Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCED PRIVACY PRESERVATION WITH PERTURBED DATA USING FEATURE SELECTION

[1]V.S. PRAKASH, [2]Dr. A. SHANMUGAM

[1] Assistant Prof., Department of Computer Applications, Bannari Amman Institute of Technology, Sathyamangalam – 638 401, Tamil Nadu, India.
[2] Principal, Bannari Amman Institute of Technology, Sathyamangalam – 638 401, Tamil Nadu, India

E-mail: [1]prakashfuture@gmail.com [2]principal@bitsathy.ac.in

ABSTRACT

In data mining applications, privacy plays an imperative role. This has triggered the development of many privacy preserving data mining techniques. To facilitate privacy preservation in data mining or machine learning algorithms over horizontally partitioned or vertically partitioned data, many protocols have been proposed using SMC and various secure building blocks. Our previous works focused on preserving privacy by adapting individually adaptable perturbation model, which enables the individuals to choose their own privacy levels. But the downside is that it does not discover the computational results for privacy properly. This paper proposed a feature selection with privacy preservation in multi-partitioned dataset. Data can be sealed for privacy by perturbation technique as pseudonym name. In multi-partitioned data evaluation, it creates classification of data and selection of feature for data mining decision model which construct the structural information of model in this paper. The purpose of gain ratio method has taken in this paper to enhance the privacy in multi-partitioned data set. All features don't require protecting the confidential data for best model. The data representation for privacy preserving data mining has taken to increase the data mining technique to construct finest model without breaking the privacy individuals. An experimental evaluation is conducted to estimate the performance of the proposed enhanced privacy preservation with perturbed data using feature selection [EPPDFS] in multi-partitioned datasets demonstrated by diverse experiments conducted on both synthetic and real-world data sets.
Keywords: *Privacy Preserving, Data Mining, Perturbed Data, Feature Selection*

1. INTRODUCTION

Clustering is the procedure of identifying groups inside high-dimensional databases, supported with relationships, with negligible information of their organization. Conventional clustering algorithms achieve it over central databases; nevertheless, current applications need datasets dispersed among numerous sites. Consequently, in dispersed database environments, all dispersed data must be rigorous on an essential site previous to affecting traditional algorithms.

There are two separate states that insist the requirement for carrying out cluster analysis in a dispersed way. The first happens when the quantity of data to be examined is comparatively great, which command a substantial computational effort, which at times is even not viable, to realize this task. The best choice, then, is dividing data, collect them in a dispersed way and combine the results. The second happens when data is obviously dispersed among numerous geographically dispersed units and the cost connected to its centralization is very high.

Certain recent applications seize databases so huge, that it is not probable to remain them integrally in the focal memory, even utilizing robust machines.

i) Keeping data in a derived memory and grouping data subsets independently. Limited results are kept and, in a subsequent stage, are collected to group the entire set;

ii) Employing an incremental clustering algorithm, in which all elements is independently passed to the main memory and connected to one of the offered clusters or owed in a novel cluster. The outcomes are reserved and the element is not needed, so as to provide space to the other one;

iii) Using similar completion, in which numerous algorithms process concurrently on stored data, rising efficiency.

There is a sequence of boundaries which obstruct the operation of conventional data mining methods on distributed databases. The strategy usually taken, the combination of all dispersed databases in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a middle unit, pursued by algorithm application, and is robustly disproved, as in these cases, it is significant to get into deliberation some subjects, that is: the prospect of survival of parallel data with diverse names and layouts differences in data structures, and variances among one and another database.

Alternatively, incorporation of numerous databases in a distinct location is not recommended when it is made of huge databases. If an organization maintained databases to apply data mining algorithms, this procedure might command huge data transmission, which might be slow and expensive. Furthermore, any alteration that might happen in distributed data, in case, addition of novel information or modification of those previously offered will have to be reorganized along with the essential database. This needs a very compound data updating approach, with excess of information transmission in the system.

In cases in which the data set is combined and wants to be splitted in subsets, owing to its size, two strategies are used: horizontal and vertical partitioning (Figure 1). The first strategy is more utilized and comprises in horizontally splitting database, generating consistent data subsets, so that every algorithm controls on diverse records allowing for, but, the similar set of attributes. Another strategy is vertically splitting the database, generating mixed data subsets; in this point of view, each algorithm activates on the similar records, selling, but, with diverse attributes.

	1	2	3	n

	1	2	3	n

Figure 1 Horizontal And Vertical Partitioning

Both horizontal and vertical database partitioning are general in numerous areas of research, mostly in environments with dispersed systems and/or databases, to which viable application fits into it. The method how data is separate in a database environment depends on a sequence of aspects which not at all times observe the clustering analysis. Operational requirements might openly provide authority in the structure of data distribution and data mining algorithms must be robust enough to manage with these restrictions.

In our research work, we plan to present an effective and efficient cluster based privacy preserving data perturbation technique to mine Multi-partitioned data sets. Multi-partitioned data comprises of both vertical and horizontal data sets which is current demand of e-business and e-commerce data mining environment. In e-business data mining models, privacy becomes a key issue in preserving individual's data on product / service transactions. However the transparency and exposure of the product / service increase the volume of transaction to more new and existing clients. To evaluate a trade of between data privacy and transparency of individual's data, data perturbation technique is presented with validation and authentication. Gaussian distribution model is appropriated for data perturbation to preserve the private data of the individual's data authenticity for sharing is provided with respective individuals along with its validated period of sharing. However in the multi-partitioned data distribution, data perturbation raised ambiguity between vertical and horizontal partitions of the data. To overcome the ambiguity, we plan to introduce divisive k-neighbor clusters for multi-partitioned data sets.

2. LITERATURE REVIEW

Some effort has been made to address the problem of privacy preservation in data clustering. The paper [1] discovers the outlook of using multiplicative subjective projection matrices for secrecy conserving distributed data mining. In [2], proposed a structure that permits universal alteration of unique data using randomized data perturbation technique and the modified data is then offered as a conclusion of client's query through cryptographic technique. Privacy-preserving data mining (PPDM) distress the predicament of realizing data mining tasks [3] devoid of any straight admittance to the exclusive data sets, as the providers preserve isolation on their data. A protocol [4] can be employed to sustain such investigates in a privacy-sensitive manner. In this work, we plan to provide privacy preservation scheme for perturbed data in multi-partitioned datasets.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Data objects that have been spitted into clusters employing k-means clustering are troubled by processing geometric alterations on the clusters in such a way that the object association of every cluster and direction of objects inside a cluster stay behind the same [5]. In addition, the effectiveness of the dataset must be measured as the conversion takes place. Such data conversion crisis such that privacy customary must be assembled and the efficacy must be optimized is an NP-hard crisis. In [6], the author proposed an estimate algorithm for the data conversion crisis. The focused data processing addressed in this paper is categorization employing connection rule, or associative categorization.

Privacy-preserving data mining (PPDM) is one of the current inclinations in privacy and security investigation. Current advances in data compilation, data distribution and associated technologies have instated a novel period of research where offered data mining algorithms should be reassessed from a diverse point of view, this of confidentiality protection. The paper [7] discovered all the features of privacy concerns in data mining, particularly connected with clustering, and gives a method for privacy preserving grouping with a theoretical banking situation. Here the author proposed a representation for grouping horizontally partitioned or central data sets using a easy PCA based alteration approach. The Naïve Bayes categorization has been employed as of its applicability in case of evaluation dataset [8].

Along with the existing privacy preserving techniques, data anonymization presents a effortless and efficient way to defend the responsive data. Nevertheless, in most of the connected algorithms, data particulars are misplaced and the outcome dataset is far less instructive than the unique one [9]. In topical years of data mining requests, an efficient method to protect privacy is to anonymize the dataset that comprise private information before being unconfined for mining [10].

In [11], the author discovers a serious susceptibility of existing data perturbation algorithms that an opponent might develop to reinstate other users' private information (e.g., mean, variance and the sharing of unique data) from the agitated data, since all the participants distribute the similar noise allocation. Perturbation technique is a very significant method in privacy preserving data mining. In [12], defeat of information versus conservation of privacy is recognized a trade off amongst the users in it.

3. PROPOSED ENHANCED PRIVACY PRESERVATION WITH PERTURBED DATA USING FEATURE SELECTION

The proposed work is efficiently designed to enhance the privacy and security of perturbed data present in the multi-partitioned datasets. The proposed enhanced security mechanism for perturbed data in multi-partitioned using feature selection is processed under four different phases. The first phase describes the process of sharing of multi-partitioned data with the users. The second phase describes the process of clustering the multi-partitioned data with the divisive k-neighbor clustering procedure. The third phase describes the process of enhancing the applicability of perturbed data in multi-partitioned dataset. The fourth phase depicts the process of enhancing the security and privacy of perturbed data by selecting the features in the datasets. The architecture diagram of the proposed enhanced privacy preservation with perturbed data in multi-partitioned datasets is shown in fig 3.1.

Multi-partitioned data set consists of data which have been divided from any logical database. The data has been partitioned into two types: horizontally data partition and vertically data partition. Horizontal partitioning engages setting diverse rows into diverse tables. A general outline of vertical partitioning is to divide dynamic data from static data in a table where the dynamic data is not used as often as the static. Generating a view for the two tables re-establish the unique table with a performance penalty, though the performance will increase when accessing the static data e.g. for statistical analysis.

From the fig 3.1, it is being observed that the privacy preservation is achieved by adapting the adaptable perturbation model. Before adaptation of privacy preservation scheme, divisive k means clustering is applied to the multi-partitioned dataset which begin with individual, inclusive cluster and, at every step, divide a cluster until only singleton clusters of entity points stay behinds. In this case, we call for to choose, at every step, which cluster to divide and how to carry out the split. The divisive k means cluster is efficiently used for privacy preservation mechanism which overcome the issue of data perturbation technique raises the ambiguity among the clustered datasets results in unreliability. The divisive K-means is supported the idea that an axis point can symbolize a cluster. After clearing the ambiguity of the partitioned dataset, the adaptable perturbation model is used to enhance the privacy preservation scheme among the partitioned datasets.

Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

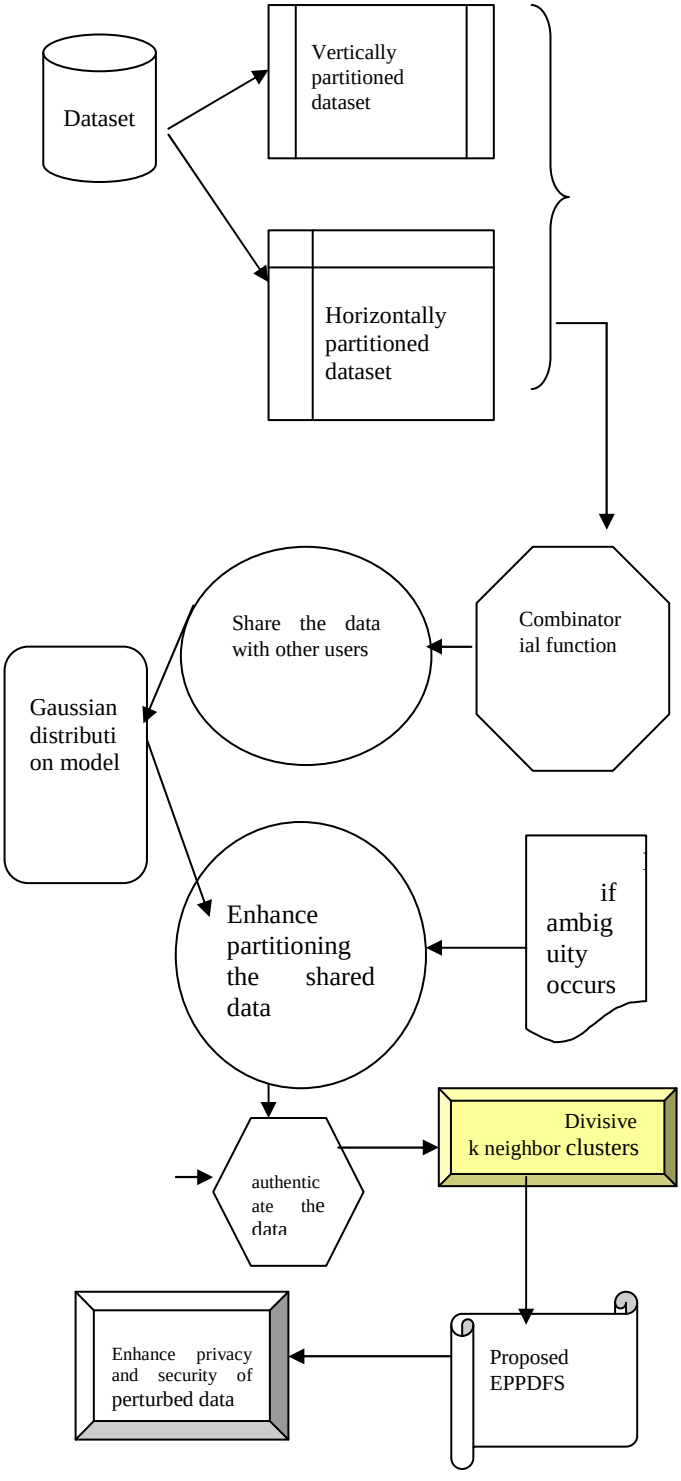

Figure 3.1 Architecture Diagram Of The Proposed EPPDFS

3.1 Sharing The Multi-Partitioned Dataset Using Combinatorial Function

The critical design of data perturbation is to modify the data so that valid individual data values cannot be improved, while preserving the value of the data for statistical summaries. Since the data doesn't react the real values of private data, even if a data item is linked to an individual that individual's privacy is not violated. If user A and User B wants to partition the data in the database, out first work presented an appropriate technique in a horizontal and vertical manner. To partition the dataset horizontally, gradient descent model is adapted here. For vertical partition of data in the dataset, kth element vector technique is employed. After partition the data either horizontally or vertically from the database, they have some sets of data to share it with other users in a safe and secure manner. That is, no other third party should involve in this sharing data between the users. Then users can partition the data from the database in a vertical manner. They received some sets of vertically partitioned data sets.

To share the data sets with different users along with privacy preservation, our first work presented a technique named Data Perturbation technique which is used to share the data from different users and unite all those data to get one complete true data sets. The combinatorial function is used to preserve the data sets which are to be shared among the users and it does not allow the third party members to seize the data. The combinatorial function will allow the users to share both the horizontal and vertical partitioned data sets to share with different users.

3.2 Clustering The Partitioned Dataset With Divisive K-Neighbor Clusters

Here, in this work, we are going to present a divisive k-means clustering to remove the ambiguity occurred in multi-partitioned dataset. The alternative of hierarchical clustering is called *top-down clustering* or *divisive clustering*. We established with all datasets in one cluster. The cluster is divided using a flat clustering algorithm. This practice is functioned recursively until every dataset is in its individual singleton cluster.

K-Means algorithms are admired and extensively used clustering methods. They divide the data to diminish the principle:

Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$E = \sum_{j=1}^{K} \sum_{x_i \in S_j} d^2(x_i, s_j) \quad \text{..... eqn 1}$$

Where K is the amount of clusters,

$s_j = \frac{\sum_{x_i \in S_j} x_i}{|S_j|}$ s_j is the axis of cluster of S_j

d (a,b) is the Euclidean distance

Divisive K-Means clustering algorithms regularly occupy deciding an arbitrary primary division or centers, and continually re-computing the centers supported on division and then re-evaluating the division based on the centers. Such process can be established to congregate to a restricted minimum, whereas the crisis of recognizing the inclusive minimum is NP-hard. We suggest a hierarchical divisive K-Means algorithm that reduces the same principle as the standard K-Means with clustering process planned as a hierarchical process.

For a given set of N items to be grouped, and an N*N distance (or resemblance) matrix, the procedure of divisive hierarchical clustering is this:

1. Establish by conveying every item to a cluster, so that if you enclose N datasets, you now enclose N clusters, each comprising just one item. Let the distances (resemblances) among the clusters the similar as the distances (similarities) among the items they include.
2. Discover the contiguous (most similar) pair of clusters and combine them into a particular cluster, so that now you contain one cluster less.
3. Calculate resemblances between the new group and each of the old groups.
4. Reiterate steps 2 and 3 until all items are grouped into a distinct cluster of size N. (*)

Divisive clustering initiates from the top, indulging the entire dataset as a cluster. It constantly divides a present cluster (a leaf node in a binary tree) until the amount of clusters achieves a pre-defined value K, or some other ending measures are met.

3.3 Proposed Feature Selection Scheme For Privacy Preservation

The proposed EPPDFS works on perturbed data which can achieve data mining assignment as if it processes on the unique data. But it can never understand the outcome or class concerning the data. It can state only the outcomes to all partitioned data in the data sets. As all the partitioned data only recognizes the outcome of their calculation on their data, then it conserves their security on their data. The proposed EPPDFS work gathers the data with the pseudonym technique for privacy preservation of perturbed data.

Usually pseudonym technique is utilized in perturbed method which is sealed by individual procedures of perturbed data. So the renewed groups of data values enclose only the pseudonym technique which is useful to examine the perturbed data in the dataset i.e., data miner who cannot infer any genuine values. It creates the noise of data in the perturbed data declares the genuine data changed to any other type of definite data or genuine database is altered into customized (perturbed) database. The perturbed data can be created by analyzing the queries based on the original datasets.

Each set of features and its values of sub-features have processed with both pseudonym set of feature and pseudonym values of sub-feature. The fig 3.2 shows the original database with perturbed data which was helpful for research work. The unique and perturbed database are both sustained by the system

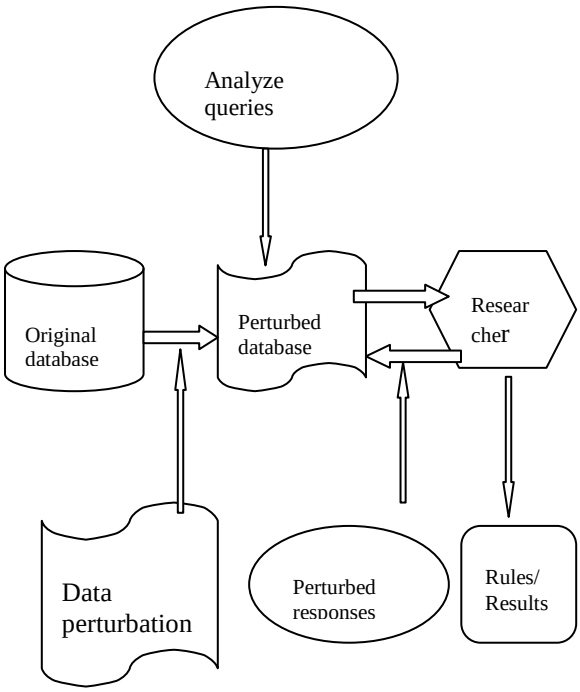

Fig 3.2 Perturbed Data Model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Let DA is the Dataset and C is the class set of database. When the entire dataset split into number of partition to distinguish the distinctive identifier feature value then the information vital to categorize data set DA supported on this partitioning would be info (DA) = 0. As, information gain on this set of features will be upper limit on all pure partition then there is no vital of categorization by such partitioning. The gain ratio which endeavors to evade this trouble by utilizing split information which is described below. The split information of a feature X is accepted as SI_X which is processed by

$$SI_X(DA) = \sum_{j=1}^{p} \frac{|DA_j|}{|DA|} \log_2\left(\frac{|DA_j|}{|DA|}\right)$$

…….. eqn 1

where DAj is the jth separation of D which fit in to feature X. The average values of each product of number of tuples in diverse partition are regarding the total number of tuples in D. It differentiates the measures of categorization information from information gain supported on similar partitioning. The gain ratio is described as below.

$$Gainratio(X) = \frac{Gain(X)}{SI(X)}$$

…….. eqn 2

Let D_i are the partition data of D and Cj are the different classes. Let P_i is the possibility of subjective tuples in D which fits in to class C_j. The estimated information of D for numerous classes is described by

$$Info(DA) = \sum_{j=1}^{allclasses} P_i \log_2(P_i) \quad \text{……..}$$

eqn 3

Here info (D) is the estimated information of D to recognize the class label. But the precise categorization is considered by individual feature after dividing the data as

$$Info_X(DA) = \sum_{j=1}^{no.ofpartitiones} \frac{|DA_j|}{|DA|} * \inf o(DA_i)$$

…….. eqn 4

where $\frac{|DA_j|}{|DA|}$ is the weight of the i th separation of data and $Info_x(DA)$ is the estimated information vital to categorize the tuples. The information gain is described as the difference among info (DA) and info $_X$(DA) ie

$$Gain(X) = \inf o(DA) - \inf o_X(DA_i) \text{.....}$$

eqn 5

The feature containing maximum gain ratio is chosen as the splitting feature. But the information on the perturbed dataset should not be zero, since the ratio turn out to be unbalanced. So when the set of perturbed data are putted their features, it might be dishonored the split information when the gain ratio is considered. The restrictions for this dimension is that the information gain of the analysis of preferred must be large, i.e., as a minimum of the standard tests examined with the perturbed datasets. The function of privacy should sustain by every peer which are taking part in the network. The diverse colluding sets of perturbed data might break their dimension to choose best feature.

The pseudo code below describes the process of the proposed EPPDFS.

Step 1: Assemble the occurrences of perturbed data from pseudonym name.

Step 2: Exchange pseudonym data to genuine set of data

Step 3: Based on the distribution of classes, the partitioned data are sent to individual class

Step 4: In every class, the #instances are marked as its individual condition

Step 5: Partitioned set of data are marked with both feature set and classes.

Step 6: Process the gain ratio method for feature collection

Step7: Sorting the values for best feature selection to enhance the privacy of perturbed data.

Step 8: Preserve the privacy for best feature data set.

With the above steps, the proposed EPPDFS provides a security for perturbed data in data mining applications. The next section describes the experimental evaluation of the proposed EPPDFS scheme and compared the results with the existing approaches.

5. EXPERIMENTAL EVALUATION

The proposed enhanced privacy preservation with perturbed data using feature selection [EPPDFS] has been implemented in Java. The experiments were run on an Intel P-IV machine with 2 GB memory and 3 GHz dual processor CPU. We are going to compare the proposed enhanced privacy preservation with perturbed data using feature selection [EPPDFS] with our previous works and an existing technique or single partitioned datasets. Using combinatorial function,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the datasets are partitioned effectively as such horizontally or vertically. So, the scalability of the products/services became less. After that, the strength of the data set to be shared remains identical from the beginning of the dividing process employing divisive k means clustering algorithm. After successfully removing the ambiguity occurred over dataset, in this work, we efficiently presented feature selection process for preserving privacy in perturbed data to enhance the privacy in multi-partitioned datasets. The performance of the proposed enhanced privacy preservation with perturbed data using feature selection [EPPDFS] is measured in terms of

i. Privacy level,
ii. Adversary attack rate and
iii. Authenticity

6. RESULTS AND DISCUSSION

The proposed enhanced privacy preservation with perturbed data using feature selection is reliably made for enhancing the privacy preservation in multi-partitioned dataset. The proposed EPPDFS allowed the users to share their file with other users by achieving a safe and secure communication. An experimental evaluation has also been carried out with benchmark dataset to estimate the performance of the proposed EPPDFS. The below table and graph describes the performance of the proposed EPPDFS and compared the results with an existing techniques for single partitioned datasets and with our previous works.

Table 5.1 User Density Vs. Privacy Level

User density	Privacy level (%)			
	Proposed EPPDFS	CPPDP	*CF*	Existing Technique for SPD
20	*70*	*56*	*42*	*37*
40	*76*	*52*	*50*	*43*
60	*79*	*69*	*56*	*48*
80	*84*	*73*	*62*	*54*
100	*90*	*76*	*67*	*60*

The above table (table 5.1) describes the privacy level of the users in the environment with their shared data in it. The privacy level of the proposed enhanced privacy preservation with perturbed data using feature selection is compared with an existing techniques for single partitioned datasets and with our previous works CPPDP [cluster based privacy preserving data perturbation technique] and CF [Combinatorial function].

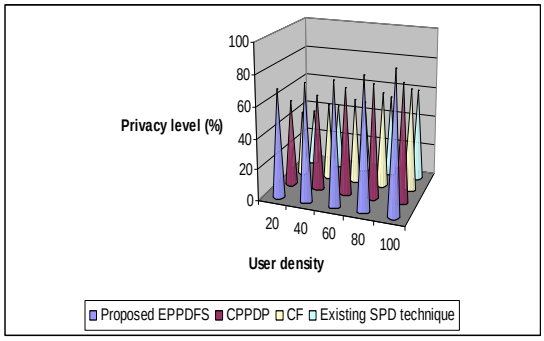

Figure 5.1 User Density Vs. Privacy Level

Fig 5.1 describes the privacy level of the users in the environment with their shared data in it. Normally, multi-partitioned data contains both vertical and horizontal data sets which are present command of e-business and e-commerce data mining background. In e-business data mining representations, privacy turns into an issue in preserving individual's data on product / service transactions. In the proposed EPPDFS, the privacy level of the individuals' data is efficiently preserved and processed over by adapting the feature selection process. In our previous works, the dataset in the database are efficiently partitioned over both horizontally and vertically, and shared with the users' involved in the transaction. While partitioning the dataset, there is a great chance for the ambiguity raised in the partitioned dataset. To remove the repeatability of data, the k-divisive algorithm is presented. Both our previous work does not much concentrated on the privacy preservation. So, compared with existing techniques for single partitioned datasets and with our previous works CPPDP [cluster based privacy preserving data perturbation technique] and CF [Combinatorial function], the proposed EPPDFS provide a high privacy level for individuals' data on partitioned datasets. The variance in privacy levels is 50-60% high in the proposed EPPDFS.

Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5.2 No. Of Perturbed Data Objects Vs. Adversary Attack Rate

No. of perturbed data objects	Adversary attack rate (%)			
	Proposed EPPDFS	CPPDP	CF	Existing Technique for SPD
10	6	11	15	23
20	10	16	19	28
30	13	19	23	34
40	15	22	27	39
50	18	26	31	45

The above table (table 5.2) describes the adversary attack rate of accessing unauthenticated data and the efficiency of a privacy preservation of data in the network. The adversary attack rate of the proposed enhanced privacy preservation with perturbed data using feature selection is compared with an existing techniques for single partitioned datasets and with our previous works CPPDP [cluster based privacy preserving data perturbation technique] and CF [Combinatorial function].

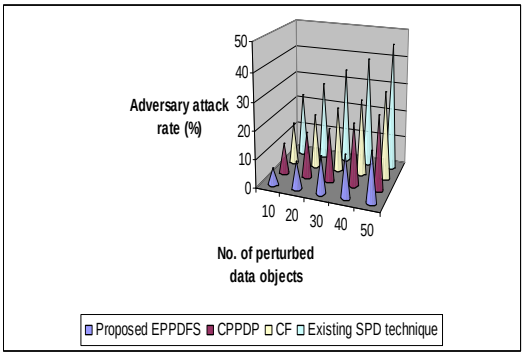

Figure 5.2 No. Of Perturbed Data Objects Vs. Adversary Attack Rate

Fig 5.2 describes the Adversary attack rate of accessing unauthenticated data of the users involved in the transaction of message among different users based on the number of perturbed objects present. In the proposed EPPDFS for privacy preservation, we have implemented successfully individually adaptable perturbation techniques which followed the preservation of privacy in multi-partitioned dataset. Since each objects/data in the dataset are clustered until each item in the dataset is clustered with a singleton cluster size, there is a less chance of multi-partitioned data set to be hacked by the adversaries. And also the privacy level of individuals' data is determined based on the two phase perturbation model by sampling the perturbed data objects. So, in the proposed EPPDFS, the adversary rate of accessing the multi-partitioned data in the dataset is very less compared to an existing techniques for single partitioned datasets and with our previous works CPPDP [cluster based privacy preserving data perturbation technique] and CF [Combinatorial function]. The variance in the adversary attack rate in the proposed EPPDFS is 40-50% low.

Table 5.3 No. Of Partitioned Data Vs. Authenticity

No. of Partitioned data	Authenticity (%)			
	Proposed EPPDFS	CPPDP	CF	Existing Technique for SPD
10	58	40	30	24
20	67	43	37	28
30	73	49	43	34
40	80	54	48	40
50	87	59	52	46

The above table (table 5.3) describes the authenticity of partitioned data in the dataset. The authenticity level of the proposed enhanced privacy preservation with perturbed data using feature selection is compared with an existing techniques for single partitioned datasets and with our previous works CPPDP [cluster based privacy preserving data perturbation technique] and CF [Combinatorial function].

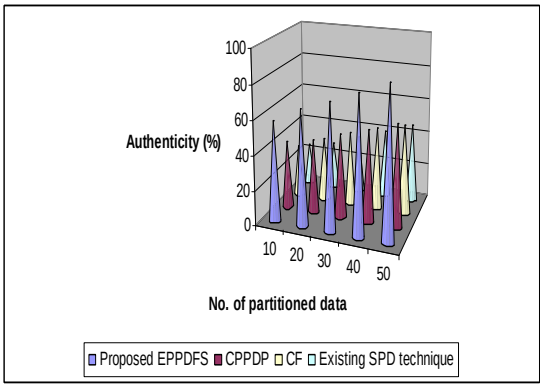

Figure 5.3 No. Of Partitioned Data Vs. Authenticity

Fig 5.3 describes the authenticity of partitioned datasets. When the partitioning data set size increases, the authenticity factor of the data sets should be high to enhance the partitioning process. The proposed EPPDFS provides high authenticity

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

rate when compared to existing techniques for single partitioned datasets and with our previous works CPPDP [cluster based privacy preserving data perturbation technique] and CF [Combinatorial function].

7. CONCLUSION

In the proposed EPPDFS, the data miner can find huge amount of data for making classification model. The classification of individual instances usually preserves more information. The data mining processing work can derive the feature selection using gain ratio technique for best feature as framework. The ordering of feature set has made from data mining framework. The sensitive and non-sensitive feature help to derive the classification of data model for privacy preservation of data at data miner. To share the data with other users, we first partition the datasets effectively in both horizontal and vertical manner. Then clustering process is done efficiently and enhanced the privacy preservation scheme by adapting the two phase perturbation model. Compared to an existing Combinatorial function (CF) for multi-partitioned dataset, the proposed enhanced privacy preservation with perturbed data using feature selection [EPPDFS] outperforms well and an experimental evaluation has been carried over with bench data sets obtained from popular e-business / e-commerce sites. (Amazon, e-bay etc.,) and estimated the performance of the proposed enhanced privacy preservation with perturbed data using feature selection in terms of privacy level, adversary attack rate and authenticity.

REFERENCES

[1] Kun Liu, Hillol Kargupta, et. Al., "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining" ,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 1, JANUARY 2006.

[2] P.Kamakshi , Dr.A.Vinaya Babu, "Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 4, APRIL 2010, ISSN 2151-9617

[4] Ying peng Sang, Hong Shen et. Al., "Effective Reconstruction of Data Perturbed by Random Projections", IEEE TRANSACTIONS ON COMPUTERS, VOL. 61, NO. 1, JANUARY 2012

[6] Jaideep Vaidya, et. Al., "Privacy-Preserving Kth Element Score over Vertically Partitioned Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 2, FEBRUARY 2009

[3] Dhiraj, S.S.S. et. Al., 'Privacy preservation in k-means clustering by cluster rotation', IEEE Region 10 Conference TENCON 2009 – 2009

[4] Natwichai, Juggapong , "An approximation algorithm for privacy preservation of associative classification", International Conference on Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI-CON), 2010

[5] Keshavamurthy, B.N. et. Al., 'Privacy-preserving Naive Bayes classification using trusted third party and different offset computation over distributed databases', 1st International Conference on Parallel Distributed and Grid Computing (PDGC), 2010

[6] Weijia Yang et. Al., "Knowledge Reserving in Privacy Preserving Data Mining", Second International Symposium on Intelligent Information Technology Application, 2008. IITA '08.

[7] Deivanai, P. et. Al., "A hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data", International Conference on Recent Trends in Information Technology (ICRTIT), 2011

[8] Fan Zhang et. Al., 'Data perturbation with state-dependent noise for participatory sensing', Proceedings IEEE INFOCOM, 2012

[9] Li Liu et. Al., "The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data", Sixth IEEE International Conference on Data Mining Workshops, 2006. ICDM Workshops 2006.

[10] Banu, R.V. et. Al., 'Preservation of Data Privacy Using PCA Based Transformation', International Conference on Advances in Recent Technologies in Communication and Computing, 2009. ARTCom '09.

[11] Li Liu , Murat Kantarcioglu et. Al., 'The applicability of the perturbation based privacy preserving data mining for real-world data', Science direct on Data & Knowledge Engineering 65 (2008) 5–21

[12] Keke Chen et. Al., "Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 20, NO. 12, DECEMBER 2009