Correlation-Based Feature Subset Selection for Privacy Preserving Data Mining

Jintu Ann John[1], Neethu Maria John[2]

Department of Computer Science and Engineering, Mahatma Gandhi University, Mangalam College of Engineering

Email id: jintuannjohn@gmail.com

9747825486

Abstract— In recent year's privacy preservation in data mining has become an important issue. The main aim is to protect the sensitive information in data while extracting knowledge from large amount of data**.** The data from different sources have to collaborate effectively while maximizing the utility of collected information. With the advance of the information age, data collection and data analysis have exploded both in size and complexity. The attempt to extract important patterns and trends from the vast data sets has led to a challenging field called Data Mining. When a complete data set is available, various statistical, machine learning and modeling techniques can be applied to analyze the data. Usually, data are distributed across multiple sites. The data warehousing approach has been used to mine distributed databases. It is used to collect data from all the participating sites and are stored at a centralized warehouse. However, many data owners are unwilling to share their data with others due to privacy and confidentiality concerns. So it is a limitation to perform mutually beneficial data mining tasks. To overcome this Privacy-Preserving Data Mining has emerged. The research of Privacy-Preserving Data Mining is aimed at bridging the gap between collaborative data mining and data confidentiality. It involves many areas such as statistics, computer sciences, and social sciences. The exisiting cryptographic techniques for privacy preservation are not very effective for providing privacy preservation on large scale datasets.So random decision trees are used to generate equivalent and accurate models with much smaller costs.The server sends data to multiple clients after doing dataset partitioning.In this paper the dataset is partitioned based on the correlation between the features.This method is more accurate than the existing vertical partitioning of dataset.

Keywords— Data mining; privacy preserving data mining; feature subset selection; feature clustering; Random Decision Tree; homomorphic encryption; security.

INTRODUCTION

Data mining is the process of mining information from large databases. Building and applying any data mining model generally assumes that the underlying data is freely accessible. But it is true. Privacy and security factors may limit the sharing or centralization of data. Privacy-preserving data mining has emerged as an effective method to solve this problem [2]. Distributed solutions have been proposed that can preserve privacy while still enabling data mining. Privacy preservation means, Protecting specific individual values, breaking the link between values and the individual they apply to, protecting source, etc. This paper aims for a high standard of privacy: Not only individual entities are protected, but to the extent feasible even the schema are protected from disclosure.

The server sends data to multiple clients.The distributed data have to be protected.The data should be send only after the partitioning of dataset.The dataset partitioning is done according to the correlation between the features.The similar features are belongs to the same class and forms the clusters.Then construct Random Decision Tree for each clusters.The Random Decision Tree[1] approach is more accurate and efficient than existing cryptographic techniques.

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features.

The Correlation-Based Feature Subset Selection algorithm works in four steps.

- The irrelevant features are removed.
- Construct Minimum Spanning Tree using Prim's Algorithm.
- Features are divided into clusters.
- The most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features.

Features in different clusters are relatively independent.The features belongs to a particular cluster are similar. To ensure the efficiency of Correlation-Based Feature Subset Selection algorithm,we adopt the efficient minimum-spanning tree (MST) clustering method. After the partitioning of dataset construct Random decision tree.Random decision tree algorithm constructs multiple decision trees randomly. When constructing each tree, the algorithm picks a remaining feature randomly at each node expansion without any purity function check . A categorical feature (such as gender) is considered remaining if the same categorical feature has not been chosen previously in a particular decision path starting from the root of tree to the current node. Once a categorical feature is chosen, it is useless to pick it again on the same decision path because every example in the same path will have the same value (either male or female). However, a continuous feature (such as income) can be chosen more than once in the same decision path. Each time the continuous feature is chosen, a random threshold is selected.

A Random decision tree stops growing any deeper if one of the following conditions is met:

- A node becomes empty or there are no more examples to split in the current node.
- The depth of tree exceeds some limits.
- Each node of the tree records class distributions.

Then apply Homomorphic encryption to the leaf nodes of RDT.This is done for providing more privacy and security.Homomorphic encryption is a form of encryption that allows computations to be carried out on ciphertext, thus generating an encrypted result ,which,when decrypted matches the result of operations performed on the plain text.

The rest of the paper is organized as follows: in Section 1,we describe the related works. In Section 2, we present the Correlation-based feature subset selection algorithm. In Section 3, reports extensive experimental results to support the proposed Correlation-Based Feature Subset Selection algorithm . Finally, in Section 4, we summarize the present study and draw some conclusions.

II RELATED WORK

In [1] ,the privacy preserving data mining is done by generating Random Decision Tree framework. Here, the dataset partitioning is done by vertical and horizontal partitioning of dataset. In both cases, according to the slowest machine the time delay takes place. So when a new slowest machine came the time increases. Privacy and security concerns can prevent sharing of data, derailing data mining projects. Distributed knowledge discovery, if done correctly, can alleviate this problem. In this paper, we tackle the problem of classification. We introduce a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data [6] distributed over two or more parties. A completely random decision tree algorithm [2] that achieves much higher accuracy than the single best hypothesis and is comparable to boosted or bagged multiple best hypotheses. The advantage of multiple random tree is its training efficiency as well as minimal memory requirement. Data is distributed in various sites that need to be mined in a secure manner without revealing anything except the results of mining. Privacy-preserving horizontal distributed classification techniques [7] where multiple sites collaborate and broadcast the mining results. However in the process, no information about either the data maintained in the sites or data obtained during computation is divulged. Two protocols are presented to construct a Privacy Preserving Naïve Bayesian classifier using the Pailler's homomorphic encryption techniques.

Advances in computer networking and database technologies have enabled the collection and storage of large quantities of data, also the freedom and transparency of information flow on the Internet has heightened concerns of privacy. Nowadays the scenario of one centralized database that maintains all the data is difficult to achieve due to different reasons including physical, geographical restrictions and size of the data itself. The data is normally maintained by more than one organization, each of which aims at keeping its information stored in the databases privately, thus, privacy-preserving techniques and protocols are designed to perform data mining on distributed data when privacy is highly concerned. Cluster analysis [13] is a frequently used data mining task which aims at decomposing or partitioning a usually multivariate data set into groups such that the data objects in one group are most similar to each other. It focuses on arbitrarily partitioned data [8] which is a generalization of horizontally partitioned data and vertically partitioned data along with Shamir's Secret Sharing Schemes which was designed with the goal of achieving complete privacy for secure computation and communication between different parties**.**

Distributed clustering algorithm DK-Means is developed, which improves the K-Means algorithm, that is, the site in the clustering process does not require transferring large amounts of data objects, only needs to send the clustering centers as well as the total number of clusters of data objects, which reduces data traffic of the distributed clustering process so as to improve operating efficiency. This strategy [4] would be applied on the database into two different parties, recursively produce k cluster centers from each of the party, and then merge these 2k centers into the k final centers.

III PROPOSED ALGORITHM

A. Correlation-Based Feature Subset Selection Algorithm

The proposed method can be able to identify and remove the irrelevant and redundant information. Good feature subsets contain features that are highly correlated with the class, yet uncorrelated with each other.

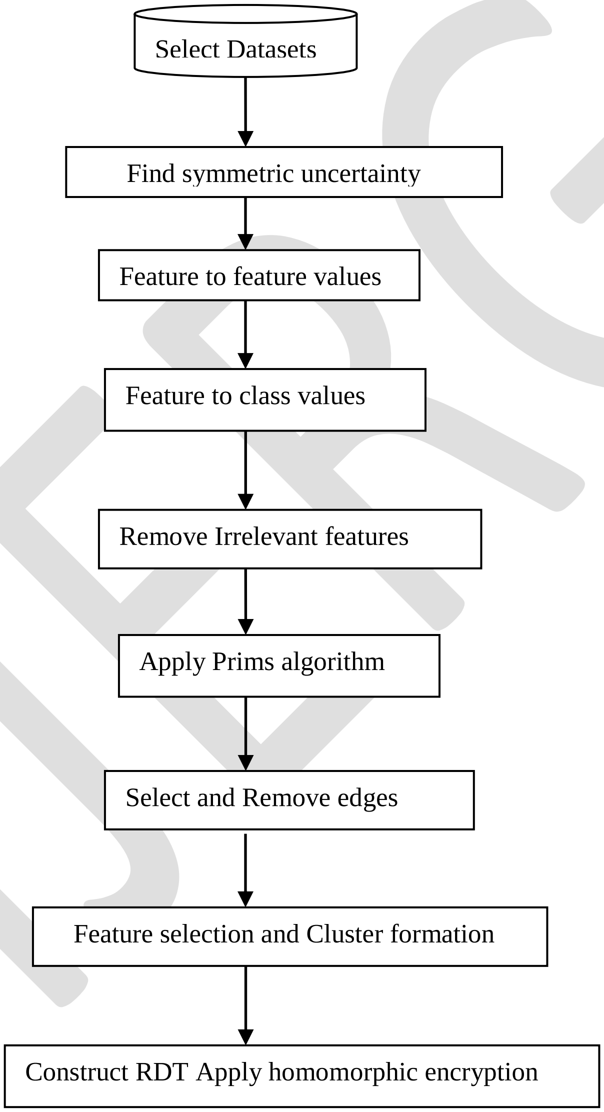
Select Datasets
Find symmetric uncertainty
Feature to feature values
Feature to class values
Remove Irrelevant features
Apply Prims algorithm
Select and Remove edges
Feature selection and Cluster formation
Construct RDT Apply homomorphic encryption

Fig 1: Framework for Correlation-Based Feature Subset Selection Algorithm

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected. In our proposed algorithm, redundant feature elimination can be done by:

1) the construction of the minimum spanning tree from a weighted complete graph;

2) the partitioning of the MST into a forest with each tree representing a cluster;

3) the selection of representative features from the clusters.

The symmetric uncertainty can be defined as:

$$\mathrm{SU(A,B)} = \frac{2*Gain(A|B)}{H(A)+H(B)}$$

Where,

1.H(A) is the entropy of a discrete random variable A.

$$\mathrm{H(A)} = -\sum_{a\epsilon \mathrm{A}} p(a)\log_2 p(a)$$

p(a) is the prior probabilities for all values.

2.Gain(A|B) is the amount by which the entropy of B decreases.

$$\begin{aligned}\mathrm{Gain(A|B)} &= \mathrm{H(A)} - \mathrm{H(A|B)}\\ &= \mathrm{H(B)} - \mathrm{H(B|A)}\end{aligned}$$

H(A|B) is the conditional entropy.

IV PSEUDO CODE

```
Step 1:Removes Irrelevant features.
Step 2:Contruct Minimum Spanning Tree.
Step 3:Using Prims Algorithm to generate the minimum Spanning Tree.
Step 4:Tree partition.
Step 5:Remove reduntant  edges.
Step 6:Select representative features.
Step 7:Cluster formation.
```

	P1		**P2**		
	Oulook	**temperature**	**Humidity**	**windy**	**play**
P1	Sunny	hot	High	weak	no
	Sunny	hot	High	strong	no
	Overcast	hot	High	weak	no
	Rainy	mild	High	weak	no
	Rainy	cool	Normal	strong	no
	Overcast	cool	Normal	strong	yes
P2	Sunny	mild	High	weak	no
	Sunny	cool	Normal	weak	yes
	Rainy	mild	Normal	weak	yes
	Sunny	mild	Normal	strong	yes
	Overcast	mild	High	strong	yes
	Rainy	mild	High	strong	no

Table 1: Sample weather dataset

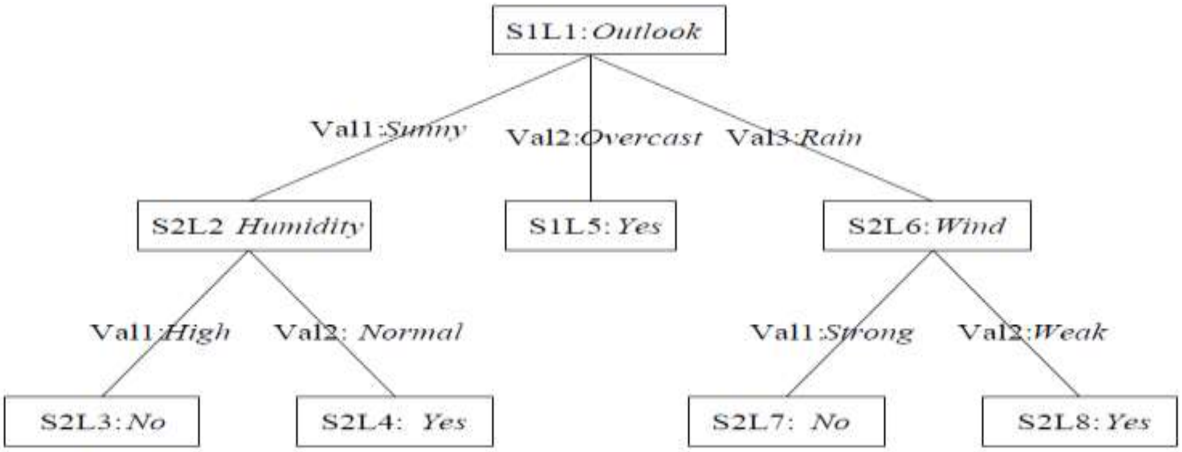

Fig 2: Random Decision Tree for weather dataset.

V SIMULATION RESULTS

The proposed method is more accurate than the existing method.In vertical partioning of dataset is partitioned vertically.The dataset partitioning is not based on any factor.It is partitioned vertically without considering any facts.But in the proposed system the dataset is partitioned based on the similarity of the features.The more correlated features are belongs to the same class.Then if there are redundant features it should have to eliminate.Then clusters are formed.The features in each clusters are highly correlated.But the clusters are not similar to each other.

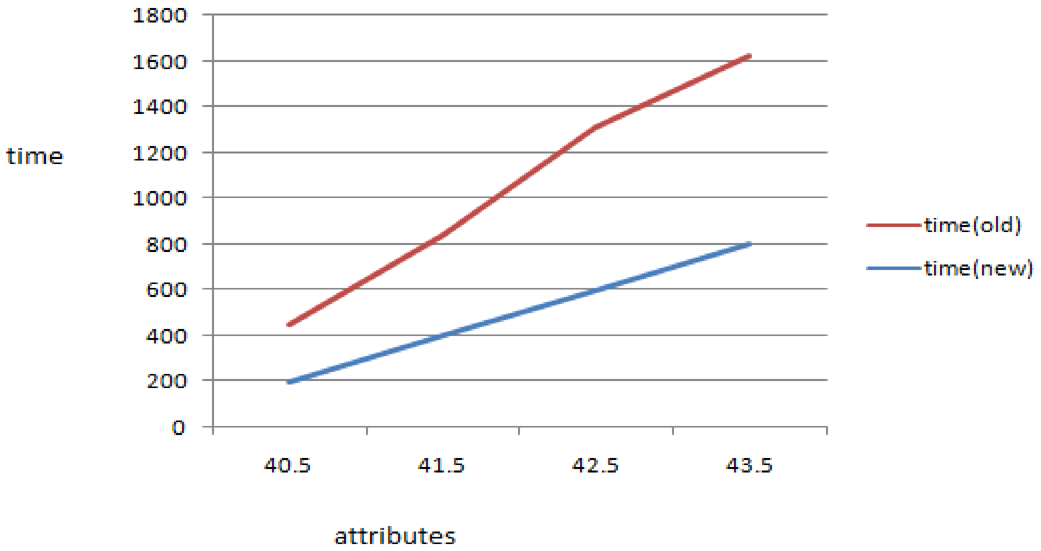

Fig 3: Performance Evaluation

VI CONCLUSION

In this paper, we have presented a novel correlation-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree , and 3) selecting representative features. Then generates the clusters.The dataset partitioning based on this approach is more efficient and accurate than the existing vertical partitioning.In the proposed algorithm, a cluster consists of features. Then construct Random Decision Tree based on these clusters.Since RDTs can be used to generate equivalent, accurate and better models with much smaller cost, we have proposed distributed privacy-preserving RDTs. Our approach leverages the fact that randomness in structure can provide strong privacy with less computation.Then apply homomorphic encryption to provide more security. The RDT algorithm scales linearly with data set size, and requires significantly less time than existing cryptographic approaches.

REFERENCES:

[1] Jaideep Vaidya, Basit Shafiq, Wei Fan, Danish Mehmood, and David Lorenzi,"A Random Decision Tree Framework for Privacy Preserving Data Mining", IEEE Transactions On Dependable And Secure Computing, Vol. 11, No. 5, September/October 2014

[2] J. Vaidya, C. Clifton, and M. Zhu, "Privacy-Preserving Data Mining", Advances in Information Security first ed., vol. 19,Springer-Verlag, 2005.

[3] W. Fan, H. Wang, P.S. Yu, and S. Ma, "Is Random Model Better?On Its Accuracy and Efficiency," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 51-58, 2003

[4] G. Jagannathan and R.Wright, "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data",In 11th KDD, pages 593–599, 2005.

[5] W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," Proc. IEEE Int'l Conf. Data Mining Workshop on Privacy, Security,and Data Mining, pp. 1-8, Dec. 2002.

[6] J. Vaidya, C. Clifton, M. Kantarcioglu, and A.S. Patterson,"Privacy-Preserving Decision Trees over Vertically Partitioned Data," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 3,pp. 1-27, 2008.

[7] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037,Sept. 2004.

[8] G. Jagannathan and R.N. Wright, "Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,pp. 593-599, Aug. 2005.

[9] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03),Nov. 2003.

[10] J. Souza, "Feature Selection with a General Hybrid Algorithm,"PhD dissertation, Univ. of Ottawa, 2004.

[11] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," J. Machine Learning Research, vol. 10,no. 5, pp. 1205-1224, 2004.

[12] C. Sha, X. Qiu, and A. Zhou, "Feature Selection Based on a New Dependency Measure," Proc. Fifth Int'l Conf. Fuzzy Systems andKnowledge Discovery, vol. 1, pp. 266-270, 2008.

[13] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification,"J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

[14] M. Dash and H. Liu, "Consistency-Based Search in Feature Selection," Artificial Intelligence, vol.151, nos. 1/2, pp. 155-176, 2003.