International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 01
January 2018

Preserving Privacy of Pre and Post Mining Using Balanced Constraint Measure Algorithm on Temporal Datasets

B. Laxmikantha & Dr.Arvind Kumar Sharma
Research Scholar, OPJSU,Churu,Rajasthan,India.
Associate professor, Dep of CSE, OPJSU,Churu,Rajasthan,India

ABSTRACT

Recently, it is observed that data mining technique may come across two problems- potential discrimination and potential privacy violation. Discrimination occurs as a result of use of discriminatory datasets for data mining tasks. Privacy violation occurs if a person's sensitive information is displayed to an unauthorized entity as a result of data mining tasks. Use of privacy preserving techniques to make data privacy protected can affect the amount of discrimination caused. It is important to study the relation of privacy and discrimination in the context of data mining. In this paper, we are trying to propose a method in which privacy preserving technique can be used to prevent discrimination and we can make the original data both privacy protected and discrimination-free.

Keywords

Discrimination discovery, discrimination prevention, privacy preserving techniques

1. INTRODUCTION

Data mining is widely utilised in the medical examination field to extract a useful report from patient data. Today, the individual data can refer to humans as well as data mining algorithms supported by companies and hospitals, and leading to significant privacy threats to individuals. In recent years, various algorithms have been proposing for modifying or transforming data to preserve privacy. The data can be processed by the privacy-preserving algorithms become less efficient compared to the original data. And To retain the fidelity of the data and also implement sufficient privacy, a dynamic, balanced algorithm has to be developed.

1.1 A DATA MINING

Data Mining (Divya et al. 2013) is an essential and stimulating area of research for extracting meaningful knowledge from massive data sets. Data Mining is growing popular today in healthcare because there is a need for an efficient analytical methodology for detecting unknown and valuable information from health data. The Data Mining provides several benefits in healthcare such as fraud detection in health insurance, providing medical solutions to patients at lower cost, the discovery of causes of diseases and description of medical treatment methods. The data produced by health organisations (details regarding hospitals, patients, medical claims, treatment costs, etc.) is difficult to analyse (as it is very vast and complex) for making the critical decision concerning patient health. So, there is a need to generate a powerful tool for analysing and selecting relevant information from this complex data. The interpretation of health data improves healthcare by improving the performance of patient management tasks. And The outcome of Data Mining technologies is to provide advantages to healthcare organisations for grouping patients having similar diseases or health issues so that healthcare methods offer effective treatments. It can also be useful for prognosticating the length of stay of patients in a hospital, for medical examination and planning for efficient knowledge system administration. Recent technologies are being used in medical field to enhance the healing services cost-effectively. A Data Mining techniques are also used to analyse various factors that are responsible for diseases, e.g. type of food, different working environments, education levels, living conditions, availability of potable water, health care services, cultural, environmental and agricultural factors. Data Mining Algorithms (Aura Conci et al. 2002) have a considerable potential to produce a knowledge-rich environment to improve the quality of clinical decisions significantly.

1.2 THE PRIVACY PRESERVING ALGORITHMS

Various methods are available for the privacy-preserving data mining to include randomisation, k-anonymization and distributed privacy-preserving data mining. The randomisation method: In the randomisation method for privacy-preserving data mining, noise is added to the data to mask the attribute values of records (Agarwal 2002). The sound combined is sufficiently large so that individual record values cannot be recovered. And Therefore techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed to extract useful information from these total distributions.

The k-anonymity model and l-diversity: These avoid the possibility of indirect identification of records from public databases. And This is because combinations of record attributes can be used to identify individual files precisely. In the k-anonymity method, the granularity of data representation is reduced by using techniques such as generalisation and suppression. This granularity is reduced sufficiently so that any given record maps on to at least k other documents in the data. The l-diversity model (Gaoming Yang et al. 2013) was designed to handle some weaknesses in the k-anonymity model since protecting identities to the level of k-individuals is not the same as preserving the identical sensitive values, especially when raw values are homogeneous within a group. And To do so, the concept of intra-group diversity of sensitive values is promoted within the anonymisation scheme (Machanavajjhala et al. 2006).

1.2.1 Privacy Preservation Distributing

Often, users may wish to derive aggregate conclusions from data sets which are partitioned across entities. Such partitioning may be horizontal (when the records are assigned to multiple objects) or vertical (when the attributes are spread across various substances). While the data partners may prefer not to share their entire data sets, they may consent to limited information sharing after implementing a variety of protocols. The overall desired result is to maintain privacy for each entity while deriving satisfactory total results over the entire data.

2. LITERATURE SURVEY

2.1 INTRODUCTION

Information mining systems have broadly used for various business exercises and different assembling organizations crosswise over numerous industry segments. The crude information is shared or sharing the removed data in a type of tenets it turns into a pattern among business organizations, as it should be a commonly advantage method for expanding efficiency for all gatherings included. The issue of shielding delicate learning mined from databases. The touchy learning is spoken to by an extraordinary gathering of affiliation rules called delicate affiliation rules. These principles are fundamental for key choice and must stay private (i.e., the standards are private to the organization or association owning the data).Data proprietors need to know ahead of time some information (governs) that they need to secure. Such guidelines are principal in basic leadership, so they should not be found.

In the primary area of this part displays distinctive methodologies for the protection safeguarding in the information bases(pre mining),the next segment which manages the security conservation in the successive example mining and affiliation lead stowing away and in last segment manages the adjusting the security conservation and the learning revelation.

2.2 A SURVEY ON PRIVACY PRESERVING DATA MINING

Jian Wang et al (2009) plans to emphasize a few protection safeguarding information mining innovations obviously and afterward continues to investigate the benefits and deficiencies of these advances. In the current years the advances in the innovation has prompt absorption of colossal measure of information. These information can be put away and utilized for a different purposes. They might be prepared in such a way it might prompt abuse of information .this force a distinct fascination in the region of security safeguarding in information mining.

Security Preserving Data mining Analysis is (Lindel et al 2000) an amalgamation of the information of heterogeneous clients without unveiling the private and vulnerable subtle elements of the clients. They proposed randomized reaction methods to comprehend the DTPD (Building Decision Tree on Private Data)problem. The essential thought of the randomized reaction is to scramble the information such that the focal place can't tell with probabilities superior to a predefined edge whether the information from a client contain honest data or false data. It presented the buildup approach which develops contained bunches in the informational collection and after that produces pseudo information from the measurements of these groups. The requirements to the bunch are characterized in the terms of sizes of the groups which are picked in a route in order to protect k-secrecy (Berberidis et al 2005). Usage of different strategies like information change, applications to save security, cryptographic techniques, and higher dimensionality challenges.

2.3 AN EFFICIENT ALGORITHM FOR PRIVACY PRESERVING TEMPORAL PATTERN MINING

Various research works uncover that the idea of information mining for the most part manages the extraction of possibly valuable data from substantial accumulations of information for an assortment of useful territories, for example, client relationship administration, showcase bin investigation, and bio-informatics (Ali et al 2007). Information digging can be utilized for anticipating and examining the medicinal records of healing centers in a town, e.g. potential flare-ups of irresistible illnesses, investigation of client exchanges for statistical surveying applications, and so on. It is in this manner basic to advance successful sharing of restorative information for significant coordinated effort inside the human services group and with different gatherings, (for example, inquire about organizations, pharmaceutical and insurance agencies), in order to improve the quality and viability of medicinal services arrangements. For instance, a healing facility may need to share clinical records from its autonomous databases to an exploration foundation trying to make another medication or assess another treatment.

3. PROPOSED SYSTEM

In this proposed work, the privacy is applied on the medical database by the data owner, i.e. the hospital. Hospitals (data owners) may wish to share their medical databases with other organizations: drug manufacturers, medical insurance companies or other hospitals. The data owners need to maintain the privacy on the patients' medical database .The privacy patterns on patients' medical data are customized according to the organizational need of the users. The data owner chooses which parts of the data are to be anonymized. Here, the privacy was applied by converting patient name and disease name to numbers by ASCII code, and privacy was also applied on sensitive diseases, frequent diseases, seasonal diseases and geographical diseases. After this, the prefix span algorithm was used to mine the importance diseases. The results of the mining did not compromise patients' privacy and also retained adequate meaningful data for users.

4. RESULTS

The experimental results of the proposed privacy-preserving algorithm are described in this section. The proposed algorithm is evaluated in terms of running time, memory usage, and variations on significant diseases. The generated sanitized database is evaluated in terms of Knowledge Discovery and Information Loss. The proposed methodology is compared with the earlier privacy preserving algorithms (Arumugam et al 2013 and Arumugam et al 2014).

4.1 Experimental Design

The proposed approach is implemented using java (jdk 1.7). The experimentation of the technique was carried out using a synthetic medical database with a dual core processor PC machine with 2 GB main memory running in 32 bit version of Windows 7 Operating System. The synthetic medical dataset generated in this proposed system contains four attributes: patient name, geographical location, disease name, and disease duration. The medical dataset consists of 10000 numbers of data.

4.2 Evaluation of Running Time

In this section the time requirement for each algorithm— (Arumugam et al 2013, Arumugam et al 2014) and proposed algorithm are evaluated to create the sanitized database from the original database.

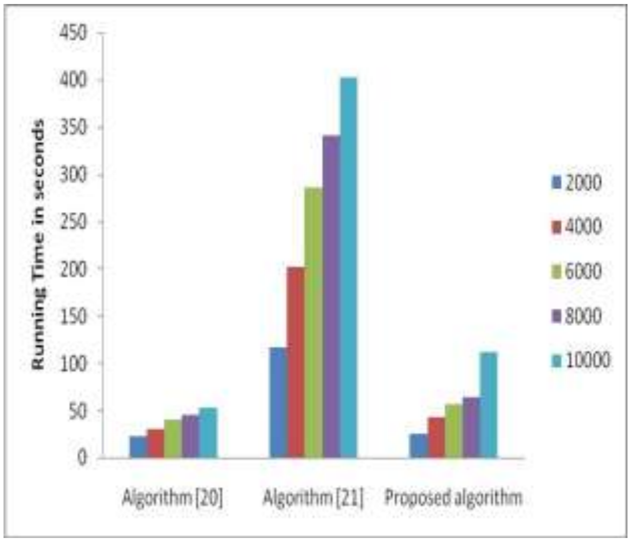

Figure 4.2 Evaluation of running time

Figure 4.2, which represents the evaluation of running time, shows that the algorithm (Arumugam et al 2014) requires the longest execution time for the sanitization process. This is because it generates the sequential rule from the original database, and then applies the sanitization process for significant diseases using the sequential rule. As the proposed algorithm does not generated the sequential rules from the original database, it requires less running time compared to the algorithm (Arumugam et al 2014). But because the proposed algorithm performs the sanitization process for the whole set of significant diseases, it takes longer than algorithm (Arumugam et al 2013), which performs the sanitization process for one significant disease at a time and does not check the KD and PF criteria.

4.3 Evaluation of Memory Usage

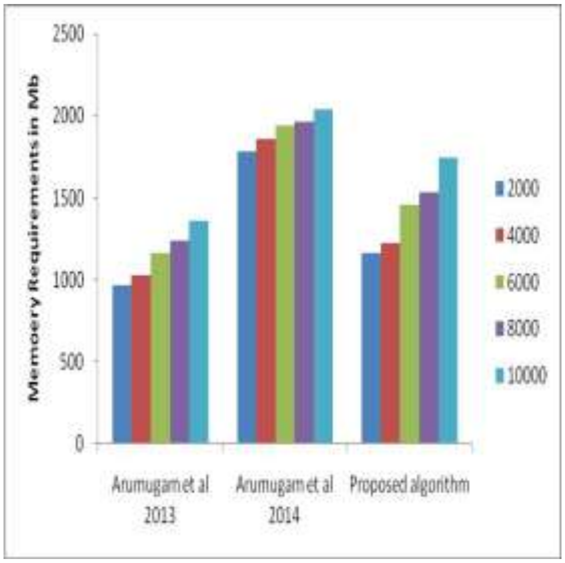

Figure 4.3 Evaluation of memory requirements

Figure 4.3, which speaks to the assessment of memory utilization, demonstrates that calculation (arumugam et al 2014) requires the greatest memory for the purification procedure. This is on the grounds that it produces the successive run from the first database (which requires more memory space to store the created consecutive standards), at that point applies the sterilization procedure for huge ailments utilizing the consecutive manage As the proposed calculation does not create the successive principles from the first database, it requires less memory space than calculation (arumugam et al 2014).

In any case, as the proposed calculation plays out the sterilization procedure for the entire arrangement of huge infections, it needs more memory space than calculation (arumugam et al 2013), which plays out the cleansing procedure for one noteworthy malady at any given moment and it doesn't check the KD and PF criteria. Here, the normal memory space necessity for calculation (arumugam et al 2013) in the assessment procedure is considered

4.4 Evaluation of Number of Diseases Changed

Figure 5.4, which represents the evaluation of number of diseases affected by the sanitization process, shows that algorithm (arumugam et al 2013) has more number of diseases affected by the sanitization process. This is because it makes sanitization process for each significant disease (sensitive, frequent, seasonal and geographical) separately.

As the proposed algorithm directly removes the significant diseases from the original database, it has greater number of diseases affected by the sanitization process than algorithm (arumugam et al 2014), which applies the sanitization process on the generated sequential rules.

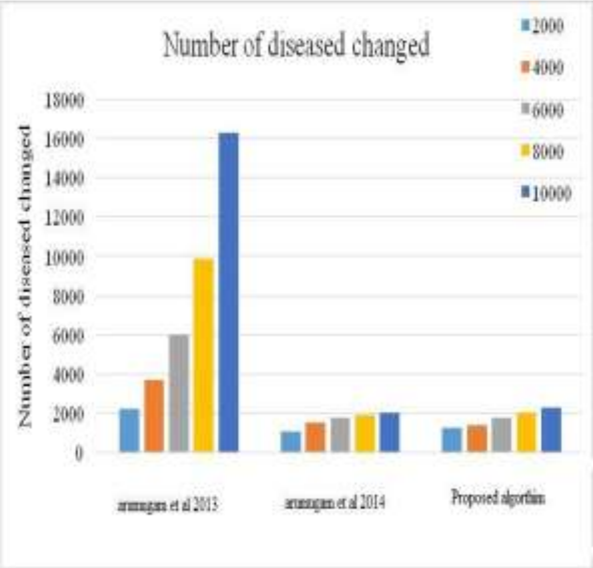

Figure 4.4 Evaluation of number of diseases changed

4.5 Evaluation of Significant Diseases Changed

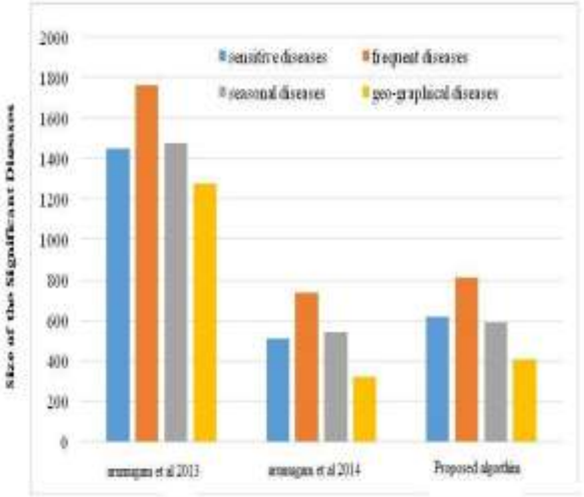

Figure 4.5 Evaluation of each significant disease using different algorithms

This section evaluates each significant disease for the privacy preserving algorithms (arumugam et al 2013, arumugam et al 2014 and proposed algorithm). The input data is taken as 5000 (constant) for each algorithm and the significant diseases affected by each algorithm are evaluated. Figure 6.5 (Evaluation of each significant disease using different algorithms) confirms that frequent diseases are affected more than the other diseases in all algorithms. Algorithm (arumugam et al 2013), which removes every significant disease from the original database (i.e. the released sanitized database) will not have any significant diseases.

Instead of sharing the database, algorithm (arumugam et al 2014) releases the sanitized sequential rules by removing the significant diseases from the generated sequential rules since the number of affected significant diseases are less when compared with the result of algorithm (arumugam et al 2013) and the proposed algorithm. The proposed algorithm has found more removable significant diseases compared to algorithm (arumugam et al 2014).

5. DISCUSSION

This work has exhibited a calculation that effectively creates a cleaned database which keeps up the best possible harmony between the Information Privacy and Knowledge Discovery. In the wake of creating the noteworthy infections from the first database, the first database is changed over into cleaned database through the proposed arbitrary disinfection process. Accordingly the produced sterilized database was assessed on the criteria of learning revelation (KD) and security factor (PF) by looking at the noteworthy ailments created from the first database and the disinfected database. On the off chance that the KD and PF esteems in the purified database fulfill the client determined edge esteems, it (cleaned database) is prepared for distributing, else the proposed calculation rehashes the disinfection procedure until the KD and PF esteems are fulfilled.

6. CONCLUSION

Data mining techniques are applied across various domains, irrespective of the fields of application. Applying mining techniques yields huge benefits for users but can also extract private information of an individual or company (secret information). Privacy preserving algorithms are deployed to protect data secrecy and individuality. In recent scenarios, balancing of knowledge discovery is another issue in data mining applications, because the privacy preserved knowledge becomes worthless to the user due to excessive data loss. Earlier, privacy preserved data mining algorithms focused only on hiding or masking data, without focus on the quality of knowledge discovery. Previous attempts were either in pre- or post-mining process with different data sets; the scope of the knowledge discovery was very less and information loss was too high. The main objective of the present study is to provide a generic approach for data owners to protect the privacy of individuals (data owners) and increase the scope of knowledge discovery either in pre- and post-mining. This thesis gives the hospital (data owner) decide to control over its data protection using either pre- or post-mining principles (or both).

REFERENCE

[1] Jingjing Qiang, Bing Yang, Qian Li & Ling Jing 2011, „Privacy-preserving SVM of horizontally partitioned data for linear classification", Image and Signal Processing (CISP), 2011 4th International Congress, vol. 5, pp.2771- 2775.

[2] Jingquan Li & Michael Shaw, J 2012, „Safeguarding the Privacy of Electronic Medical Records", Cyber Crime: Concepts,

Methodologies, Tools and Applications, vol. 2, no. 1, pp. 232-249.

[3] Kantarcioglu, M & Clifton, C 2004, „Privacy-preserving distributed mining of association rules on horizontally partitioned data‟, Knowledge and Data Engineering, IEEE Transactions, vol.16, no. 9, pp. 1026 -1037.

[4] Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam, Majedi Sampson Pun & Adepele Williams

[5] 2009, „A Data Privacy Taxonomy‟, In Proceedings of the 26th British

[6] National Conference on Databases, pp. 42-54.

[7] Keng-Pei Lin & Ming-Syan Chen 2010, „Privacy-Preserving Outsourcing Support Vector Machines with Random Transformation‟, In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 363-372.

[8] Latanya Sweeney 2002, k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557-570.

[9] Lindell Y & Pinkas B 2000, „Privacy preserving data mining‟, Advances in Cryptology, pp. 36-54.

[10] Marcia Angell 2009, „Drug Companies & Doctors: A Story of Corruption‟, The New York review of books, vol. 56, no.1.

[11] Maryam Khan 2010, „Medical Tourism: Outsourcing of Healthcare‟,

[12] International CHRIE Conference-Refereed Track, pp. 23.