© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Rotation Perturbation Technique for Privacy Preserving in Data Stream Mining

Kalyani Kathwadia[1], Aniket Patel[2]
Department of Information Technology Silver Oak College of Engineering And Technology Ahmedabad, Gujarat, India

ABSTRACT

Datasets is very challenging task in the systems. It is real processing. Data mining technique classification is one of the most important technique, in this paper is to classify the data as improve the classification accuracy, we have used ensemble model for classification of data. Randomization process added to privacy sensitive data after next process reconstruction to the main data from the perturbed data. Principal Component Analysis (PCA) is used to preserve the variability in the data. Rotation transformation can enlarge the increase the base classifiers and improve the accuracy of the ensemble classifier. In this paper, we analyses a rotation perturbation technique for PCA find eigenvector, load line plot and Zscore-Normalization method using to dimension in stream mining.

Keywords: Data mining, Classification, Privacy, PCA, Z score–Normalization

I. INTRODUCTION

The Data Mining is the system of examining large pre-existing database in order to develop new information. Data mining algorithm like clustering, classification work on this data and provide crisp information for analysis [1]. In recent year data mining as a powerful data analysis tool in many areas and wide application with the development of data base technology and network technology, a large number of useful data, which contains much individual privacy information has been gain in various fields, such as condition information, customer preferences, personal background information etc. Privacy is becoming an increasingly important issue in data mining applications that deal with health care, security, financial, behavioural, and other types of sensitive data [11]. Data Stream is concerned with extracting knowledge structure in models and patterns continuous streams of information. Data stream can data sizes many times greater than recollection, and can be extended to objection actual time applications by machine information or data mining [14].

Privacy preserving is one of the most extensive investigation fields in the data protection field and it has become in the protected conversion of intimate data. A number of algorithmic techniques have been designed for privacy preserving data mining. The ongoing privacy preserving data mining procedure are classified based on distortion, association rule, hide association rule, taxonomy, clustering association, classification outsourced, data mining scattered.it is used to easily protect individual privacy in data sharing. Privacy preserving data mining has become it allows sharing of privacy sensitive data for analysis purposes. Privacy preserving data mining have three types of methods: Reconstruction based method, Heuristic based method, and Cryptographic based method. In this paper we are using Reconstruction based method.

There are three types of reconstruction based approaches: Data perturbation, Data randomization, Data Swapping. Reconstruction based approaches generate privacy appreciative database by extracting sensitive data from the original database. Reconstruction based techniques perturb the original data to achieve privacy preserving perturbing the data for privacy preserving is very conducive technique used by reconstruct the distributions at an aggregate level in order to perform the mining.

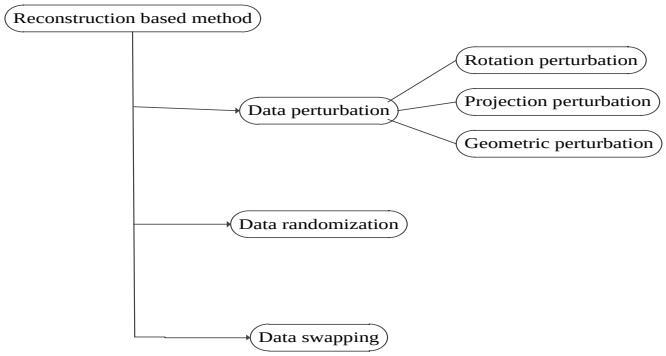


Figure 1. Reconstruction Based Method

II. BACKGROUND AND RELATEDWORK

In this paper, we are using data perturbation techniques. This section presents Rotation perturbation, projection perturbation, Geometric data perturbation. The classification of datasets has come under heated discussion among in recent years. The accuracy of combined classifiers increased and processing time reduced

DATA PERTURBATION

Data perturbation three section it is a rotation perturbation, projection perturbation, geometrics perturbation. Data perturbation approaches can be grouped into two main categories the probability distribution approach and value distortion approach [11]. The work in proposed classification data perturbation. The data reconstructs to original data from its classification models.

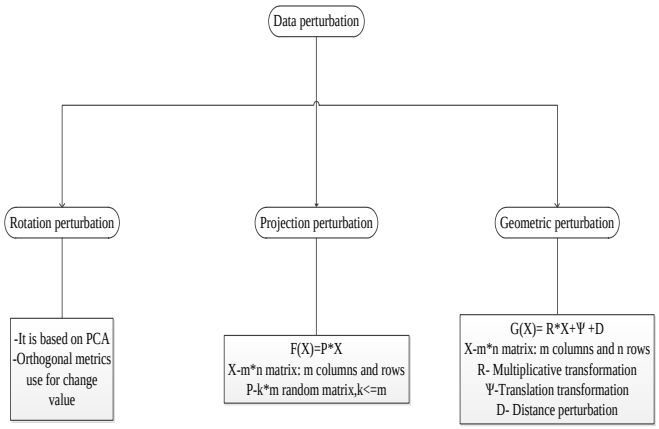


Figure 2. Classification Data Perturbation Techniques

Level-1 Rotation Perturbation: Rotation perturbation is used for classification and clustering for privacy preservation. Rotation perturbation is used in based on principal component analysis (PCA). It is orthogonal metrics use for change values. It is also used geometrics data transformation as G(x) = RX. Where R is a rotation metrics X is original dataset. Rotation perturbation is also distance preserving.

Level-2 Projection perturbation: In projection perturbation we project the set of original data from original space to random space. Suppose F(X) = P*X, Where X is m*n metrics: m columns and n rows, P is random metrics.is applied to original dataset X for the perturbation.

Level-3 Geometric perturbation: Geometric perturbation is used in privacy preserving in collaborative data mining.it is most widely used. Many popular data mining models are steady in geometrics perturbation. K- Nearest neighbour classifier, linear classifier, support vector machine classifier are steady means classifier with geometrics data perturbation has almost same accuracy as the original data. Suppose G(X) = R*X+ Ψ + D, Where X is m*n metrics, R is multiplicative transformation, Ψ translation transformation, D is distance perturbation.

PRINCIPAL COMPONENTS ANALYSIS (PCA)

PCA is a numerical process that uses an ethical metamorphosis to convert a set of measurement of probably revised variables into a set of values of linearly incorrect variable called PCA. A method of analysis which involved finding the linear combination of a set of variables that has maximum variance and removing its effect repeating this successively. PCA is used to perturb the multidimensional data into dimensions. It is used to reduce dimensionality of the dataset, Find patterns in high dimensional data, Visualize data of high dimensionality. PCA application are Text processing, Image processing, Speech recognition, Recommendation engine. In this paper we find principal components using eigenvector and eigenvalue.

Now we are selecting some data set D into N*N metrics and we find eigenvectors and eigenvalues.

An eigenvector of an N * N matrix A is a nonzero vector X such that AX = λX for some scalar λ. A scalar λ is called an eigenvalue of A if there is a nontrivial solution X of AX = λX; X is an eigenvectors comparable to λ.

λ is an eigenvalue of N * N matrix A if and only if the equation:

$$(A-\lambda I)X = 0 \text{ …….. (1)}$$

The A - λI has the form

$$A-\lambda I=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$$

$$= \begin{bmatrix} a_{11-\lambda} & a_{12} & a_{13} \\ 0 & a_{22-\lambda} & a_{23} \\ 0 & 0 & a_{33-\lambda} \end{bmatrix}$$

This λ equals one of the entries a_{11}, a_{22}, and a_{33} in A After compute eigenvector VI. Gaussian elimination mathematical technique is used.

Loading line plot is a multivariate model for viewing the PCA. Each variables coefficient to the new data.

Loading = $\sqrt{\lambda_1} * eig_1$

λ_1 Eigenvalues, eig_1 Eigenvectors

NORMALIZATION

In Data Mining, Normalization is scaling techniques or a mapping technique or a pre-processing stage. The technique which provides linear transformation on original range of data is called Min-Max Normalization. Also technique keeps relationship among original data is called Min-Max Normalization. Min – Max Normalization is a normalization scheme which linearly change completely X to Y = (X - Min)/ (max –min), where min and max are the minimum and maximum values in X, where X is the set of observed values of X. it can be easily seen that when X=min, then Y=0. In this paper we are using Z score normalization it is very useful statistic because it allows us to calculate the probability of a score occurring within our normal distribution and enables us to compare two score that are from different normal distributions. Z score equation is:

$$\text{Z score} = x-\mu / \sigma;$$

Where μ is mean, x is data, σ standard deviation. In this equation using first we find values. After we get data then we are putting in numerical data of normalization. Standardized data is part of the derivation process similar data received in various formats is transformed to a common format that enhances the comparison process for example contains directions. Standard deviation σ;

$$\sigma = \sqrt{\frac{\Sigma(x-X)^2}{n-1}}$$

Where x is data value, X is mean value, n is number of value. Database Normalization is the process of organizing the attributes and relations of database to reduce data redundancy and improve data integrity.

III. PROCEDURE

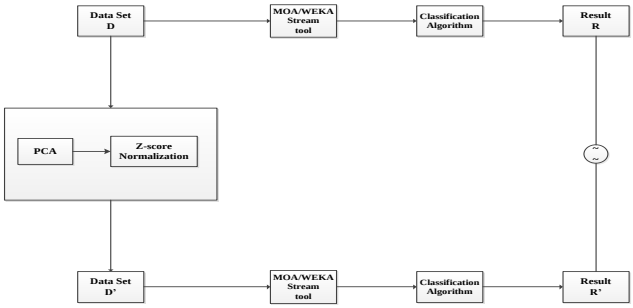


Figure 3. Proposed System

Procedure: Privacy Preserving Data Mining by PCA and Normalization.
Input: Data set D.
Intermediate Result: Perturbed data stream D'.
Output: Classification, Result R' of data stream D'.
Steps: -
Step 1: Data Set D.
(It is represent n*n metrics)
Step 2: Calculate Eigenvalue.
(det (A-λi) =0)
Step 3: Compute Eigenvectors using Gaussian elimination value. (A-λi I)
Step 4: Loading Line Plot
$\sqrt{\lambda_1} * eig_1$
Step 5: Create perturb dataset D' by using z-score normalization in original dataset D. (z-score=$\frac{x-\mu}{\sigma}$)
Step 6: Apply Classification algorithm in data stream.

Step 7: Get Result.

IV. RESULT AND DISCUSSION

In our implemented on java NetBeans we using WEKA tool and bank dataset used. An compared original dataset and modified dataset. For classification algorithms using in this data that is NaiveBayes and J48.

Table 1. Table Dataset D and D' Bank marketing for NaiveBayes

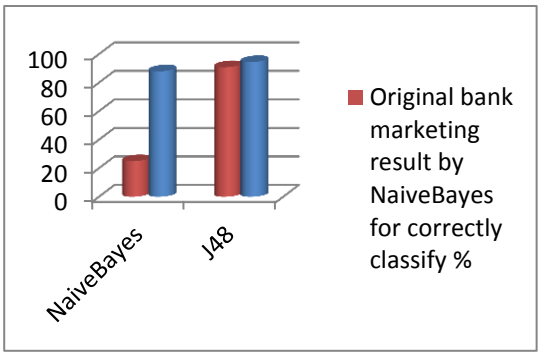
For Original Dataset	Measure	For Modified Dataset
weka.classifiers .bayes.NaiveBa yes	Scheme	weka.classifiers .bayes.NaiveBa yes
bank-full1	Relation	out3 (kk.csv)
44242	Instances	44242
Age, balance, duration	Attributes	Age, balance, duration
0.35 seconds	Time taken to build model	0.03 seconds
24.9695%	Correctly Classified Instances in %	88%
75.0305%	Incorrectly Classified Instances in %	12%
0.4304	Kappa statistic	0.7569
0.1458	Mean absolute error	0.2011
0.2976	Root mean squared error	0.3197
24.8936 %	Relative absolute error	100%
95.5333 %	Root relative squared error	100%
96.9486 %	Coverage of cases (0.95 level)	95.0296 %
67.7772 %	Mean rel. region size (0.95 level)	92.8058 %
a b <-- classified as 36857 2568 | a = no 2398 2419 | b = yes	Confusion Matrix	a b <-- classified as 44154 1849 | a = no 2428 2424 | b = yes

Table 2. Table Dataset D and D' Bank marketing for J48

For Original Dataset	Measure	For Modified Dataset
weka.classifiers .trees.J48 -C 0.25 -M 2	Scheme	weka.classifiers .trees.J48 -C 0.25 -M 2
bank-full1	Relation	out3 (kk.csv)
44242	Instances	44242
3	Attributes	3
0.01	Time taken to build model	0.01
90.8707 %	Correctly Classified Instances	94.5%
9.1293 %	Incorrectly Classified Instances	5.5%
0.476	Kappa statistic	0.9514
0.1207	Mean absolute error	0.114
0.2699 %	Root mean squared error	0.1235
86.66 %	Relative absolute error	25.03%
51.9514 %	Root relative squared error	45.358%
a b <-- classified as 37966 1459 | a = no 2580 2237 | b = yes	Confusion Matrix	a b <-- classified as 44148 1849 | a = no 2588 2342 | b = yes

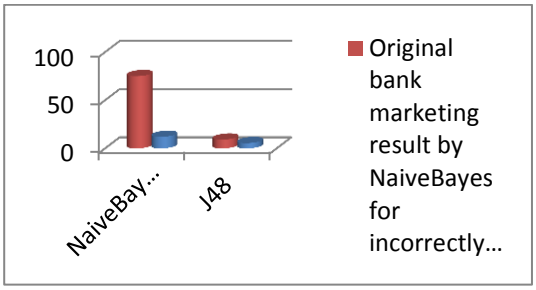
- Original dataset and modify dataset result by NaiveBayes and J48 for correctly instance:

DATA	NAIVEBAYES	J48
Original	24.9695	90.8707
Modify	88	94.5



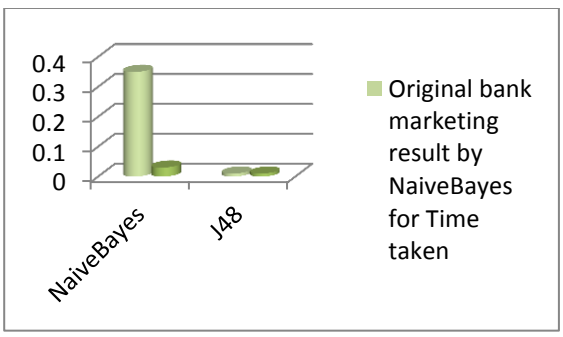
- Original dataset and modify dataset result by NaiveBayes and J48 for Incorrectly instance:

DATA	NAIVEBAYES	J48
Original	75.0305	9.1293
Modify	12	5.5



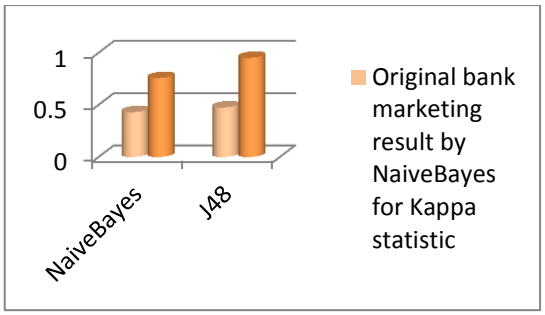
- Original dataset and modify dataset result by NaiveBayes and J48 for time taken build model:

DATA	NAIVEBAYES	J48
Original	0.35	0.01
Modify	0.03	0.01



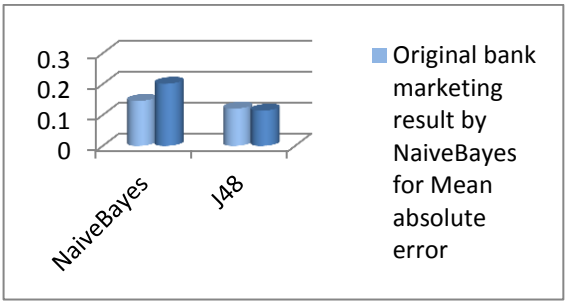
- Original dataset and modify dataset result by NaiveBayes and J48 for kappa statistic:

DATA	NAIVEBAYES	J48
Original	0.4304	0.476
Modify	0.7569	0.9514



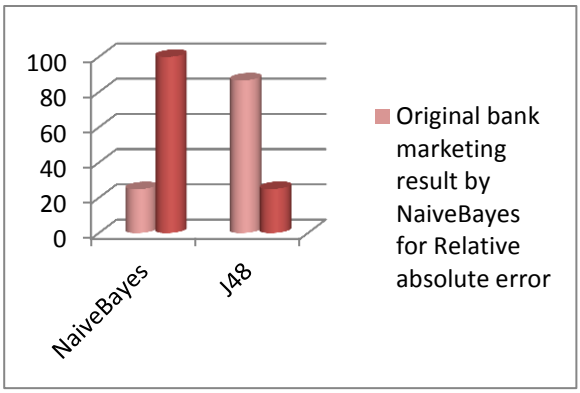
- Original dataset and modify dataset result by NaiveBayes and J48 for mean absolute error:

DATA	NAIVEBAYES	J48
Original	0.1458	0.1207
Modify	0.2011	0.114



- Original dataset and modify dataset result by NaiveBayes and J48 for relative absolute error:

DATA	NAIVEBAYES	J48
Original	24.8936	86.66
Modify	100	25.03



V. CONCLUSION

Data classification is very important in the field of data mining. In this paper we discussion in data perturbation techniques in rotation perturbation based on PCA. Approaches can be useful find perturbation technique to maintain privacy of data can improve in WEKA tool and improve classification accuracy. And checked the correctly classify, incorrectly classify, time taken, kappa statistic, mean absolute error, relative absolute error in our system.

VI. REFERENCES

[1]. N P Nethravathi, Prasanth G Rao, P Deepa Shenoy, Venugopal K R, Indramma M."CBTS: correlation based transformation strategy for privacy preserving data mining" (IEEE), 2015.

[2]. Julius Adebayo, Lalana Kagal. "A Privacy Protection for Procedure for Large Scale Individual Level Data" (IEEE), 2015.

[3]. C.Gokulnath, M.K.Priyan, Vishnu Ballan." Preservation of privacy in data mining by using PCA based perturbation technique" (IEEE), 2015.

[4]. Yingpeng Sang, Hong Shan, and Hui Tian. "Effective reconstruction of data perturbed by random projections" (IEEE) 2012.

[5]. Shuheng Zhou, Katrina Liggett, Larry Wasserman." Differential privacy with compression" (IEEE) 2009.

[6]. M.Parvathy, K. Sundarakantham, S. Mercy Shalinie, and C.Dhivya. "An efficient privacy

protection machism for recommendation using hybrid transformation technique" (IEEE) 2014.

[7]. R.viidya Banu, N.Nagaveni." preservation of data privacy using PCA based transformation" (IEEE) 2009.

[8]. Zhen Lin, Jig Wang, Lian Liu, Jun Zhang."Generalized random rotation perturbation for vertically partitioned data sets" (IEEE) 2009.

[9]. Liming Li, Qishan Zhang."Privacy preserving clustering technique using hybrid data transformation method" (IEEE), 2009.

[10]. Li Liu, Murat kantarcioglu, and Bhavani Thuraisingham."The application of the perturbation model based privacy preserving data mining for real-word data" (IEEE) 2006.

[11]. Kun Liu, Hillol Kargupta, senior Member, IEEE, and Jessica Ryan. “Random projection based multiplicative data perturbation for privacy preserving distribution data mining" (IEEE) 2006.

[12]. Keke Chen, Ling Liu. "Privacy preserving data classification with rotation perturbation" (IEEE), 2005.