, personal loan, age, gender. Similarly hospital, university dataset having 9 attributes, 750 instances contains sensitive attributes diseases, gender, age, nationality, placement details, and percentage.

The following Bar diagram shows privacy preserving data mining by CRT (Chinease Remainder Theorem) approach execution time in seconds for different data sizes in KB's. As the size of the data increases gradually, the execution maintains same amount of time, with little changes accordingly.

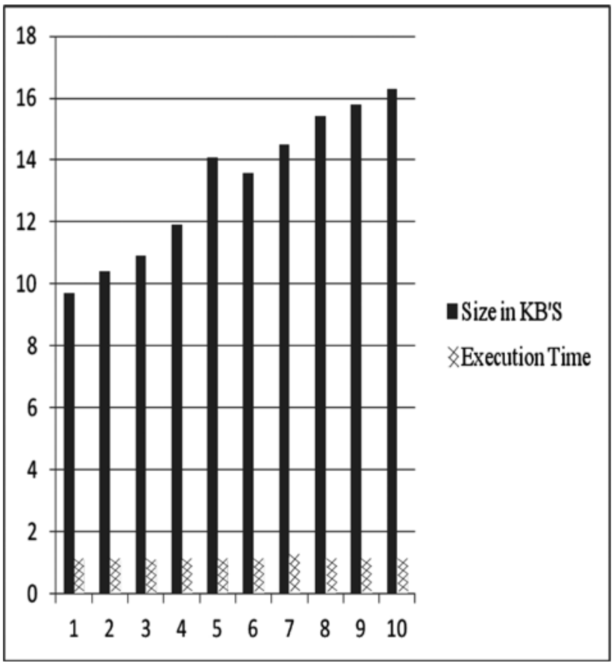

Fig. 1. CRT Bar diagram

The subsequent three diagrams provides fuzzy classification of privacy preserving data mining results based on diverse queries, threshold values and data sizes.

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

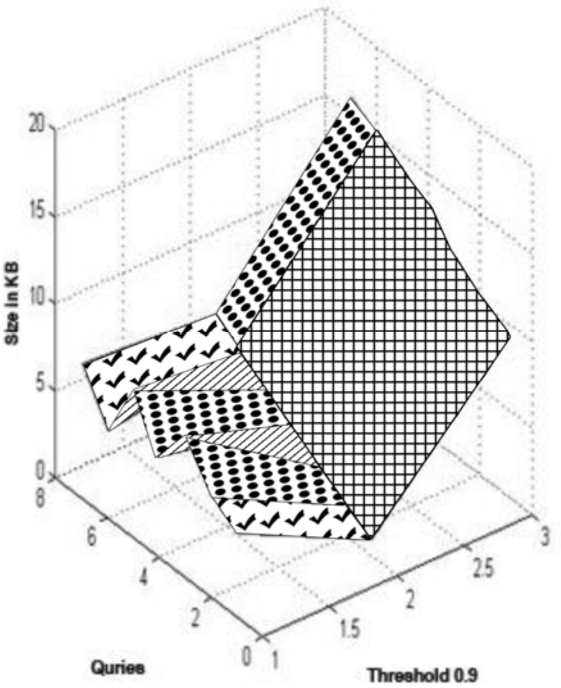

Fig. 2. Fuzzy classification with threshold 0.9

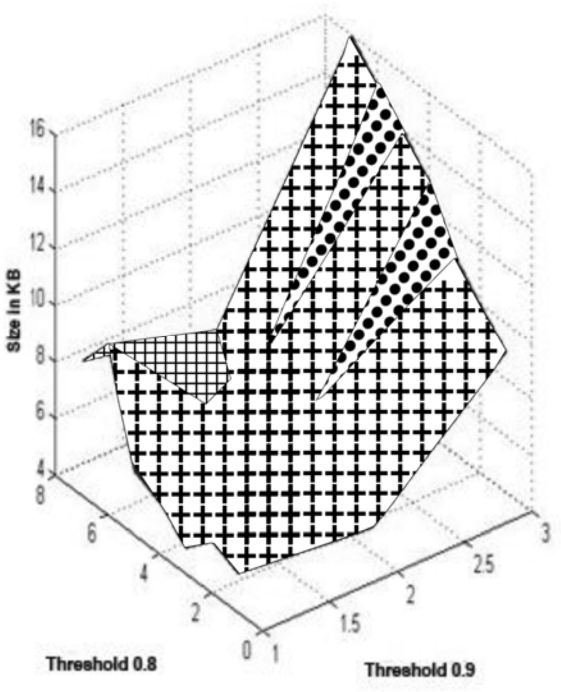

Fig. 3. Fuzzy classification with threshold 0.8 and 0.9

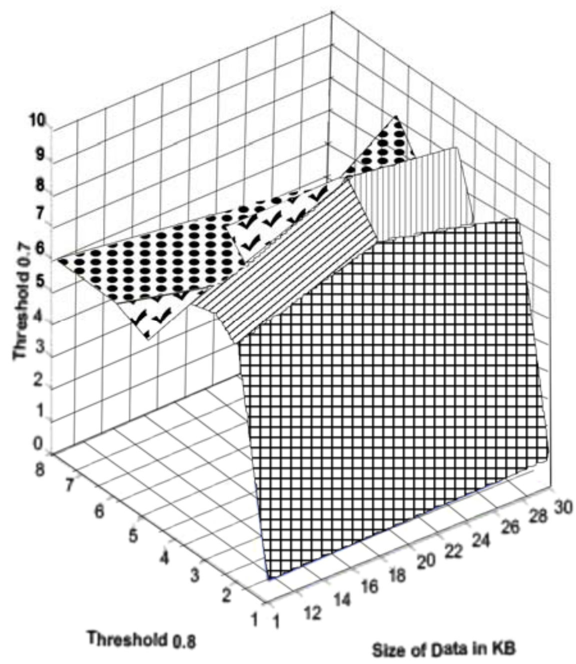

Fig. 4. Fuzzy classification with threshold 0.7 and 0.8

The experiment represents understanding of observations that made by different threshold values.

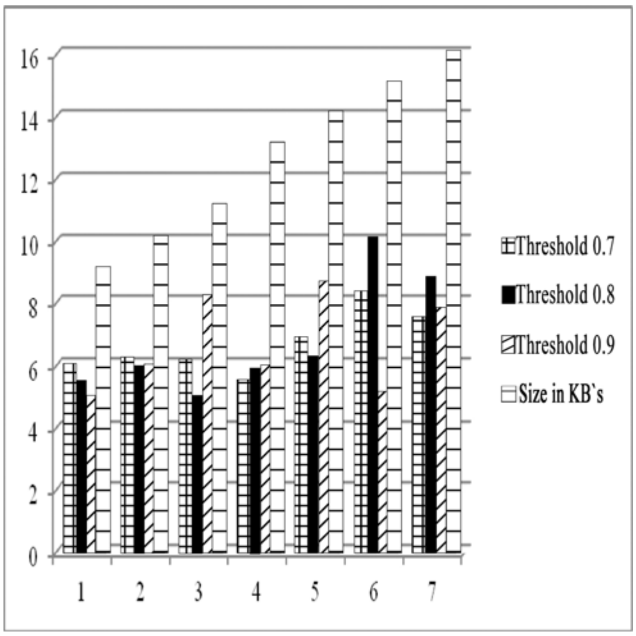

Fig.5. Fuzzy classification with threshold 0.7, 0.8 and 0.9

The results that are depicted in the following figure evaluation of both the CRT and fuzzy approaches.

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

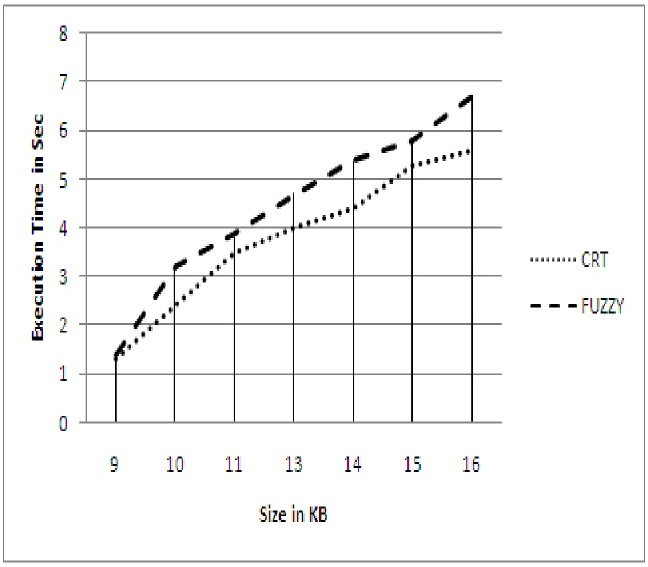

Fig. 6. Result analysis of CRT and Fuzzy

Privacy preserving MFI document clustering dealt with 200 documents with keywords data mining, privacy preserving, clustering, classification, feature extraction having measures minimum support value is 40 and the maximum support value is taken as 50 and the minimum length of 3 and the maximum length should be 3, Then the MFI's (maximal frequent item sets) according to jaccard similarity were considered to perform hierarchical document clustering using Tanagra for our project are as follows.

1	__warehouse /\ frequent_itemsets /\ __association	59.0
2	_analysis /\ feature_extraction /\ __association	60.5
3	data_mining /\ feature_extraction /\ frequent_itemsets	59.0
4	data_mining /\ feature_extraction /\ __association	59.5
5	data_mining /\ frequent_itemsets /\ __association	60.0
6	__predicting /\ classification /\ feature_extraction	61.0
7	__predicting /\ classification /\ __association	63.5
8	__predicting /\ feature_extraction /\ frequent_itemsets	62.0
9	__predicting /\ feature_extraction /\ __association	65.5
10	__predicting /\ frequent_itemsets /\ __association	65.0
11	classification /\ feature_extraction /\ frequent_itemsets	63.0
12	classification /\ feature_extraction /\ __association	64.5
13	classification /\ frequent_itemsets /\ __association	63.5
14	feature_extraction /\ frequent_itemsets /\ __association	67.5

Fig. 7. Maximal frequent item sets with support value

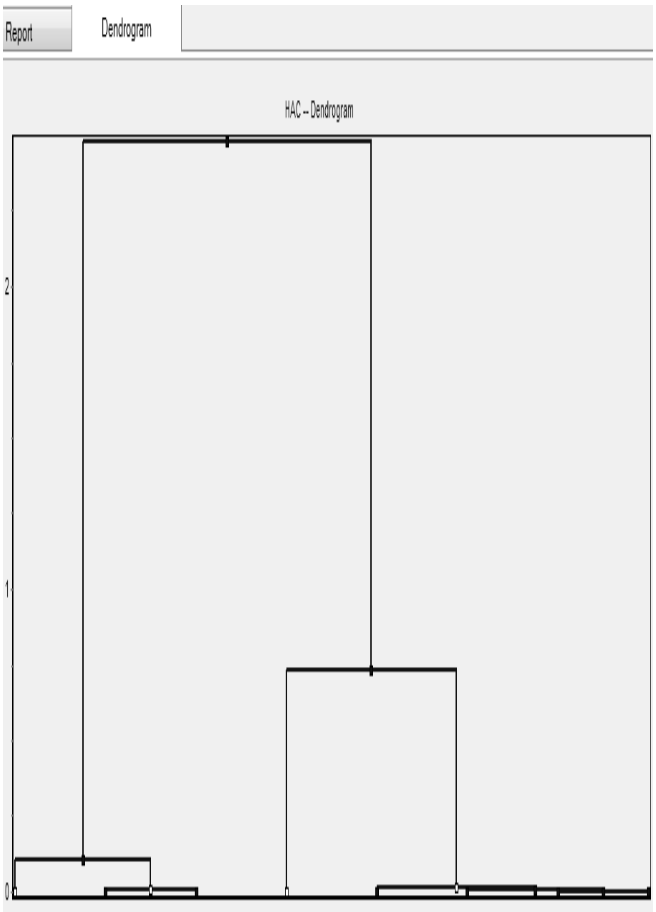

Fig.8.Hierarchical document clustering (HAC) dendogram

IV. CONCLUSTION AND FUTURE SCOPE

The amount of digital Information generated by fast growing technology and science concerned in storing, retrying, sharing resources availability of data processing and securing the data increasing everywhere, however it is in the dark concerning how data is shared and utilized. Privacy preserving data mining initiated new direction and got serious attention by researchers. we try to solve the problem to some extent by proposed novel methods. In this paper we present and evaluate comparative based analytical results of privacy preserving data mining algorithms empirically that extent to protects the sensitive data. we observe the computation time, communication cost to share the sensitive data among parties will be reduced and provides trade of between data privacy, utility.

we need to consider privacy as a whole organizational social issue, not only a technological view. organizations need to provide better services to citizens what they want safely and securely. However, people also want to familiar how their personal information is being used, who has gain access to it, and what it objectives, appropriate control over the sensitive data. Privacy preserving data mining field is expected to flourish.

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

ACKNOWLEDGMENT

The authors would like to extend their gratitude to the anonymous reviewers and who continuously supported to bring out this technical paper. I would like to thank my advisor, Dr. A. Demerara, Director of Academic Audit Cell, Jawaharlal Nehru Technological University Hyderabad, for his professional assistance and remarkable insights. I would like to express my gratitude to Prof. G. Narsimha for helping me take the first steps in the research area and finally, special thanks to my mother and father for providing the moral support to me. we are pleased to acknowledge our sincere thanks to our beloved chairman sri.V. Vidya Sagar and Director prof. S.R.K. Paramahamsa at VVIT for valuable cooperation.

REFERENCES

[1] Wu. Xindong, "Data mining: An ai perspective," in Intelligent Informatics, 2003, pp. 23.
[2] Robert Nisbet, John, Gary Miner, "Handbook of Statistical Analysis and Data Mining Applications," in ISNN: 1932-1872, Publisher: John Wiley & Sons, 2013.
[3] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, Vol. 29, no. 2, pp. 439 -450, 2000.
[4] Chris Clifton and Don Marks, "Security and privacy implications of data mining," in *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp.15-19, Citeseer,1996.
[5] David Navetta, "Legal Implications of Big Data: A Primer" in *ISSA Journal*,| March 2013.
[6] Dakshi Agrawal and Charu C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in Proceedings of the twentieth *ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp.247-255, 2001.
[7] J Atallah Mikhail and Du.Wenliang, "Secure multi-party computational geometry," in *Proc. Algorithms and Data Structures, Springer*, pp. 165-179, 2001.
[8] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the twenty- second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 211-222, 2003.
[9] Alexandre Ev_mievski, "Randomization in privacy preserving data mining" in *proc. ACM Sigkdd Explorations* Newsletter, 4(2), pp.43-48, 2002.
[10] R. J. Bayardo, R. Aggrawal, "Data privacy through optimal k-anonymization" in *proceedings of ICDE conference*, 2005.
[11] Charu C. Aggarwal, "Privacy-Preserving Data Mining: Models And Algorithms". IBM T. J. Watson Research Center.
[12] P.Rajesh, G.Narsimha, "Fuzzy based privacy preserving classification of data streams," in *proceeding of CUBE (ACM) conference*, pp. 784 -788, 2012.
[13] P.Rajesh, G.Narsimha, "Privacy preserving MFI based similarity measure for hierarchical document clustering," in *proceedings of cornell university*, arXiv:1207.2900, 2012.
[14] P.Rajesh, G.Narsimha, "Privacy preserving data mining using CRT theorem". In *Proceedings of ICECCS (springer) conference*, 2012.
[15] Isaac Cano, Susana Ladra, and Vicen_c Torra, "Evaluation of information loss for privacy preserving data mining through comparison of fuzzy partitions," in *IEEE Int. Conf. on Fuzzy Systems*, pp. 1-8, 2010.
[16] Elisa Bertino, Dan Lin, and Wei Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms".
[17] Dasgupta, Chen, Kosara, "Measuring Privacy and Utility in Privacy-Preserving in Visualization," in Journal compilation, The Eurographics Association and Blackwell Publishing Ltd, 2013.
[18] Nirmal Thapa, "Context Aware Privacy Preserving Clustering And Classification," in Theses and Dissertations--Computer Science, University of Kentucky, 2013.
[19] Andrew Frank and Arthur Asuncion, "UCI machine learning repository", 2010. URL http://archive. ics. uci. edu, 2011.
[20] Elisa Bertino, and Igor Nai Fovino, "A Framework for Evaluating Privacy Preserving Data Mining Algorithms". in Data Mining and Knowledge Discovery, 11, pp.121–154, 2005, DOI: 10.1007/s10618-005-0006-6.