Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Privacy Preserving Clustering Based on Fuzzy Data Transformation Methods

M. Naga lakshmi[1]
[1]*Research Scholar: Dept of Computer Science, S.P.M.V.V, Tirupati, Andhra Pradesh, INDIA*

K Sandhya Rani[2]
[2]*Professors: Dept of Computer Science, S.P.M.V.V, Tirupati, Andhra Pradesh, INDIA*

Abstract: Knowledge extraction process poses certain problems like accessing sensitive, personal or business information. Privacy invasion occurs owing to the abuse of personal information. Hence privacy issues are challenging concern of the data miners. Privacy preservation is a complex task as it ensures the privacy of individuals without losing the accuracy of data mining results. In this paper, fuzzy based data transformation methods are proposed for privacy preserving clustering in centralized database environment. In case one, a fuzzy data transformation method is proposed and various experiments are conducted by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular fuzzy membership function, Gaussian fuzzy membership function to transform the original dataset. In case two, a hybrid method is proposed as a combination of fuzzy data transformation approach specified in case one and Random Rotation Perturbation (RRP). The experimental results proved that the proposed hybrid method yields good results for all the member functions which are used in case one.

Keywords: Privacy Preservation, Clustering, Fuzzy Membership Function, Random Rotation Perturbation, Data Transformation.

I. INTRODUCTION

Data storage is growing at a phenomenal rate. Data users are interested to extract useful information from these large amounts of data. Data mining is becoming an increasingly important tool that can transform the data into useful knowledge. Data mining tasks provide accurate information for decision making. Some of the data mining tasks are classification, prediction, association rules and clustering. Clustering is a well known problem in statistics and engineering and widely used in various applications including biology, medicine, marketing etc. Clustering is the process of grouping the set of items by finding similarities between the data according to characteristics found in the data. There are two types of privacy concerns when mining the databases. Firstly, disclosure of personal or sensitive data existed in the databases. Secondly, sensitive patterns should not be disclosed when mining is performed on the shared data. New technologies are required for protecting sensitive information in electronic commerce [1]. Privacy preserving data mining is a new research area, which allows mining useful information while preserving privacy of individuals. Privacy issues are considering in two situations. They are centralized database environment and distributed environment. In centralized environment, database is available in single location. In this environment, privacy preserving data mining techniques are used to hide sensitive data of individuals. In distributed database environment, data is distributed to multiple sites. In this environment privacy preserving data mining techniques are applied for integrating the data from multiple sites, without revealing the privacy of individuals.

Fuzzy sets were introduced by zadeh in 1965 [2] to represent uncertainty, vagueness and provides formalized tools for dealing with the impression intrinsic to many problems. Fuzzy sets perform a gradual assessment of the input dataset by using fuzzy membership function. Fuzzy logic has been considered as an attractive method for data distortion which can reduce the information loss. In this paper two fuzzy data transformation methods are proposed for privacy preserving clustering in centralized database environment. In method one, a fuzzy data transformation method is proposed which uses fuzzy membership functions to transform the original dataset. In method two, a hybrid method is proposed as a combination of fuzzy data transformation approach specified in case one and Random Rotation Perturbation (RRP). Related works of privacy preserving clustering is discussed in the following section.

II. RELATED WORK

In [3], the authors conducted privacy surveys about general privacy, consumer privacy, medical privacy and created privacy indexes to summarize the results and discussed the trends in privacy. Privacy issues in Hippocratic databases are discussed and identify the technical challenges, problems in designing such databases and suggested some approaches that may lead to solutions in [4]. Authors in [5] addressed the problem of protecting the underlying attribute values when sharing the data for clustering and proposed a novel spatial transformation method called rotation based transformation to achieve privacy. Hybrid data transformation approach for privacy preserving clustering is presented in [6], by adopting geometric transformation methods to modify the sensitive numerical data using translation data perturbation, scaling data perturbation, rotation data perturbation, reflective data perturbation in centralized database

© 2013, IJARCSSE All Rights Reserved

environment. Double reflecting data perturbation and rotation data perturbation based hybrid data transformation approach for privacy preserving clustering is proposed in [7]. Privacy preserving clustering approach through cluster bulging has been presented by authors in [8]. The authors in [9] proposed random response method of geometric transformation for privacy preserving clustering in centralized database environment. In [10], a fuzzy based approach is proposed for privacy preserving clustering. The authors used fuzzy membership function to transform the original dataset in order to preserve the privacy of individuals. In [11], random rotation perturbation approach and framework of random rotation perturbation for privacy preserving classification is proposed. The authors also presented a multi-column privacy model to address the problems of evaluating privacy quality for multidimensional perturbation.

III. PROPOSED METHODS

In this section two fuzzy based methods are proposed for privacy preserving clustering. In method one, a fuzzy based data transformation approach is proposed and various experiments are conducted by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular membership function, Gaussian membership function to transform the original dataset. In method two, a hybrid method is proposed as a combination of fuzzy data transformation with various membership functions specified in case one and Random Rotation Perturbation (RRP). In another experiment novel additive perturbation approach is applied on original dataset to obtain the distorted dataset for comparison purpose.

A. Fuzzy based Data Transformation

Data distortion is the process of hiding sensitive data values without loss of information. A fuzzy transformation method distorts the sensitive numerical attributes using built in fuzzy membership functions such as Z-shaped fuzzy membership function (Zmf), Triangular fuzzy membership function (Trimf), Gaussian fuzzy membership function (Gaussmf). Table 1 shows the algorithm for fuzzy based data transformation method. The input to the fuzzy transformation method is a dataset D consists of sensitive attribute data in m rows and n columns. The input data D is transformed by initially suppressing the identifier attributes and distort the dataset using fuzzy membership function.

TABLE 1: ALGORITHM FOR FUZZY BASED DATA TRANSFORMATION METHOD

Input: (a) Dataset D consists of sensitive attribute data in m rows and n columns.

(b) Fuzzy membership functions such as Zmf, Trimf, and Gaussmf.

Output: Distorted Datasets and each D′ consist of m rows and n columns.

Begin

(1) Suppress the identifier attributes

(2) For each membership function (Zmf, Trimf, Gaussmf)

(3) For each sensitive attribute in D do

(4) Transform the attribute using selected fuzzy membership function.

(5) End For

(6) End For

(7) Release the all distorted datasets for clustering analysis.

End

B. Hybrid Method

A privacy preserving clustering technique is introduced in order to achieve the dual goal of privacy and utility. A hybrid method combines the strength of existing techniques and gives better results when compared to the single data perturbation method. This method consists of a combination of the two techniques namely fuzzy data perturbation and Random Rotation Perturbation (RRP). The important characteristic of RRP is preserving the geometric properties of the dataset. So the distorted dataset is clustered with similar accuracy when clustering is performed on original dataset. In this method, the original dataset is transformed using fuzzy data transformation method described in the table 1, which will be given as input for RRP method to obtain the final distorted dataset. The following table displays the algorithm for proposed hybrid method.

© 2013, IJARCSSE All Rights Reserved

TABLE 2: ALGORITHM FOR HYBRID METHOD

Input: (a) Original Dataset D consists of sensitive attribute data of size m x n.

(b) Fuzzy membership functions such as Zmf, Trimf, and Gaussmf.

Output: Distorted datasets and each D″ consist of size m x n.

Begin

(1) Suppress the identifier attributes.

(2) For each membership function (Zmf, Trimf, Gaussmf)

(3) For each sensitive attribute in D do

(4) Transform the attribute using fuzzy membership function.

(5) End For

(6) Generate an n × n rotation matrix R randomly.

(7) Obtain the final distorted dataset D″ = D × R.

(8) End For

(9) Release the distorted dataset D″ for clustering analysis.

End

C. Novel Additive Perturbation Technique

A review of novel additive perturbation technique [12] is given in this section for privacy preserving data mining. This technique is used to modify the given input dataset in order to hide the highly sensitive information. The additive data perturbation technique is designed for distributed environment where a data owner wants to transform the input data extracted from group of parties. The transformed data is used to perform the data mining operations such as clustering, classification. The following algorithm in Table 3 explains the additive data perturbation.

TABLE 3: ALGORITHM FOR NOVEL ADDITIVE DATA PERTURBATION

1) Data owner acquire the input data from multiple parties by giving queries

2) Identify the sensitive data items and perform additive data perturbation on the selected values by adding small amount of noise to protect the values of sensitive data items.

3) To enhance the privacy protection of additive data perturbation, perform swapping on the perturbed dataset obtained in step 2.

4) Release the final distorted dataset to perform data mining operations like classification, clustering.

Experimental results of the proposed method are discussed in the next section

IV. IMPLEMENTATION OF PROPOSED METHODS

The proposed methods are implemented to evaluate the performance of data distortion methods. Experiments are conducted for implementing the proposed fuzzy data transformation approach and hybrid method on three real life datasets obtained from UCI [13]. Wine data set with thirteen attributes and 178 records, Iris data set with four numerical attributes and 150 records, Credit-g dataset with 5 numerical attributes and 1000 instances are considered. Experiment one, a fuzzy data transformation method as given in Table 1 is conducted to transform the original dataset by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular membership function, Gaussian

© 2013, IJARCSSE All Rights Reserved

membership function. Experiment two, a hybrid method as given in Table 2 is conducted by combining fuzzy data transformation approach specified in experiment one and Random Rotation Perturbation (RRP).The well-known k-means clustering algorithm is used to measure the clustering quality. When the data is transformed, the clusters in the original dataset should be equal to the clusters in the distorted dataset. WEKA (Waikato Environment for Knowledge Analysis) software [14] is used to test clustering accuracy of the original and modified data base. The effectiveness is measured by misclassification error [5]. The misclassification error, denoted by M_E, is measured as follows.

$$M_E = \frac{1}{N}\sum_{i=1}^{k}(|Cluster_i(D)| - |Cluster_i(D')|)$$

In the above formula
N - Number of points in the original dataset.
K - Number of clusters.
$Cluster_i$ (D) - Number data points of the ith cluster of the original data set.
$Cluster_i$ (D′) - Number of data points of the ith cluster of the transformed dataset.

The following table shows the M_E values obtained for the proposed fuzzy data transformation method using Z-shaped fuzzy membership function (Zmf), Triangular membership function (Trimf), Gaussian membership function (Gaussmf) and novel additive method on three datasets. Higher M_E values indicates lower clustering quality where as lower M_E show the higher clustering quality. The experiments are conducted 10 times and M_E value is taken as an average of 10. The following table displays the misclassification error values of fuzzy data transformation approach.

TABLE 4: MISCLASSIFICATION ERROR RATES OF FUZZY DATA TRANSFORMATION

Data Distortion Methods	Iris	Wine	Credit-g
Zmf	0.09746	0.0892	0.1242
Trimf	0.15	0.17	0.2223
Gaussmf	0.1731	0.1785	0.2164
Novel Additive	0.196	0.199	0.2432

When comparing the misclassification error values in Table 4, it is confirmed that the proposed fuzzy data transformation methods which uses the three membership functions gives the lower misclassification error for all the three datasets. Among three membership functions Z-shaped fuzzy membership function gives the lower misclassification error.

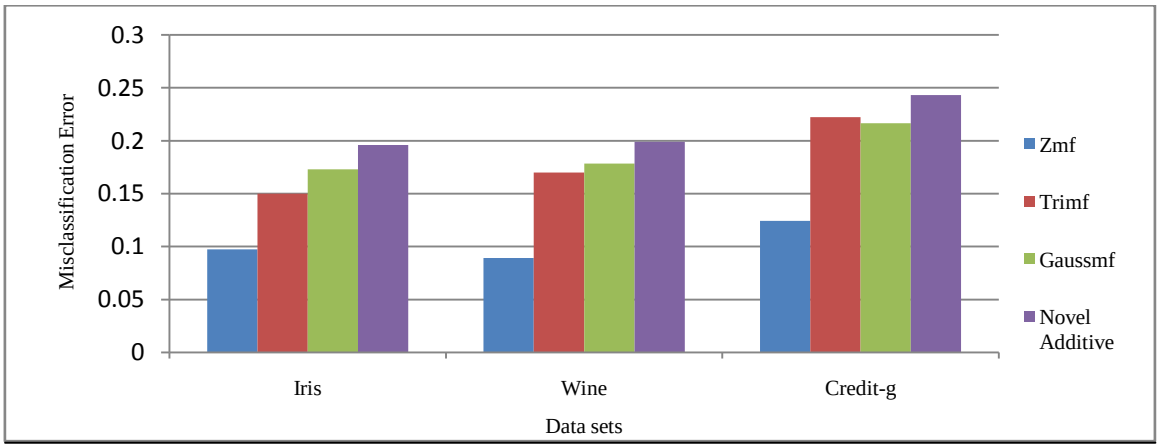


Fig. 1: Comparison of Misclassification Error Values Data Transformation Approaches

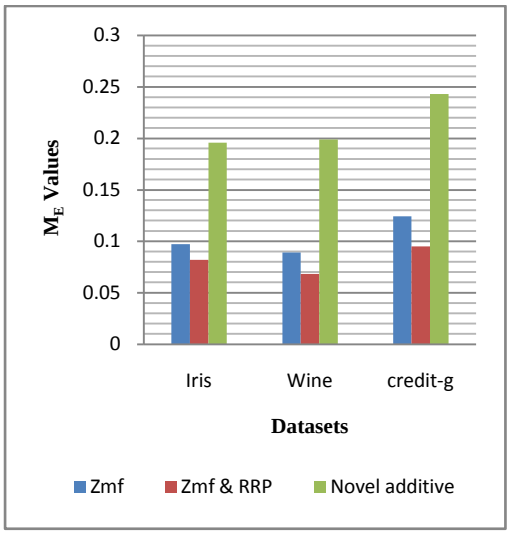
As the results shown in Fig. 1, it clearly indicates that for all the three datasets, the proposed fuzzy data transformation method gives the lower misclassification error for all the three member functions. These results proved that fuzzy data transformation methods gives higher utility than the novel additive data perturbation method. The following table shows the misclassification error values of the proposed hybrid method.

© 2013, IJARCSSE All Rights Reserved

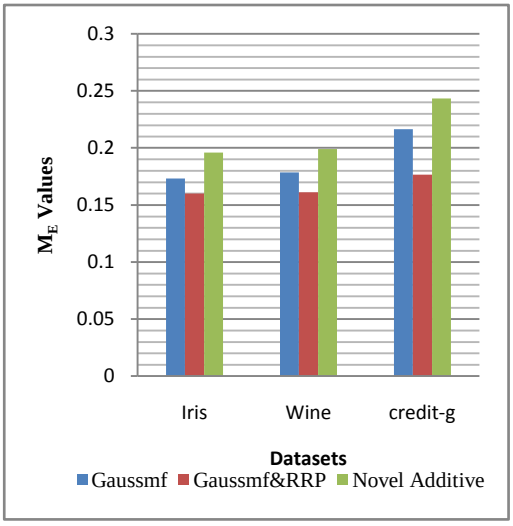
TABLE 5: MISCLASSIFICATION ERROR RATES OF HYBRID METHOD

Data Distortion Methods	Iris	Wine	credit-g
Zmf	0.09746	0.0892	0.1242
Hybrid method (Zmf & RRP)	0.08199	0.06853	0.0951
Trimf	0.15	0.17	0.2223
Hybrid method(Trimf & RRP)	0.1333	0.1477	0.2
Gaussmf	0.1731	0.1785	0.2164
Hybrid method (Gaussmf& RRP)	0.1599	0.1612	0.1765
Novel Additive	0.196	0.199	0.2432

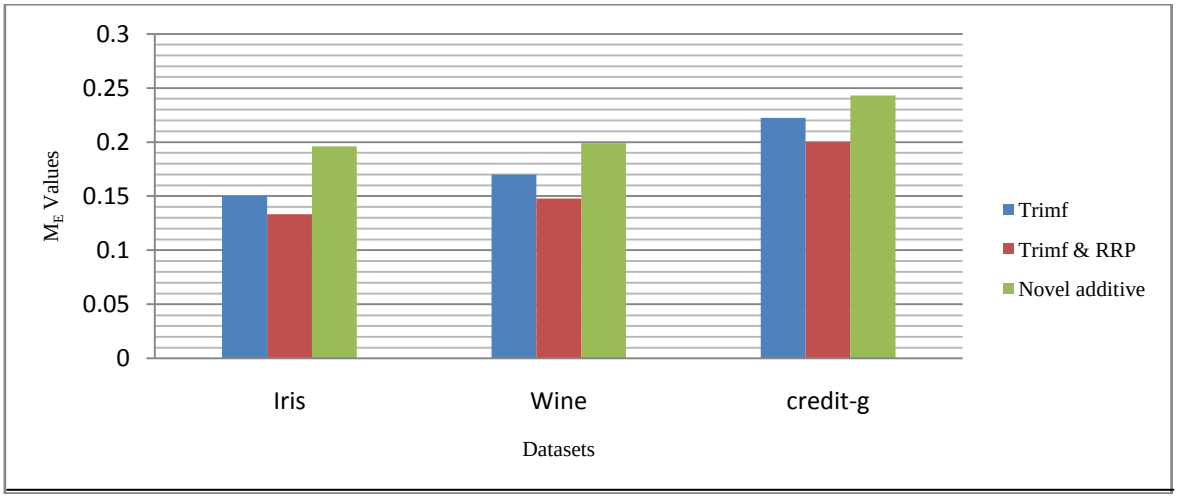
Table 5 illustrates the misclassification error values calculated for the proposed methods and novel additive data perturbation method. When comparing the misclassification error values of the fuzzy data transformation method with hybrid method, it clearly indicates that the proposed hybrid method provides lower misclassification error values for all the three datasets. Hence the hybrid method provides better clustering quality and the transformed dataset that is generated with the hybrid method looks very different from the original dataset, which preserves the privacy of individuals. The following graphical representation shows the effectiveness of the proposed methods.



(a)

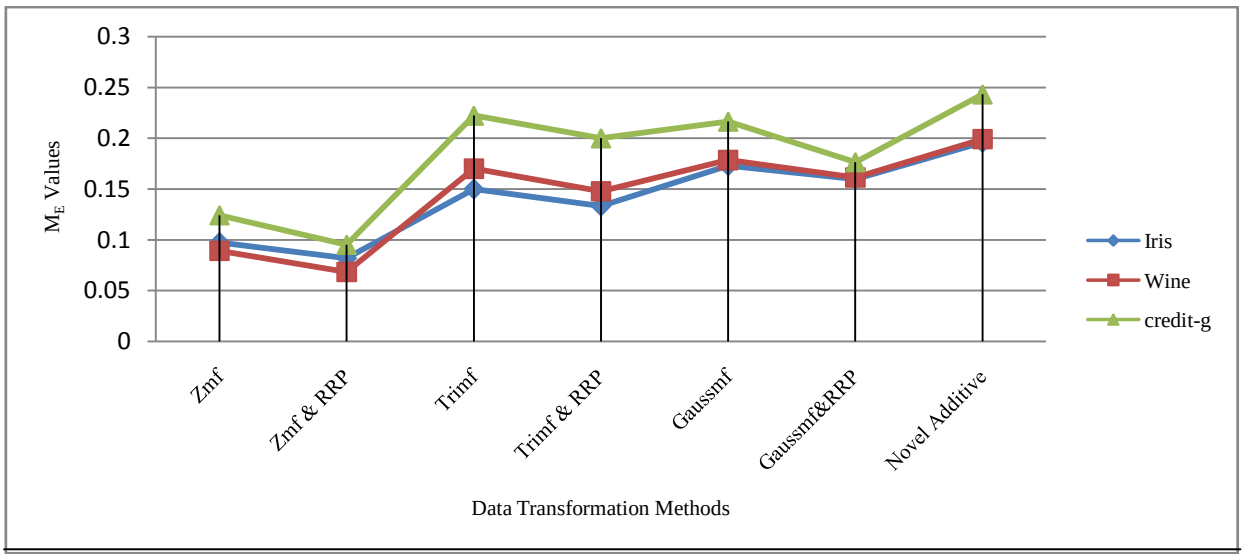


(b)



(c)

© 2013, IJARCSSE All Rights Reserved



(d)

Fig. 2: Comparison of Misclassification Error Values of fuzzy Data Transformation Methods

The misclassification error values of fuzzy data transformation method, hybrid method using Z-shaped fuzzy membership function and novel additive perturbation methods are depicted in Fig. 2(a). Among these methods hybrid method gives lower misclassification error. So the hybrid method effectively preserves the clustering accuracy and also provides privacy.

Fig. 2(b) shows the misclassification error values of the novel additive perturbation method and triangular membership function used in fuzzy data transformation method and hybrid method. Among these methods, hybrid method gives lower misclassification error, provides higher data utility when compared to the fuzzy data transformation method and novel additive method. The misclassification error values of fuzzy data transformation method, hybrid method based on Gaussian membership function and novel additive method are illustrated in Fig. 2(c). It clearly reveals that hybrid method yields lower misclassification error.

The misclassification error values of fuzzy data transformation methods and hybrid method for three membership functions such as Z-shaped, triangular, Gaussian on three datasets are shown n Fig. 2(d). It clearly indicates that the proposed hybrid method gives lower misclassification error for all the three member functions and for all the three datasets.

V. CONCLUSION

Privacy places a vital role for organizations when the data consists of sensitive information and which is shared among different users. The problem protecting individual privacy while releasing the data for clustering analysis is considered in this paper. Random rotation is one of the popular approaches for data perturbation and it can preserve privacy without affecting the accuracy for clustering analysis. Two methods are proposed in order to address this problem. Method one is a fuzzy based transformation approach that uses Z-shaped fuzzy membership function, Triangular membership function and Gaussian membership functions for data transformation. Experiments are conducted on three real life datasets from UCI and the results proved that the proposed method satisfying the privacy constraints as well as retains the clustering quality. To enhance the privacy preservation, method two which is a hybrid method is proposed by adopting the techniques fuzzy data transformation approach as specified in method one and random rotation perturbation. Experiments on three real life datasets reveal that, hybrid method is efficient for data utilization as well as privacy preservation.

REFERENCES

[1]. L. Korba, “Privacy in distributed electronic commerce”, in Proceedings of the 35 Hawaii International Conference on System Science, Hawaii, and January 2002.

[2]. Zadeh, “Fuzzy sets”, Information and control, vol 8, pp.338-353, 1965.

[3]. A.F.Westin, Freebies and privacy: what net users think Technical report, Opinion Research Corporation, July1999 Available from http://wwwprivacyexchange.org/iss/surveys/sr990714.htm?

[4]. R. Agrawal, J. Kiernan, R. Srikant and Y Xu, “Hippocratic Databases”, Proc of the 28th Int’l Conf. on Very Large Databases (VLDB’02), 2002, pp. 143-154.

[5]. S. R. M. Oliveira, and O. R.Zaiyane, “Achieving Privacy Preservation When Sharing Data for Clustering”, In Proceedings of the Workshop on Secure Data Management in a Connected World, in conjunction with VLDB’2004.Toronto, Ontario, Canada, pp. 67–82, 2004a.

[6]. S.R. M. Oliveira, O.R. Zaiyane (2003), “Privacy Preserving Clustering by Data Transformation”, in proceedings of 18th Brazilian Conference on Databases.

© 2013, IJARCSSE All Rights Reserved

[7]. Liming Li, Qishan Zhang, "A Privacy preserving Clustering Technique Using Hybrid Data Transformation Method", In proceedings of 2009IEEE international conference of grey systems and intelligent services.

[8]. Mohammad Ali Kadampur, D.V.L.N Somayajulu, S.S. Shivaji Dhiraj and Shailesh G.P. Satyam, "Privacy preserving clustering by cluster bulging for information sustenance", of the 4th International Conference on Information and Automation for Sustainability(ICIAfS'08), Colombo, Sri Lanka, December 2008.

[9]. Jie Liu,Yifeng XU,Harbin, "privacy preserving clustering by Random Response Method of Geometric Transformation", In proceedings of Fourth international conference on internet computing for science and engineering 2009.

[10]. B. Karthikeyan, G. Manikandan, V.Vaithiyanathan, A Fuzzy based approach for Privacy Preserving Clustering, Journal of Theoretical and Applied Information Technology, 31st October 2011. Vol. 32 No.2.

[11]. K.Chen, and L.Liu, A Random Rotation Perturbation Approach to Privacy Data Classification, In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp.589-592, 2005.

[12]. P. Kamakshi, A.Vinaya Babu, " A Novel Framework to Improve the Quality of Additive Perturbation Technique", In proceeding of International Journal of Computer Applications, Volume 30, No. 6, September 2011.

[13]. UCI Machine Learning Repository http://archive .ics.uci.edu/ml/datasets.html.

[14]. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

© 2013, IJARCSSE All Rights Reserved