PRIVACY-PRESERVING DATA MINING OF MEDICAL DATA USING DATA SEPARATION–BASED TECHNIQUES

Gang Kou [1]***, Yi Peng***[1*]***, Yong Shi***[1, 2]***, and Zhengxin Chen***[1]

[*1]*College of Information Science & Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA*
Email:{gkou, ypeng, yshi, zchen}@mail.unomaha.edu

[2]*Chinese Academy of Sciences Research Center on Data Technology & Knowledge Economy, Graduate University of the Chinese Academy of Sciences, Beijing 100080, China*

ABSTRACT

Data mining is concerned with the extraction of useful knowledge from various types of data. Medical data mining has been a popular data mining topic of late. Compared with other data mining areas, medical data mining has some unique characteristics. Because medical files are related to human subjects, privacy concerns are taken more seriously than other data mining tasks. This paper applied data separation-based techniques to preserve privacy in classification of medical data. We take two approaches to protect privacy: one approach is to vertically partition the medical data and mine these partitioned data at multiple sites; the other approach is to horizontally split data across multiple sites. In the vertical partition approach, each site uses a portion of the attributes to compute its results, and the distributed results are assembled at a central trusted party using a majority-vote ensemble method. In the horizontal partition approach, data are distributed among several sites. Each site computes its own data, and a central trusted party is responsible to integrate these results. We implement these two approaches using medical datasets from UCI KDD archive and report the experimental results.

Keywords: Classification, Privacy-preserving data mining, Medical data mining, Vertically partitioned data

1 INTRODUCTION

Data mining or knowledge discovery in databases (KDD), which focuses on the extraction of useful knowledge from large amount of data, has steadily attracted researchers and practitioners from various fields. As early as 1989, when the first KDD workshop was held in Detroit, Michigan, privacy issues have been brought up. This is an especially important issue in medical data mining. Medical data are normally privacy-sensitive. Compared with other data mining areas, medical data mining has some unique characteristics. Cios and Moore (2002) organized these characteristics into four groups: heterogeneity of medical data; ethical, legal, and social issues; statistical philosophy; and special status of medicine. Privacy and security of human data is one of the ethical, legal, and social issues. Many countries have enacted laws to protect data privacy. For instance, U.S. federal rules set guidelines to conceal individual patient identifiers (Cios & Moore, 2002). At the same time, data mining researchers also suggest methods and techniques to protect data privacy. Currently, there are three major classes of privacy-preserving techniques: data obfuscation, summarization, and data separation (Clifton, 2002).

The objective of this paper is to apply data separation-based techniques to preserve privacy in classification of medical data. We take two approaches to protect privacy: vertical partition and horizontal partition. In the vertical partition approach, each site uses a portion of the attributes to compute its results, and the distributed results are assembled at a central trusted party using majority-vote ensemble method. In the horizontal partition approach, data are distributed among several sites. Each site computes its own data, and a central trusted party is responsible to integrate these results. We implement these two approaches using two medical datasets from UCI Machine Learning repository: Wisconsin prognostic breast cancer dataset and heart-disease dataset.

This paper is structured as follows. The next section explains why and how we use vertical and horizontal separation techniques to protect privacy of medical data. The third section describes the classification experiments. The last section concludes the paper.

2 PRIVACY-PRESERVING MEDICAL DATA MINING: DATA SEPARATION TECHNIQUES

A good way to explain data separation-based privacy-preserving techniques is to use examples. In this section we use two scenarios to illustrate how vertical and horizontal techniques can be applied to protect medical data privacy. Vertical separation techniques are used when a data owner wants a third party to analyze data for him/her. A data owner may be a hospital or a medical center. Though the third party or parties involved in the data mining process is trusted, the data privacy will be more reliably guarded if we vertically partition the data. The basic idea of vertical separation is that only the data owner has the entire dataset and each third party has only a portion of the dataset. Hence data privacy is protected. Take the Wisconsin prognostic breast cancer dataset as an example. Let's say the data owner decides to ask several third parties to analyze the data. There are nine variables in this dataset. We can remove one variable at a time and create nine sub-datasets. Each sub-dataset has only eight variables. These sub-datasets are analyzed at nine separate sites, and the results are returned back the data owner. The data owner then can run a majority-vote procedure to obtain the final classification results. Figure 1 illustrates this process.

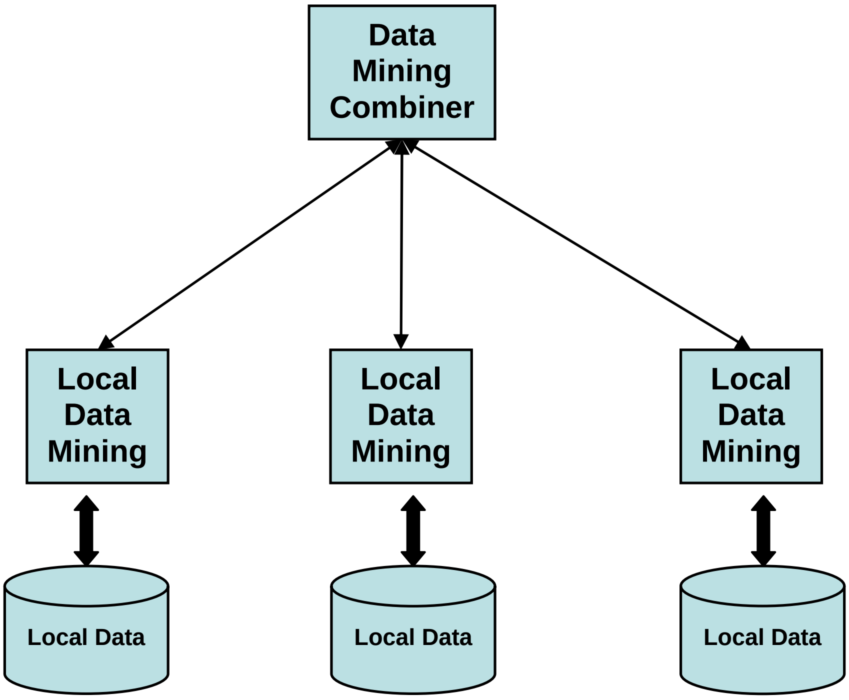

Figure 1. Private Distributed Data Mining (Copied from Clifton 2002)

Horizontal separation techniques are used when the datasets are distributed among multiple data owners. Each data owner has the ability to analyze his/her data. The problem is that each dataset has limited data objects and therefore can not produce high-quality classifiers. In other words, classifiers generated by small datasets often lack generalizability. On the other hand, data owners often do not want to share their data. The heart-disease dataset from UCI Machine Learning repository belongs to this situation. This dataset was collected from several locations, such as Cleveland Clinic Foundation, Hungarian Institute of Cardiology, and Long Beach V.A. Medical Center. Each data source has limited data records. Horizontally partitioned data allow data owners to keep their own data and increase the classification accuracy by sharing classifiers, such as classification rules, among data owners. Using multiple classifiers, data owners can normally increase their classification accuracy (e.g., use majority-vote method) and need not to share their actual data with others.

The next section describes the datasets, the experimental procedures, and classification results.

3 MEDICAL DATA CLASSIFICATION EXPERIMENTS

3.1 Vertical Data Separation Experiment

To implement the vertical and horizontal data separation techniques, we select two datasets from UCI repository: Wisconsin prognostic breast cancer dataset and heart-disease dataset (UCI Machine Learning repository, 2006, Bennett & Mangasarian, 1992). The Wisconsin prognostic breast cancer dataset is used for vertically partitioned data analysis, and the heart-disease dataset is used for horizontally partitioned data analysis.

The Wisconsin prognostic breast cancer dataset has 699 records and 9 variables. These records belong to either benign or malignant class. As mentioned in section 2, we created nine different sub-datasets by removing one variable at a time. Therefore, we got 9 sub-datasets with each sub-dataset having only eight variables. These sub-datasets were classified separately using See5 (Rulequest Research 2003) software with adaptive boosting and 10-fold cross validation. The results were integrated using the Privacy-preserving classification (vertical) process, as follows:

```
Privacy-preserving Classification (Vertical) Process

Input: The Medical dataset M = { $M_1, M_2, M_3, \cdots, M_n$ }, each of the medical records has m attributes

Output: Average classification accuracies for benign and malignant of the dataset in 10-fold cross-validation; scores for all records; decision trees ensemble.

Step 1 Generate m subsets with one different attribute removed from M at each time.
Step 2 Training each subset with See5 with adaptive boosting and 10-fold cross validation to get m decision trees $D_1, D_2, D_3, \cdots, D_m$ .
Step 3 Ensemble the final decision function D= { $D_1, D_2, D_3, \cdots, D_m$ }, via majority vote of the m decision trees from step 2.
Step 4 Classify M by the final decision function.
END
```

For comparison purposes, we also classify the whole dataset (i.e., with 9 variables) using See5 with adaptive boosting and 10-fold cross validation. The results are summarized in Figure 2.

Malignant error indicates the percentage of malignant records that have been misclassified as benign. Benign error indicates the percentage of benign records that have been misclassified as malignant. On the X axis, number 1 through 9 refers to the classification results of each sub-datasets; “All” refers to the classification result using the entire dataset; “MV” refers to the majority-vote classification result. Figure 2 tells us that the classification result using the whole dataset is better than using sub-datasets and the majority-vote result, except for number 2. The majority-vote result is slightly higher than the average of 9 sub-datasets for malignant class (3.32% vs. 3.59%) and slight lower than the average of 9 sub-datasets for benign class (2.84% vs. 2.6%). To summarize, using vertical data separation techniques, we can protect data privacy but classification accuracy is somewhat sacrificed.

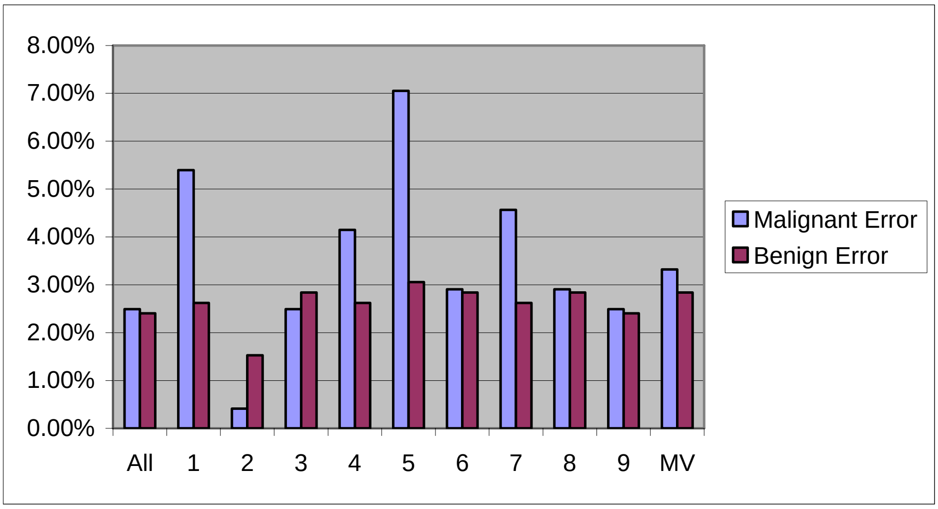

Figure 2. Wisconsin Prognostic Breast Cancer Dataset Classification Results

3.2 Horizontal Data Separation Experiment

The heart-disease dataset has 797 records and 13 variables. These records belong to either heart-disease or normal class. This data was collected from Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital of Zurich, and Long Beach V.A. Medical Center. The subset from Zurich was dropped because it is highly imbalanced. These datasets have the same set of data variables but different number of records: Cleveland has 303 records; Hungarian set has 294 records; and Long Beach set has 200 records. Each dataset is classified separately using See5 with adaptive boosting and 10-fold cross validation. The results of three datasets are integrated using the Privacy-preserving classification (horizontal) process:

Privacy-preserving Classification (Horizontal) Process

Input: The Medical dataset from r different sources, M^1 = { $M_1^1, M_2^1, M_3^1, \cdots, M_{n1}^1$ }, M^2 = { $M_1^2, M_2^2, M_3^2, \cdots, M_{n2}^2$ },…, M^r = { $M_1^r, M_2^r, M_3^r, \cdots, M_{nr}^r$ }, each of the medical records has m attributes

Output: Average classification accuracies for Normal and Heart-disease of the dataset in 10-fold cross-validation; scores for all records; decision trees ensemble.

Step 1 Establish training r datasets with See5 with adaptive boosting and 10-fold cross validation to get r decision trees $D_1, D_2, D_3, \cdots, D_r$.

Step 2 Assemble the final decision function D= { $D_1, D_2, D_3, \cdots, D_r$ }, via majority vote of the r decision trees from step 1.

Step 3 Classify all r datasets by the final decision function.

END

For comparison purposes, we also classify the combined dataset (include all three datasets) using See5 with adaptive boosting and 10-fold cross validation. The results are summarized in Figure 3.

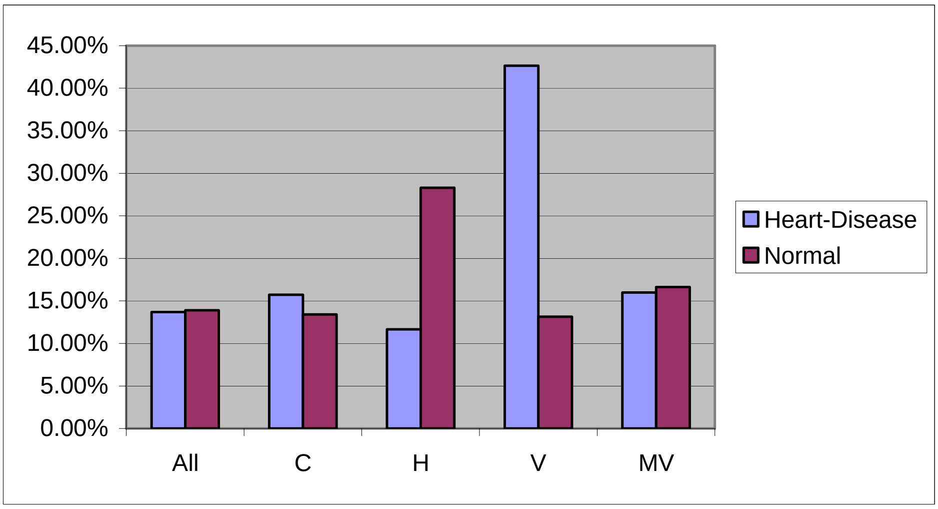

Figure 3. Heart-disease Dataset Classification Results

Heart-disease indicates the percentage of heart-disease records that have been misclassified as normal. Normal indicates the percentage of normal records that have been misclassified as heart-disease. On the X axis, C, H, and V refer to Cleveland dataset, Hungarian dataset, and Long Beach dataset, respectively; "All" refers to the classification result using the combination of three datasets; "MV" refers to the majority-vote classification result. Figure 3 tells us that the classification result using the combined dataset is better than using individual dataset and the majority-vote result. The majority-vote result is better than the average of three individual dataset for both classes (heart-disease: 15.99% vs. 23.35%; normal: 16.63% vs. 18.28%). To summarize, using horizontal data separation techniques, we can both protect data privacy and achieve fairly high classification accuracy.

4 CONCLUSION

Privacy-preservation is an important issue in medical data mining. This paper investigates data separation techniques in medical data classification. The experiments demonstrate that data separation techniques can not only protect data privacy, but also increase classification accuracy sometimes (e.g., horizontally partitioned data).

The techniques used in our experiments are straightforward, and there is much room for improvement. For instance, in the vertically partitioned data situation, simply removing one or several variables from datasets does not ensure that data can not be traced to an individual record. In such a case, more sophisticated methods and techniques are required.

5 ACKNOWLEDGMENTS

This research has been partially supported by Key Project #70531040,#70472074, National Natural Science Foundation of China; 973 Project #2004CB720103, Ministry of Science and Technology, China; and BHP Billiton Co., Australia. The authors would like to thank the principal investigator responsible for the data collection at each of the following institution for Heart-disease dataset: Hungarian Institute of Cardiology. Budapest: Andras Janosi, University Hospital, Zurich, Switzerland: William Steinbrunn, University Hospital, Basel, Switzerland: Matthias Pfisterer, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano.

6 REFERENCES

Cios, K. J. & Moore, G. W. (2002) Uniqueness of medical data mining, *Artificial Intelligence in Medicine*, Vol. 26, Issue 1-2, 1-24.

Clifton, C. (2002) Privacy, Security, and Data Mining, presented at the combined conference 13th European Conference on Machine Learning (ECML'02) and 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, 19-23 August.

Bennett, K. P. & Mangasarian, O. L. (1992) "Robust linear programming discrimination of two linearly inseparable sets", *Optimization Methods and Software* 1, 23-34, Gordon & Breach Science Publishers.

Rulequest Research (2003). Retrieved April 29, 2006, from http://www.rulequest.com/see5-info.html

UCI Machine Learning repository (2006) Retrieved April 29, 2006, from http://www.ics.uci.edu/~mlearn/databases/