Privacy Preserving in Data Mining by Normalization

Syed Md. Tarique Ahmad
Dept. of Computer Science
Pacific University, Udaipur
Rajasthan, India

Shameemul Haque
Dept. of Computer Science
King Khalid University, Abha,
KSA

Prince Shoeb Khan
Dept. of Computer Science
King Khalid University, Abha,
KSA

ABSTRACT
Extracting previously unknown patterns from massive volume of data is the main objective of any data mining algorithm. In current days there is a tremendous expansion in data collection due to the development in the field of information technology. The patterns revealed by data mining algorithm can be used in various domains like Image Analysis, Marketing and weather forecasting. As a side effect of the mining algorithm some sensitive information is also revealed. There is a need to preserve the privacy of individuals which can be achieved by using privacy preserving data mining. In this paper we use min- max normalization approach for preserving privacy during the mining process. We clean the original data using min- max normalization approach before publishing. For experimental purpose we have used k- means algorithm and from our results it is obvious that our approach preserves both privacy and accuracy.

Keywords
Clustering, K- Means, Accuracy, Privacy, Min-Max Normalization, Normalization.

1. INTRODUCTION
Since we are in an era of information explosion, it is very important to be able to find out useful information from massive amounts of data. Consequently, various data mining techniques have been developed. Data mining is often applied to fields such as marketing, sales, finance, and medical treatment. Besides, the rapid advance in Internet and communications technology has led to the emergence of data streams. Due to the consecutive, rapid, temporal and unpredictable properties [1,2] of data streams, the study of data mining techniques has transformed from traditional static data mining to dynamic data stream mining.

In recent years, enabled by the rapid development of various telecommunication technologies, many companies have improve their competitive edge by forming strategic alliances or information outsourcing, one after another. Consequently, many companies frequently expose private data while engaging in data analysis activities, which has led to grave threats to data privacy. For example, online marketing companies usually employ information technology outsourcing with a data mining company for cluster mining, in order to earn greater profits and to find the best target groups of customers. Therefore, how to preserve private data without disclosure while obtaining an accurate mining result in the process of mining will become increasingly difficult, which in turn has led to the development of Privacy- Preserving Data Mining techniques. Nonetheless, traditional Privacy-Preserving Data Mining is not applicable in a data stream environment which requires dynamic updating. For example, for a massive amount of income data, the execution efficiency of traditional methods can no longer respond to user demand. Furthermore, the potential infinite number of data streams plus limited memory space has constrained the traditional methods from obtaining the mining result with accuracy. In view of the above-mentioned issues, studies on Privacy-Preserving Data Stream Mining in recent years have become one of the important issues in the field of data mining.

Several privacy preserving algorithms have been proposed and are used nowadays. In this paper, we propose a new method using min-max normalization for preserving data through data mining. In general, min- max normalization is used as a preprocessing step in data mining for transformation of data to a desired range. Our purpose is to use it for preserving privacy through data mining. We use K- means clustering to validate the proposed approach and validate for accuracy.

The rest of the paper is organized as follows: Section 2 provides an overview of literature review carried out in clustering techniques; Section 3 elaborates the implementation of min- max normalization and K- mean clustering techniques in our proposed system. Experimental results and simulations are tabulated and compared in Section 4 and finally in Section 5, we arrive to an overall conclusion from our work.

2. LITERATURE SURVEY
The study of Privacy-Preserving Data Mining techniques started extensively since 2000 [3], covering development ap-proximately in two categories: Perturbation-Base technique [3, 4] and Secure Multi-Party Computation Base technique [5, 6]. The main idea of Perturbation-Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data. Geometric Data Transformation Methods (GDTMs) [7] is one simple and typical example of data perturbation technique, which perturbs numeric data with confidential attributes in cluster mining in order to preserve priva-cy. Nonetheless Kumari et al. [4] proposed a privacy preserving clustering technique of Fuzzy Sets, transforming confidential attributes into fuzzy items in order to preserve privacy. Furthermore, the largest issue encountered when implement-ing a perturbation technique is the inaccurate mining result from a perturbed data. In view of this issue, the technique of Random-data perturbation introduced by Agrawal and Skrikant [3] was the first study addressed. Whereas the technique derives the original data distribution using a random noise for data distribution, and constructs a result similar to the original data, it finally use this similar result to execute mining. This method could construct a more accurate data mining model, while reducing mining errors. In addition, usually the perturbation technique that has higher privacy preservation comes with a lower level of mining accuracy, whereas most of the perturbation techniques today belong to the one-size-fits-all and are relatively inflexible. To resolve this issue, Liu and Thuraisingham [8] developed the two-phase perturbation technique which frames different intervals

according to differ-ent user demand, and directly obtain sample data from a specific interval to derive the original data distribution. In the study on Secure Multi-Party Computation Base technique, Vaidya and Clifton [9] proposed the method of privacy pre-serving clustering technique over vertically partitioning data, whereas data with different attributes and different locations are considered as the same data set, all data could perform K-means under preserving privacy. On the contrary, Meregu and Ghosh [6] proposed the method of privacy preserving cluster mining over horizontally data partitioning, whereas it is framework of "Privacy-preserving Distributed Clustering using Generative Model." In this framework, each data independently owns an individual source, using local data to train generative models, and delivers model parameters to the central combiner responsible for model integration, hence avoiding direct contact between data source and combiner in order to accomplish privacy preserving through this method.

Among the cluster mining algorithms, K-means is one of the most popular and well-know methods mainly due to its simple concept, easy implementation and comprehensible mining result. Although the method has its own drawbacks [10], most of the existing data stream clustering algorithm are nonetheless developed based on studies of this method. In literature [11, 12], a machine learning algorithm names, Very Fast machine Leaning (VFML) has been proposed, whereas this method depends on determining an upper boundary to be applied as data items test in each step of the algorithm. Subse-quently, Very Fast K-Means (VFML) clustering and Very Fast Decision Tree (VFDT) classification techniques have been de-veloped based on the concept of VFML, and applied on the data stream of artificial and real network. On the other hand, Ordonez [13] developed an incremental K-means algorithm to improve the problems of clustering binary data streams with Kmeans. Incremental K-means not only real-time processing and artificial datasets, but simplification of data processing for binary data could also eliminate the need for data normaliza-tion. The concept of this algorithm is based on the updating cluster center and weight immediately following examining a batch of data, in order to perform fast clustering. Furthermore, Aggarwal et al. [14] proposed another CluStream which is applicable in data stream clustering, using summarized statis-tical information of data streams to cluster according to the user desired cluster numbers. On the other hand, Gaber et al. [15] has developed a Lightweight Clustering algorithm to handle high speed data stream. This algorithm is based on the concept of Algorithm Output Granularity, which is mainly used to adjust the minimal boundary value of distance among datasets representing different clusters, then controls the out-put-input ratio according to available resources, and to output a combined clustering result when the memory space is full. More recently, Yang and Zhou [16] further developed an HCluStream data stream clustering algorithm which processes combined attributes based on CluStream algorithm in order to solve the weakness of inability to perform non-numerical data mining by CluStream.

A shearing based data transformation approach was proposed by Manikandan et al. [17] for achieving privacy. Sheared data depends on the noise value. If the noise is more the user may easily identify that the data is a modified one and not original.

3. PROPOSED SYSTEM

In this paper we put forward an approach for privacy preserving using min- max Normalization.

3.1 Min-Max Normalization

Min-max normalization performs a linear transformation on the original data. For mapping a value, v of an attribute A from range [min_A, max_A] to a new range [new_{minA}, new_{maxA}], and the computation is given by.

$$\frac{v - \min_A}{\max_A - \min_A}\left(new_{\max A} - new_{\min A}\right) + new_{\min A}$$

Where v is the new value in the required range.

The main advantage of Min- Max normalization is that it preserves the relationships between the original data values [18].

Table 1 is the sample data set used for experiment. Table 2 is the corresponding normalized values for the 'Age' attribute after applying min- max normalization. The steps involved in our approach can be summarized in the form of a procedure as shown below. Fig. 1 shows the below diagram for the proposed system.

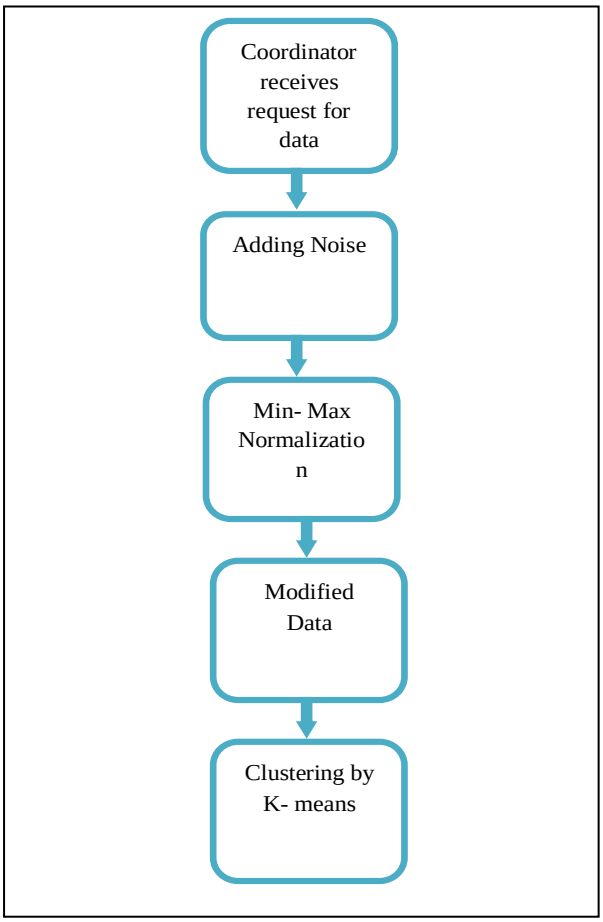

Fig 1: Flow Diagram for proposed system.

3.2 Procedure Steps

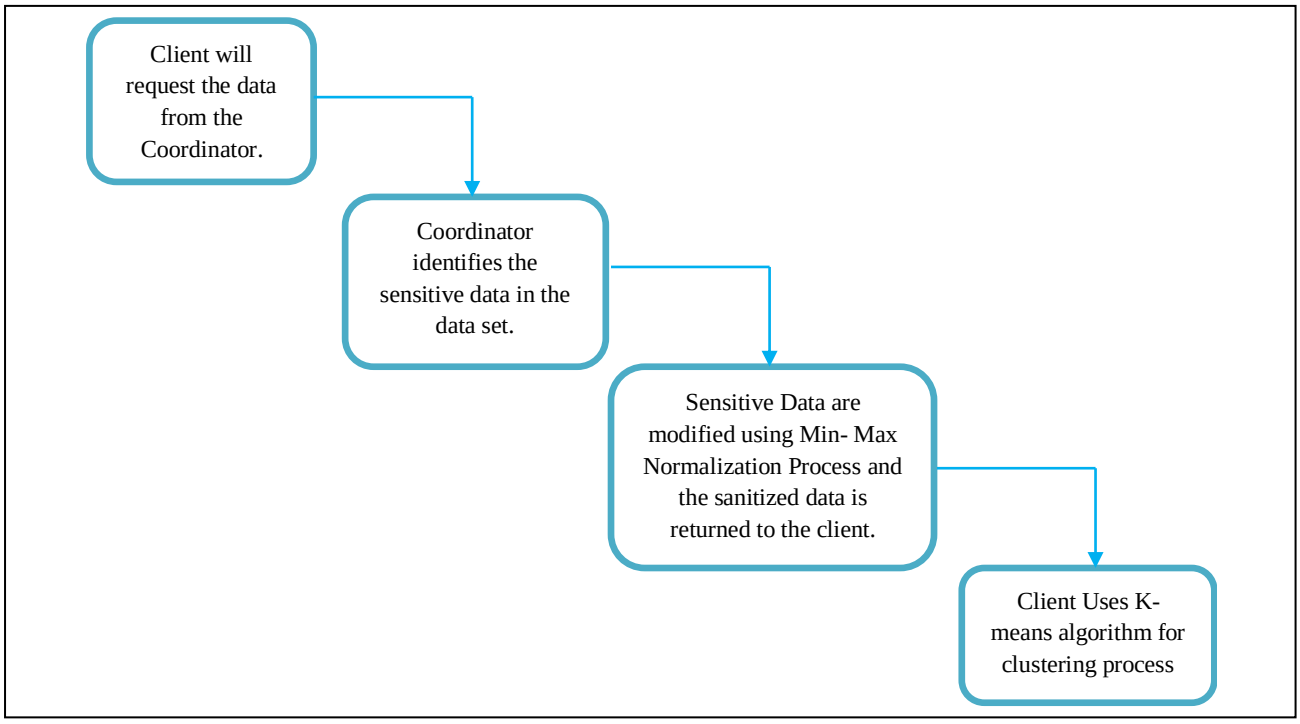

Fig 2: Step by Step Procedure

Table 1. Original data

S No.	Name	Age	Sex
1	Salman	2	M
2	Sonia	10	F
3	Bushra	20	F
4	Rajeev	25	M
5	Sharmi	12	F
6	Dhoni	30	M
7	Abu	20	M

Table 2. Normalized Data

S No.	Name	Age	Sex
1	Salman	10	M
2	Sonia	33	F
3	Bushra	62	F
4	Rajeev	76	M
5	Sharmi	39	F
6	Dhoni	90	M
7	Abu	62	M

3.3 Clustering Technique

In our methodology, in order to check for the effectiveness of min- max normalization on the data partitioning techniques, we used K- means clustering algorithm. In K- means, the objects are clustered based on attributes into 'n' number of clusters where 'n' is a positive integer. The central idea of this Clustering is to minimize the sum of squares of the distance between data and corresponding cluster centroid in that data set. The clustering process must be carried out until it gets stabilized. Then, the objects are grouped based on the inter-relative distance among each object and the centroid.

4. RESULTS AND SIMULATIONS

In this paper, we have used min- max normalization to achieve privacy and accuracy during data mining and accuracy is tested using K- means clustering. Here the computations for min- max normalization of sample data, k-means clustering and effectiveness calculations are carried out in C++.

We have also tested the efficiency of our approach on a real time dataset, "adult-dataset" from UCI data repository [19]. This data set comprises of about 32561 records with 12 attributes namely age, work class, education, marital status, occupation, relationship, race, sex, capital gain and capital loss, hours per week and native country. For experimental purpose, we used only age as the key attribute to carry out normalization in our work.

The clustering of data before and after normalization for 2-clusters is given in the figures Figure 3 and Figure 4 respectively. Similarly, those for 3-clusters are provided in the figures Figure 5 and Figure 6. Table 3 and Table 4 describe the clustering of data before and after min- max normalization for 2-clustering and 3-clustering respectively.

Table 5 and Figure 7 summarize the comparisons among Fuzzy S- Shaped approach, Shearing Noise addition approach and our proposed min- max normalization approach.

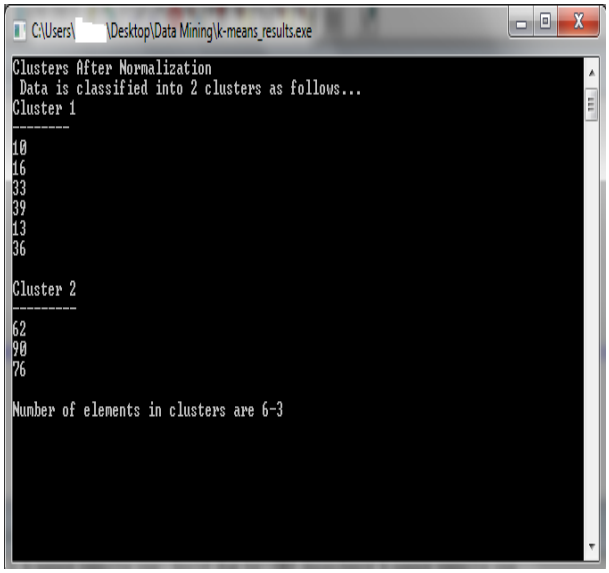

Fig 3: Snapshot for 2 clusters After Normalization.

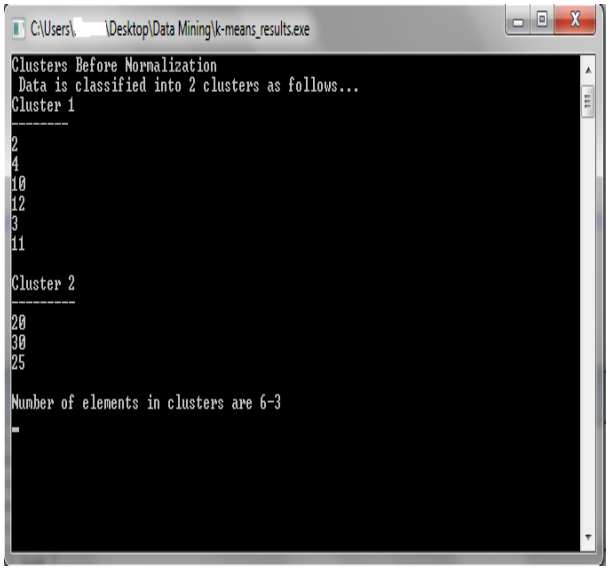

Fig 4: Snapshot for 2 clusters before Normalization.

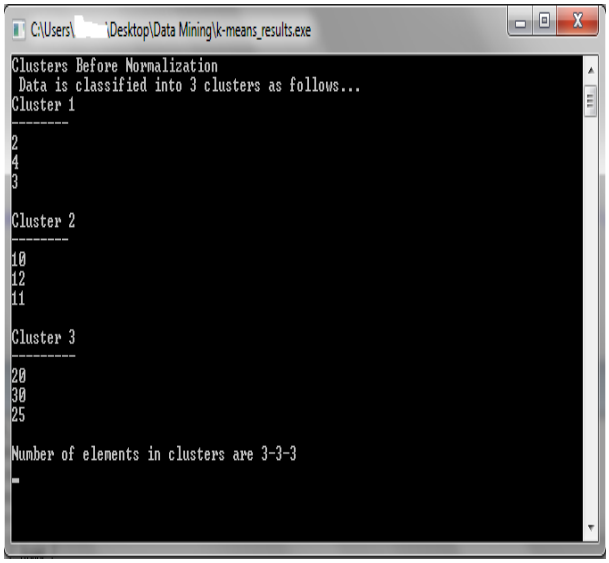

Fig 5: Snapshot for 3 clusters before Normalization.

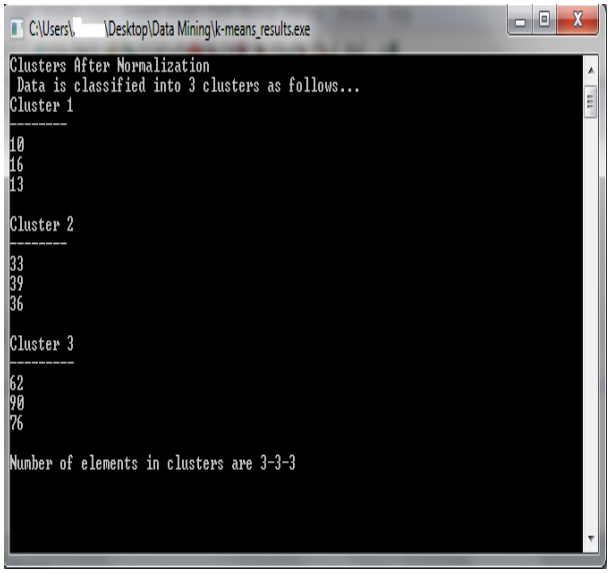

Fig 6: Snapshot for 3 clusters after Normalization.

Table 3. Results of 2 clusters

K=2	Cluster 1	Cluster 2
Before Normalization	{2,4,10,12,3,11}	{20,30,25}
After Normalization	{10,16,33,39,13,36}	{62,90,76}

Table 4. Results of 3 clusters

K=3	Cluster 1	Cluster 2	Cluster 3
Before Normalization	{2,4,3}	{10,12,11}	{20,30,25}
After Normalization	{10,16,13}	{33,39,36}	{62,90,76}

Table 5. Results of 3 clusters

Original Data	Min-Max Normalization	Fuzzy S-Shaped	Shearing Noise (=10)
2	10	0	22
3	13	0.0025	33
4	16	0.0102	44
10	33	0.1632	110
11	36	0.2066	121
12	39	0.2551	132
20	62	0.7449	220
25	76	0.9362	275
30	90	1	330

From the above comparisons we observe that data transformation approach based on shearing, scales the values and scatters them over a large range. On the other hand, Fuzzy approach based on S- Shaped membership function narrows down the range of values to (0, 1). The produced results of both the approaches prove evidently the duplication of data for privacy preservation. Our approach overcomes this limitation as the normalized values lie in the same range as the actual range of the attribute. Thus, the distortion of data

for sake of privacy will not be revealed to the analyst or data-users, at the same time preserving privacy. Fig. 3 shows the comparison graph.

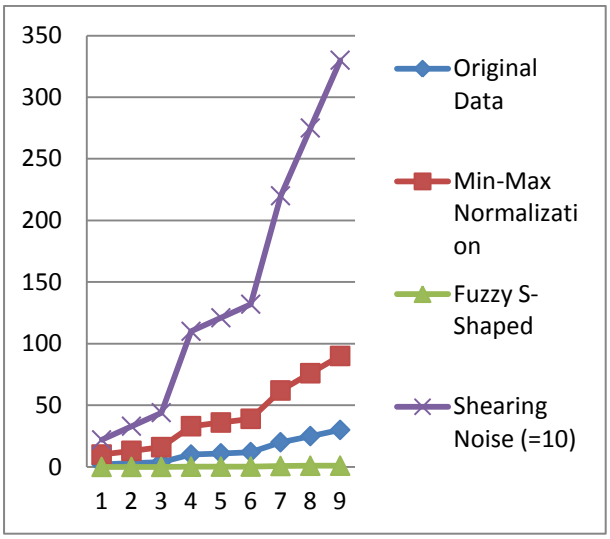

Fig 5: Comparison Graph.

5. CONCLUSION

In this paper we have dealt with min- max normalization based data transformation to preserve data privacy. This approach transforms the original data to privacy- preserved data maintaining the inter relative distance among the data. Experiments have proven that performing k- means clustering on the distorted data produces same clustering results as original data. Thus we have succeeded in achieving both accuracy and privacy. We have tested the technique for numerical data set. The future scope of this paper is to extend the same over categorical data.

6. REFERENCES

[1] Cohen, E. and Strauss, M., "Maintaining Time Decaying Stream Aggregates," Proceedings of the 22th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, San Diego, California, U.S.A., pp. 223233 (2003).

[2] Chang, J. H. and Lee,W. S., "Finding Recent Frequent Itemsets Adaptively over Online Data Stream," Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., U.S.A., pp. 487492 (2003).

[3] Agrawal, R. and Srikant, R., "Privacy-Preserving Data Mining,", Proceeding of the ACM SIGMOD Conference on Management of Da-ta, Dallas, Texas, U.S.A., pp. 439-450 (2000).

[4] Kumari, P. K., Raju, K. and Rao, S. S., "Privacy Preserving in Cluster-ing Using Fuzzy Sets," Proceedings of the 2006 International Confer-ence on Data Mining, Las Vegas, Nevada, U.S.A., pp. 26 29 (2006).

[5] [20] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M. Y., "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations Newsletter, Vol. 4, pp. 28 34 (2002).

[6] Meregu, S. and Ghosh, J., "Privacy-Preserving Distributed Clustering Using Generative Models," Proceedings of the 3th IEEE International Conference on Data Mining, Melbourne, Florida, U.S.A., pp. 211-218 (2003).

[7] Oliveira, S. R. M. and Zaïane, O. R., "Privacy Preserving Clustering by Data Transformation," Proceedings of the 18th Brazilian Sympo-sium on Databases, Manaus, Brazil, pp. 304 318 (2003).

[8] Liu, L. and Thuraisingham, B., "The Applicability of the Perturbation Model-Based Privacy Preserving Data Mining for Real-World Data," Proceedings of the 6th IEEE International Conference on Data Min-ing, Hong Kong, China, pp. 507 512 (2006).

[9] Vaidya, J. and Clifton, C., "Privacy-Preserving KMeans Clustering over Vertically Partitioned Data," Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Da-ta Mining, Washington, D.C., U.S.A., pp. 206 215 (2003).

[10] Chen, T. S., Lin, C. C., Chiu, Y. H. and Chen, R. C., "Combined Den-sity-Based and Constraint-Based Algorithm for Clustering," Journal of Donghua University, Vol. 23, pp. 36 38 (2006).

[11] Hulten, G., Spencer, L. and Domingos, P., "Mining Time-Changing Data Streams," Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francis-co, California, U.S.A., pp. 97 106 (2001).

[12] Domingos, P. and Hulten, G., "Mining High-Speed Data Streams," Proceedings of the Association for Computing Machinery 6th Inter-national Conference on Knowledge Discovery and Data Mining, Bos-ton, U.S.A., pp. 71 80 (2000).

[13] Ordonez, C., "Clustering Binary Data Streams with K-means," Pro-ceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, California, U.S.A., pp. 12 19 (2003).

[14] Aggarwal, C., Han, J., Wang, J. and Yu, P. S., "A Framework for Clus-tering Evolving Data Streams," Proceedings of the 29th International Conference on Very Large Data Bases, Berlin, Germany, pp. 81-92 (2003).

[15] Gaber, M. M., Krishnaswamy, S. and Zaslavsky, A., "On-Board Min-ing of Data Streams in Sensor Networks," Springer, Berlin Heidel-berg, Germany, pp. 307-335 (2005).

[16] Yang, C. and Zhou, J., "HClustream: A Novel Approach for Cluster-ing Evolving Heterogeneous Data Stream," Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, China, pp. 682-688 (2006).

[17] Manikandan G, Sairam N et al. "Privacy preserving clustering by shearing based data transformation", Proceedings of International Conference on Computing and Control Engineering. (2012).

[18] Han J, and Kamber M, "Data mining-concepts and techniques", 2nd Edn. San Francisco: Morgan Kaufmann Publishers. (2006).

[19] http://archive.ics.uci.edu/ml/datasets.html.