Asian Journal of Information Technology 15 (10): 1531-1538, 2016
ISSN: 1682-3915
© Medwell Journals, 2016

Privacy Preserving K-Anonymization Clustering Approach for Reducing Information Loss

[1]G. Chitra Ganabathi and [2]P.Uma Maheswari
[1]Department of Computer Science and Engineering,
CMS College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[2]Department of Computer Technology, Anna University, Chennai, India

Abstract: The k-anonymity problem has recently drawn considerable interest from research community and a number of algorithms have been proposed. We are also considered the problems of existing methods and addressed such limitations with the aid of our proposed k-anonymization technique. We have intended to propose using a clustering based K-Anonymous method. The primary goal underlying our approach is that the k-anonymization problem can be considered as a clustering problem. Intuitively, the k-anonymity requirement will be generally transformed into a clustering problem, where it is required to discover a set of clusters each of which contains at least k records. Moreover, our proposed research will be reduced the Information Loss. The records will be initially collected and which will be analyzed for two different attributes: Numerical attributes; categorical attributes. For these two kinds of attributes, we will then find the distance of records, separately. Followed by the distance calculation, each record will be evaluated and the Information Loss (IL) in each and every record will be obtained by our proposed research. This will be performed based on the anonymity value for the input dataset. Now, we will obtain the loss of information in each record and from this we can cluster out the minimum IL record. The clustering will be done by an Adaptive Particle Swarm Optimization based Fuzzy C-Means (APSO-FCM) Clustering algorithm. Fuzzy C-Means (FCM) is one of the clustering algorithms used to make a group of data into clusters, in which one data can be allocated for two or more clusters. In FCM, the objective function will be optimized with Particle Swarm Optimization (PSO) algorithm. The PSO algorithm imitates the social characters shown by swarms of animals. In this algorithm, a point in the search space which is a possible solution, is called a particle. The group of particles in a specific iteration is called ‘swarm’. While looking out for food, the birds are either scattered or go collectively before they find out the place where they are able to locate the food. While the birds are on the search for food moving from one location to another, there is often a bird which is able to smell the food effectively, in other words, the bird is discernible of the location where the food is likely to be found having superior food resource data. As they tend to convey the data, particularly the excellent data at any time while looking for the food from one location to another, attracted by the excellent data, at the end, the birds will throng at the location where there is strong possibility for locating food. Thus, the clustering of nearly minimum of IL records will be gained by our proposed research. The proposed approach will be implemented in MATLAB and planned to be evaluated using various databases. Our proposed research will be made better performance evaluation results with reduced Information Loss.

Key words: Adaptive particle swarm optimization, fuzzy C-Means Clustering algorithm, information loss, numerical attributes, categorical attributes

INTRODUCTION

Data mining is a technique used to extract un-known patterns from large volume of data. The term data mining refers to the non-trivial extraction of valid, implicit, potentially useful and ultimately understandable information in large databases with the help of the modern computing devices (Malik *et al*., 2012). Many data mining techniques have been developed. They are association rule mining, classification and clustering (Lakshmi and Ks, 2012). Among many data mining techniques, association rule mining is receiving more attention to the researchers to find correlations between items or items sets efficiently (Lakshmi and Ks, 2012).

Corresponding Author: G. Chitra Ganabathi, Department of Computer Science and Engineering,
CMS College of Engineering and Technology, Coimbatore, Tamil Nadu, India

Data mining as a powerful data analysis tool has made tremendous contributions in many areas and has the wide applications prospect (Zhang and Bi, 2010). And, it allows us to analyze personal data or organizational data, such as customer records, criminal records, medical history, credit records, etc (Dung *et al*., 2010). It is based on the artificial intelligence, machine learning and statistical techniques and can analyze the original data, make the inductive reasoning, excavated from a potential model, predict customer behavior (Yun *et al*., 2010). Most existing data mining algorithms are carried out under the assumption that all the data could be available at a single central site (Zhu *et al*., 2009).

Privacy preserving data mining (PPDM) has become a hot spot in data mining field (Peng *et al*., 2010). It has become increasingly popular because it allows sharing of privacy-sensitive data for analysis purposes (Wang *et al*., 2009). In privacy preserving data mining, the goal is to perform data mining operations on sets of data without disclosing the content of the sensitive data (Wu, 205). Privacy preserving data mining addresses the need of multiple parties with private inputs to run a data mining algorithm (Peng *et al*., 2010). The data owners would like to mine on their collective data but in a way that no information about their private data sets is available to other participants (Blanton, 2011). The main consideration in privacy preserving data mining is two fields. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms, should also be excluded because such a knowledge can equally well compromise data privacy as we will indicates (Vassilios *et al*., 2004). Specifically, with distributed data sources and consideration of privacy protection for all participants, sensitive information is not allowed to be disclosed between them in a cooperative research (Wang *et al*., 2007).

In privacy preserving of data mining must safeguard from divulging sensitive data during publication of individual data. To maintain privacy, a number of techniques have been proposed for modifying or transforming the data (Mandapati *et al*., 2013). We can simply divide these methods into two main categories. The first category of methods modifies data mining algorithms so that they may carry out data mining operations on distributed datasets without knowing the exact values of the data or without directly accessing to the original datasets and the second category, several randomization-based data distortion methods focus on perturbing the whole dataset or the confidential parts of the dataset using certain distribution of random noises. (Peng *et al*., 2010). The success of privacy preserving data mining algorithms is measured in terms of its performance, data utility, level of uncertainty or resistance to data mining algorithms etc (Malik *et al*., 2012).

Literature review: Several techniques were proposed by various authors for privacy preserving data mining and a few of them are explained below: data mining is emerging as one of the key features of many business organizations. Privacy becomes an important factor in data mining. One of the most important tasks of data mining is to find patterns in data. Vijayarani *et al*. (2010) have proposed that the sensitive information was not revealed after mining. But the data quality was important such that no false information was provided and the privacy was not jeopardized. They have also analyzed the problem of building privacy preserving algorithms for one category of data mining techniques was the association rule mining.

Li *et al*. (2012b) have proposed a privacy preserving data mining which addressed the problem of developing accurate models about aggregated data without access to precise information in individual data record. In their setting, the more trusted a data miner was the less perturbed copy of the data it could access. They have proved that their solution was robust against diversity attacks with respect to their privacy goal. Their solution has prevented them from jointly reconstructing the original data more accurately than the best effort using any individual copy in the collection. And their solution has allowed a data owner to generate perturbed copies of its data for arbitrary trust levels on-demand. This feature has offered data owner's maximum flexibility.

Consider a scenario in which the data owner has some private or sensitive data and wants a data miner to access them for studying important patterns without revealing the sensitive information. Privacy-preserving data mining aims to solve that problem for which (Bhaduri *et al*., (2011) have proposed a randomly transforming the data prior to their release to the data miners. Here, they have discussed nonlinear data distortion using potentially nonlinear random data transformation and they have shown how it could be useful for privacy-preserving anomaly detection from sensitive data sets. They have shown how their general transformation could be used for anomaly detection in practice for two specific problem instances: a linear model and a popular nonlinear model using the sigmoid function. They have also analyzed the proposed nonlinear transformation in full generality. A main contribution of the researchers was the discussion between the invertibility of a transformation and privacy preservation.

Li *et al.* (2012a) have proposed that a technique called slicing which partitioned the data both horizontally and vertically. They have shown that slicing preserves better data utility than generalization and could be used for membership disclosure protection. Another important advantage of slicing was that it could handle high-dimensional data. They have shown how slicing could be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the l-diversity requirement. Their experiments have also demonstrated that slicing could be used to prevent membership disclosure. And, the results of the experiments have shown that slicing preserved better data utility than generalization and that was more effective than bucketization in workloads involving the sensitive attribute.

Wang *et al.* (2012) have proposed a protocol which allowed individual meters to report the true electricity consumption reading with a pre-determined probability. They have also proposed a novel method to protect the privacy of customers with smart meters while assuring the capability to load serving entities to estimate the current electricity consumption. And they have also provided a privacy-preserving approach to utilizing smart meters at end users levels. Load Serving Entities (LSE) could reconstruct the total electricity consumption of a region or a district through inference algorithm but their ability of identifying individual user's energy consumption pattern was significantly reduced.

Dunning and Kresman (2013) have proposed a new algorithms for assigning anonymous IDs were examined with respect to trade-offs between communication and computational requirements. The new algorithms were built on top of a secure sum data mining operation using Newton's identities and Sturm's theorem. An algorithm for distributed solution of certain polynomials over finite fields has enhanced the scalability of the algorithms. Markov chain representations were used to find statistics on the number of iterations required and computer algebra given a closed form results for the completion rates.

Many applications are benefited from data sharing, especially data statistics and data mining. But as the shared data may contain private information of data owner, it has a high risk of revealing data owner's privacy. Data obfuscation is proposed to gain a balance between data privacy and data usability. But, it is hard for the present obfuscation schemes to remain the usability of data in a fine-grained level. Yang *et al.* (2013) have proposed a data obfuscation scheme that adds an accurate "noise" to the original data to protect the privacy while keeping the numeral characteristics of data unchanged in different levels. Besides, the scheme could also lower the impact on data mining. Furthermore, by allocating different keys to users, different users had different permissions to access to data. The experiments have shown that their scheme obfuscates date correctly, efficiently and securely.

Problem definition: Data mining is a general technique to regain and determine valuable hidden knowledge and information from private data. This has direct to concerns that private data may be breached and mishandled. As a result, it is essential to guard individual's own data through some privacy preserving approaches. The common problems in existing privacy preserving approaches are given below:

- The difficulty of distributing privately held data so that the person's who are the subjects of the data cannot be discovered has been researched broadly
- The database with the tuple data does not be preserved secretly
- In some of the existing systems, another person can straightforwardly access database
- In the existing (Vijayarani *et al.*, 2010) privacy preserving data mining based on association rule method, there is a chance to provide a false information if the data quality is important. So they need the data mining techniques for securing both data and knowledge
- Slicing was introduces as a new approach to privacy preserving data publishing (Li *et al.*, 2012b). In slicing method, each attribute was in exactly one column. The method, "overlapping slicing" has made the attributes as duplicates by providing the attributes in more than one column. So, the results of the method have shown that the random grouping was not very effective
- The drawback of privacy-preserving data obfuscation scheme (Yang *et al.*, 2013) is that the obfuscated data was order sensitive, i.e., if the sort of obfuscated data varies, the retrieving procedure will be unsuccessful. This system has required a large number of keys which leads to make improvement in the key management
- One major drawback is that the privacy preservation of individuals when data is shared for clustering is very complex

- The protection of the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis is hard to achieve
- State-of-art solutions for the K-Anonymous problem undergo from high information loss mainly due to dependence on pre-defined generalization strategies or total order inflicted on every attribute domain

These are the main drawbacks of various existing works which motivate us to do this research on privacy preserving data mining.

MATERIALS AND METHODS

The primary goal underlying our approach is that the k-anonymization problem can be considered as a clustering problem. Intuitively, the k-anonymity requirement will be generally transformed into a clustering problem, where it is required to discover a set of clusters each of which contains at least k records. Moreover, our proposed research will be reduced the Information Loss. The records will be initially collected and which will be analyzed for two different attributes: Numerical attributes; categorical attributes. For these two kinds of attributes, we will then find the distance of records, separately. Followed by the distance calculation, each record will be evaluated and the Information Loss (IL) in each and every record will be obtained by our proposed research. This will be performed based on the anonymity value for the input dataset.

Now, we will obtain the loss of information in each record and from this we can cluster out the minimum IL record. The clustering will be done by an Adaptive Particle Swarm Optimization based Fuzzy C-Means (APSO-FCM) clustering algorithm. Fuzzy C-Means (FCM) is one of the clustering algorithms used to make a group of data into clusters in which one data can be allocated for two or more clusters. In FCM, the objective function will be optimized with Particle Swarm Optimization (PSO) algorithm. PSO algorithm imitates the social characters shown by swarms of animals. In this algorithm, a point in the search space which is a possible solution, is called a particle. The group of particles in a specific iteration is called 'swarm'. While looking out for food, the birds are either scattered or go collectively before they find out the place where they are able to locate the food. While, the birds are on the search for food moving from one location to another, there is often a bird which is able to smell the food effectively, in other words, the bird is discernible of the location where the food is likely to be found, having superior food resource data. As, they tend to convey the data, particularly the excellent data at any time while looking for the food from one location to another, attracted by the excellent data, at the end, the birds will throng at the location where there is strong possibility for locating food. Thus, the clustering of nearly minimum of IL records will be gained by our proposed research.

Distance calculation of input records: The records will be initially collected and which will be analyzed for two different attributes:

- Numerical attributes
- Categorical attributes

For these two kinds of attributes, the distance of records is calculated, separately. Followed by the distance calculation, each record is evaluated and the Information Loss (IL) in each and every record is obtained. Then the loss of information in each record and from this the minimum IL record will be clustered (Fig. 1).

Clustering input records using Fuzzy C Means algorithm: Input Records (R_n) need to be clustered for

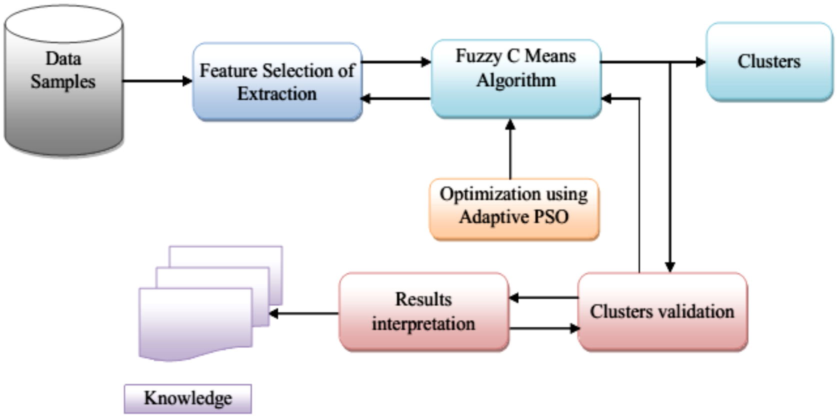

Fig. 1: Architecture of the proposed methodology

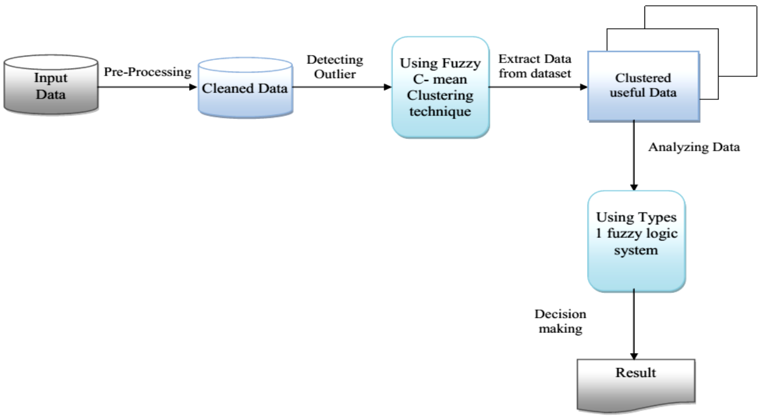

Fig. 2: Working mechanism of fuzzy C-means clustering technique

applying the optimization algorithm. Based on the energy (r) level of each input record (R_n), grouping of Records (R_n) is made employing the Fuzzy C-Means Technique. Fuzzy C-Means (FCM) is a technique of clustering which permits one piece of data to belong to two or more clusters. In pattern recognition, this technique is often applied. FCM is applied based on minimization of the subsequent objective function (Fig. 2):

$$O = \sum_{r=1}^{N}\sum_{j=1}^{c}\left[\mu^{m}{}_{ij}(x_i - c_j)^2\right] \tag{1}$$

Where:

m = Any real number >1
u_{ij} = The degree of membership of x_i in the cluster j
x_i = The ith of d-dimensional measured data
c_j = The d-dimension center of the cluster
||*|| = Any norm stating the comparison between any measured information and the center

Fuzzy partitioning is performed through an iterative optimization of the objective function illustrated above, with the revise of membership u_{ij} and the cluster centers c_j by:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left|x_j - c_j\right|}{\left|x_j - c_k\right|}\right)^{\frac{2}{m-1}}} \tag{2}$$

$$c_j = \frac{\sum_{i=1}^{N}\mu^{m}{}_{ij}x_i}{\sum_{i=1}^{N}\mu^{m}{}_{ij}} \tag{3}$$

This iteration will stop when $\{(\mu_{ij}^{(k+1)}\text{-}\mu_{ij}^{k})\}<t$ where t is a termination criterion among 0 and 1 whereas k is the iteration step. This procedure congregates to a local minimum or a saddle point of O. The objective function is optimized using adaptive particle swarm optimization algorithm. As a result the Records (R_n) are grouped according to their energy (r) level. To attain competent grouping result, the objective function of FCM is optimized by means of Adaptive Particle Swarm Optimization (APSO).

Adaptive Particle Swarm Optimization (APSO): Population based search algorithm is known to be Particle Swarm Optimization (PSO). It is formed to pretend the manners of birds in hunt for food on a cornfield or fish school. The technique can competently find optimal or near optimal solutions in large search spaces. There are two dissimilar kinds of versions are employed according to PSO. The first is “individual best” and the second is “global best”.

The “pbest”: It is the individual best selection algorithm by evaluating each individual position of the particle to its own best position pbest, only. The data about the other particles is not employed in this pbest.

The “gbest”: It is the worldwide best selection algorithm which acquires the global information by making the movement of the particles contains the position of the best particle from the whole swarm. In addition, every particle exploits its experience with previous incidents in terms of its own best solution.

Disadvantages of PSO: The inertia weight factor is stable in a single generation. The search direction is not obvious and offers slow convergence.

As a result, Adaptive Particle Swarm Optimization (APSO) technique is employed offering more precise clustering result where the inertia weight is as well considered. Now, the association rules were optimized by means of APSO. The working process of the APSO is specified in Fig. 3.

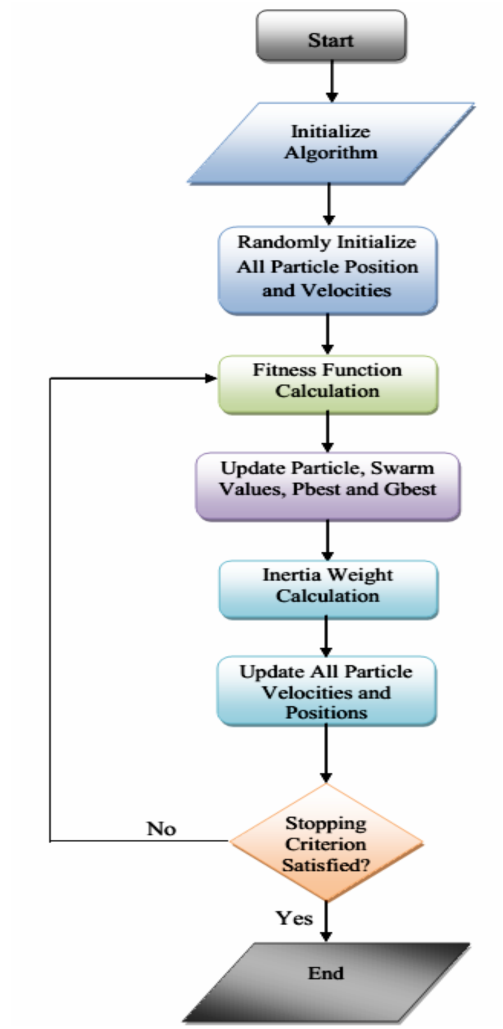
Start
Initialize Algorithm
Randomly Initialize All Particle Position and Velocities
Fitness Function Calculation
Update Particle, Swarm Values, Pbest and Gbest
Inertia Weight Calculation
Update All Particle Velocities and Positions
Stopping Criterion Satisfied?
No
Yes
End

Fig. 3: Working matchine of APSO

FCM objective function optimization using APSO

Swarm initialization: For a population size u, produce the particles arbitrarily.

Define the fitness function: According to the present population, the fitness function selected should be applied for the constraints. Here eqn. (4) is the fitness function.

$$\text{Fitness} = \sum_{i=1}^{n} x_1 + x_2 +,\dots, x_n \tag{4}$$

The gb and pb initialization: At first, the fitness value computed for each particle is set as the Pbest value of each particle. Among the pbest values, the best one is chosen as the value

Velocity computation: The novel velocity is computed by means of the below Eq. 5:

$$V_i^{d+1} = w^d V_i^d + c_1.r_1.(pb_i^d - x_i^d) + c_2.r_2.(gb^d - x_i^d) \tag{5}$$

Where:

c_1, c_2 = Constants with the value of 2.0
w^d = Inertia weight
c_1, c_2 = Independent random numbers generated in the range [0.1]
V_i^d = Velocity of ith particle i
x_i^d = Current position of the particle
pb_i^d = Best fitness value of the particle at the current iteration
gb^d = Best fitness value in the swarm

$$x_i^d = x_i^d + \delta V_i^d \tag{6}$$

Inertia weight calculation:

$$w^d = (w_{mx} - w_{mn}) \times (m_{it} - d) / m_{it} + w_{mn} \tag{7}$$

Where:

w_{mx} = Maximum inertia weight
w_{mn} = Minimum inertia weight
w_n = Maximum number of iteration

Swarm updation: Work out the fitness function again and revise the pb and gb values. If the novel value is better than the previous one, substitute the old by the current one. And furthermore choose the best pb as the gb.

Criterion to stop: Prolong till the solution is good enough or maximum iteration is reached. Thus, the input records were clustered with less information loss by means of optimization using adaptive PSO.

RESULTS AND DISCUSSION

This study delineates the results, we obtained for our recommended technique. This section contains the experimental setup and dataset description, comparative analysis of our recommended technique with the existing technique based on the time taken to update the number of records we give and the information loss we obtained for our recommended technique cluster sizes.

Experimental setup and dataset description: Our recommended technique is implemented in matlab version R2013 that has the system configuration as Intel i5 processor with 4GB RAM and 500GB hard disk. The dataset we used for our experimentation is adult dataset and it is taken from UCI machine learning repository (Lakshmi and Ks, 2012). To analyze the performance of k-anonymity, the adult dataset is a benchmark dataset. We have removed some of the attributes along with the missing records from the adult dataset.

Table 1: Time taken to update the cluster with various sizes

Time taken (sec)			
Cluster = 3		Cluster = 5	
Existing	Proposed	Existing	Proposed
292	733	273	823
351	735	360	765
480	845	599	798
665	853	626	875
846	923	648	812
680	912	745	875
682	846	760	850
903	1103	767	976
1535	858	1713	790
1981	1020	1903	1002

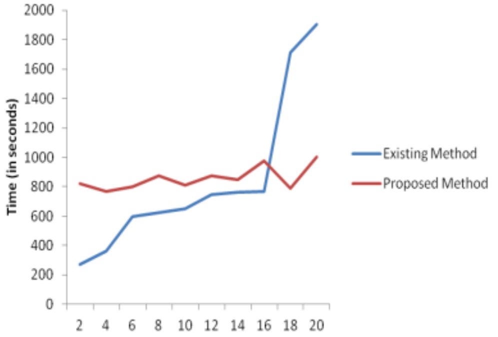

Fig. 4: Time taken to update when cluster is 3

Comparative analysis: We have compared our recommended technique with the existing technique (Zhang *et al*., 2013) based on the updating time by adding different number of records. We compared the time taken to update for different cluster values (Table 1).

In Fig. 4, we compared our work with the existing technique (Zhang *et al*., 2013) based on time taken to update when we set the k value as three. Here, when we update two records the time taken to update in the existing technique is 0.292 sec and it is 0.733 sec for our recommended technique. When we update four records, the time taken to update is 0.351 seconds for the existing work and it is 0.73 sec for our recommended technique; Likewise, when the records we given to update is sixteen, the time taken is 0.903 sec for the existing technique and it is 1.103 sec for our recommended technique and when the records we given to update is twenty, the time taken is 1.981 sec for the existing technique and it is 1.020 sec for our recommended technique.

In Fig. 5, here when we add two records, the time taken to update is 0.273 sec for the existing work and 0.823 sec for our recommended technique and when we add twenty records, the time taken to update for the existing technique is 1.602 sec and for our recommended technique, it is 1.002 sec.

Table 2 shows the information loss we obtained for our recommended technique by varying the k values.

Table 2: Information Loss for different cluster values

K-value	Information loss
3	16.54
5	16.41

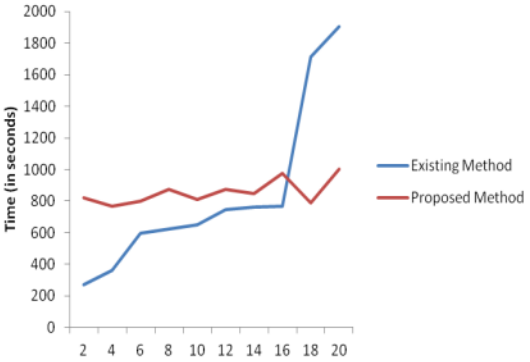

Fig. 5: Time taken to update when cluster is 5

Here, when we set the cluster value as 3, the information loss we obtained for our recommended technique is 16.54 and when we set the cluster value as 5, the information loss is 16.41.

CONCLUSION

The primary intension of the research is efficient clustering of input records with less information loss with the combination of Fuzzy c means and APSO optimization technique. Our suggested methodology had two main stages such as:

- Fuzzy c means clustering
- Adaptive PSO

Initially the input group of records is separated in to two main groups, numerical attributes and categorical attributes. Then, the distance from each record will be calculated. Then, information loss is calculated. In order to obtain less information loss the input records are clustered using Fuzzy C-Means algorithm. The objective function of Fuzzy C-Means algorithm will be optimized using adaptive particle swarm optimization algorithm. From the experimental results, we observed that our proposed hybrid system gives better results when compared to the other existing methods.

REFERENCES

Bhaduri, K., M.D. Stefanski and A.N. Srivastava, 2011. Privacy-preserving outlier detection through random nonlinear data distortion. Syst. Man Cybern. Part B Cybern. IEEE. Trans., 41: 260-272.

Blanton, M., 2011. Achieving full security in privacy-preserving data mining. Proceedings of the 2011 IEEE Third Inernational Conference on Social Computing, Privacy, Security, Risk and Trust (PASSAT), October 9-11, 2011, IEEE, Boston, Massachusetts, USA., ISBN: 978-1-4577-1931-8, pp: 925-934.

Dung, L.T., H.T. Bao, N.T. Binh and T.H. Hoang, 2010. Privacy preserving classification in two-dimension distributed data. Proceedings of the 2010 Second International Conference on Knowledge and Systems Engineering (KSE), October 7-9, 2010, IEEE, Hanoi, Vietnam, ISBN: 978-1-4244-8334-1, pp: 96-103.

Dunning, L.A. and R. Kresman, 2013. Privacy preserving data sharing with anonymous ID assignment. Inf. Forensics Secur. IEEE. Trans., 8: 402-413.

Lakshmi, N.M. and H.R. KS, 2012. Privacy preserving association rule mining in horizontally partitioned databases using cryptography techniques. Int. J. Comput. Sci. Inf. Technol., 3: 3176-3182.

Li, T., N. Li, J. Zhang and I. Molloy, 2012a. Slicing: A new approach for privacy preserving data publishing. Knowl. Data Eng. IEEE. Trans., 24: 561-574.

Li, Y., M. Chen, Q. Li and W. Zhang, 2012b. Enabling multilevel trust in privacy preserving data mining. Knowl. Data Eng. IEEE. Trans., 24: 1598-1612.

Malik, M.B., M.A. Ghazi and R. Ali, 2012. Privacy preserving data mining techniques: current scenario and future prospects. Proceedings of the 2012 Third International Conference on Computer and Communication Technology (ICCCT), November 23-25, 2012, IEEE, Allahabad, India, ISBN: 978-1-4673-3149-4, pp: 26-32.

Mandapati, S., R.B. Bhogapathi and R.B. Chekka, 2013. A hybrid algorithm for privacy preserving in data mining. Int. J. Intell. Syst. Appl., 5: 47-53.

Peng, B., X. Geng and J. Zhang, 2010. Combined data distortion strategies for privacy-preserving data mining. Proceedings of the 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), August 20-22, 2010, IEEE, Chengdu, China, ISBN: 978-1-4244-6539-2, pp: 572-576.

Vassilios, S.V., B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004. State-of-the-art in privacy preserving data mining. SIGMOD Record, 33: 50-57.

Vijayarani, S., A. Tamilarasi and R. SeethaLakshmi, 2010. Privacy preserving data mining based on association rule-a survey. Proceedings of the 2010 International Conference on Communication and Computational Intelligence (INCOCCI), December 27-29, 2010, IEEE, Erode, India, ISBN: 978-81-8371-369-6, pp: 99-103.

Wang, J., Y. Luo, Y. Zhao and J. Le, 2009. A survey on privacy preserving data mining. Proceedings of the 2009 First International Workshop on Database Technology and Applications, April 25-26, 2009, IEEE, Wuhan, China, ISBN: 978-0-7695-3604-0, pp: 111-114.

Wang, S., L. Cui, J. Que, D.H. Choi and X. Jiang *et al*., 2012. A randomized response model for privacy preserving smart metering. Smart Grid IEEE. Trans., 3: 1317-1324.

Wang, W., B. Deng and Z. Li, 2007. Application of oblivious transfer protocol in distributed data mining with privacy-preserving. Proceedings of the First International Symposium on Data, Privacy and E-Commerce ISDPE, November 1-3, 2007, IEEE, Chengdu, China, ISBN: 978-0-7695-3016-1, pp: 283-285.

Wu, C.W., 2005. Privacy preserving data mining with unidirectional interaction. Proceedings of the IEEE International Symposium on Circuits and Systems ISCAS, May 23-26, 2005, IEEE, Yorktown Heights, New York, USA., ISBN: 0-7803-8834-8, pp: 5521-5524.

Yang, P., X. Gui, F. Tian, J. Yao and J. Lin, 2013. A privacy-preserving data obfuscation scheme used in data statistics and data mining. Proceedings of the 2013 IEEE 10th International Conference on High Performance Computing and Communications, Embedded and Ubiquitous Computing (HPCC_EUC), November 13-15, 2013, IEEE, Zhangjiajie, China, pp: 881-887.

Yun, L., L.X. Cheng and Z. Feng, 2010. Application of data mining in intrusion detection. Proceedings of the 2010 International Conference on Computer Application and System Modeling (ICCASM), October 22-24, 2010, IEEE, Taiyuan, China, ISBN: 978-1-4244-7235-2, pp: 153-155.

Zhang, X. and H. Bi, 2010. Research on privacy preserving classification data mining based on random perturbation. Proceedings of the 2010 International Conference on Information Networking and Automation (ICINA), October 18-19, 2010, IEEE, Kunming, China, ISBN: 978-1-4244-8104-0, pp: 173-178.

Zhang, X., C. Liu, S. Nepal and J. Chen, 2013. An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud. J. Comput. Syst. Sci., 79: 542-555.

Zhu, Y., L. Huang, W. Yang, D. Li, Y. Luo and F. Dong, 2009. Three new approaches to privacy-preserving add to multiply protocol and its application. Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining WKDD, January 23-25, 2009, IEEE, Moscow, Russia, ISBN: 978-0-7695-3543-2, pp: 554-558.