International Journal of Scientific & Engineering Research, Volume 7, Issue 8, August-2016
ISSN 2229-5518

Privacy and Data Mining

Shilpa M.S
Dept. of Computer Science
Mohandas College of Engineering and Technology
Anad,Trivandrum
shilpams333@gmail.com

Shalini.L
Dept. of Computer Science
Mohandas College of Engineering and Technology
Anad,Trivandrum

***Abstract*—Privacy preserving data mining deals with hiding an individual's sensitive identity without sacrificing the usability of data. It has become a very important area of concern but still this branch of research is in its infancy .People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. The basic idea of privacy preserving data mining is to ensure that data mining algorithms are implemented effectively without compromising the security of sensitive information contained in the data. In addition a brief discussion about certain privacy preserving techniques are also presented.**

***Index Terms*—Mining, Information Extraction, Big data , Privacy approaches.**

I. INTRODUCTION

As computers are becoming the backbone of science and economy enoramous quantities of machine readable documents become available. Almost 75% of business information lives in the form of text. The enoramous amount of information cannot s be simply used for processing by computers which typically handle text as simple sequence of character strings.The major challenge is identifying useful information from this enoramous amount of structured and unstructured data without revealing the secret of sensitive information contained in the data. The main consideration of PPDM is two-fold. Firstly, Sensitive raw data which tends to reveal an individual's private information should not be used for mining. Secondly, Privacy preserving data publishing and data sharing must be ensured. The evolvement of data mining has lead to serious impact on the privacy Data mining technologies initially helped the users in accessing and reducing large amounts of information. The percentage of difficulty in addressing privacy issues with respect to data mining was increased by the following:

• The cost of data mining tools is less while its availability is high.
• Most of the data is digitized and it is impossible for the humans to manually preprocess the data.
• Aggregation of data is increased.
•The readily available nature of data mining tools to extract patterns that go beyond actual data and its ability to predict the repetitive nature of patterns.

Many applications make use of data warehouses as central repositories. The major concern with aggregating such personal information and mining it is that personal profiles of individuals can be created using information held in systems. In recent years different methods have been proposed so as to preserve privacy. Data Mining in health care will be mainly discussed in this paper.

DATA MINING IN HEALTHCARE

Data mining holds great potential for the health care industry. It enables health systems to systematically use data and analytics to identify inefficiencies and identify best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently.

Researchers and doctors have been using medical datasets for research.This research has played a critical role in medical progress.

The field of healthcare compliance is in the midst of a sea change leading to wide use of healthcare data mining. Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amounts of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. The decisions rests with health care professionals.

The main goal of developing such a system is to allow medical data to be shared in a way that preserves patient's privacy and data utility. Patient data includes registration data(eg. Contact information, SSN) , Demographics(eg. Date of birth, gender, race),diagnosis codes, genomic information, medication and allergies, immunization status, laboratory test results, radiology images.

II. PRIVACY PRESERVING DATA PUBLISHING

Privacy preserving data publishing is ensured by the system. Initially different levels were set and depending on the ranking of each user the rules were applied.

Age	Sex	Zipcode	Disease
22	M	47906	dyspepsia
22	F	47906	flu
33	F	47905	flu
52	F	47905	bronchitis
54	M	47302	flu
60	M	47302	dyspepsia
60	M	47304	dyspepsia
64	F	47304	gastritis

Figure: Original data

The original data is assumed to be a private table consisting of multiple records.To make the data table satisfy the required privacy model, one can apply the following anonymization operations.

Generalization: In order to perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This significantly reduces the data utility of the generalized data.

Age	Sex	Zipcode	Disease
[20-52]	*	4790*	dyspepsia
[20-52]	*	4790*	flu
[20-52]	*	4790*	flu
[20-52]	*	4790*	bronchitis
[54-64]	*	4730*	flu
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	gastritis

Figure: Generalized table

Bucketization: We show the effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. Our experiment results show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data.

Age	Sex	Zipcode	Disease
22	M	47906	flu
22	F	47906	dyspepsia
33	F	47905	bronchitis
52	F	47905	flu
54	M	47302	gastritis
60	M	47302	flu
60	M	47304	dyspepsia
64	F	47304	dyspepsia

Figure: Bucketized table

Multiset-based Generalization:We observe that this multiset-based generalization is equivalent to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but break the association between them within one bucket.

Age	Sex	Zipcode	Disease
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	dysp.
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	bron.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	flu
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	gast.

Figure: Multiset-based generalized table

One-attribute-per-Column Slicing : We observe that while one-attribute-per-column slicing preserves attribute distributional information, it does not preserve attribute correlation, because each attribute is in its own column. In slicing, one group correlated attributes together in one column and preserves their correlation. For example, in the sliced table shown in Table correlations between Age and Sex and correlations between Zip code and Disease are preserved. In fact, the sliced table encodes the same amount of information as the original data with regard to correlations between attributes in the same column.

Age	Sex	Zipcode	Disease
22	F	47906	flu
22	M	47905	flu
33	F	47906	dysp.
52	F	47905	bron.
54	M	47302	dysp.
60	F	47304	gast.
60	M	47302	dysp.
64	M	47304	flu

Figure: One-attribute-per-column slicing

Slicing: Another important advantage of slicing is its ability to handle high-dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub-table with a lower dimensionality. Slicing is also different from the approach of publishing multiple independent sub-tables in that these sub-tables are linked by the buckets in slicing.

(Age,Sex)	(Zipcode,Disease)
(22,M)	(47905,flu)
(22,F)	(47906,dysp.)
(33,F)	(47905,bron.)
(52,F)	(47906,flu)
(54,M)	(47304,gast.)
(60,M)	(47302,flu)
(60,M)	(47302,dysp.)
(64,F)	(47304,dysp.)

Figure: Sliced table

III) PRIVACY PRESERVING ASSOCIATION RULE MINING

Apriori algorithm: To mine association rules across two databases where columns in the table are at different sites, splitting each row apriori algorithm is used. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. But the algorithm has disadvantages:

(1) It assumes that the transaction database is memory resident.
(2) It requires many database scans.

So a new algorithm called Fast Distributed Mining (FDM)algorithm was adopted. The algorithm was implemented with encryption. In the server side file is encrypted using RC4 algorithm which is then decrypted at the client side using the RC4 algorithm.

The FDM (Fast Distributed Algorithm for Data Mining) algorithm has the following distinguishing characteristiques:

(1) Candidate set generation is Apriori-like. However, some interesting properties of locally and globally frequent itemsets are used to generate a reduced set of candidates at each iteration, this resulting in a reduction in the number of messages interchanged between sites.

(2) After the candidate sets were generated, two types of reduction techniques are applied, namely a local reduction and a global reduction, to eliminate some candidate sets from each site.

(3) To be able to determine if a candidate set is frequent, the algorithm needs only O(n) messages for the exchange of support counts, where n is the number of sites from the distributed system. This number is much less than a disrect adaptation of Apriori, which would need O(n^2) messages for calculating the support counts.

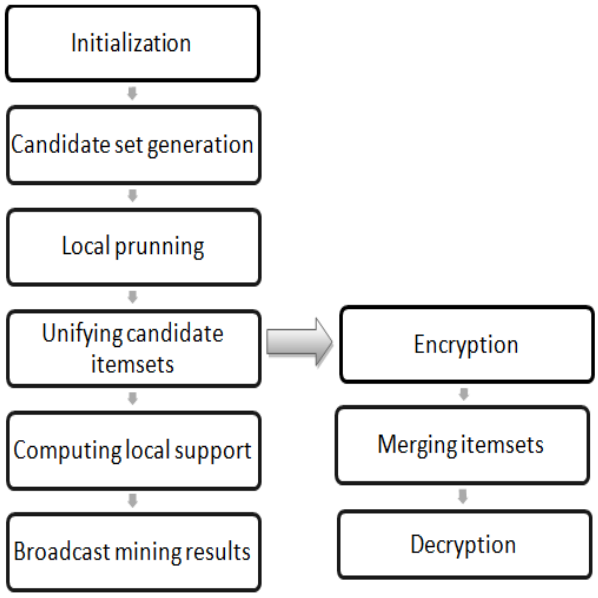

Figure: Architecture diagram for FDM.

IV. RESULT ANALYSIS

The analysis concludes that privacy preserving association rule mining out performs all other privacy preserving techniques including anonymization techniques. techniques.

Figure: Graphical analysis

V. CONCLUSION

In today's world,preserving the privacy is a major concern. People are very much concerned about their sensitive information and donot wish to share them. Thesurvey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From the analysis, it has been found that there is no single technique that is consistent in all domains. All methods have different efficiency in performing depending on the type of data as well as the type of application or domain. But still from the analysis, it can concluded that Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique since it provides encryption

of sensitive data. On the other hand Data Perturbation helps to preserve data and hence sensitivity is maintained .In future, we want to propose a hybrid approach of these techniques. In this article, a brief introduction to the field of Privacy preserving data mining was given. The main aim of privacy preserving data mining is developing certain algorithms to hide or provide privacy to certain sensitive information so that they cannot be accessed by unauthorized parties or intruder Privacy and accuracy in case of data mining is a pair of ambiguity and so succeeding one can lead to adverse effect on other. In this scenario an effort was made to review a good number of existing PPDM techniques. Another data mining algorithm could be used called Fast Distributed Mining(FDM) algorithm. FDM algorithm involves generation of candidate set which is similar to that of apriori algorithm. But it uses properties of local and global frequent itemsets to generate reduced set of candidates at each iteration. After the generation of candidate sets reduction strategies are applied to eliminate some candidate set from each side Finally, it can be concluded that with FDM algorithm there is significant reduction in the number of candidate sets and message size .The algorithm needs only O(n) messages to determine whether the candidate set is frequent or not and it is so much less than other mining algorithms.

VI. REFERENCES

[1] J. Han, M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers.

[2] Charu C. Aggarwal, Philip S. Yu "Privacy-Preserving Data Mining Models and algorithm"advances in database systems 2008 Springer Science, Business Media, LLC.

[3] Vaidya, J. & Clifton, C. W, "Privacy preserving association rule mining in vertically partitioned data," In Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada 2002.

[4] Ahmed HajYasien. Thesis on "PRESERVING PRIVACY IN ASSOCIATION RULE MINING" in the Faculty of Engineering and Information Technology Griffith University June 2007.

[5] R. Agrawal and R. Srikant. "Privacy Preserving Data Mining",ACM SIGMOD Conference on Management of Data, pp: 439-450,2000.

[6] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology, 15(3), pp.36-54, 2000.

[7] Aris Gkoulalas-Divanis and Vassilios S. Verikios, "An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine, 2010.

[8] Stanley, R. M. O. and R. Z Osmar, "Towards Standardization in Privacy Preserving Data Mining", Published in Proceedings of 3rd Workshop on Data Mining Standards, WDMS' 2004, USA, p.7-17.

[9] McCallum, Andrew; Nigam, Kamal (1998). "A comparison of event models for Naive Bayes text classification". AAAI-98 workshop on learning for text categorization.

[10] Zhang, Harry. "The Optimality of Naive Bayes". FLAIRS2004 conference.

[11] Rish, Irina (2001). "An empirical study of the naive Bayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.

[12] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). "Tackling the poor assumptions of Naive Bayes classifiers". ICML

[13] Tjong Kim Sang, Erik F.; De Meulder, Fien (2003). "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". CoNLL.

[14] Jenny Rose Finkel; Trond Grenager; Christopher Manning (2005).
"Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling" 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363–370.