Modified Privacy Preserving Data Mining System for Improved Performance

[1]Harshada A. Deshpande, [2]Harshali P.Patil

[1,2]Dept. of Computer Engineering, Thakur College of Engg. & Tech. Mumbai, India

Abstract

Privacy of information and security issues now-a-days has become the requisite because of big data. A novel framework for extracting and deriving information when the data is distributed amongst the multiple parties is presented by Privacy Preserving Data Mining (PPDM). The concern of PPDM system is to protect the disclosure of information and its misuse. Major issue with PPDM that exists is to use the coherent data mining algorithm for preserving privacy of data. Various PPDM techniques have been proposed till now. One of them is DPQR (Data Perturbation & Query Restriction) algorithm, which is implemented only on Boolean data. In proposed approach, the SVD (Singular value decomposition) data perturbation technique is applied for data modification; discretization of raw data is done and generates the perturbed/distorted data. SVD technique improves the level of privacy protection by providing higher degree of data distortion. The algorithm is applied on perturbed data with association rule mining and Hamiltonian matrix concepts to find out frequent itemsets. By this the confidential data is preserved. The performance metrics for the approach are privacy level, efficiency, scalability and data quality. The time required for calculating matrix inversion is reduced by using Hamiltonian matrix. Main performance metrics is the privacy level and the privacy preserving degree is improved by dimensional reduction based perturbation i.e. multi dimensional perturbation (SVD and NMF) data perturbation techniques. The novelty of the proposed method is, it is being applied on numeric data and expected to achieve comparable parameters as shown on Boolean data.

Keywords

PPDM, DPQR, Data Modification, SVD

I. Introduction

Data mining is the process of extracting hidden information from large database into understandable manner. It is emerging as one of the key features of many business organizations. Because of recent advancement in hardware technology, ability to store personal data about users increased, and continuous refinement of data mining algorithm lead to misuse of data, hence the problem mining data preserving its privacy arise.

Privacy Preserving Data Mining (PPDM) is known as the new era of research in data mining. The Privacy Preserving Data Mining (PPDM) algorithms deals with extraction of hidden predictive information from large databases and preserve sensitive information from divulgence or inference. Three philosophical approaches used in PPDM research are: (1) data hiding, (2) rule hiding, (3) Secure Multiparty Computation (SMC). In order to preserve privacy, various data transformation methods are used for privacy computations. Various methods for privacy preserving have been proposed; still a lot of research work is being carried out.

Various data mining techniques are association, classification, clustering and regression. Finding patterns in data is the most important tasks of data mining. Association Rule mining technique is considered here as the problem of building privacy preserving algorithms. To preserve confidentiality of data, various techniques for modifying or transforming the data are proposed. A survey on some of the techniques used for privacy-preserving data mining may be found in [5]. The main objective is to preserve confidentiality of data, as extracting important information from large databases is achieved by data mining. The main purpose of this method is security of database and keeping the utility and certainty of mined rules at highest level.

II. Literature Review

There has been a remarkable progress of using association rules mining in privacy preserving of data making it an important research branch of data mining [4]. The slicing concept for privacy preserving data publishing was proposed by Tiancheng Li et al [2] used data anonymization with the help of bucketization. Pui k. Fong and Jens H. Weber-Jahnke proposed an approach utilized for decision tree mining that protects centralized sample data sets [3] which provided privacy through data set complementation and decision-tree building process.

Rizvi et al. first proposed the MASK (Mining Associations with Secrecy Constraints) algorithm [6]. In MASK algorithm privacy preserving data mining was achieved using randomized disturbance and reconstruction of distribution for association rules. Its practical applications in day to day life were limited because of low time-efficiency of reconstructing original support. S. Agrawal et al proposed EMASK (Efficient Mask) which improved the time efficiency [7]. The exponential complexity of reconstructing original support was not achieved in EMASK. Andruszkiewicz proposed a new optimization technique MMASK (Modified MASK) [8], to decrease time complexity than EMASK. But in all the algorithms proposed the privacy preserving degree is low which are based only on the data perturbation approach. Haoliang Lou, et al proposed DPQR algorithm (Data Perturbation and Query Restriction) method based on improved MASK algorithm, integrated two strategies data perturbation: data discretization, data transformation and add noise to raw data, Query restriction: applying data hiding, partitioning and sampling [1]. By implementing these two strategies privacy preserving degree has been improved. The method to calculate the privacy preserving degree is also given. The main drawback of this approach is it is only suitable for Boolean data and future research is to make the algorithm suitable for numerical and other type of data.

III. Proposed Method

A. Methodology of PPDM

The methodology of PPDM System is that it converts data (Raw data) into perturbed/distorted data. The frequent item sets are acquired from perturbed data by applying data mining techniques; here association rule mining is applied. The PPDM algorithm generates the frequent item sets which are further used by data miners to extract information; by this the privacy of original data is preserved.

B. Modified Technique

In existing DPQR algorithm by matrix block method recursive relations between adjacent inverse matrix are found, hence reducing the time complexity for calculating M_n^{-1}. Further the algorithm is implemented on test data and the privacy preserving degree is calculated with minimum support 0.3%. The DPQR algorithm is suitable only for Boolean Data set.

To carry forward the future scope of [1] and overcome the issues I would modify the PPDM (DPQR) algorithm to implement it on the numerical and other type of data. The test data considered is the medical data in numeric form. Further, implement the algorithm on test data from other domain and calculate the privacy preserving degree from the algorithm. So, this methodology considers the numeric data in matrix form. In order to apply Data perturbation or add noise in data, the data distortion can be done by using Singular Value Decomposition (SVD) or Non-negative Matrix Factorization (NMF) as they both provide the higher degree of data distortion. Moreover, these SVD and NMF techniques provide high-level accuracy in results mined data as well as high-level of privacy [9]. By using Hamiltonian Matrix concept the time complexity of calculating M_n^{-1} is reduced, as it uses complex conjugate method. After this the algorithm will be applied on the distorted data and by using association rule mining the frequent item sets can be found. Refer Fig. 1.

IV. Expected Output

The expected output of the algorithm is the frequent item sets from the data set where the data is preserved. By applying data perturbation technique on the data and using association rules mining frequent itemsets are searched from perturbed data. The privacy preserving degree will also be calculated. The performance of privacy preserving data mining algorithm will be measured on the basis of its performance parameters such as privacy level, efficiency, scalability, resistance to data mining algorithm. Compare the running/execution times of the algorithm for two data sets.

Efficiency, is the ability of a algorithm implied on all the resources is to execute with good performance in terms privacy preserving of data. Efficiency improved significantly.

Scalability, evaluates the efficiency of relevant information which is mined while ensuring privacy is the trend of a PPDM algorithm for increasing sizes of the data.

Data Quality, which evaluate the impact of the sanitization on the data- base DQ. By using SVD data utility of original data is maintained.

Privacy level, to estimate the degree of uncertainty, for checking whether sensitive information is hidden and can be predicted by correlation between dimensions.

V. Conclusion

Development of PPDM became necessity for the privacy concerns of personal information by various institutions over the ever-increasing gathering ability of data. Preserving Data Mining (PPDM) presents a novel framework for extracting and deriving information when the data is distributed amongst the multiple parties. In this paper, new modified technique is proposed which is suitable on numeric datasets. The modified algorithm uses SVD perturbation technique and applies it on numerical data which provides higher degree of privacy preservation of data. The aim data modification technique and data mining used is that after mining the private data and knowledge remains secured.

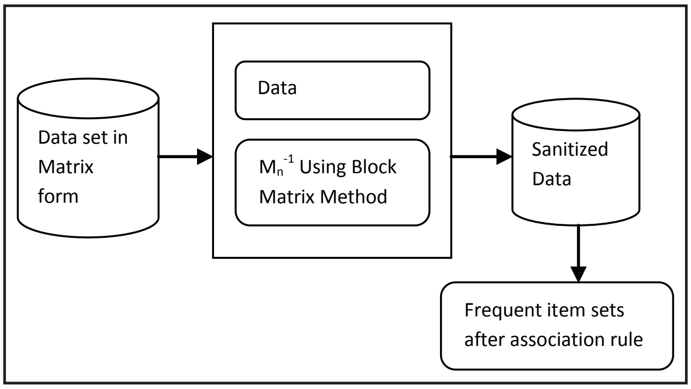
Data set in Matrix form
Data
M_n^{-1} Using Block Matrix Method
Sanitized Data
Frequent item sets after association rule

Fig. 1: Modified PPDM

References

[1] Haoling Lou, Yunlong Ma, Feng Zang, Min Liu, Weiming Shen,“Data Mining for Privacy Preserving Association Rules Based on Improved MASK Algorithm”, Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.

[2] Tiancheng Li, Ninghui Li, Ian Molloy,“Slicing: A new approach for Privacy Preserving Data Publishing”, IEEE Transactions on Knowledge and Data Engineering, Vol. 24. No. 3, March 2012.

[3] Pui K. Fong, Jens H. Weber-Jahnke,“Privacy Preserving Decision Tree Learning Using Unrealized Data Sets”, IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 3, Feb. 2012.

[4] S. Vijayarani, Dr. A. Tamilarasi, R. Seetha Lakshmi “Privacy Preserving Data Mining Based on Association Rule-A survey”, Proceedings of the International Conference on Communication and Computational Intelligence 2010.

[5] S.V. Vassilios, B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel, T. Yannis,“State of the Art in Privacy Preserving Data Mining”, Published in SIGMOD Record, Vol. 33, No.1, 2004, pp. 50-57.

[6] S.J.Rizvi, J. R. Harita,“Maintaining data privacy in association rule mining”, International Conference on Very Large Data Bases, Hong Kong, 2002, pp. 682-693.

[7] S. Agrawal, V. Krishnan, J. R. Haritsa,“On addressing efficiency concerns in privacy-preserving mining”, International Conference on Database Systems for Advanced Applications, Jeju Island, 2004, pp. 113-114.

[8] P. Andruszkiewicz,“Optimization for MASK scheme in privacy preserving data mining for association rules”, International Conference on Rough Sets and Intelligent Systems Paradigms, Warsaw, 2007, pp. 465-474.

[9] MohammadReza Keyvanpour, Somayyeh Seifi Moradi, “Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification- based Framework”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2 Feb 2011.

Harshada A. Deshpande, received her graduate degree in Computers. Currently she is a M.E. Scholar studying in Thakur College of Engineering & Technology, Mumbai. Her M.E. project is based on Data Mining and it's Security.

Harshali P. Patil, is an Assistant Professor at the Department of Computer Engineering, Thakur College of Engineering & Technology,Mumbai. Her research interest is in the area of Data Mining. At TCET additionally, she is also a ME Coordinator where she motivates the students for planning and execution of the research work. She has worked on many research based projects that involve innovation.