I.J. Computer Network and Information Security, **2014, 3, 52-58**
Published Online February 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2014.03.07

Classification via Clustering for Anonym zed Data

Sridhar Mandapati
Associate Professor, Dept. of Computer Applications, R.V.R & J.C College of Engineering, Guntur, India.
Email:mandapati_s@yahoo.com

Dr. Raveendra Babu Bhogapathi
Professor, Dept. of Computer Science & Engineering, VNR VJIET, Hyderabad, India.
Email:rbhogapathi@yahoo.com

Dr. M.V.P.C.Sekhara Rao
Professor, Dept. of Computer Science & Engineering, R.V.R & J.C College of Engineering, Guntur, India.
Email:manukondach@gmail.com

Abstract — Due to the exponential growth of hardware technology particularly in the field of electronic data storage media and processing such data, has raised serious issues related in order to protect the individual privacy like ethical, philosophical and legal. Data mining techniques are employed to ensure the privacy. Privacy Preserving Data Mining (PPDM) techniques aim at protecting the sensitive data and mining results. In this study, the different Clustering techniques via classification with and without anonym zed data using mining tool WEKA is presented. The aim of this study is to investigate the performance of different clustering methods for the diabetic data set and to compare the efficiency of privacy preserving mining. The accuracy of classification via clustering is evaluated using K-means, Expectation-Maximization (EM) and Density based clustering methods.

Index Terms — Privacy Preserving Data Mining (PPDM), Classification, Clustering, K-means, EM, Density based.

I. INTRODUCTION

Now-a-days a vast amount of people data is collected by the corporate, individuals and government. It is resulting in a gigantic amount of sensitive records that describe people's financial interests, health issues, activities and demographics [1]. These records often violate the privacy of individuals if they are published. By using the data mining tools, the important information of individuals can be extracted and could be used for different purposes. But, there is a serious need to preserve individual's confidential and sensitive information.

To preserve the privacy of individuals, many techniques were proposed in the literature [2, 3, 4, 5, 6, 7, and 8]. During the data mining process methods of data transformation are used to maintain the privacy of the data and the data is anonym zed to preserve privacy. These methods reduce the representation of the granularity for privacy. This leads to information loss in the data or algorithms' effectiveness and a trade-off between information loss and privacy. *K*-anonymity is a popular anonymization approach [4, 5, 6, and 8]. A set of data comply with protection of *k*-anonymity if each individuals record stored in the released data set can't be distinguishable from at least *k*-1 individuals, whose data also appear in the data set. The probability of guarantying an individual protection based on the released data in the data set does not exceed 1/*k*.

The common methods for de-identification of the data in *k*-anonymity algorithms are Generalization and Suppression [5, 6]. The major drawback of *k*-anonymity, recognized by several authors, is that they can't prevent attribute disclosure. To overcome this problem, *l*-diversity is defined [2]. The method requires that each equivalence class has at least *l* well-represented values for each sensitive attribute. While *l*-diversity protects against sensitive attribute disclosure, it has limitations [7, 9] like – Skegness and Similarity. To overcome this problem, closeness and slicing are defined [7, 10]. *T-closeness* method requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. While slicing performs the partitioning of the data both vertically and horizontally this achieves better data utility than the generalization, and can be used for membership disclosure protection. The main advantage of slicing is to protect membership disclosure, and it can handle high-dimensional data. To overcome the drawback of Generalization another important method ANGEL that requires an effective generalization in privacy protection is defined [11]. It is applicable to any monotonic principles (e.g. *l*-diversity, *t*-closeness etc.); with its superiority (in correlation preservation) especially obvious when tight privacy control must be enforced.

Achieving privacy using clustering techniques is not intensively studied as the other data mining techniques [12]. This research investigates the accuracy of

Copyright © 2014 MECS

classification via clustering of the data with and without anonymization and the efficiency of privacy preserving mining to extract the desired patterns from the dataset is studied. The accuracy of classification via clustering is evaluated using *K*-means, EM, and Density based clustering methods. The rest of the study is organized as follows: a theoretical based work is presented in section 2, the methodology is outlined in section 3, section 4 gives the experimental results and conclusion and future research are outlined in section 5.

II. RELATED WORKS

Aggarawal et al [13] proposed a new method of data record anonymization, where data records of quasi-identifiers are first clustered, and then cluster centres are published. To ensure the privacy of the data records, ensure that each cluster must contain no less than a pre-specified number of data records. Compared with the *k*-anonymity this method has much more general choices for centres of clusters. In many situations, it releases further information without compromising privacy constraint. Clustering is done with a constant-factor approximation algorithm. This is the new set of anonymization algorithms where performance is independent of the anonymity parameter *k.* The extended algorithm remains unclustered to allow point of fraction due to few outlier points which thus increases the cost of anonymization. To ensure the data publishing for accurate analysis and to be more effective, a new approximation algorithm for new clustering objectives was proposed. It could be applicable in the other clustering scenarios also.

Jiuyong Li et al [14] proposed to achieve *k*-anonymity by Clustering in Attribute Hierarchical Structures. In this, *k*-anonymity is viewed as clustering with a constraint of the minimum number of objects in every cluster. Generalization is defined as the distance between tuples, to characterise distortions and properties of distance. Ji-Won Byun et al [15] proposed an important requirement to minimizing information loss for ensuring the data anonymization. For this clustering techniques were proposed, and it was observed that data records which are naturally similar to each other should be part of the same equivalence class. The clustering problem is referred as the k-member clustering problem and it is proved that the problem is NP-hard and a greedy heuristic, the complexity of which is in O (n^2) is presented. Thus, a suitable metric to measure the loss of information for the anonymized data which works for both categorical and numeric data was proposed.

Aris Gkoulalas-Divanis et al [16] proposed a novel clustering based framework to anonymize the transactional data. This provides a basis for designing algorithm that allows publishing data with less information loss and can persuade a wide range of privacy requirements. Based on this, they developed PCTA, a generalization based algorithm to construct anonymizations that incur a small amount of information loss underneath many different privacy requests. Experiments with datasets confirm that PCTA significantly performs well than the current state-of-the-art algorithms in data utility and efficiency.

M. Ercan Nergiz et al [17] discussed thoughts on *k*-anonymization. *K*-Anonymity provides privacy protection by making sure that data cannot be found of an individual. Any identifying information, in a *k*-anonymous dataset, occurs in *k* tuples. Recently, varied kinds of algorithms with different assumptions and restrictions have been proposed, to achieve optimal and practical *k*-anonymity, with diverse metrics to measure quality. To assess, a family of clustering-based algorithms which are more flexible and which attempts to improve precision by disregarding the restrictions of user-defined Domain Generalization Hierarchies is proposed. The evaluation approaches with respect to cost metrics show that metrics may behave differently with different algorithms and may not correlate with some applications' accuracy on output data.

Lopez M.I et al [18] proposed a classification via clustering approach to calculate the final marks in a university course basing on forum data. The aim is twofold: to find out if student involvement in the course forum can be an interpreter of the final marks for the course and also to examine whether the classification via clustering approach can get similar accuracy to traditional classification algorithms. The experiments were supported by using real data from first-year university students. Numerous clustering algorithms using this approach were compared with other classification algorithms in guessing whether students pass or fail the course basing on their Model forum usage data. Results demonstrate that the Expectation-Maximisation (EM) clustering algorithm produces results comparable to those of the best classification algorithms, particularly when using only a set of selected attributes.

Mrutyunjaya Panda et al [19] proposed a novel classification via clustering method for anomaly based network intrusion detection system. They presented a methodology to recognize the attacks during the regular activities in a system. Also they proposed a novel classification via sequential Information Bottleneck (sIB) clustering algorithm for an effective anomaly based network intrusion detection model.

III. MATERIALS AND METHODS

(A) Dataset:

A total of 9 attributes with 768 tuples is taken from the Pima Indians Diabetes Database. A tested positive (1) or tested negative (0) class label is provided to each tuple. The number of tested positives and tested negatives includes 268 and 500 respectively. The attributes of dataset are all of numeric type. The attributes are 'preg – number of times pregnant', 'plas – plasma glucose concentration a 2 hours in an oral glucose tolerance test', 'pres – diastolic blood pressure', 'skin – triceps skin fold thickness', 'insu – 2-hour serum insulin', 'mass – body mass index', 'pedi – diabetes

Copyright © 2014 MECS

pedigree function' , 'age – age' and 'class – class variable (0 or 1)'. The age attribute of diabetes dataset is anonymized using the principle of *k*-anonymization. Table 1 and 2 shows the original and the transformed attribute data. The dataset is classified using 10 fold cross validation the original and the *k*-anonymized dataset.

TABLE 1: Original diabetes dataset

preg	age	class
6	50	tested_positive
1	31	tested_negative
8	32	tested_positive
1	21	tested_negative
0	33	tested_positive
5	30	tested_negative
3	26	tested_positive
10	29	tested_negative
2	53	tested_positive
8	54	tested_positive
4	30	tested_negative
10	34	tested_positive
10	57	tested_negative
1	59	tested_positive
5	51	tested_positive

TABLE 2: The k-Anonymous diabetes dataset

preg	age	class
6	OLD	tested_positive
1	Middle	tested_negative
8	Middle	tested_positive
1	Young	tested_negative
0	Middle	tested_positive
5	Middle	tested_negative
3	Young	tested_positive
10	Young	tested_negative
2	OLD	tested_positive
8	OLD	tested_positive
4	Middle	tested_negative
10	Middle	tested_positive
10	OLD	tested_negative
1	OLD	tested_positive
5	OLD	tested_positive

(B) K-means Clustering:

It is one of the simple and quick unsupervised learning algorithms that solve the well-known clustering problem adapted to many problem domains. In order to form the cluster, a similarity (or distance) measurement must be established first. In the literature various similarity measurements exist, one of the most commonly used metrics for clustering problems is the Euclidean distance [20, 21]. A small distance between two objects implies a strong similarity whereas a large distance implies a low similarity. In an n-dimensional space of features, Euclidean distance can be calculated between objects' x' and 'y' as follows:

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

with 'n' being the number of features in each object. The K-Means algorithm partitions objects from a data set into a number of K disjoint subsets. For each cluster, the partitioning algorithm maximizes the homogeneity within the cluster by minimizing the square-error. The formula for the square error is:

$$S(E) = \sum_{i=1}^{K}\sum_{j=1}^{n} \left| dist(x_j - c_i) \right|^2 \quad (2)$$

The square error is estimated as the distance squared between each object 'x' and the centre (or mean) of its cluster. Object 'c' represents the respective centre of each cluster.

The square error is minimized by K-Means using the following algorithm. The centres of the K clusters are initially chosen randomly from within the subspace. The objects in the data set are then partitioned into the nearest cluster. K-Means iteratively computes the new centres of the clusters that are formed and then repartitions them based on the new centres. The K-Means algorithm continues this process until the membership within the clusters stabilizes, thus producing the final partitioning.

(C) EM Algorithm:

The Expectation-Maximization (EM) Algorithm is an iterative algorithm that starts with an initial estimation for 'θ' and iteratively modifies the Maximum Likelihood (ML) of the observed data [20, 22]. This algorithm works best in a situation where the data is incomplete or can be thought of as incomplete. It is typically used with the mixture of models like Gaussian mixtures. The following is the procedure for the EM:

1. **Initialization step**: initialize the algorithm with a guess θ^0 i.e.,
$$\theta_k^0 = (\mu_1^0, \mu_2^0, \ldots, \mu_k^0) = \mu_k^0 \quad (3)$$
Where 'k' is the Gaussian mixture of the current component, θ^0 is the initial estimation and 'μ' is the mean.
2. **Expectation step**: Estimate the expected values of the hidden variable Z_{ij} using the current hypothesis
$$\theta^t = (\mu_1^t, \mu_2^t, .., \mu_k^t)$$

Copyright © 2014 MECS

$$E(z_{ik}) = \frac{exp\left[-\frac{\left(x_i-\mu_k^t\right)^2}{2\sigma^2}\right]}{\sum_{j=1}^{k}\exp[-\frac{\left(x_i-\mu_j^t\right)^2}{2\sigma^2}]} \quad (4)$$

where 't' is the no.of iterations, $E(z_{ik})$ is the expected value for the hidden variable, 'k' is the dimension, 'σ' is the standard deviation.

3. **Maximization step**: Update the hypothesis $\theta^{t+1} = (\mu_1^{t+1}, \mu_2^{t+1}, .., \mu_k^{t+1})$ using from step 2.

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} E(z_{ik})x_i}{\sum_{i=1}^{n} E(z_{ik})} \quad (5)$$

4. **Convergence step**: if $\|\theta^{t+1} - \theta^t\| < \varepsilon$, otherwise, go to step 2.

The parameters of the model are the hidden variables. The mean and standard deviation of Gaussian mixtures are the hidden variables in this case.

(D) DBSCAN Algorithm:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is generally used as density based algorithm [2]. It utilizes the concept of density connectivity and density reachability. Density based clustering algorithm has played a vital role in finding nonlinear shapes structure based on the density, and it is good for the outlier analysis. The following is the procedure for the DBSCAN algorithms.

Algorithm:

Let X=**{**x_1, x_2,...., x_n} be the set of data points. Basically, two parameters are required to form the clusters in this algorithm. They are: 'ε' (Eps) and minimum number of points (MinPts).

1) Start with an arbitrary starting point that has not been visited.
2) Extract the neighbourhood of this point using 'ε' (All points which are within the 'ε' distance are neighbourhood).
3) If there are sufficient neighborhoods about this point then clustering process is started, and the point is noted as visited or else the point is labelled as noise (Later this point can become a part of the cluster).
4) When a point is set up to be a part of the cluster then its 'ε' neighbourhood is also part of the cluster, and the procedure from step 2 is reiterated for all 'ε' neighbourhood points. This is done until all points in the cluster are found.
5) Next, a new unvisited point is repossessed and processed, leading to the detection of a further cluster or noise.
6) The process is continued until all the points are marked as visited.

IV. RESULTS AND DISCUSSION

The effectiveness of the clustering algorithms is calculated using accuracy. The accuracy can be defined as the ratio of the number of correctly classified instances to the total number of examined instances. The accuracy is calculated as follows:

$$Accuracy = \frac{Number\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances} \times 100 \quad (6)$$

To evaluate the performance of with and without anonym zed data, three classifications via clustering algorithms are used i.e. K-means, EM and Density based. WEKA toolkit is used to analyze the Meta classifier performance for clustering algorithms. The accuracy of the classification via clustering is depicted in the Table 3, and the same is plotted in the Figure 1. From the figure, it is observed that the accuracy does not change considerably and is within the manageable limits except for EM. But the point here is the number of clusters is fixed i.e. only two clusters. For this reason the data is not getting the high accuracy values. If the number of clusters is changed, the accuracy may also change. The main aim of this study is to evaluate the performance using clustering techniques and to confirm that clustering can also be applicable to the anonym zed data. The results are somewhat confirming that like, classification, clustering is also applicable to privacy preserving data mining.

TABLE 3: Classification via clustering accuracy

Techniques Used	Accuracy
K-means without anonymization	64.84%
K-means with anonymization	64.71%
EM without anonymization	37.63%
EM with anonymization	30.98%
Density Based without anonymization	65.36%
Density Based with anonymization	65.62%

The performance of dataset is identified by various parameters like kappa statistic, Mean Absolute Error (MAE) and Receiver Operating Curves (ROC) and is tabulated in Table 4. The Root Mean Square Error (RMSE) is shown in Figure 2. The precision, recall and F-measure are tabulated in Table 5 and are shown in Figure 3 and Figure 4.

Copyright © 2014 MECS

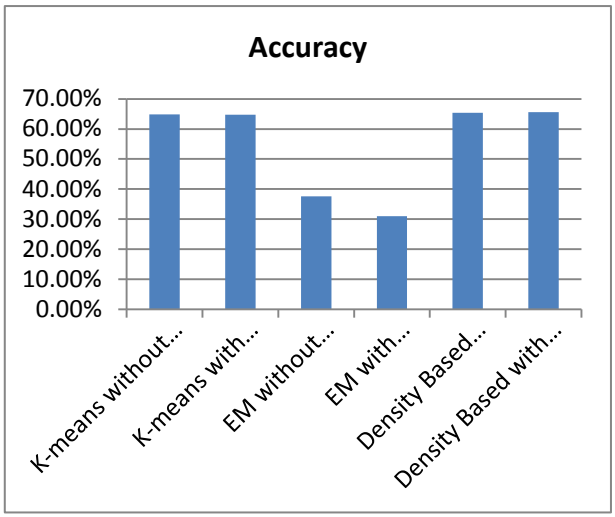

Figure 1: Accuracy of classification via clustering

TABLE 4: Kappa Statistic, MAE, ROC

Classification via clustering Algorithm	Kappa Statistic	MAE	ROC
K-means without anonymization	0.214	0.3516	0.605
EM without anonymization	0.2766	0.3529	0.648
Density based without anonymization	0.3858	0.28990	0.634
K-means with anonymization	0.2435	0.3619	0.605
EM with anonymization	0.2364	0.3464	0.618
Density based with anonymization	0.3037	0.3438	0.665

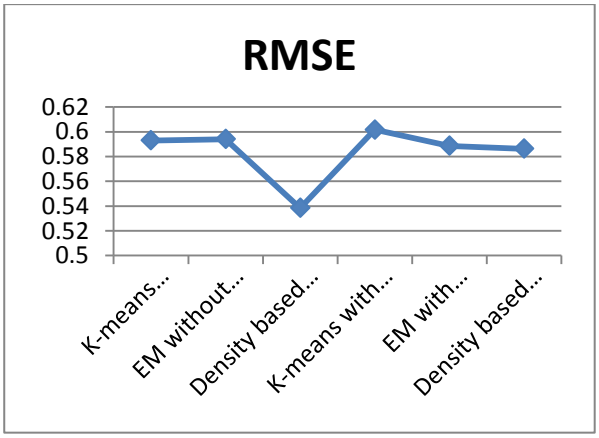

Figure 2: RMSE

TABLE 5: Precision, Recall and F-measure

Classification via clustering Algorithm	Precision	Recall	F-measure
K-means without anonymization	0.643	0.648	0.645
EM without anonymization	0.782	0.710	0.726
Densitybased without anonymization	0.653	0.654	0.653
K-means with anonymization	0.678	0.647	0.655
EM with anonymization	0.766	0.638	0.668
Density based with anonymization	0.694	0.656	0.664

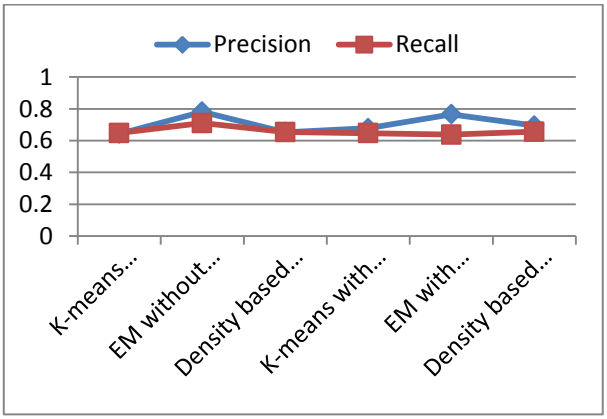

Figure 3: Precision and Recall

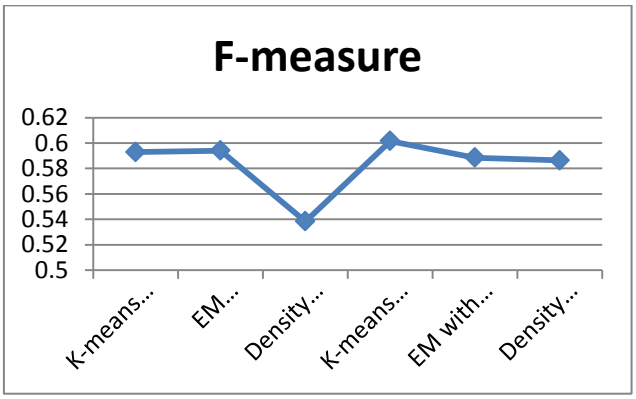

Figure 4: F-measure

V. CONCLUSION

In this study, it was attempted to compare the accuracy of classification via clustering of k-means, EM and Density based with and without anonymized dataset.

Copyright © 2014 MECS

The Diabetic dataset was used for evaluating the performance and the dataset was *k*-anonymized. The experimental results demonstrate that the accuracy is not diminished due to anonymization of the data. So with the help of the WEKA toolkit it is confirmed that the performance of classification via clustering for the anonymized data is improved.

REFERENCES

[1] M. Last et al., "Improving accuracy of classification models induced from anonymized datasets", Inform.Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.07.034.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,"l-Diversity: Privacy Beyond k-Anonymity", Proc. Int'l Conf. Data Engineering (ICDE), pp. 24, 2006.

[3] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity", In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 49–60, 2005.

[4] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity", In Proc. Int'l Conf. Data Engineering (ICDE), pp. 25, 2006.

[5] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", In International journal on uncertainty, Fuzziness and knowledge based systems, 10(5), pp.571 – 588, 2002.

[6] L. Sweeney, "k-anonymity: a model for protecting privacy", In International journal on uncertainty, Fuzziness and knowledge based systems, 10(5), pp. 557 – 570, 2002.

[7] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and -Diversity", In Proc. Int'l Conf. Data Engineering (ICDE), pp. 106115, 2007.

[8] P. Samarati, "Protecting respondents' identities in microdata release", In IEEE Transactions on Knowledge and Data Engineering, pp.13(6):1010–1027, 2001.

[9] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation", In VLDB '06: Proceedings of the 32nd international conference on Very large data bases, pages 139–150. VLDB Endowment, 2006.

[10] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy,"Slicing: A New Approach for Privacy Preserving Data Publishing", In IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 3 MARCH 2012.

[11] Yufei Tao, Hekang Chen, Xiaokui Xiao, "ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication", In IEEE transactions on knowledge and data engineering, Vol. 21, no. 7, July 2009.

[12] Ali Inan, Selim V. Kaya, Yu cel Saygın , Erkay Savas, Ayc a A. Hintoglu, Albert Levi, "Privacy preserving clustering on horizontally partitioned data", Elsevier Data & Knowledge Engineering, pp. 646–666, 2007.

[13] G. Aggarwal, T. Feder, K. Kenthapadi, A. Zhu, R. Panigrahy, and D. Thomas, "Achieving anonymity via clustering in a metric space", In PODS '06: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2006.

[14] Jiuyong Li, Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Jian Pei, "Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures", Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science Volume 4081, pp 405-416, Springer, 2006.

[15] J.-W. Byun, A. Kamra, E. Bertino, and N. Li., "Efficient k-anonymization using clustering techniques", In Internal Conference on Database Systems for Advanced Applications (DASFAA), 2007.

[16] Aris Gkoulalas-Divanis, Grigorios Loukides,"PCTA: Privacy-constrained Clustering-based Transaction Data Anonymization", 4th International Workshop on Privacy and Anonymity in the Information Society, pp. 5, ACM, 2011.

[17] M.E. Nergiz and C. Clifton, "Thoughts on k-Anonymization", In Proc. 22nd Int'l Conf. Data Eng. Workshops (ICDEW '06), pp. 96, 2006.

[18] M.I. Lopez, J.M Luna, C. Romero, S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums", In EDM, pp. 148-151. www.educationaldatamining.org, (2012).

[19] M. Panda and M. Patra, "A novel classification via clustering method for anomaly based network intrusion detection system", In International Journal of Recent Trends in Engineering, pp.1–6, 2009.

[20] A.K.Jain and and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, Englewood Cliffs, USA, 1988.

[21] J. Erman, M. Arlitt and A. Mahanti, "Traffic classification using clustering algorithms", In SIGCOMM-06 workshops, sept.11- 15, Pisa, Italy.pp.281-286.ACM Press, 2006.

[22] The Expectation Maximization Algorithm. http://www.cs.unr.edu/~bebis/mathmethods/EM/lecture.pdf. Sept. 25, 2004.

Authors' Profiles

Sridhar Mandapati obtained his masters in Computer Applications from S.V University, Tirupathi. He is currently working as Associate Professor in the Department of Computer Applications at R.V.R. & J.C College of Engineering, Guntur. He has 14 years of teaching experience. At present he is pursuing Ph.D. from Acharya Nagarjuna University, Guntur. He has seven internal publications and attended several national and international conferences. His areas of research interest include Data Mining, Information Security and Image Processing.

Dr. Raveendra Babu Bhogapathi obtained his Masters in Computer Science and Engineering from Anna University, Chennai and Ph.D. in Applied Mathematics from S.V University, Tirupati. He is now working as Professor in the Department of Computer Science and Engineering, VNR VJIET, Hyderabad. He has 28 years of teaching experience. He has more than 30 International and National publications to his credit. His research areas of interest include Data Mining, Image Processing, Pattern Analysis and Information Security.

Copyright © 2014 MECS

Dr. M.V.P.Chandra Sekhara Rao obtained his Masters in Computer Science and Engineering from JNTU, Kakinada. He received his Ph.D. in Computer Science and Engineering from JNTU, Hyderabad. At present he is a Professor in the Department of Computer Science and Engineering at R.V.R & J.C College of Engineering, Guntur. He has 17 years of teaching experience. He has nine international publications and attended several international conferences. His area of research interest is Data Warehouse and Data Mining.

How to cite this paper: Sridhar Mandapati, Raveendra Babu Bhogapathi, M.V.P.C.Sekhara Rao,"Classification via Clustering for Anonymization Data", IJCNIS, vol.6, no.3, pp.52-58, 2014. DOI: 10.5815/ijcnis.2014.03.07

Copyright © 2014 MECS