Privacy Preserving RFE-SVM for Distributed Gene Selection

Fodé Camara[1], Mouhamadou Lamine Samb[1], Samba Ndiaye[1] and Yahya Slimani[2]

**[1] Department of Mathematics, Cheikh Anta Diop University
Dakar, Senegal**

**[2] University Department of Computer Science, Faculty of Sciences of Tunis
1060 Tunis, Tunisia**

Abstract

The support vector machine recursive feature elimination (SVM-RFE) is one of the most effective feature selection methods which has been successfully used in selecting informative genes for cancer classification. This paper extends this well-studied algorithm to the privacy preserving distributed data mining issue. For gene selection over multiple patient data from different sites, we propose a novel RFE-SVM method which aims to learn global informative gene subset to get the highest cancer classification accuracy, with limits on sharing of information. We experiment it using Leukemia bio-medical dataset. The experimental results show that it can provide good capability of privacy preserving and generates a set of attributes that is very similar to the set produced by its centralized counterpart.

Keywords: *Privacy Preserving, Gene selection, Distributed Data Mining, RFE-SVM, Cancer Diagnostic.*

1. Introduction

Recently, advances in computing, communications and current hardware technologies have made it possible to collect and store large amounts of data in digital form. For example, high throughput data acquisition technologies have resulted in gigabytes of gene expression data being gathered at steadily increasing rates in biological and bioinformatics sciences. This increasing ability to track and collect large amounts of data has created tremendous opportunities for knowledge-based detecting patterns.

Medical databases are often ideal candidates for large scale, and thus candidates for possibly distributed data mining applications. In fact, data mining over multiple data sources has become an important practical problem with applications in different areas. Due to the sensitive characteristics of personal health records, privacy concern is taken more seriously than other data mining applications. For example, different bioinformatics companies may wish to coordinate themselves in knowing aggregate trends. However, due to privacy concerns, their medical records cannot be brought together. Then, privacy preserving data mining (PPDM) over horizontally partitioned data can be used to achieve this.

This paper applied a privacy preserving gene selection for cancer classification over multiple patient data from different sites. For selecting relevant genes in this case, we propose a novel RFE-SVM algorithm. We experiment it using Leukemia bio-medical dataset. The experimental results show that it can provide good capability of privacy preserving and generates a set of attributes that is very similar to the set produced by the traditional RFE-SVM algorithm.

The remainder of this paper is organized as follows: In section 2, we briefly review the privacy preserving distributed data mining problem. Section 3 provides some background on the RFE-SVM algorithm and the secure multi-party problem. In section 4, we present our privacy preserving RFE-SVM approach. In Section 5, we describe the experiments. In Section 6, we analyze the experiment results. Finally, Section 7 concludes with a discussion of the contributions of our proposal and current research plans.

2. Related work

In contrast to the centralized model, the Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the data mining results from all the data across these distributed sources. Since this different sources of information are often relating to human subject, many questions concerning their privacy are raised. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends

without leaking the trends of their individual stores. As another example, we consider a center for disease control which may want to use data mining to identify trends and patterns in disease outbreaks, such as understanding and predicting the progression of a flu epidemic. Insurance companies have considerable data that would be useful for such a task, but privacy considerations prevent them from releasing the data. An alternative is that each organization performs local operations on its site; this produces intermediate data that can be used to obtain the data mining results, without revealing the private information at each site.

There are many variants of this problem, depending on how the data is distributed, what type of data mining we wish to do, and what constraints are made on shared information.

In all these alternatives, we are in front of two contradictory problems. How to solve the point of conflict between the desire to find a model starting from the union of all these databases, and the right which has an individual to preserve information relating to his privacy?

The main proposal to solve the problem of Privacy Preserving Distributed Data Mining is the Secure Multi-party Computation. A Secure Multi-party Computation (SMC) problem deals with computing any function on any input, in a distributed network where each participant holds one part of the inputs, while ensuring that no more information is revealed to a participant in the computation than its owner input and the output of the function. Secure two party computation was first investigated by Yao [1, 2] and was later generalized to multi-party computation [3]. For example, in a 2-party setting, Alice and Bob may have two inputs x and y, and may wish to both compute the function $f(x, y)$ without revealing x or y to each other. This problem can also be generalized across k parties by designing the k arguments function $h(x_1,.., x_k)$.

This approach was introduced into the data mining community for the first time by Lindell and Pinkas in [4]. It allows two different entities to build a decision tree without none of entities being able to know something about the other. Since, many techniques were suggested in the literature. We can divide those techniques in two groups: (i) distributed algorithms over vertically partitioned data; and (ii) distributed algorithms over horizontally partitioned data. There is a horizontally partitioned when the different sites may have different sets of records containing the same attributes, and a vertically partitioned when the different sites may have different attributes of the same sets of records.

The problem of distributed privacy preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations. A broad overview of the intersection between the fields of cryptography and privacy-preserving data mining may be found in [5]. Clifton et al. [6] give a survey of multi-party computation methods.

3. Preliminaries

We start this section with a subsection summarizing the RFE-SVM algorithm. Then, we continue with preliminaries on secure multi-party computation.

3.1 The RFE-SVM algorithm

The well-studied RFE-SVM algorithm [7, 8] is a wrapper feature selection method which generates the ranking of features using backward feature elimination. It was originally proposed to perform gene selection for cancer classification [7]. Its basic idea is to eliminate redundant genes and yields better and more compact gene subsets. The features are eliminated according to a criterion related to their support to the discrimination function, and the SVM [8] is re-trained at each step. RFE-SVM is weight-based method, at each step; the coefficients of the weight vector of a linear SVM are used as the feature ranking criterion. The RFE-SVM algorithm [7] can be broken into four steps:

1. Train an SVM on the training set;
2. Order features using the weights of the resulting classifiers;
3. Eliminates features with the smallest weight;
4. Repeat the process with the training set restricted to the remaining features.

3.2 The Secure Multiparty Computation problem

Consider a set of parties who do not trust each other, nor the channels by which they communicate. Still, the parties wish to correctly compute some common function of their local inputs, while keeping their local data as private as possible. This, in a nutshell, is the problem we wish to solve, privacy-preserving data mining, is a special case of the secure multiparty computation problem. Before proposing algorithm that preserves privacy, it is important to define the notion of privacy. The framework of secure multiparty computation provides a solid theoretical underpinning for privacy [3]. The key notion is to show that a protocol reveals nothing except the results.

4. Proposed Approach

4.1 Problem Definition

An inherent tension lies between using medical records for legitimate clinical research and concerns about patient privacy. Consider the example of two different bioinformatics companies. They want to coordinate in knowing relevant genes for cancer classification over the union of their patient data. Due to privacy concerns, their medical records cannot be brought together. In this framework we propose a privacy-preserving distributed RFE-SVM which aims to protect the privacy of patients while maintaining researchers' ability to analyze globally specific genes.

In [9], Fang and al. proposed architecture with a good capability of privacy preserving for decision tree learning. We extend this work to the Feature Selection field. To our knowledge, this work is the first attempt to understand different aspects of using RFE-SVM algorithm over distributed data.

4.2 Security tools

Computation on encrypted data does not make sense unless the encryption transformation being used has some homomorphic properties. To define our distributed RFE-SVM algorithm, we use the additive homomorphic encryption and decryption scheme defined in [9].

The encryption scheme is as follows:

- The algorithm uses a large number r, such that $r=p\times q$, where p and q are large security prime numbers.
- Given x, which is a plaintext message, the encrypted value $y=E_p(x) = x+p \ (mod\ r)$.

The decryption scheme is as follows:

- Given y, which is a ciphertext message, we use the security key p to recover plaintext $x=D_p(y) =y \ mod\ p$.

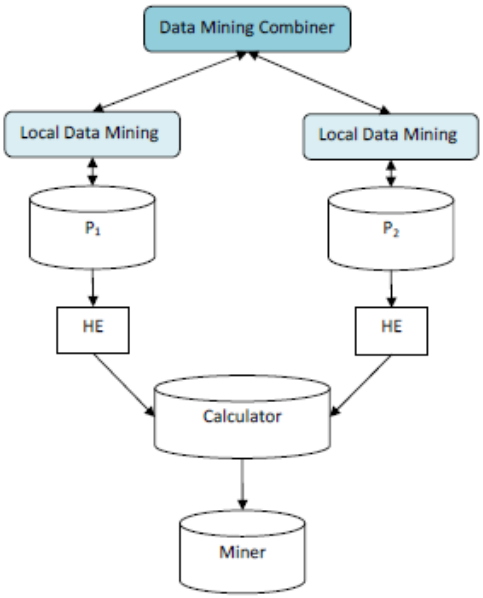

Fig. 1 Secure Multiparty Computation Architecture.

4.3 Algorithm

Assume that there are two parties named P_1 and P_2 which respectively has m_1 and m_2 sample records, and want jointly selecting the most relevant genes for cancer classification. As described in architecture (c.f. Figure 1), our privacy preserving algorithm is composed of three parts (Algorithm 1, 2, and 3).

5. Experimental studies

5.1 Experimental Set Up

The experiment was conducted with dual core 2.20 GHz with 4.00 Go of memory on windows platform, and we implemented the distributed algorithm using Java Agent DEvelopment (Jade) framework [10] and the Weka API [11].

Input: Local training set and the primary key *pk*
Output: Encrypted ranking scores
Step 1: Initialize the survived Gene subset $G_S = \{a_{1,..,} a_n\}$, with a_i the attribute at position *i*.
Step 2: Repeat until all features are ranked:
(a) Train a SVM on the local training set with genes in G_S.
(b) Compute *w*, the weight vector of the resulting local classifier
(c) Compute c_j, the ranking scores for genes in G_S: $c_j=(w_i)^2$
(d) Send $Enc_{pk}(c_j)$ to the Calculator
(e) Receive the smallest ranking score in $P_1 \cup P_2$ from the Miner, where P_1 and P_2 are the two sites.

Algorithm. 1 Pseudocode of algorithm performed by the two parties

Input: All the encrypted ranking scores from P_1 and P_2
Output: The sum of the encrypted ranking scores
Step 1: Receive all $T_1[j]=Enc_{pk}(c_j)$ from P_1 and all $T_2[j]=Enc_{pk}(c_j)$ from P_2
Step 2: Compute $T[j]= T_1[j]+ T_2[j]$
Step 3: Send the array *T* to the *Miner*

Algorithm. 2 Pseudocode of algorithm performed by the Calculator

Input: An array *T*, which contains the sums of the encrypted ranking score
Output: The gene with the smallest ranking score
Step1: Receive *T* from Calculator
Step2: Decrypt each *T[i]* using the security key *pk*.
Step3: Find the gene with the smallest ranking score and send it to P_1 and P_2.

Algorithm. 3 Pseudocode of algorithm performed by the Miner

5.2 Dataset

To demonstrate real practicality of our approach, we ran experiments on Leukemia bio-medical dataset. The Leukemia dataset consists of samples from patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). It initially contains expression levels of 7129 genes taken over 72 samples. Training dataset (Train), given to select genes and adjust the weights of the classifiers, consists of 38 samples (27 ALL and 11 AML). Also an independent test set is provided to estimate the performance of the classifiers. It contains 34 samples (20 ALL and 14 AML).

In our experiments, we only use the 1000 informative genes. To select these, we use '*InfoGainAttribute-Eval*' algorithm and Search Method "*Ranker*" of Weka API[11].

The same testing dataset is used in the two sites, and we partition the training dataset into two parts:

- In site P_1: The local training dataset (Train1) contains 19 samples (14 ALL and 5 AML).
- In site P_2: The local training dataset (Train2) also contains 19 samples (13 ALL and 6 AML).

This horizontally distributed training dataset (Train) has the following propriety: Train1 $\cup$ Train2=Train where $\cup$ denoted the set union operation.

6. Results and discussion

From the horizontally distributed data described above, we conduct experiment to evaluate the performance of our method. Then we compared its performance with the traditional RFE-SVM which is often considered as one of

the best gene selection algorithms in the literature [7]. The comparison results are shown in Figures 2, 3 and Tables 1, 2, 3. The Figure 2 and the Table 1 display the classification accuracy, varying with the number of genes, between traditional RFE-SVM (non-privacy preserving approach) and our approach (privacy preserving approach in distributed feature selection). Table 2 shows the selected gene subsets and their classification accuracy obtaining by using the traditional RFE-SVM approach and our approach, respectively, from which we can see that although the subset returned by RFE-SVM is smaller than that returned by our privacy preserving algorithm, the performances of RFE-SVM and PPD RFE-SVM classifiers are the same.

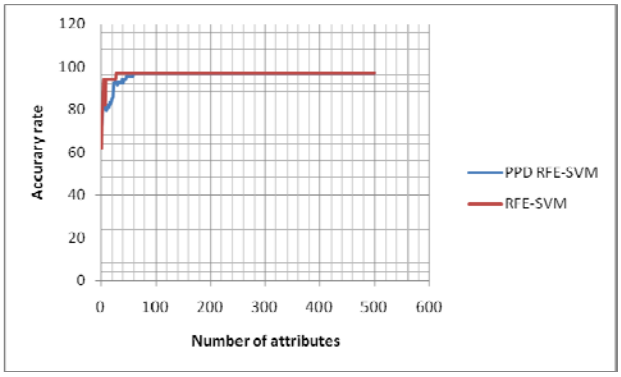

Fig. 2 The classification accuracy varying with the number of genes.

It is obvious that our privacy preserving process increases the complexity computational. But computation cost does not constitute really a problem because there are several parallel architectures.

Table 1: Accuracy rate for different values of N

N	5	10	20	30	40	50	60	1000
RFE-SVM	0.88	0.94	0.94	0.97	0.97	0.97	0.97	0.97
PPD RFE-SVM	0.81	0.81	0.85	0.91	0.93	0.95	0.97	0.97

Table 2: Performance comparisons between RFE-SVM and PP RFE-SVM

Algorithm	***The smallest number of genes***	***Accuracy rate***
RFE-SVM	28	0.97
PPD RFE-SVM	60	0.97

The Figure 4 and 5 highlight the effectiveness of our distributed gene selection method. Figure 4 shows how informative genes generated by our privacy preserving are similar to those returned by the traditional RFE-SVM. At least 60% of genes are the same. We also see that the top 5 relevant gene subsets obtained are the same. In Figure 5, we display the 10 informative genes selected by the traditional RFE-SVM and our privacy preserving RFE-SVM. Note that the optimal *k* relevant gene subset is not unique, because of the combinatorial nature of the gene selection problem.

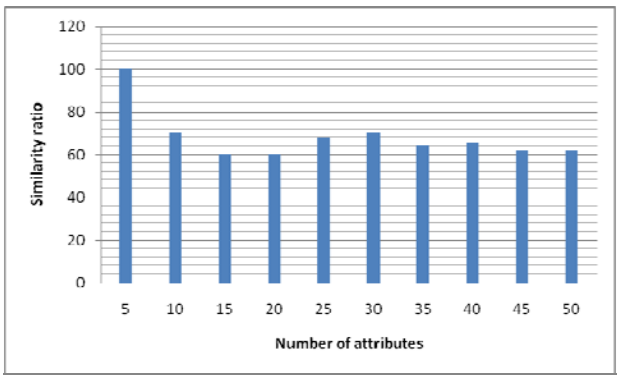

Fig. 3 Similarity between subsets returned by traditional RFE-SVM and PPD RFE-SVM

Table 3: The 10 gene subsets.

Algorithm	Selected gene subset	Accuracy
RFE-SVM	{4847, 1834[a], 1779[a], 6539[a], 5772[a], 461[a], 5039[a], 3847[a], 2001, 6184}	0.94
PPD RFE-SVM	{1834[a], 1779[a], 1745, 6539[a], 5772[a], 4499, 461[a], 5039[a], 3847[a], 1394}	0.81
[a] The genes present in the two subsets		

7. Conclusions and Future Work

In this work, we proposed a privacy-preserving RFE-SVM for distributed gene selection. It aims to select, over multiple data sources, the smallest informative gene subset to get the highest cancer classification accuracy. To our knowledge, this paper is the first attempt to understand different aspects of using RFE-SVM algorithm over

distributed data. Our experimental results show that our approach has a good capability of privacy preserving, accuracy and efficiency.

In the future, we plan to run experiments on others bio-medical datasets. We also plan to investigate the possibility of reducing the size of smallest informative gene subset while keeping the classification accuracy.

References

[1] A.C.C. Yao, Protocols for secure computations, Proc. of the 23rd Annual IEEE Symposium on Foundations of Computer Science, Chicago, Illinois, November 1982, pp. 160-164.

[2] A.C.C. Yao, How to generate and exchange secrets, Proc. of the 27th Symposium on Foundations of Computer Science (FOCS), Toronto, Canada, October 1986, pp. 162-167.

[3] W. Du, M.J. Atallah, Secure multi-party computation problems and their applications: a review and open problems, New Security Paradigms Workshop, Cloudcroft, New Mexico, September 2001, pp. 11-20.

[4] Y., Lindell, B. Pinkas, Privacy Preserving Data Mining, In Advances in Cryptology - CRYPTO 2000, pp. 36-54. Springer-Verlag, August 20-24 2000.

[5] B. Pinkas, Cryptographic Techniques for Privacy Preserving Data Mining, ACM SIGKDD Explorations 4(2) (2002).

[6] C. Clifton, M. Kantarcioglu, Tools for privacy preserving distributed data mining, SIGKDD Explorations 4(2), 28-34 (2003).

[7] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering (TDKE), 17(4), pp. 491-502, 2005.

[8] Y. Tang, Y. Q. Zhang, Z. Huang, Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(3), pp. 365-381, 2007.

[9] W. Fang, B. Yang, D. Song, Preserving Private Knowledge In Decision Tree Learning, Journal of computers, 5(5), May 2010.

[10] JADE, Java Agent Development framework, http://jade.cselt.it/

[11] H. W. Ian and F. Eibe. Data Mining Pratical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, October 1999.