SW-SDF based privacy preserving data classification

Kiran.P,Kavya N. P.
Research scholar, V T U,Belgaum, Karnataka, India
kiranmys@rediffmail.com
Prof.,R N S I Technology,Bangalore, Karnataka, India
npkavya@yahoo.com

ABSTRACT

The core objective of privacy preserving data mining is to preserve the confidentiality of individual even after mining. The basic advantage of personalized privacy preservation is that the information loss is very less as compared with other privacy preservation algorithms. These algorithms how ever have not been designed for specific mining algorithms. SW-SDF personalized privacy preservation uses two flags SW and SDF. SW is used for assigning a weight for the sensitive attribute and SDF for sensitive disclosure which is accepted from individual. In this paper we have designed an algorithm which uses SW-SDF personal privacy preservation for data classification. This method ensures privacy and classification of data.

General Terms

Data Mining, Security.

Indexing terms

DBB tree index

Academic Discipline And Sub-Disciplines

Education

SUBJECT CLASSIFICATION

Computer science

TYPE (METHOD/APPROACH)

Quasi-Experimental;

1. INTRODUCTION

Privacy preserving data mining (PPDM) is a novel approach in data mining which preserves the privacy of the individual or company related information even after mining process[1,8,9]. The data supplier is usually referred as data publisher and miner as data recipient. Usually both belong to different organization. The main goal of data publisher is to publish data in a way which ensures privacy. Different methods that have been proposed include data anonymization, data swapping and randomization of the record values.

Data anonymization is an approach in which original values of the records are generalized in such a way that duplicate records are created. The major data anonymization techniques that were proposed in literature are k-anonymity, l-diversity and t-closeness. In k-anonymity[4] each record must have at least k-1 instances of it. This suffers from homogeneity attack in which each block contains same sensitive values. l-diversity [5] requires that each record must have multiple instances such that it contains l well represented sensitive values. The drawback of this approach is that it suffers from Similarity Attack. This happens when all the sensitive values in a quasi group are distinct but refer the same meaning. t-closeness [6] requires each of the quasi group to have distribution of sensitive value equal to the distribution of published data.

SW-SDF personal privacy preservation[2] uses sensitive weight to differentiate between sensitive attribute values. SW=0 is assigned to sensitive value which does not require privacy and SW=1 for sensitive value which require privacy. A statistical based method or clustering method can be used for identification of sensitive value[3]. An additional flag SDF is used or disclosure of the record. For all records whose SW=0 , SDF=0 is initialized. For SW=1, SDF is accepted from user. SDF=0 indicates a record which is not sensitive and does not require privacy. SDF=1 indicates record which are sensitive and requires privacy. A DBB tree indexing technique is used for faster retrieval of records and QIDB creation. Each QIDB can be generalized to the same level.

In this paper we address issues related to privacy preserving data classification. The goal of mining is to retrieve information from large amount of data which can be used for decision making [7]. Data classification is a data mining function that classifies data into discrete ones using tree structured hierarchy. This can be subsequently used for prediction and generating rules. Many real world applications require privacy preservation classification. For example consider patient information having the details pid, pname, page, pzipcode and pdisease. In this pid and pname are called identifiers and are removed by data publisher before giving it to data recipient. Page and pzipcode are called quasi identifiers since they have a reference in some external database like voters database. pdisease is sensitive attribute since the record owner is not ready to reveal his identity. The resultant data base will contain page,pzipcode and pdisease. This can be used by the data recipient for classification. The classification may include identification of dependent values of page and pzipcode for the pdisease like 'heart disease'. This classification must not reveal the identity of the individual and company related information. Our solution is to avoid disclosing of sensitive data beyond the source by generalization. The resultant published data can still be used for constructing the data classifier approximately equivalent to that of the original classifier using SW-SDF personal privacy.

2. Related work in privacy preservation

Privacy preservation was initially applied towards data recipient end. Models using modification or data swapping was done for ensuring privacy without considering how the data was used. This was the major disadvantage of most of the approaches initially [8, 9]. In literature most of the privacy preservation was initially implemented for randomization technique which was modified for the rest of the approaches. In [10] privacy preservation for association rule was defined for distributed environment. Each site holds attribute of each transaction and two-party algorithm was used to identify the frequent itemset. A cryptographic approach was used in [11] to retrieve information where data was partitioned horizontally it adds little overhead to the mining task. Algebraic technique for identification of association rules was proposed in [12] which assumes multiple data providers and a single data miner.

Clustering is an unsupervised technique for classifying data to a labeled representation. Many techniques were also proposed for privacy preservation data clustering[13,14]. In [13] method for k-means clustering was indicated. It assumes that different sites contain different attributes for a common set of entities. Each site learns the cluster of each entity, but learns nothing about the attributes at other sites. Arbitrary partitioned data which is a generalization of horizontal and vertical data was proposed in [14]. It also gave the representation of protocol for k-means clustering. Privacy preserving classification was defined for individual technique. In [15] author has defined an algorithm for ID3 over horizontally portioned data involving two parties. Secure privacy preserving algorithm for support vector machine (SVM) classification over vertically partitioned data was indicated in[16]. PCDS method [17] extends the process of data streams classification to achieve privacy preservation. It uses perturbation to ensure privacy. Protocols to develop a Naïve Bayes classifier[18] on both vertically as well as horizontally partitioned data were indicated.

3. Data classification using ID3

Author [20] introduced the concept of classification algorithm ID3 and is indicated in Algorithm 1. The input is records containing finite set of attributes. One of the attribute among this given set contains class attributes. The decision tree in ID3 is built using top-down approach. In the first iteration it finds the attribute which best classifies the data considering the target class attribute. Once the attribute is identified in the given set of attributes algorithm creates a branch for each value. This process is continued until all the attributes are considered. In order to calculate which attribute is the best to classify the data set information gain is used. Information gain is defined as the expected reduction in entropy. Entropy of a data set of tuples S is defined in equation 1.

$$\mathrm{entropy}(s) = \sum_{i=1}^{d} -p_i \log_2 p_i \qquad (1)$$

Where d is the different classes that are there in each attribute. pi is the actual number attribute which belong to a particular class i. The information gain of an attribute A is then defined in equation 2.

$$\mathrm{gain}(S, A) = \mathrm{entropy}(S) - \sum_v \frac{|S_v|}{|S|} \mathrm{entropy}(S_v) \qquad (2)$$

with Sv the subset of S with tuples having value v for attribute A.

```
--------------------------------------------------------------------------------------------------------------------------------

Algorithm 1 The ID3 Algorithm

Input: A set of attributes, C class attribute, S data set of tuples

Output: Decision tree

--------------------------------------------------------------------------------------------------------------------------------

If ( empty(S))
Then
Return false;
else if ( check_class_value(S))
 then  // entire attribute class has the same value
Return specific_class_value().
elseif ( empty(A))
then
            // all the attributes have been classified
```

```
        Return most_frequent_value(S)
else
        // different classes exists
        // find the information gain among the given set of A
// find the attribute a which has the maximum information gain
highest_information_gain(S,A)
// based on the selected attribute a
Let a has m values for S
Partition S in m parts S(attr1), ..., S(attrm) such that attr1, ..., attrm are the different
values of a.
Return a tree with root a and m branches labeled atr1...attrm, such that
branch i contains ID3(A − {a} ,C, S(attri)).
end if
--------------------------------------------------------------------------------------------------------------------------------------
```

4. Existing system

Most of the existing approach use randomization for privacy preserving. The block diagram of existing randomization approach is shown in figure 1. The overall execution can be divided into two phases first the data publisher end and second is the data recipient. In data publisher side the client systems provide data. Data is either horizontally partitioned or vertically partitioned. Randomization is applied to the contents in such a way that it ensures the privacy and there by classification. Data join is done on the resultant data that is provided by multiple locations after randomization. This data is passed to the data recipient. The data recipient retrieves the data distribution and passes it to the data classifier. The main drawback is that it has been proven that randomization provides very less privacy [19]. Efficiency of classification using randomization decreases for the miner if the data is more.

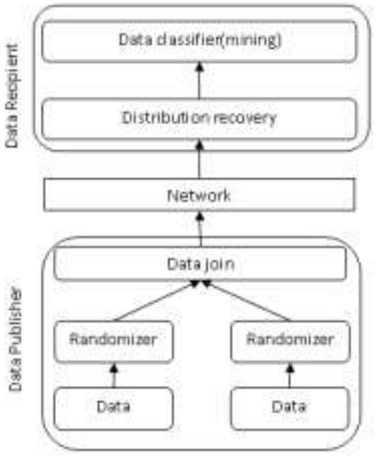

Fig: 1. Randomization approach

5. Proposed system

SW-SDF based privacy preserving data classification uses the basic concepts of SW-SDF[2]. The classification algorithm has been designed using the previous ID3 representation. The basic structure of the proposed system is shown in figure 2.

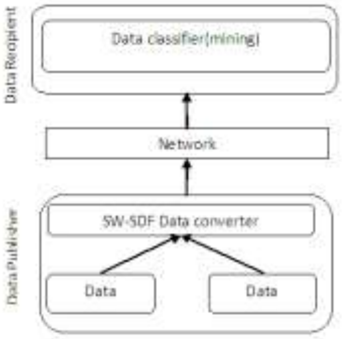

Fig: 2. SW-SDF based privacy preserving data classification

In the data recipient end data is accepted from multiple locations. The relation may be divided in to horizontal or vertical representation. The joined data is passed on to SW-SDF data converter which modifies the data in such a way that privacy and classification is retained.

6. Notation for SW-SDF data converter

Let us assume relation T is made of n records represented as T={t1,t2,t3,…,tn}. Quasi attribute is represented as t.q and sensitive attribute as t.s .For classification t.s is a categorical attribute and t.q is a non categorical attribute. it also contains two additional information t.sw representing sensitive weight and t.sdf for disclosure. let us assume that SW is calculated for the sensitive attribute and a value of SW=1 is assigned to an attribute which actually requires privacy. For these records SDF is accepted. SDF=1 indicate that the record owner is ready to disclose his identity. SDF=0 indicates the record owner which is not ready to disclose his identity. In our representation we use only SDF for data converter and is indicated as t.sdf. SW-SDF data converter is shown in Algorithm 2.

```
------------------------------------------------------------------------------------------------------------------

Algorithm 2 SW-SDF_data_converter

Input: set of records T, with q attributes

Output: Decision tree

------------------------------------------------------------------------------------------------------------------

If ( empty(T))
Then
Return false;
else if ( check_class_value_withSDF(T))
then  // entire attribute class has the same value and none of the records is sensitive
insert_dbb_tree(T)  // no change with the records
        elseif (empty(q))
                //no attribute for classification
                If ( Check_for_sensitive(T) ) then
                Generalize_rec(T);
                endif
        else
                //there is a combination of SDF=0 and SDF=1
highest_information_gain(T,q)
Let q has m values for T
Partition T in m parts T(attr1), ..., T(attrm) such that attr1, ..., attrm are the different
values of q.
insert_dbb_tree with root q and m branches labeled atr1...attrm, such that
branch i contains SW-SDF_data_converter(q − {a} ,T(attri)).
```

```
end if
---------------------------------------------------------------------------------------------------------------------
```

Division base bit (DBB) tree indexing technique has been used for the construction of the node. Insert_dbb tree uses the data structure shown in figure 3. It contains the entire record Q, sdf value S and Flag F. it also has an index for its parent node and an index which points to multiple records. The resultant structure after insertion is shown in figure 4.

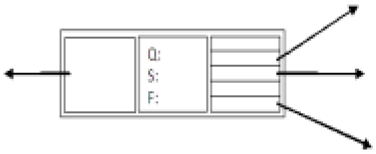

Fig: 3. Data structure of DBB tree

Bit representation is used for comparing one record with the other. The initial branching is represented by the information gain. For example let us assume a patient database shown in table 1. Zipcode and age are non categorical attribute. Disease is a categorical attribute.

Table 1. Patient data base

Label	Zip code	Age	Disease	SDF
A	56212	20	Flu	1
B	56212	60	Heart	0
C	56212	20	Flu	1
D	56212	80	Heart	1
E	56212	80	Flu	1
F	57213	20	Flu	1
G	57213	60	Heart	0
H	57213	80	Heart	1
I	57213	20	Flu	1
J	58211	20	Flu	1
K	58211	30	Flu	1
L	58211	80	Heart	0

To start with T contain all the records. Information gain for each a quasi attribute is calculated, for the above example information gain(age)= 0.70 and information gain(zipcode)= 0.05. In this example age is considered as the root node for division. Age is compared for creation of the sub nodes of the tree. Record a and c both have the same age so a bit value of 00 is used as an index for record a to point to record c. For the next record f the same comparison is done once again the difference of bit is zero so it will check wither there is already an index of zero. In this example there is an index already present so the record c will point to f. the structure of insertion is shown in figure 4. The labeled representation of all the records after insertion using the structure is shown in figure 5.

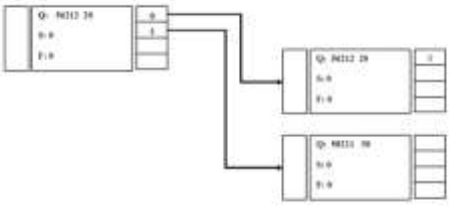

Fig: 4. Structure of a tree after insertion

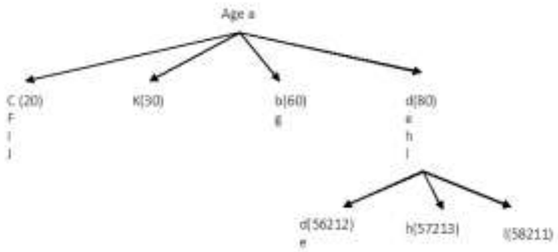

Fig: 5. Label representation of DBB tree

```
--------------------------------------------------------
Algorithm 3 check_class_value_withSDF
Input : T records
Output: Boolean
                    n=size(T)
                    for i=1 to n
                              if t.sdf=0  then
                                        return False;
                                        Break ;
                              Endif
                    Endfor
                    Return True
--------------------------------------------------------
```

Dept first search is used for retrieving records after the identification of the node from where it must be searched. In check_class_value_withSDF algorithm we find whether all the nodes referenced by the node to be searched contains SDF=1. Reorganization of the tree happens when there is a combination of SDF=0 and SDF=1 in a particular node where insertion is to be done. The reorganization is based on the information gain obtained from the rest of the non categorical attributes. In this example node with 80 ie d, e, h and l gets reorganized to indicated the new nodes in the hierarchy. If it is a leaf node Generalize_rec is invoked. This algorithm identifies the QIDB block for generalization. For example record b must not be disclosed. It searches for records in its parent node. The first criteria is wither the record has not been used which is checked by flag f and it must be non sensitive ie SDF=1. Generalize_rec also checks whether the frequency distribution of QIDB equal to the overall distribution. The QIDB block will have b, f and g. Based on the bit representation it gets generalized. The resultant information will have privacy because data is generalized and the decision tree is also retained.

7. Experimental results

The efficiency of our algorithm was realized using the standard American data set of 400 records. The details of the data base included age, education, marital status and occupation. value generation hierarchy was created for each of the quasi attributes. A single sensitive attribute disease was used. SW was identified by using statistical approach. The resultant is shown in table 2. Frequency distribution of sensitive attribute is shown in table 3.

Table 2. Resultant SW for sensitive attribute disease

Disease	SW
Heart	1
Flu	0

Table 3. Frequency distribution block

Disease	Distribution
Heart	0.33
Flu	0.67

The initial representation of decision tree is shown in figure 6 and the resultant SW-SDF based privacy preserving data classification is shown in figure 7.

Fig: 6. Decision tree before SW-SDF based privacy preserving data classification

Fig:7. Decision tree after SW-SDF based privacy preserving data classification

8. Conclusion and future work

SW-SDF based privacy preserving data classification ensures the privacy of the individual by using two flags SW and SDF. SW is automatically calculated using statistical inference and according SDF is accepted form user. The classification algorithm indicated works efficiently for creating decision and information loss is also less. In this paper we have used a DBB tree indexing technique for faster retrieval and classification. Future work includes developing SW-SDF for other mining methods. In our example we have considered only one sensitive attribute, methods must be defined for multiple sensitive attributes. Developing new measures for SW-SDF personalized privacy preservation data mining methods can be investigated. Improvisation of the current method incorporating different releases can be done.

REFERENCES

[1] .Kiran P, S Sathish Kumar and Kavya N P, Modelling Extraction Transformation Load embedding Privacy Preservation using UML. In International Journal of Computer Applications (IJCA), vol. 50, No. 6, July 2012.

[2] .Kiran P and Kavya N P, SW-SDF based personal privacy with QIDB-Anonymization method. International Journal of Advanced Computer Science and Applications (IJACSA), vol 3, issue 8, August 2012.

[3] .Kiran P, S Sathish kumar, Hemanth S and Kavya N P, Assignment of SW using statistical based data model in SW-SDF based personal privacy with QIDB-anonymization method, In Proc of second IEEE International conference on Parallel, Distributed and Grid computing Dec 06-08 2012,pp. 816 - 821.

[4] .L. Sweeney, “k-Anonymity: A Model for Protecting Privacy”. Int’l J.Uncertain. Fuzz., vol. 10, no. 5, pp. 557-570, 2002.

[5] .Machanavajjhala A, Gehrke J, Kifer D and Venkitasubramaniam M, “l-diversity: Privacy beyond k-anonymity”. In Proceedings of the 22nd IEEE International Conference on Data Engineering(ICDE), 2006.

[6] .Ninghui Li , Tiancheng Li , Suresh Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l–Diversity”. ICDE Conference, 2007.

[7] .J. Han and M. Kamber, Data Mining Concepts and Techniques. Morgan Kaufmann, 2001.

[8] .Lindell, Y., Pinkas, B.: Privacy preserving data mining. J. Cryptol. 15(3), 177–206 (2002)

[9] .R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD

conference on Management of Data, Dallas, TX, May 14-19 2000. ACM.

[10] .Vaidya, J.,Clifton,C.: Privacy preserving association rule mining in vertically partitioned data. In: The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, pp. 639–644. (2002).

[11] .Kantarcıo˘ glu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. Knowl. Data Eng. 16(9), 1026–1037 (2004)

[12] . N. Zhang, S. Wang, and W. Zhao, “A new scheme on privacy preserving association rule mining,” in Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Springer Verlag, 2004.

[13] .Vaidya J., Clifton C (2003) Privacy-preserving k-means clustering over vertically partitioned data. In: The 9th ACM Lin, X., Clifton, C., Zhu, M.: Privacy preserving clustering with distributed EM mixture modeling. Knowl. Inf. Syst. 8(1), 68–81 (2005)

[14] .Jagannathan,G., Wright,R.N.: Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In: Proceedingsof the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, pp. 593–599 (2005)

[15] . M.-J. Xiao, L.-S. Huang, H. Shen, and Y.-L. Luo. Privacy preserving id3 algorithm over horizontally

partitioned data. In Sixth International Conference on Parallel and Distributed Computing Applications

and Technologies (PDCAT’05), pages 239–243. IEEE Computer Society, 2005.

[16] . H. Yu, J. Vaidya, and X. Jiang. Privacy-preserving svm classification on vertically partitioned data. In

Proceedings of PAKDD ’06, volume 3918 of Lecture Notes in Computer Science, pages 647 – 656. Springer-Verlag, January 2006.

[17] . Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, Privacy-Preserving Classification of Data Streams, Tamkang Journal of Science and Engineering, Vol. 12, No. 3, pp. 321-330 (2009).

[18] . Jaideep Vaidya, Murat Kantarcıoglu and Chris Clifton, Privacy-preserving Naïve Bayes classification, The VLDB Journal (2008) 17:879–898.

[19] . Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar, In Proceedings of the Third IEEE International Conference on Data Mining (ICDM’03), pp 99-106.

[20] J. R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.