Privacy Preserving Data Mining in Bio Medical Databases

D. Aruna Kumari, T. Mohan Krishna and G. Sai Pallavi

Department of Electronics and Computers, K L University, India

Abstract: Biomedical involves the applications of the natural sciences, especially the biological and physiological to clinical medicine. It is a discipline in biological, medicine to improve human health by integrating the medical usages to help the clinical practices. This can be done through the advancement of medical sciences and the database management system which creates huge number of databases in the medical world. Data mining is one of the emerging research area for extracting and analysing the knowledge from the raw bio medical data which helps to discover and manage the large heterogeneous data. This data can be used to make the work of doctors slightly less. The advancements in this field is not only limited to this sector but can be implemented into various sectors.

Key words: Randomization • Weka • Privacy

INTRODUCTION

Data mining unveils the data which is hidden in the database, but owners will not be happy if that hidden data is not confidential and they feel very uncomfortable if this data was revealed to the public. This Problem motivates Cardiac/Cancer/Diabetic patients medical records and it is used to analyse and diagnose various health conditions of a patient. Hospital management or Hospital Administration people will recognize that this kind of information storage about patient’s health is very useful and valuable for making decisions and for other data mining purposes without violating the privacy of individuals.

There are two main considerations in privacy preserving data mining

- Sensitive data of individuals like names, addresses should be altered or suppressed from the original database. So that person’s private data remains private even after mining process.
- Sensitive knowledge i.e. the hidden knowledge which is extracted from a database should also be excluded, because such knowledge can equally well compromise data privacy.

In this research, it is proposed to study various privacy preserving data mining techniques by comparing the metrics called as accuracy of data mining result and privacy of sensitive data. It is proposed to apply PPDM on healthcare systems, Privacy-Preserving Data Mining (“PPDM”) allows health organizations to extract useful hidden patterns without revealing individually identifiable information.

Generally the data is entered and maintained into Cardiac/Cancer/Diabetic patients medical records and it is used to analyse and diagnose various health conditions of a patient. Hospital management or Hospital Administration people will recognize that this kind of information storage about patient’s health is very useful and valuable for making decisions and for other data mining purposes without violating the privacy of individuals.

All healthcare organizations gather personal and health data of patients using software driven electronic health record systems. As a result, various critical health data of a patient gets distributed and stored in different databases of various healthcare organizations. Patients change healthcare providers when they relocate, change jobs, or switch insurance companies. As a result, healthcare providers need a technology that unambiguously identifies patients and provides a complete health history for each patient. Such a technology enables healthcare providers in delivering the best diagnosis and treatment decisions. Using such a technology, healthcare providers can save the cost of duplicating and gathering already existing health data.

Corresponding Author: D. Aruna Kumari, Department of Electronics and Computers, K L University, India.
Health organizations from around the world conduct research-oriented surveys to collect the statistical data about health of the people. Third-party health organizations or government authorities implement appropriate health schemes and plans using this statistical data that is received from various healthcare providers.

Before healthcare providers provide data, they must refine the data in terms of normalization (numbers), privacy (like patient name and address) and suppress the sensitive data.

Privacy-Preserving Data Mining (PPDM) extracts useful information for third-party organization without revealing individually identifiable information. PPDM provides the extracted data in a highly user-friendly format. One can configure PPDM to implement privacy and data suppression at various levels.

One of the technique for privacy preservation is randomization. It was first introduced by Agarwal and Srikanth[4]. It works on the principle of adding random noise to the original data. The randomization is effective when it is used at data-collection time. Another technique to privacy preserving data mining is to swap the values across the fields in a tuple. Data values are not obscured at all, but data inference problem is more.

Aggregation is the another method for privacy preserving data mining where k-number of records of a data set will be replaced by a representative record. The values of representative records are formed by talking the average value of all values. The problem with this method is more information loss.

One more way to achieve privacy preserving data mining is to suppress all sensitive data. In this method one can take exact attribute and it can be replaced with less informative value.

PPDM implements data suppression using the process of k-anonymization. A larger value of k provides greater the privacy protection. The k-anonymization process is superior to other de-identification approaches that simply remove certain identifiers. These approaches are prone to linkage attacks that combine the subject data with other publicly available information to reidentify the data subjects. The k-anonymization technique prevents such linkage attacks, while preserving the integrity of the de-identified data.

**Privacy Preserving in Data Mining:** Privacy preserving data mining is an area of data mining to protect sensitive information. It gives a new ray to the field of datamining without interpreting the underlying data. Classification of privacy preserving data mining is given below [ref].

![Fig. 1: Classification of privacy preserving data mining](image)

<table>
<thead>
<tr>
<th>Table 1: Before generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: After generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>[15,30]</td>
</tr>
<tr>
<td>[35,55]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Before suppression</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAME</td>
</tr>
<tr>
<td>SAI</td>
</tr>
<tr>
<td>DEEPTHI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4: After suppression</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAME</td>
</tr>
<tr>
<td>SAI</td>
</tr>
<tr>
<td>DEEPTHI</td>
</tr>
</tbody>
</table>

**Related Work:** Privacy preserving data mining has many applications and to secure the individual privacy we have many methods like anonymization, data swapping, randomization.

**Anonymization:** In the model the many attributes in the dataset may be in-conjunction with the other values that are used to uniquely identify the records. For example if the fields like DOB and ZIPCODE are used to uniquely identify the records by removing the sensitive fields from the values. The main idea in this model is to reduce granularity that is not to uniquely identify the Kth record from the (K-1) set of records. Generalization and suppression are two techniques implemented in anonymization method which are used to secure the individual data record. In generalization, original value is masked and displays the value within the interval. Suppression represents the sensitive values are masked by (*))[4-6].
Data Swapping: Additive and multiplication of noise are not only the distortion techniques. Data swapping is another for value distortion. Data swapping is dependent on the values of the neighboring data unlike randomization which is implemented on the independent data. In this technique values of the records are interchanged and original data is not revealed to the researchers and privacy is preserved. Drawback of data swapping is the accurate results are not shown.

Randomization: This algorithm is used over the centralized data mining and allows the secure to sensitive data of an individual[2][4]. For example, in hospital databases it has the patients related values like name, age, address, disease, contact. To apply the randomisation on this data we used the WEKA tool. We add some randomised value to the attributes which we want and we get modified dataset.

Experimental Results: To implement the randomization technique we used the software called “WEKA”(Waikato Environment for Knowledge Analysis). This has the applications like Explorer, Experimenter, Knowledge flow, simple CLI. In explorer we follow steps like preprocess, classification, clustering, association, attribute selection, visualization. In Preprocessing the data is choosen and modified. In classification train and test learning schemes that classify or perform regression. In clustering the data is divided into clusters. In association we learn association rules for the data. Attributes selection, selects the most relevant attributes in the data. In visualization it views an interactive 2D plot of the data. This tool helps us to get the precise outputs. We take the example of the patient data related to diabetes which consists the attributes like preg, age, class. This sensitive information is modified to nearest neighbouring value which is not seen when the researchers want to research on particular area.

Clustering: The Clustering means grouping of objects or set of objects in such a way that objects in the same group can be called as cluster are similar to each other than other clusters which are differ. It can be considered the most important unsupervised learning problem; which deals with the unlabelled data and making it into a proper structure. Another definition of clustering could be “the process of organizing objects into groups in some way where members are similar”. Therefore a collection of objects which are coherent internally, but clearly dissimilar to the objects belonging to others.

Homogeneity: Objects Belonging To The Same Cluster Are Similar To each other. Separation: Objects Belonging To Different Clusters Are Dissimilar To each other.

Three Elements are 1) The Set of objects, 2) The Set of Attributes
In this Project we deal with the data which is not structured, available in hospitals that contains information of the patients and their details of their sickness, medicines used for their treatment, Time span for their health recovery. For a better analysis and to get help in clinical practises we data mine the data by clustering technique. This work helps to predict the disease occurrences in the people and can be able to judge what precautions to be taken and can be helpful to avoid them.

In this work, We data mine the data available into different cluster like depending on their disease occurred. This help to identify the patients easily and number of people effected for diseases categorized.

Again we analyse the data in the different clusters of diseases. In it we categorize the data regarding ages of the patients. Through this we can judge that which age group people are effected for particular disease and medicines to be used for aged group, dosages to be given to particular group and can predict the time to recover their health.
Fig. 5: Classification of random tree classifier

Fig. 6: Classifier output
Fig. 7: Dividing the dataset into clusters

Fig. 8: Comparison of datasets
In the above diagrams, we categorized the analyzed into different groups as per the age of the patients. This helps to know the clear details regarding the patients and their issues with disease and the treatment given to the patients is observed, time taken for their recovery is also observed.

The process is done because the treatment for patients may be differ from patients to patients depending on their age and health factors. The medicine usage may be different. The dosages can be differ with seriousness of the disease. The effects occurred by the disease. The reasons for the disease occurrence can be easily understood.

For example the treatment used for a child and adults may be differ, because the immune system and the age factors, growth of body, physical conditions are essential. The procedure for an adult is independant with the child and the medical dosages used for them is separate. the administrators in a hospital need to decide on the amount of supplies and number of staff and free beds required for an upcoming month. To make this decision, the administrators require an accurate prediction of the number of patients to expect during the coming month and an approximation of how long each patient will remain in the hospital. As another example, the federal and provincial health administrators need to decide whether a disease outbreak is in progress and if so, what preventive measures will be most effective against a disease.

Through this work, the clinical practitioner’s may be benefited and can treat the patients efficiently with less time. They can provide appropriate medicine required accroding to the data provided. They can predict the disease stage. They can estimate the severeness of the disease. The performance benchmarking for clinicians by utilizing the same clinical data that is used for making healthcare decisions in improving the quality of healthcare delivered to the patients.

CONCLUSION

From the results above stated it is clearly observed how the randomization algorithm is being implemented on the dataset that is taken and the results are properly analysed how much privacy is being done on the dataset by choosing the algorithm.

ACKNOWLEDGMENT

We would like to thank Dr.D.Arunkumari for useful discussions and suggestions related to this work.

REFERENCES

5. Methods and Techniques to Protect the Privacy Information in Privacy Preservation Data Mining N.Punitha R.Amsaveni.
6. Vector Quantization For Privacy Preserving Clustering In Data Mining D.Aruna Kumari, Dr.Rajasekhar Rao and M.Suman.