ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

ANALYSIS OF CLUSTERING METHODS USING BICAT FOR PRIVACY PRESERVING IN BIGDATA APPLICATIONS

C.Surya[1], R.Kavitha[2]
M.E.(Mobile and Pervasive Computing), Anna University, Trichy, TamilNadu[1]
csurya714@gmail.com
Assistant Professor - CSE/IT Department, Anna University, Trichy, TamilNadu[2]
rkavitha.autt@gmail.com

ABSTRACT— Data mining is the process of exploration and analyzing large datasets in order to discover meaningful patterns and rules. With the help of data mining technologies, one may able to access any kind of information from the database. But the continuous development of data mining technologies brings serious threats to the sensitive information of the individuals. Hence, conserving the privacy of the sensitive attributes is a must before the data is being published. While publishing the data, releasing the aggregate information is usual, but not the individual information. Thus, the process of hiding an individual's sensitive information without sacrificing the usability of data is termed as Privacy Preserving Data Mining (PPDM). The main objective of the PPDM is to develop an algorithm for modifying the original data so that the privacy should remain even after the mining process. Therefore, from the perspective of the individuals, it is essential to develop a hybrid algorithm which undergoes protecting the sensitive attributes of the individuals. The classifier is employed to determine the attributes which is highly sensitive from a large dataset. Hence, this paper presents the experimental study on the performance of identifying the sensitive attributes with various classifiers and thereby to provide privacy while clustering using BicAT.

KEYWORDS— privacy preserving, sensitive attributes, classifiers, fuzzy clustering, BicAT.

I. INTRODUCTION

Data mining is defined as the process of extracting patterns, associations and relationships among the data involving the model creation and the concluded result will become knowledge. As a knowledge discovery process, it involves the data cleaning, data integration, data selection or reduction, data transformation, pattern discovery, pattern evaluation and knowledge presentation. It also involves the process of analyzing data from different perspectives and summarizing it into useful information.

The growing technologies of data mining involve serious threats to the sensitive information of the individuals. After mining process, one may able to reveal any data about an individual which falls into pitfall of data mining. Releasing the aggregate information is usual meanwhile the individual information is not. Hence, conserving the privacy of sensitive attributes is considered to be a must before publishing the data. Privacy can be achieved through several methods such as data hiding/masking, data suppression, generalization, anonymization, perturbation, encryption, randomization, condensation, fuzzification, secure multi-party computation, etc.

The primary goal of the privacy preserving data mining (PPDM) is to hide sensitive information before being published. It is the process of extracting relevant knowledge from large amounts of data while at the same time protecting the sensitive attributes. Data privacy is different from data security whereas the data privacy is suitably defined as the appropriate use of data and the data security ensures whether the data are accurate and reliable or not. Also the privacy of the attributes should not affect the utility of the attributes. Mostly, the privacy and utility are fundamentally in tension with each other. We can achieve perfect privacy by not releasing any data, but this solution has no utility. Also, preserving should be done without comprising the usability of the data.

There are number of methods used for preserving the privacy of the sensitive attributes while clustering and classification. Some of the methods are cryptographic algorithms, noise addition and data swapping. All of these methods bring a bit of complexity in the algorithm and increase the processing time. Our main is to reduce this processing time and at the same time to provide an optimum solution to the problem of privacy preserving. Therefore, the fuzzy logic is applied to preserve the individual's information while revealing the details in public. The first thing should have to undergo is identifying the sensitive attributes and thereby to provide privacy on fuzzy clustering or bi-clustering.

The rest of the paper deals with the literature review, related works, experimental setup, experimental procedure, results and discussion and conclusion of this study.

II. LITERATURE REVIEW

This section reviews the literatures on the various privacy preserving methods that are carried out by different researches. Dimitrios Karapiperis and Vassilios S.Verykios approached the LSH (Locality-Sensitive Hashing) technique for identifying

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

candidate record pairs which have undergone an anonymization transformation. When data to be matched is deemed to be sensitive, PPRL (Privacy-Preserving Record Linkage) techniques should be employed. No sensitive information in a record should be disclosed to parties other than the owner [1]. Gaofeng Zhang, Xiao Liu and Yun Yang carried out noise obfuscation by utilizing the noise data. Noise service requests can be generated and injected into real customer service requests so that the malicious service providers would not able to distinguish which requests are real ones. In the process of noise generation, noise generation probabilities determine which kinds of noise requests should be generated and the noise injection intensity decides how many noise requests should be generated [2]. Wanchun Dou, Xuyun Zhang, Jianxun Liu and Jinjun Chen approached the HireSome-II concept in order to enhance the credibility of a composition plan. The evaluation of a service is promoted by some of its QoS history records, rather than its advertised QoS values. HireSome-II can protect the cloud privacy, as a cloud is not required to unveil all its transaction records [3].

Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan and Yong Ren investigated various approaches that can help to protect sensitive information. It intimates the privacy concerns about different levels of users in data mining application and develops corresponding approaches for privacy protection. The data provider can realize a perfect protection of their privacy by revealing no sensitive data to others, but this may kill the functionality of data mining [4]. Shweta Taneja, Shashank Khanna, Sugandha Tilwalia and Ankita proposed a novel C-Tree algorithm in which it rearranges the records of the dataset and perturb the primary attribute. This algorithm is implemented and tested on a micro data in which the sensitive data is perturbed in an efficient way which will never reveal anyone's identity [5]. Syed Md. Tarique Ahmad, Shameemul Haque and SM Faizanut Tauhid approached the fuzzy based data transformation methods in which the k-means clustering algorithm is applied on the modified data and it is found that the relativity of the data is also maintained. Fuzzy sets perform a gradual assessment of the input dataset by using fuzzy membership function [6]. Sonali M.Khairnar and Sanchika Bajpai presented an algorithm which forms the cluster and performs computation sum after checking information loss [7].

Slava Kisilevich, Lior Rokach, Yuval Elovici and Bracha Shapira proposed an algorithm in which the values are suppressed only on certain records depending on other attribute values, without the need for manually-produced domain hierarchy trees. It guarantees that the probability of identifying an individual based on the released data in the dataset does not exceed 1/k [8]. S.SelvaRathna and T.Karthikeyan proposed a method to hide fuzzy association rule using modified apriori algorithm in order to identify sensitive rules to be hidden. It provides a secure framework for privacy preserving in both vertically and horizontally distributed co-occurrence matrices. The data disclosure probability and information loss are possibly kept negligible [9]. R.Natarajan, R.Sugumar, M.Mahendran and K.Anbazhagan developed an efficient algorithm in order to provide confidentiality and improved the performance at the time when the database stores and retrieves huge amount of data. This algorithm describes the results of a blocking algorithm which reduces loss of data and minimizes the undesirable side effects by selecting the items in the appropriate transactions to change and maximize the desirable side effects [10].

III. EXPERIMENTAL SETUP AND PROCEDURE

A. *Analysis of Dataset*

The experimental setup is carried out first in order to identify the sensitive attributes. From a large dataset of the big data applications, identifying the attributes with higher sensitivity is the important one. In general, the information gain of the attributes from a dataset is calculated by using the formula as follows:

$$\text{Info}(D) = -\sum P_i \log_2 P_i \qquad (1)$$

where P_i is the probability of the class in D.

$$\text{Info}_A(D) = \sum \frac{|D_j|}{|D|} \times \text{Info}(D_j) \qquad (2)$$

And hence the gain is formulated as,

$$\text{Gain} = \text{Info}(D) - \text{Info}_A(D) \qquad (3)$$

B. *Identifying the sensitive attributes*

Thereby various kinds of classifiers are employed in order to find out the attributes with higher sensitivity. Such classifiers are Radial Basis Function Network (RBFN), Functional Trees (FT), Fuzzy Lattice Reasoning (FLR) and Multilayer Perceptron. Firstly, the attribute evaluator and search method is employed as InfoGainAtrributeEval and Ranker respectively. After selecting

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

the attributes, choose the attribute's sensitive priority order. The classifier radial basis function network (RBFN) is employed and the classified instances are tabulated for both the sensitive (prime) and non sensitive (non-prime) attributes. The same procedure is carried out for different datasets with different classifiers such as trees-FT, MISC-FLR and multi-layer perceptron and hence tabulated as follows:

TABLE I. IDENTIFYING SENSITIVE ATTRIBUTES WITH RBFN CLASSIFIER

Data sets	Information Gain – RBFN classifier	
	With Prime Attributes (Sensitive)	*Without Prime Attributes (Non-Sensitive)*
Diabetes	76.43	66.66
Iris	94.66	78
Breast Cancer	73.42	70.62
Cars	83.97	69.69
Wine	48.8	47.37
Heart Statlog	77.77	77.4
Labor	91.22	89.47
Soybean	77.89	74.56
Sick	96.12	93.63
Vote	94.94	86.2

TABLE II. IDENTIFYING SENSITIVE ATTRIBUTES WITH FUNCTIONAL TREES (FT) CLASSIFIER

Data sets	Information Gain – TREES_FT	
	With Prime Attributes (Sensitive)	*Without Prime Attributes (Non-Sensitive)*
Diabetes	74.34	69.14
Iris	96	81.33
Breast Cancer	71.32	70.27
Cars	84.36	70.07
Wine	49.14	44.31
Heart Statlog	85.55	75.18
Labor	85.96	79.47
Soybean	75.84	67.98
Sick	64.18	56.78
Vote	95.17	85.97

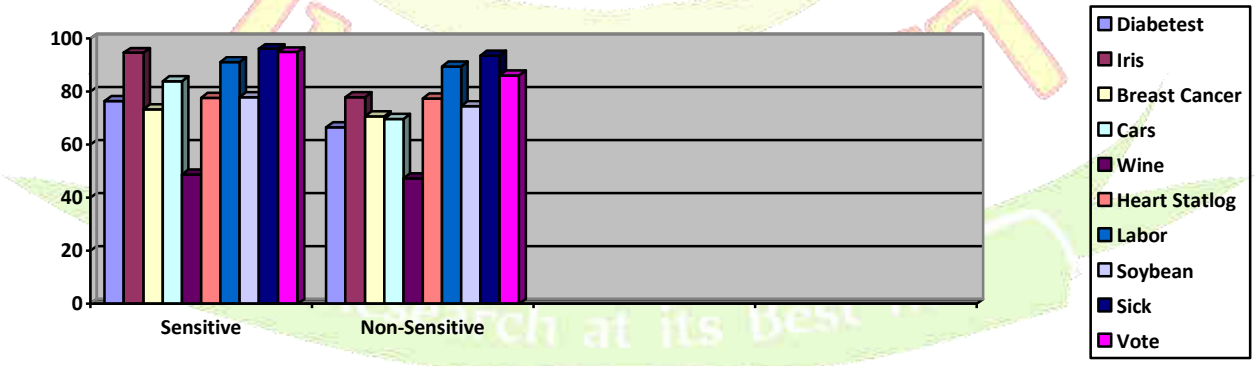

Fig.1. Sensitivity of various datasets using RBFN classifier

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

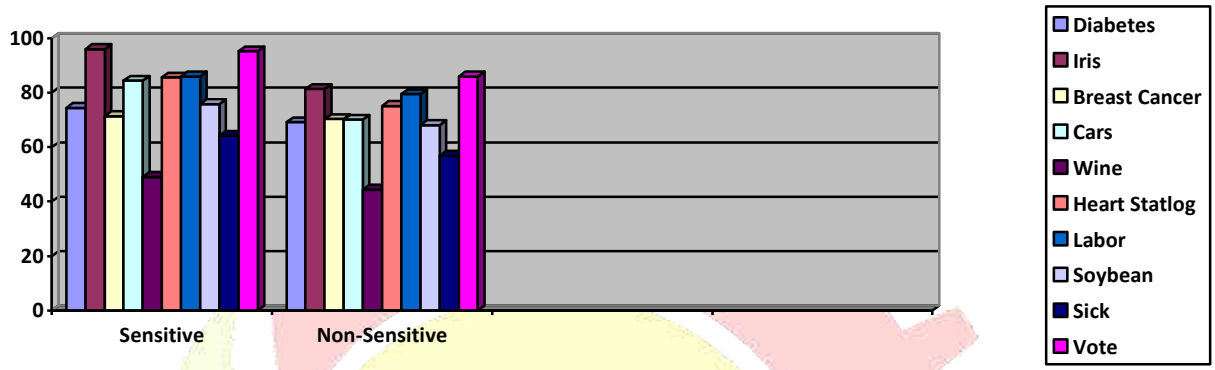

Fig.2. Sensitivity of various datasets using FT classifier

The Functional Trees (FT) classifier is the classification tree that could have the logistic regression function and this algorithm can deals with the binary variables, numeric and nominal attributes and missing values.

TABLE III. IDENTIFYING SENSITIVE ATTRIBUTES WITH FLR (FUZZY LATTICE REASONING) CLASSIFIER

Data sets	Information Gain – MISC.FLR	
	With Prime Attributes (Sensitive)	*Without Prime Attributes (Non-Sensitive)*
Diabetes	53.38	53.12
Iris	93.33	62
Breast Cancer	81.83	65.09
Cars	86.98	79.23
Wine	34.03	32.92
Heart Statlog	54.81	53.33
Labor	83.56	79.34
Soybean	73.67	68.34
Sick	63.9	60.12
Vote	90.83	88.55

TABLE IV. IDENTIFYING SENSITIVE ATTRIBUTES WITH MULTI-LAYER PERCEPTRON

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

Data sets	Information Gain – Multi-layer Perceptron	
	With Prime Attributes (Sensitive)	*Without Prime Attributes (Non-Sensitive)*
Diabetes	76.43	68.22
Iris	95.33	79.33
Breast Cancer	69.23	71.67
Cars	78.18	60.81
Wine	52.62	46.15
Heart Statlog	82.59	67.4
Labor	85.96	81.22
Soybean	77.3	70.99
Sick	64.76	61.33
Vote	94.71	82.98

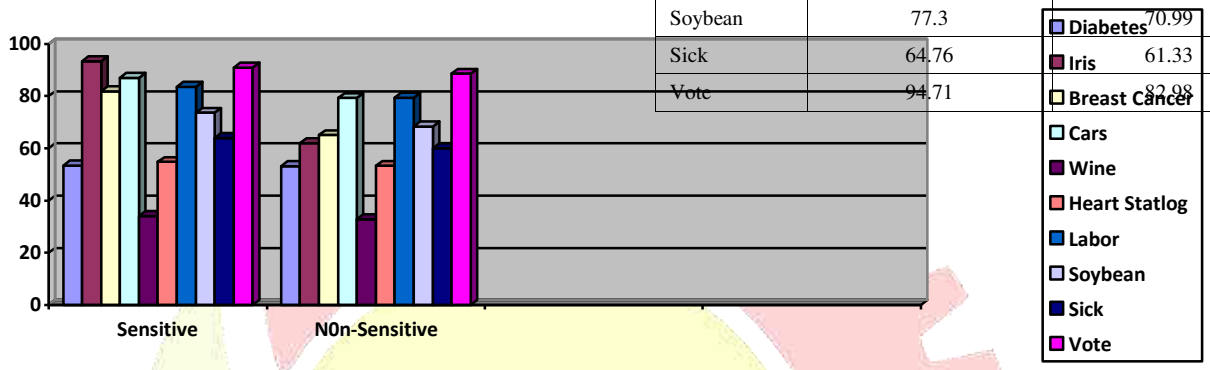

Fig.3. Sensitivity of various datasets using MISC-FLR classifier

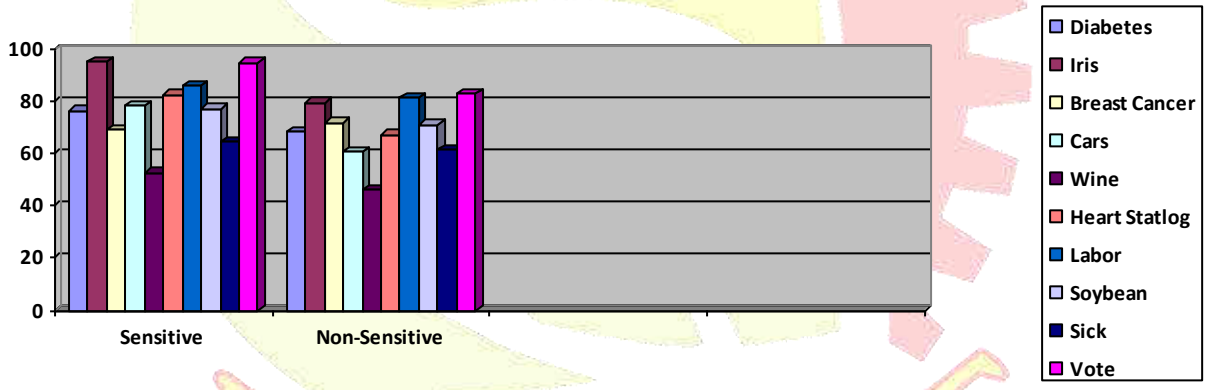

Fig.4. Sensitivity of various datasets using Multi-layer Perceptron

C. Generation of Fuzzy rules

The fuzzy rules are a collection of linguistic statements. It describes how a fuzzy inference system should make a decision regarding classifying an input. They combine two or more input fuzzy sets and associate with them an output. Fuzzy rules always written in the following form:

IF *v1* is *A1* and *v2* is *A2* and ... *vn* is *An*

THEN (*v1, v2,...,vn*) belongs to class *w*.

where *A1,A2,...,An* are input fuzzy sets and *w* is output fuzzy set. For example, one could make up a rule that says:

IF temperature is high and humidity is high **THEN** room is hot.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

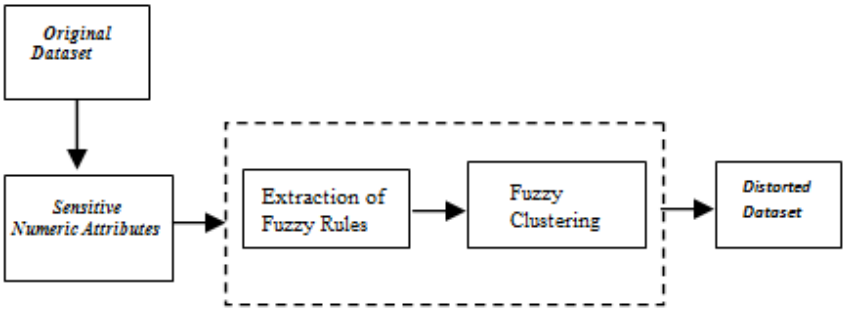

Fig.5. Fuzzy Based Approach

D. Fuzzy based Clustering

The fuzzy based approach distorts the sensitive numerical attributes using built-in fuzzy membership function. This algorithm is useful when the required number of clusters is pre-determined. It also shows that the data privacy can be maintained without compromising the accuracy of the result if features are transformed into fuzzy sets.

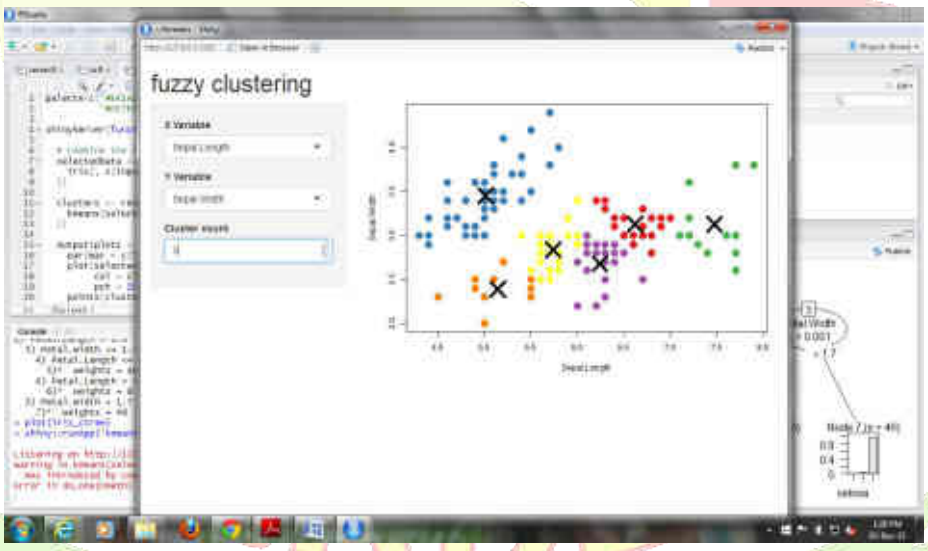

Fig.6. Fuzzy based clustering approach

E. BicAT- analysis of biclustering methods

The BicAT provides not only a framework for clustering-based data analysis that integrates various bi-clustering and clustering techniques in terms of a common graphical user interface but also the different facilities for data preparation, inspection and post-processing such as discretization, filtering of bi-clusters according to specific criteria or gene pair for constructing gene interconnection graphs. The possibility to use different bi-clustering algorithms inside a single graphical tool allows the user to compare clustering results and choose the algorithm that best fits a specific biological scenario.

IV. CONCLUSION AND FUTURE ENHANCEMENT

A. Conclusion

Due to the growing popularity of the data mining technologies, it has the possibilities to bring serious threats to the sensitive information even without the knowledge of the individuals. Hence, it is essential to develop an algorithm in order to protect the individual's sensitive data without compromising the utility of the data. Even though many algorithms have been developed to fix

this problem, the malicious service providers are still able to deuce the privacy of the individuals. To address this issue, the fuzzy based approach is carried out, by the way, the sensitive data of the individuals has been considered to be protected and it makes the third party to do statistical computation in case of any data disclosure. Also, the fuzzy clustering approach is used, in order to undergo privacy protection of the sensitive data of the individuals.

B. Future Enhancement

For an ordinary dataset, the proposed fuzzy clustering approach provides enough privacy to the sensitive attributes. While providing privacy to the sensitive data from the large datasets, it is possible to have lesser information loss. Hence, it is considered to develop an efficient algorithm in order to provide privacy based on the utility of the sensitive data.

C. Research Directions

There are several future research directions along the way of quantifying a PPDM algorithm and it's underneath application or data mining task. One is to develop a comprehensive framework according to which various PPDM algorithms can be evaluated and compared. It is also important to design good metrics that can be better reflect the properties of a PPDM algorithm and to develop benchmark databases for testing all types of PPDM algorithms.

References

[1] Dimitrios Karapiperis and Vassilios S. Verykios, "An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage", IEEE Transactions on Knowledge and Data Engineering, Vol.27, No. 4, April 2015.

[2] Gaofeng Zhang, Xiao Liu, and Yun Yang, "Time-Series Pattern Based Effective Noise Generation for Privacy Protection on Cloud", IEEE Transactions on Computers, Vol. 64, No. 5, May 2015.

[3] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan and Yong Ren, "Information Security in Big Data: Privacy and Data Mining", IEEE Access: Digital Object Identifier 10.1109/ACCESS.2014.2362522.

[4] R.Natarajan, R.Sugumar, M.Mahendran and K.Anbazhagan, "Design and Implement an Association Rule Hiding Algorithm for Privacy Preserving Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 7, September 2012.

[5] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia and Ankita, "A Hybrid C-Tree Algorithm for Privacy Preseving Data Mining", International Journal of Soft Computing and Engineering(IJSCE).

[6] Sonali M. Khairnar, Sanchika Bajpai, "Anonymization of Centralized and Distributed Social Networks by Incremental Clustering", International Journal of Computer Science and Information Technologies, Vol. 5(5), 2014, 6724-6727.

[7] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, "Efficient Multi-dimensional Suppression for k-anonymity", in Press.

[8] Syed Md.Tarique Ahmad, Shameemul Haque, SM Faizanut Tauhid, "A Fuzzy Based Approach for Privacy Preserving Clustering", International Journal of Scientific & Research, Volume 5, Issue 2, 2014, 1067 ISSN 2229-5518.

[9] S.Selva Rathna, T. Karthikeyan, "Survey on Recent Algorithms for Privacy Preserving Data Mining", International Journal of Computer Science and Information Technologies, Vol. 6(2), 2015, 1835-1840.

[10] Wanchun Dou, Xuyun Zhang, Jianxun Liu and Jinjun Chen, "HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications", IEEE Transactions on Parallel and Distributed Systems, Vol. 26, No. 2, February 2015.

[11] Sandya H.B., Hemanth Kumar P., Himanshi Bhudiraja, Susham K. Rao, "Fuzzy Rule Based Feature Extraction and Classification of Time Series Signal", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-2, May 2013.

[12] Monali Dey, Siddharth Swarup Rautaray, "Study and Analysis of Data Mining Algorithms for Healthcare Decision Support System", International Journal of Computer Science and Information Technologies(IJCSIT) ISSN 0975-9646, Vol. 5(1), 2014, 470-477.

[13] Manider Singh, "A Review on Data Mining Algorithms", International Journal of Computer Science and Information Technology Research ISSN 2348-1196, Vol. 2, Issue 2, pp: (8-14), Month: April-June 2014.

[14] Tejaswini Pawar, Snehal Kamalapur, "A Survey on Privacy Preserving Decision Tree Classifier" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 6, November-December 2012, pp. 843-847.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

[15] Yinghua Lu, Tinghuai Ma, Changhong Yin, Xiaoyu Xie, Wie Tian and ShuiMing Zhong, "Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data", International Journal of Database Theory and Application (IJDTA) ISSN: 2005-4270 Vol. 6, No. 6(2013), pp.1-18.

[16] K. Vinay Kumar, N.Sandeep Kumar and S.Vishnu Vardhan, "Privacy Protection for Dynamic Data through Anonymization", International Journal of Scientific and Research Publications, Volume 2, Issue 9, September 2012 ISSN: 2250-3153.