Prototype of Clustering and Classification Model for Privacy Preservation using Single Vector Decomposition

Richa Lodhi
Department of Information Technology
S.A.T.I
Vidisha (M.P) India

Anil Suryavanshi
Department of Information Technology
S.A.T.I
Vidisha(M.P)India

ABSTRACT
The single value decomposition technique divides the data of different parties during the process of privacy preservation. The process of single value decomposition implied in the form of clustering and classification. The combined process of clustering and classification called prototype mode for sharing privacy preservation. The utility of vector decomposition in this model is selection of data in different parties for the process of maintain the raw information. In this paper proposed a prototype model for privacy preservation and improved the efficiency of data utility and accuracy of data recovery during the process of privacy. The proposed model implements in MATLAB software and used some standard dataset for evaluation of performance. The proposed model is very efficient in compression of KPPDM model.

General Terms
Privacy preservation, Data mining, Vector Decomposition

Keywords
Clustering, classification, sample selection, PPDM

1. INTRODUCTION
Confidentiality and authentication of data is major issue in current scenario. For the confidentiality and authentication of data various technique are used such as cryptography, data Randamization, third party access control and many more method[1,2]. The conventional technique such as cryptography and other technique faced a problem of security issue in privacy preservation. Now a day's data mining technique play an important role for the privacy preservation[3,4]. For the purpose of this used rule mining technique, classification technique and clustering technique. The rule mining technique is very important role in terms of transformation. The process of transformation changes the value of minimum support and confidence. And change the order of data associated with this range and hide the information[5,6,7]. Instead of these technique used clustering and classification for the process of privacy preservation. The process of clustering and classification such as decision tree and KNN are used for this purpose. Now a day's principle of component analysis is used. The process of data privacy preservation proceeds in two different ways[8]. First act as sensitive raw data such as name, indentifiers and some other important record. And other is sensitive information mined from database using data mining algorithm[9]. The process of data mining facilities the process of algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential data can be derived from released data by unauthorized users is also commonly called the data duplication problem. Now a day's SMC play an important role in privacy preservation in concern of third party communication[10]. They believe of all parties justify by the common factor of data analysis. The protocols of SMC ensure that the communication party involve in proper manner[10]. In other words, unless proper incentives are set, current SMC techniques cannot prevent input modification by participating parties. Clustering and classification is important technique of data mining implied for grouping of data on the given model are called classes. The process of clustering done through the iteration function on the basis of decomposed value of vector around the attribute selection process. The clustered data mapped into number of classes according to their attribute. The sample selection process done through the different attribute portion process. This paper is divided into five sections. Section-1. Gives the introduction of privacy preservation and data mining. Section-2. Gives the information about related work. Proposed method in section-3.. In section-4. Discuss experimental work and finally discuss conclusion and future work in section 5.

2. RELATED WORK
In this section discuss the related work in the field of privacy preservation using data mining technique and some other technique. The data mining technique offers various algorithms for the process of privacy preservation. Association rule mining play an important role in privacy preservation. Here discuss some work along their authors. 1] In this paper author describes a Key Distribution-Less Privacy Preserving Data Mining system in which the publication of local association rules generated by the parties is presented. The association rules are securely combined to form the combined rule set using the (KDLPPDM) algorithm. The combined rule sets established are used to classify or mine the data. The results discussed in this paper compare the accuracy of the rules generated using the C 4.5 based KDLPPDM system and the C 5.0 based KDLPPDM system using receiver operating characteristics curves (ROC). [2] In this paper, they first develop key theorems, then base on these theorems, they analyze certain important privacy-preserving data analysis tasks that could be conducted in a way that telling the truth is the best choice for any participating party. they have investigated what kinds of PPDA tasks are incentive compatible under the NCC model. Based on our findings, there are several important PPDA tasks that are incentive driven. classifies the common data analysis tasks studied in this paper into DNCC or Non-DNCC categories. Most often, data partition schemes can make a difference in determining DNCC or Non-DNCC classifications. [3] This paper proposed a feature selection with privacy preservation in centralized network. Data can be preserved for privacy by perturbation technique as alias name. In centralized data evaluation, it makes data classification and feature selection for data mining decision model which make the structural information of model in this paper. The application of gain ratio technique for better performance of feature selection has taken to perform the centralized computational task. All features don't need to preserve the privacy for confidential

data for best model. The chi-square testing has taken for the classification of data by centralized data mining model using own processing unit. [4] In this paper they review on the various privacy preserving data mining techniques like data modification and secure multiparty computation based on the different aspects. Data mining is such a technique which extracts the useful information from the large repositories. Knowledge discovery in database (KDD) is another name of data mining. Privacy preserving data mining techniques are introduced with the aim of extract the relevant knowledge from the large amount of data while protecting the sensible information at the same time. [5] In this paper, they propose a generic PPDM framework and a simplified taxonomy to help understand the problem and explore possible research issues. they also examine the strengths and weaknesses of different privacy preserving techniques and summarize general principles from early research to guide the selection of PPDM algorithms. they conduct an extensive review on literature. they present a classification scheme, adopted from early studies, to guide the review process. As part of future work, they plan to apply the proposed evaluation framework to formally test a complete spectrum of PPDM algorithms. [6] In this paper author discuss about the challenges in privacy-preserving data quality assessment. A two-party scenario is considered, consisting of a client that wishes to test data quality and a server that holds the dataset. Privacy-preserving protocols are presented for testing important data quality metrics: completeness, consistency, uniqueness, timeliness and validity. For semi-honest parties, the protocols ensure that the client does not discover any information about the data other than the value of the quality metric. The server does not discover the parameters of the client's query, the specific attributes being tested and the computed value of the data quality metric. [8] This paper introduces a privacy preserving approach that can be applied to decision tree learning, without concomitant loss of accuracy. It describes an approach to the preservation of the privacy of collected data samples in cases where information from the sample database has been partially lost. This approach converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an accurate decision tree can be built directly from those unreal data sets. This novel approach can be applied directly to the data storage as soon as the first sample is collected.

3. PROPOSED METHODOLOGY

The proposed model is combination of prototype of clustering and classification process. For the process of clustering used density based clustering technique and for the process of classification used KNN. In process of clustering and classification used single point sample selection parameter. The single point parameter selection is given by the process of vector decomposition. The vector decomposition method generates two different attribute set. The decomposed attribute set generates the two different cluster group and form the index of class. And the building of class depends on the basis of sub index of cluster generates by the vector index value. Using a probability density function for the attribute bucket as total input data M, being labeled by an index j, we can write:

$$p(X|Ck) = \sum_{j=0}^{M-1} P(X|j)P(j|Ck) \dots\dots\dots\dots\dots\dots\dots (1)$$

Where p is probablity of density and M is total dataset and j is the last index of partion of buket.

Now derive the expression of cluester index for different attribute.

$$P(X) = \sum_{k=0}^{n-1} p(x|Ck)P(CK) = \sum_{j=0}^{M-1} P(X|j)P(j) \dots\dots\dots\dots\dots\dots\dots (2)$$

Here Ck represent the value of cluster index. Now we creates class index using this derivatives as

$$C(J) = \sum_{K=index}^{label} M(P)|P(Ck) \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Here C (J) represents the class of equivalence index of data. Now create two different attribute bucket as

$$\mathrm{B(c)} = \sum_{\mathrm{k=1}}^{\mathrm{a}} \sum_{\mathrm{i=1}}^{\mathrm{b}} \mathrm{(x|p)(y|p)} \dots\dots\dots\dots\dots (4)$$

Algorithm

In this phase of algorithm describe the step of prototype of clustering and classification process and vector decomposed attribute selection.

(1) Select the data from all parties for the process of attribute decomposition.

(2) Generate discrete vector value for the selection of different attribute set.

(3) Differentiate two set of index for cluster generation.

(4) Map data into cluster space for sub-index generation.

(5) CDI={e1,e2……..en}

(6) Measure the separate index value for equivalence class.

(7) EC={c1,c2……cn}

(8) Compare value at vector index

(9) Repeat iteration

(10) Then generate cluster.

(11) Set label of class C1, C2, C3.

(12) Assigned attribute of bucket.

(13) Generate classifier-Merge set of cluster & classifier with label.

(14) Calculate value of utility measure and accuracy.

(15) Published the data.

(16) Exit

4. EXPERIMENTAL RESULT

For the evaluation of performance of proposed method used MATLAB software and two well know data set are used one is breast cancer dataset and another dataset is glass dataset. These dataset obtained from UCI machine learning repository. For the validation of result used three parameter such as utility measure, CP and accuracy. This parameter shows that effectiveness of proposed method[15].

Description of evaluation parameter

UM (Utility Measures): The data utility measures assess whether a dataset keep the performance of data mining technique after the data distortion.

CP: To define change at the rank of the arrange value of the attribute.

$$CP = \sum_{i=1}^{m} |RankAvi - RankAvi|/m$$

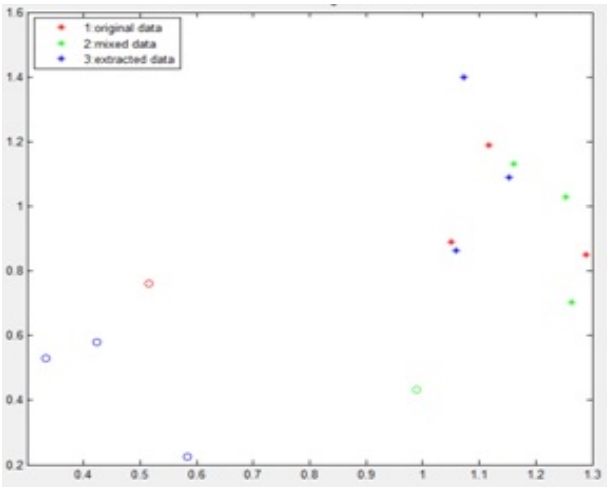

Figure 2: Shows that the data mapping and data extraction during the process of privacy preservation.

Table 1: Shows that the comparative performance parameters for the privacy preservation data classification using KPPDM and proposed method.

METHOD NAME	DATA SET NAME	UTILITY MEASURES	CP	ACCURACY
KPPDM	BREAST CANCER DATASET	5.37	80.85	93.82
	GLASS DATASET	6.84	84.44	94.32
PROPOSED METHODS	BREAST CANCER DATASET	6.75	80.85	95.82
	GLASS DATASET	6.85	84.56	96.66

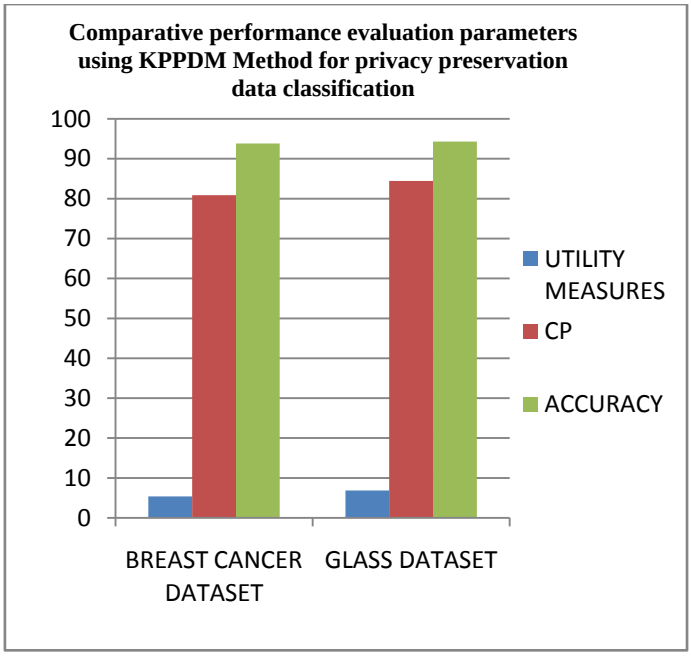

Figure 2: Shows that the comparative performance parameters for the privacy preservation data classification using KPPDM method.

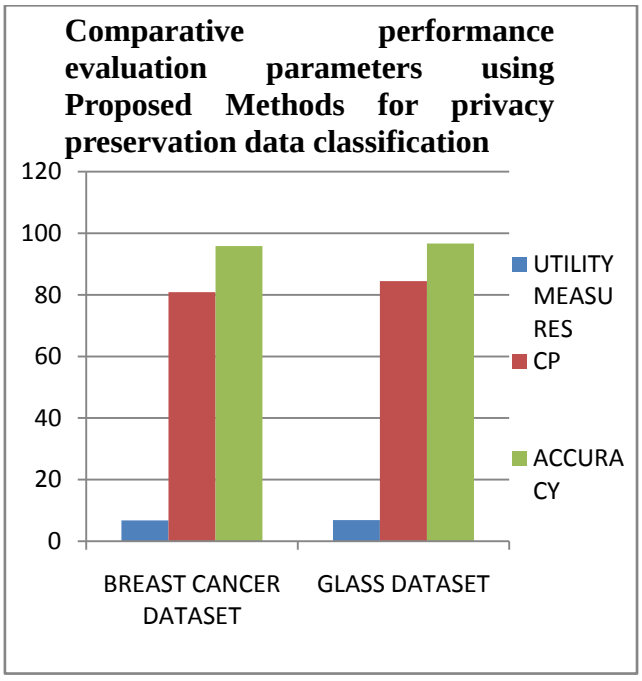

Figure 3: Shows that the comparative performance parameters for the privacy preservation data classification using Proposed method.

5. CONCLUSION AND FUTURE WORK

In this paper proposed a vector decomposition based method for privacy preservation. The proposed methods used prototype clustering and classification process for data transformation and data publishing. The process of vector decomposition used selects the different attribute of given dataset for the process of transformation. The transform data cerates the bucket of index attributes and creates all sub id for the next similar data. After the creation of index build the class for the process of classification. The process of

classification used nearest neighbor classifier (KNN). The classifier creates the number of published attribute for the recovery process. The proposed algorithm gives better result in compression of KPPDM technique. the proposed algorithm gives average classification ratio is 98%. In future used another technique for the data dimension reduction and reduces the loss of data during the process of transformation.

6. REFERENCES

[1] S KumaraSwamy,Manjula S , K R Venugopal, Iyengar S , L M Patnaik "Association Rule Sharing Model for Privacy Preservation and Collaborative Data Mining Efficiency" IEEE, 2014. Pp 1-6.

[2] Murat Kantarcioglu, Wei Jiang "Incentive Compatible Privacy-Preserving Data Analysis" IEEE, 2013. Pp 1323-1335.

[3] He manta Kumar Bhuyan , Maitri Mohant , Smruti Rekha Das "Privacy Preserving for Feature Selection in Data Mining Using Centralized Network" IJCSI, 2012. Pp 434-440.

[4] Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini Chaudhary "A Review Study on the Privacy Preserving Data Mining Techniques and Approaches" IJCST, 2013. Pp 42-46

[5] K. Srinivasa Rao, B. Srinivasa Rao "An Insight in to Privacy Preserving Data Mining Methods" CSEA, 2013. Pp 100-104.

[6] Julien Freudiger, Shantanu Rane, Alejandro E. Brito , Ersin Uzun PARC "Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing" ACM, 2014. Pp 1-9.

[7] Kumaraswamy S, Manjula S, K R Venugopal, L M Patnaik "A Data Mining Perspective in Privacy Preserving Data Mining Systems" IJCST, 2014. Pp 704-711.

[8] Beula Amalorpavam, N. Mookhambik "privacy preserving decision tree learning using unrealized data sets" IJITCSP, 2013. Pp 187-192.

[9] Ms.R.Kavitha, D.Vanathi "A Study Of Privacy Preserving Data Mining Techniques" IJCAIT, 2014. Pp 71-77.

[10] Yaping Li, Minghua Chen, Qiwei Li , Wei Zhang "Enabling Multi-level Trust in Privacy Preserving Data Mining" IEEE, 2011. Pp 1-20.

[11] Somayyeh Seifi Moradi, Mohammad Reza Keyvanpour "classification and evaluation the privacy preserving distributed data mining techniques" Journal of Theoretical and Applied Information Technology, 2012. Pp 204-210.

[12] M. Kantarcioglu, R. Nix "Incentive Compatible Distributed Data Mining" Proc. IEEE Int'l Conf. Soc. Computing/IEEE Int'l Conf. Privacy, Security, Risk and Trust, Pp 735-742, 2010.

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

[14] H. Kargupta, K. Das, and K. Liu "A Game Theoretic Approach toward Multi-Party Privacy-Preserving Distributed Data Mining" Proc. 11th European Conf. Principles and Practice of Knowledge Discovery in Databases, 2007, Pp 523-531.

[15] S. Mukherjee, H. Kargupta "Distributed Probabilistic Inferencing in Sensor Networks using Variational Approximation". J. Parallel Distrib.Comput.,2008, Pp78–