Privacy Preserving Processing of Data Decision Tree Based on Perturbation and Singular Value Decomposition

Priyank Jain[1], Dr. Manasi Gyanchandani[2], Dr.Sanyam Shukla[3], Dr.Rajesh Wadhwani[4], Lokini Rajesh[5]

Research scholar[1, 5,] Assistant Professor[2, 3, 4]
CSE Department [1, 2,3,4,5]
MANIT, Bhopal M.P India[1, 2, 3, 4, 5]

Abstract
Data mining is a set of automated techniques used to extract hidden or buried information from large databases. With the development of data mining technologies, privacy protection has become a challenge for data mining applications in many fields. To solve this problem, many privacy-preserving data mining methods have been proposed. One important type of such methods is based on Singular Value Decomposition (SVD). In the proposed algorithm, attributes are grouped according to their distance difference similarity by clustering the data set using decision tree classification. Secondly, the algorithm packetizes the attributes according to their SA value in each group. Thirdly, for each group it selects attributes from the smallest bucket and searches for a similar attributes in the attributes-1 largest buckets from the same group to create an equivalence class following the unique attribute–distinct diversity anonymization model. The proposed algorithm satisfies the utility based anonymization principle that crucial information is protected from being suppressed. Also, weights given to attributes improve clustering and give the ability to control the generalizations depth. In prototype decision tree is combination of clustering and classification technique such methods are called ensemble classifier, this new proposed method is more efficient in balancing data privacy and data utility.

Keywords:
Privacy-Preserving Data Mining, Singular Value Decomposition (SVD), Decision Tree, Clustering.

1. Introduction

Privacy preserving [1, 2] play an important role in data hiding and data security. Now in conventional technique of data hiding required cryptography technique. But now a day's data mining important tools for data privacy preserving. Data mining is a set of automated techniques used to extract hidden or buried information from large databases. The term data mining refers to the nontrivial extraction of valid, implicit, potentially useful and ultimately understandable information in large databases with the help of the modern computing devices. In the last few decades, many successful applications in data mining have been reported from varied sectors such as marketing, finance, medical diagnosis, banking, manufacturing and telecommunication. Apart from the benefits of using data per se the mining of these datasets with the existing data mining tools can reveal invaluable knowledge that was unknown to the data holder beforehand. The extracted knowledge patterns can provide insight to the data holders as well as be invaluable in tasks such as decision making [5] and strategic business planning. As a valuable technique, data mining is developing and is flourishing. But, at the same time, serious concerns have grown over individual privacy in data collection, processing and mining. And also used some data mixed technique for adaptive noise data in original data. Matrix decomposition is big role in privacy preserving in data mining classification. The types of matrix decomposition are horizontal vertical and diagonal of index data of privacy. In data mining application the utility of third party has been removed. In the process of matrix decomposition singular [16] and multiple values are involved. The singular value decomposition prevents the loss of mixed data and extracted data in decomposition of matrix.

2. Proposed Work

Sample selection maintain ratio of data between mixed data and original data during processing of classification [1, 2]. The ratio of sample selection is 1:3 by default process of sample selection. In this paper we proposed a prototype classification technique for privacy preserving technique for data classification. In prototype classification is combination of clustering and classification technique such methods are called ensemble classifier. In proposed algorithm, attributes are grouped according to their distance difference similarity by clustering the data set using decision tree classification [3]. Secondly, the algorithm packetizes the attributes according to their SA value in each group.

Thirdly, for each group it selects attributes from the smallest bucket and searches for a similar attributes in the attributes-1 largest buckets from the same group to create an equivalence class following the unique attribute–distinct diversity anonymization [4] model. If a proper attributes cannot be found in the same group, the algorithm is searching to the next group which is the most common. Finally, when an equivalence class is successfully created the attributes that belong to it are removed from the original table. This leads to better

Manuscript received May 5, 2026
Manuscript revised May 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.5.16

generalization with less information loss. In addition, by performing with great care the rare attributes the probability to suppress rare and valuable attributes is minimized. By doing so, the proposed algorithm satisfies the "utility based anonymization [4] principle that crucial information is protected from being suppressed. Also, weights given to attributes improve clustering and give the ability to control the generalization's depth [9]. Privacy preserving data mining using sample selection and singular value decomposition is the method of data-mining to preserve the data and get loss less data. The proposed method is improved sample selection and singular value decomposition over basic methods of singular value decomposition. In our model we have taken the two data sets from university of California, first is Glass and second is Abalone.

First we select one method out of four methods ss-svd, s-svd,b-svd and iss-svd. then we select source From the file.If file support the source then we have to select the data set out of two. Dataset 1 is Glass and data set 2 is Abalone. when we select data set then population has to be choose and then create variable matching preview. Then the svd method is applied on the variables. Sample selection and singular value decomposition method has different privacy parameters to preserve the privacy[17 18]:

Utility measure ,VD ,RP ,RK ,CP, CK,Accuracy. The higher the privacy ,the larger the values of VD,RP and CP and smaller the values of RK ,CK and accuracy should be high .Accuracy should be high then we get the original data after distortion. Noise is added to the data then we have to get the noise less data. We use the matrix decomposition, we arranged the values of data sets in matrx form and then decompose the matrix into two parts when we combine the matrix we should get the original matrix without any lossless data. Here we have found the complement of matrix ,inverse of matrix, rank of matix, eigen values and eigen vectors of matrix to put into the formula. When by solving these formulaes we get the values of privacy parameters whose VD,RP and CK should have the higher values and RK,CK and accuracy should be high. There are many methods of privacy preserving data mining like anonymization, randomization, cryptography etc.We have choosen the sample selection and singular value decomposition method whose sccuracy comes out to be high upto 95 %.We have made the comparision graph between all the svd method versus improved ss-svd.we can see that the accuracy of the iss-svd is higher than the other methods.It is for all parameters and which has different values for different samples.we have made the graphs for sample 1 for dataset 1 and for dataset2 and graph for sample 9 for dataset 1 and Dataset 2 in which accuracy of iss-svd in both case is higher than the other svd methods.

The section 2.1 shows proposed algorithm.

A. Proposed Algorithm

Input: A data set according to sample selection

Output : a mixed transform table data

class: E={},the set of the equivalence classes

QIC={},set of equivalence classes with similar QI sets

CIP{}, set of attributes with similar class

DIP=number of different class values in the remaining dataset

Begin

While CIP >= attribute Cluster T to m tables according QI

For i=1 to m

Bucketize attributes according SA values While

|DIPi|>=ℓ Create_equivalence_classes ()

E=E U Create_equivalence_classes()

return E

Incorporate the remaining attributes to E End

Generate equivalence class with prototype is Input: CIP

Output :E Begin

Randomly selection of a attributes tm from the smallest group

E={tm}

For p=1 until attribute-1

Select a attributes tp that minimizes the gcp

E=E U tp

Remove tp from T

Remove tm from T

Return E

End

Process of cluster generation in prototype classification

Input: data set used definedutput: QIC={}, set of tables with attributes with similar QI sets

Begin

Insert T to the decision tree classification

QIC={ QIC1, QIC2,… QICm }

return QIC End

B.Parameters Used

1. VD (Value Difference): After matrix is distributed, the value of its elements changed. The VD [13] of the datasets is represented by the relative value difference in the normal form. The VD is the ratio of the frobenence norms to the difference of A from A' to frobenence form of A.

$$VD=IA-\bar{A}I \times F/A/f$$

2. RD (Rank Difference): After a data distribution the rank of the magnitude of the data element changes too. We use several matrixes to measure the rank difference as the data element. Let, dataset A, n objects attributes, rank of i^j denotes the rank in the ascending order of the i^{th} element in attribute j^{th} [14].

$$RD = \sum_{i=1}^{m} \sum_{j=1}^{n} \left| rank^i - rankj \right| / mxn$$

If two elements have the same value, we define elements with the smaller index to have the higher rank in data set. As the rank vector for the first attribute.

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

4. RK: It represents the percentage of the elements that keep their rank of magnitude in each column after the distortion, it is computed as

$$RK= \sum^{m} \sum^{n} RK^i_j / mxn$$

Where, Avi is attributing of values.

3. Simulation Environment

It is simulating on matlab 7.8.0 and for this work we use Intel 1.4 GHz Machine. MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation Matlab is a software program that allows you to do data manipulation and visualization, calculations, math and programming. It can be used to do very simple as well as very sophisticated tasks. Image Processing Toolbox provides a comprehensive set of reference-standard algorithms and graphical tools for image processing, analysis, visualization, and algorithm development. You can perform image enhancement, feature detection, noise reduction, image segmentation, spatial transformations. Matlab is a commercial "Matrix Laboratory" package which operates as an interactive programming environment. It is a mainstay of the Mathematics Department software lineup and is also available for PC's and Macintoshes and may be found on the CIRCA VAXes. Matlab is well adapted to numerical experiments since the underlying algorithms for Matlab's builtin functions and supplied m-files are based on the standard libraries LINPACK and EISPACK. Matlab program and script files always have filenames ending with ".m"; the programming language is exceptionally straightforward since almost every data object is assumed to be an array. Graphical output is available to supplement numerical results

4. Data Set Description

Data has been taken from University of California (UCI). To perform experiment work two dataset has been taken is Glass Dataset & Abalone Dataset. The property of Dataset is given belo

Table 1: Dataset Table

	METHODS			
PARAMETERS	**SS-SVD**	**B-SVD**	**S-SVD**	**Proposed**
UTILITY MEASURE	0.16997	0.83002	-1.83003	-1.83003
VD	3.15797	2.15797	1.15797	-0.15797
RP	35	38	40	45
RK	1.9129	0.9129	-0.087	-1.087
CP	67.4523	72.4523	74.4523	67.4523
CK	2.16667	1.61667	0.16667	-0.8333
ACCURACY	80.3894	82.3894	85.3894	87.3894

5. Experiment Results

Experimental results of the proposed ISS-SVD method. It contains two datasets dataset-1-Glass and dataset-2-Abalone.We have used the SVD algorithm for our data classification.We select the SVD method and then we select one of the two datasets. We select the sample from the random samples then comes out the result of all privacy parameters. We implemented this logic in a matlab 7.8.0 and we get good result as high accuracy of data above 95%.All SVD methods have different parameters. There are seven parameters CP, CK, VD, RP, RK, utility

measure and accuracy, which should be high. Here we use singular value decomposition with sample selection. We have taken the 2 datasets from UCI (university of California).The values are taken from the datasets which are arranged in matrix form. Matrix is decomposed into two parts, when we combine the two sub matrix ,we should get the original matrix without any information/data loss. we find the complement of the matrix to apply on various parameters provided in SVD methods to get the better accuracy. The name of the proposed method is improved sample selection and singular value decomposition

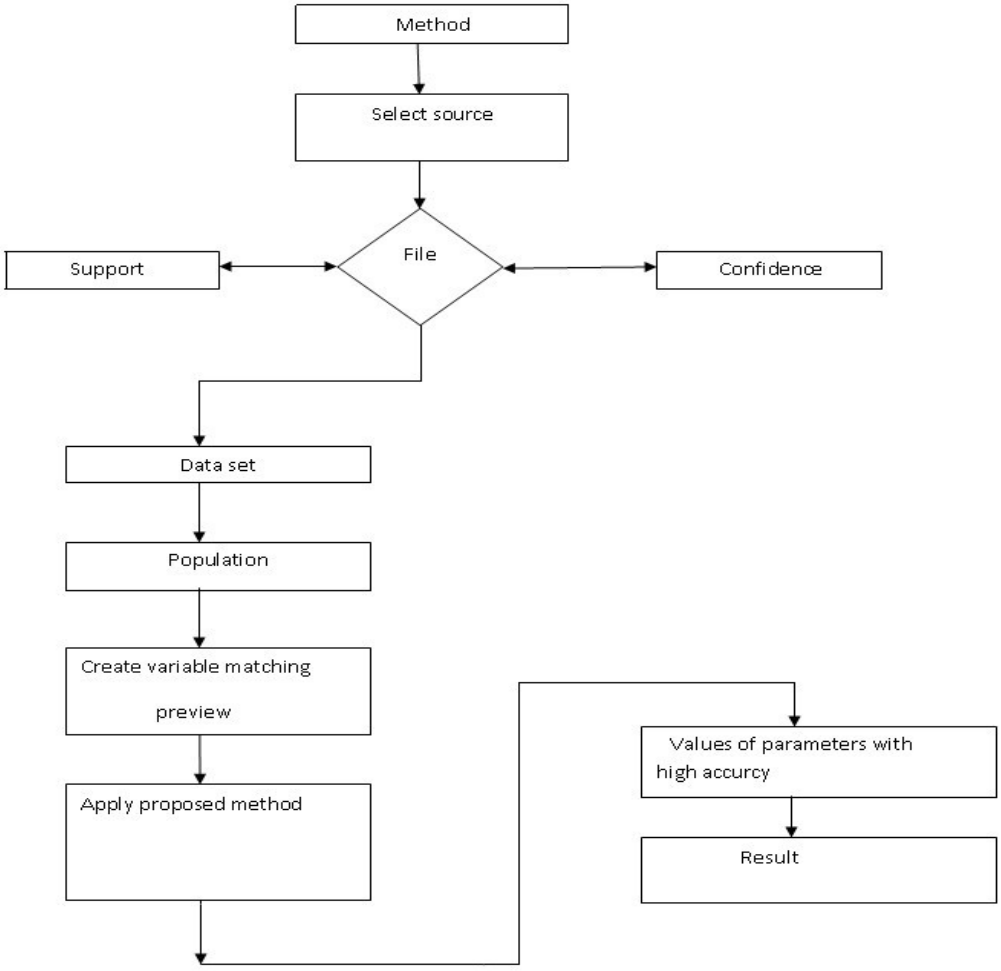

Fig 1. Hybrid model for improved Perturbation and single value decomposition method

To perform Experiment work we have applied dataset to SS–SVD method, B-SVD method, and S- SVD method & Proposed Method. Fig 2, Fig 3, Fig 4 & Fig 5 showing experiment result on various parameters.

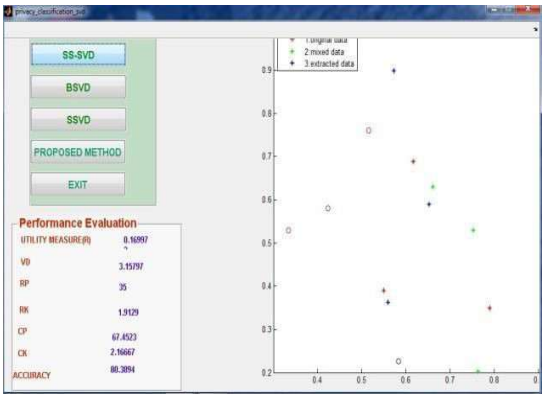

Fig 2. SS-SVD METHOD DATASET FOR SAMPLE (.1)

Experimental 1: Result of SS-SVD

We have used SS-SVD method for dataset 1 (Glass) for sample 10.we get the different values of parameters by using sample selection & single value decomposition algorithm. Here come out the results of all parameters. Accuracy rate is 80%.

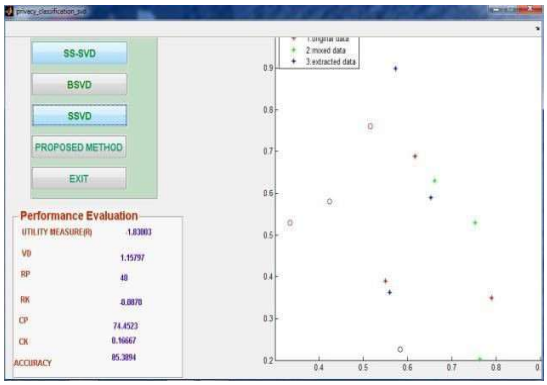

Fig 3. S-SVD METHOD DATASET FOR SAMPLE (.1)

Experimental 3: Result of S-SVD

We have used S-SVD method for dataset 1 (Glass) for sample 10.we get the different values of parameters by using sample selection & single value decomposition algorithm. Here come out the results of all parameters. Accuracy rate is 85%.

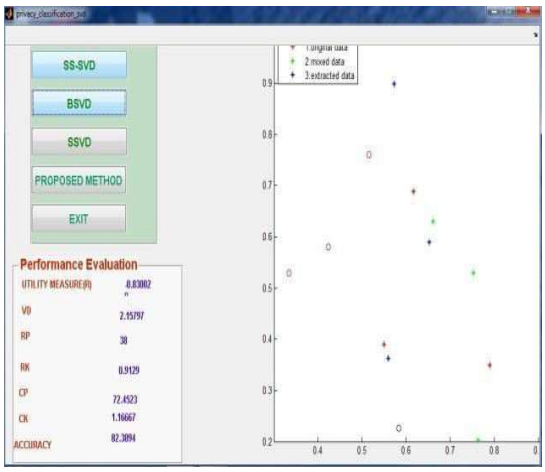

Fig 4. B-SVD METHOD DATASET 1 FOR SAMPLE (.1)

Experimental 3: Result of B-SVD

We have used B-SVD method for dataset 1 (Glass) for sample 10.we get the different values of parameters by using sample selection & single value decomposition algorithm. Here come out the results of all parameters .Accuracy rate is 82%.

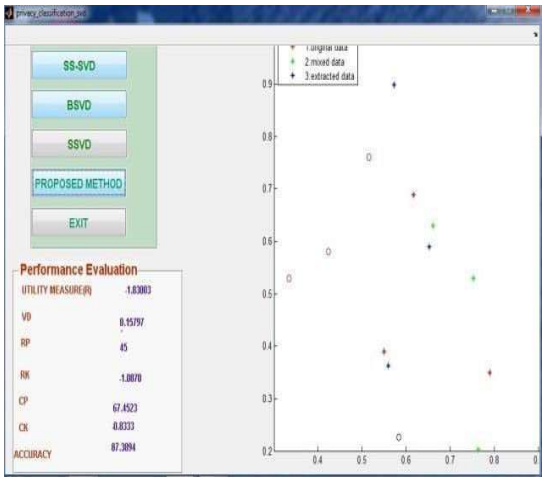

Fig 5. ISS SVD METHOD DATASET 1 FOR SAMPLE (.1)

Experimental 4: Result of ISS-SVD (proposed) method

We have used ISS-SVD (proposed) method for dataset 1 (Glass) for sample 10.we get the different values of parameters by using sample selection & single value decomposition algorithm. Here come out the results of all parameters. Accuracy rate is 87%.It is the highest accuracy among all the methods.

METHODS	UTILITY	VD	RP	R	CP	C	ACCURACY	SAMPLE
S-SVD	6.62504	10	50	9.213	84.022	8.62174	94.6895	9
ISS-SVD	6.62504	8.61304	55	8.213	84.022	7.62174	96.6895	9

Figure 6 Dataset S-SVD and ISS-SVD (proposed)

Figure 6 performs comparison between SS-SVD and ISS-SVD method on basis of various parameters used in section 2.2 like VD, RP, RK, CP, CK and Accuracy. This experiment performed in using Mat lab version 7. Graph between SS-SVD and ISS-SVD(proposed) for dataset-2 at sample 90 has two axes, first on Y-axis SS-SVD in blue color and ISS-SVD in red color are there and on X-axis seven parameters on sample 10.We can see that the accuracy of proposed method ISS-SVD is higher than SS-SVD is 96.69 %. The result is shown in figure. The Figure performs comparison between B-SVD and ISS-SVD. Graph between B-SVD and ISS-SVD(proposed) for dataset-2 at sample 90 has two axes, first on Y-axis B-SVD in blue color and ISS-SVD in red color are there and on X-axis seven parameters on sample 10.We can see that the accuracy of proposed method ISS-SVD is higher than B-SVD is 96.69%. between S-SVD and ISS-SVD(proposed) for dataset-1 at sample10 has two axes, first on Y-axis s-SVD in blue color and ISS-SVD in red color are there and on X-axis seven parameters on sample 10.We can see that the accuracy of proposed method ISS-SVD is higher than S-SVD is 87.39%. between SS-SVD and ISS-SVD(proposed) for dataset-2 at sample 90 has two axes, first on Y-axis SS-SVD in blue color and ISS-SVD in red color are there and on X- axis seven parameters on sample 10.We can see that the accuracy of proposed method ISS-SVD is higher than SS- SVD is 96.69 %.

6. Conclusions & Future Expansion

In this paper, we carries out a wide survey of the different approaches for privacy preserving data mining, and analyses the major algorithms available for each method and points out the existing drawback & *this work is extension of our previous work..* While all the purposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods. The essence of the matrix decomposition based methods is to use attribute extraction by matrix decompositions to analyze data and find and retain only the important information for data mining. These methods achieve data perturbation by removing the unimportant information for data mining. In addition to attribute extraction, sample selection also can do data analysis. In the new SS-SVD method, both sample selection and attribute extraction are used, so that the important information for data mining is found more accurately. Data mining application as privacy preserving various techniques are used such as association rule mining [6,7], clustering technique and classification technique. And also used some data mixed technique for adaptive noise data in original data. Matrix decomposition is big role in privacy preserving in data mining decision tree. Proposed approach satisfies the "utility based anonymization [4] principle that crucial information is protected from being suppressed. Also, weights given to attributes improve clustering and give the ability to control the generalization's depth [9, 10].

Future work tends to new model using different algorithm. To address this issue, we advise that the following problems should be widely studied:

(1) Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.
(2) Side-effects are unavoidable in data sanitization process. How to reduce their negative impact on privacy preserving needs to be considered carefully. We also need to define some metrics for measuring the side-effects resulted from data processing.
(3) In distributed privacy preserving data mining areas, efficiency is an essential issue. We should try to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost.

References

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.557-570,2002
[2] R. Bayardo, R. Agrawal, "Data Privacy Through Optimal k-Anonymization", In Proceedings the 21st

International Conference on Data Engineering, pp.217-228, 2005.

[3] K. Lefevre, J. Dewittd, R. Ramakrishnan, "Incognito: Efficient Full- Domain k-Anonymity", In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp.49-60, 2005.

[4] B. Fung, K. Wang, P. Yu, "Top-down Specialization for Information and Privacy Preservation", In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005

[5] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression",International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5,- 6

[6] S. Rizvi, J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[7] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", J. Am. Stat. Assoc., vol.60, no.309, pp.63-69,1965

[8] S.J. Rizvi, J.R. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th VLDB conference, pp.1-12, 2002.

[9] W. Du, Z. Zhan, "Using Randomized Response Techniques for Privacy Preserving Data Mining", In Proceedings 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.505-510, 2003.

[10] L. Guo, S. Guo, X. Wu, "Privacy Preserving Market Basket Data Analysis", In Proceedings the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp.103-114, 2007.

[11] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis Privacy- Preserving Ridge Regression on Hundreds of Millions of Record IEEE Symposium on Security and Privacy 2013

[12] Suman Jana, Arvind Narayanan A Scanner Darkly:Protecting User Privacy From erceptual Application IEEE Symposium on. Security and Privacy 2013

[13] Marina Blanton Achieving Full. Security in Privacy-Preserving Data Mining IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing pp
925-934, 2011