I.J. Modern Education and Computer Science, **2015, 5, 43-49**
Published Online May 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.05.06

Outlier Reduction using Hybrid Approach in Data Mining

Nancy Lekhi
Department of Information Technology, Chandigarh Engineering College Landran, India
Email: lekhi.nancy1@gmail.com

Manish Mahajan
Associate Professor, Department of Information Technology, Chandigarh Engineering College Landran, India
Email:cec.manish@gmail.com

Abstract—The Outlier detection is very active area of research in data mining where outlier is a mismatched data in dataset with respect to the other available data. In existing approaches the outlier detection done only on numeric dataset. For outlier detection if we use clustering method , then they mainly focus on those elements as outliers which are lying outside the clusters but it may possible that some of the unknown elements with any possible reasons became the part of the cluster so we have to concentrate on that also. The Proposed method uses hybrid approach to reduce the number of outliers. The number of outlier can only reduce by improving the cluster formulation method. The proposed method uses two data mining techniques for cluster formulation i.e. weighted k-means and neural network where weighted k-means is the clustering technique that can apply on text and date data set as well as numeric data set. Weighted k-means assign the weights to each element in dataset. The output of weighted k-means becomes the input for neural network where the neural network is the classification and clustering technique of data mining. Training is provided to the neural network and according to that neurons performed the testing. The neural network test the cluster formulated by weighted k-means to ensure that the clusters formulated by weighted k-means are group accordingly. There is lots of outlier detection methods present in data mining. The proposed method use Integrating Semantic Knowledge (SOF) for outlier detection. This method detects the semantic outlier where the semantic outlier is a data point that behaves differently with other data points in the same class or cluster. The main motive of this research work is to reduce the number of outliers by improving the cluster formulation methods so that outlier rate reduces and also to decrease the mean square error and improve the accuracy. The simulation result clearly shows that proposed method works pretty well as it significantly reduces the outlier.

Index Terms—Data Mining, Clustering, Weighted K-means, Neural Network, Outlier, and SOF

I. Introduction

There is large amount of data that available in Information Industry. This data is only useful if it converted into useful information. It is necessary to analyses this large amount of data and extracting useful information from it. Data Mining is a process of extracting the information from the huge amount of data [1]. Data mining has four major relationships. Clustering, Classification, Association, Sequential Pattern.

Cluster refers to a group of homogeneous types of objects. Cluster analysis refers to forming group of objects that are much related to each other but are greatly different from the objects in other clusters [2]. Classification is the process of discovery a model that represents the data classes or concepts. The purpose is to be able to use this model to anticipate the class of objects whose class label is unidentified. The main target of classification is to accurately calculate the value of each class variables. This classification process is divided into two steps. The first step is to build the model from the training set, i.e. randomly samples are selected from the data set. In the second step the data values are assigned to the model and verify the model's accuracy [3]. Association rules mining is to disclose the associations and relations surrounded by item sets of large data. Association rules mining is an essential part of data mining research, and association rules is the most classic approach of data mining [4]. Sequential pattern mining finds statistically suitable patterns between data. It is intimately relevant to time series mining and major case of structural data mining [5].

Outliers are impressions in data that do not comply with a well defined notion of normal behavior. Detecting outliers has essential application in data cleaning as well as in the mining of exceptional points for fraud detection, stock market analysis, intrusion detection, marketing, network sensors. Finding exceptional points among the data points is the vital aim to find out an outlier [6].

The paper is organized as follows: Section I present the introduction. In Section II related work is discussed. The clustering algorithms are discussed in Section III and Section IV. Section V presents the Outlier Detection Technique. The proposed approach with proper flow chat is given in Section VI. Section VII is for the results and discussion and Section VIII concludes the paper and

outlines the future work followed by references used in the work.

II. Related Work

K-means clustering algorithm improves by improving the initial focal point and determines the k values. Improved k-means reduce the impact of noise data in dataset improved k-means ensure that the clustering results are accurate [12].k-means works only on numeric data. New algorithm works well by modifying description of cluster center (by similarity weight and filter method) to overcome the limitation of k-means [13]. K-means algorithm cannot select variables automatically because k-means treat all variables equally in clustering process it result in poor clustering. New k-means type clustering algorithm called Weighted-k-means is introduced it can calculate variable weights automatically. But this algorithm is week to find the outlier [14]. New approach is introduced containing three methods that are clustering, pruning and computing outlier score. For clustering k-meanss algorithm is, based on some distance measure, points those are very near to center of each cluster are pruned. The points that are far from center are treated as outlier. Distance-based outlier methods have time and space complexities [15].I-CLARANS consisting three existing partition based clustering algorithms called PAM, CLARA and CLARANS also combines these algorithms with distance based method for outlier detection. This reduces computation time considerably. The I-CLARANS identifies outliers more successfully than existing algorithms [16].Hybrid approach contains cluster based approach and distance based approach. Hybrid approach reduces the time and space complexity. But there is not mention any impact on cost and performance by using hybrid technique [17]. New Approach introduced that enhanced the work by uses hybrid approach for outlier detection This method takes less computational cost and performs better than the distance based method. This Approach is only deals with numerical data not with text data or on mixture data and also performance of this approach is low [18]. The author discussed the new approach that provides the outlier detection and data clustering simultaneously. Author used two technique one is for clustering i.e. genetic k means and other for outlier detection i.e. outlier removal clustering. But this can work for large scale data of same type not for mixed type [19].

III. Weighted K-means

There are lots of methods to formulate the clusters in data mining. Weighted k-means is one of the clustering methods among them. New k-means type clustering algorithm called W-k-means can automatically calculate variable weights [7]. The variable weights produced by the algorithm measures the importance of the variable in clustering.

Algorithm

Input: n data and the number of cluster (*K*)
Output: *K* clusters
Begin

(i) Initialize the k cluster center
(ii) For loop until all data is processed

Randomly generate the weights for n number of data
And count distance from randomly generated centers

(iii) End for
(iv) Now divide the data having weights into k clusters.
End

IV. Neural Network

An Artificial Neural Network (ANN), often just called a neural network is a mathematical model or computational model based on biological neural networks. A Neural Network work on numeric values .A Neural Network Classifier is based on neural networks subsists of interconnected neurons [8]. A neuron takes positive and negative numerical values from other neurons and when the weighted sum of the numeric values is greater than a given threshold value, it activates itself. The output value of the neuron is usually a non-linear alteration of the sum of numeric values.

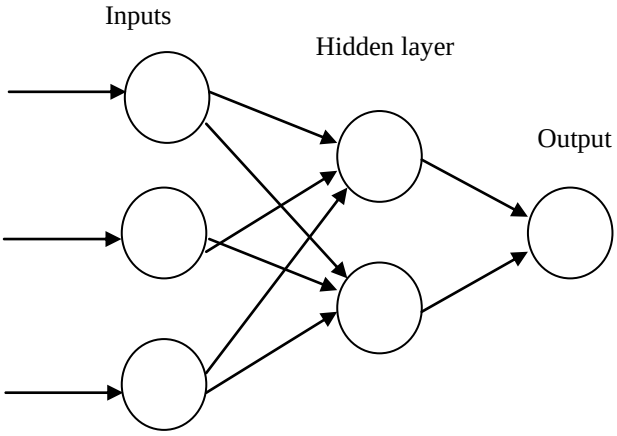

Fig. 1. Schematic representation of neural network

The layered Architecture of neural network is shown in Fig.1. Where first layer is input layer second is hidden layer and third layer of neural is output layer.

A Neural Network is defined by three types of parameters:

1. The interconnection or relative pattern between the different layers of neurons
2. The learning mechanism for updating the weights of the interconnections
3. The activation function that converts a neuron's weighted input to its output activation.

Algorithm

(i) Provide Input data(vector) to network
(ii) Create a Self-Organizing Map
(iii) Train the network
(iv) Test the network
(v) View the network

V. Outlier Detection Techniques

Outlier detection approach can be categorized into many approaches. Some of them are listed below

- Distribution based
- Distance based
- Depth based
- Cluster based
- Control chart technique
- Outlier detection integrating semantic knowledge

Outlier detection technique will be chosen based on type of data or type of outlier [9]. In this proposed work we are focusing on detecting the semantic outlier so that we choose Outlier detection integrating semantic knowledge.

A. Outlier Detection Integrating Semantic Knowledge

The records with the likewise class label should be identical with each other from the semantic knowledge that the people within group should have similar ideas [10].means that the elements in same group must have similar semantics.

Semantic outlier: A semantic outlier is a data point in the class which behaves differently with other points in same class.

VI. Proposed Work

The proposed work uses hybrid approach to formulate the cluster and Outlier Detection Integrating Semantic Knowledge to detect the semantic outliers in clusters (that formulate by hybrid approach).The main aim of proposed work is to made possible of outlier detection on text and compound data and also reduce the outlier by improving the cluster formulation rather than detecting and removing that outlier. Let us discuss the step by step procedure of proposed work.

A. Methodology-Step By Step Procedure

Step 1: Firstly initiate the dataset that contain mixed elements or data

Step 2: Apply weighted K-means on the dataset to formulate the clusters.

Weighted K-means Algorithm [11]

(i) Browse the data file that contains text, numbers and date.
(ii) According to the type of input elements divide them.
(iii) Calculate the weights using the values of (ii) step.
(iv) Assign the weights to each element.
(v) Calculate the average of weights.
(vi) Repeat the below steps until the all elements are processed
Check the element distance and according to that make the clusters
End loop
(vii) Save the output of Step (vi) in any 'xyz' file.

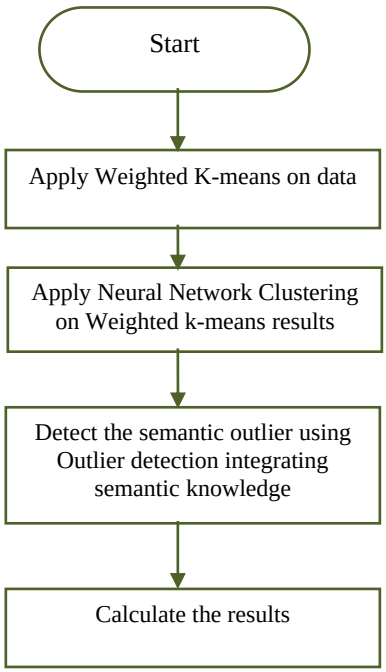

Fig. 2. Flow Chart of Proposed Work

Step 3: Clusters are created using weighted k-means now sort the cluster to easily find out the outlier if any and save the results of weighted k-means in any 'xyz' file.

Step 4: After sorting call the classifier (here output of clustering algorithm become input of classifier i.e. neural network) neural network in which we define training dataset called P and testing dataset called T.

Neural Network Algorithm

(i) load 'xyz' file
(ii) inputs = 'xyz' dataset
(iii) dimension1 = 10
(iv) dimension2 = 10
(v) net = selforgmap([dimension1 dimension2])
(vi) [net, tr] = train(net, inputs)
(vii) outputs = net(inputs)
(viii) view(net)

Step 5: Now check the Semantic outlier using Outlier Detection Integrating Semantic Knowledge from the result that is obtained from above hybrid technique (combination of weighted k-means + neural networks).

Outlier Detection Integrating Semantic Knowledge

(i) Check If there is two cluster present or not

(ii) If present then fetch each element of cluster1 one by one
(iii) Perform matching , the element that does not match to other element in same cluster treat that one as outlier
(iv) Count the outlier for cluster one and store that count in any x variable.
(v) Now pick second cluster cluster2 and fetch the element one by one
(vi) Perform step (iii) & (iv) now store the result of second cluster in y variable. And add x and y to calculate the total outlier found in clusters.

Step 6: Calculate the results and compare the proposed work results with existing work results.

VI. Results And Discussion

The simulation result shows that the proposed method performs better than that of existing method i.e. Genetic K-means. The Mean Square Error in proposed method is less as compare to that of existing method Firstly we show some of the results that obtained by proposed work in form of clusters and then we will discuss the final outcome that shows how effective the proposed method is. Then we will compare only numeric dataset results of proposed work with existing work because Genetic K-means works only upon numeric data.

Let us discuss the cluster results first that is obtained after applying the weighted k-means on numeric, text and date data containing 3000 record each.

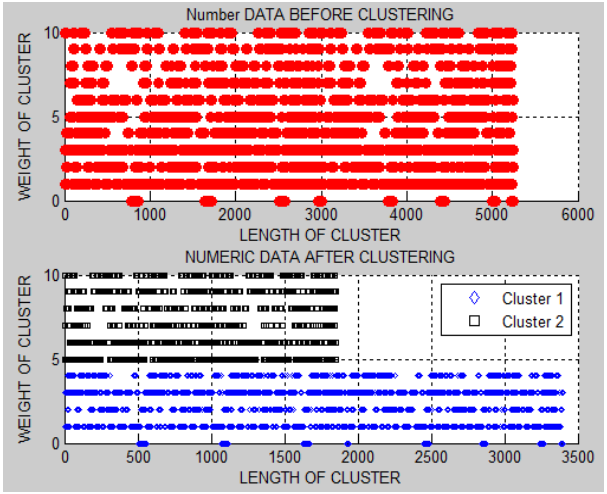

Fig. 3. Weighted k-means results for numeric data having 5250 records

After applying the weighted k-means on numeric, text and date data (5250 records) the data is divided into two clusters the red marks shown in Fig.3, Fig.4 and Fig.5 represent the total numeric, text and date data and blue and black marks represent the data divided into two clusters after apply clustering algorithm. Fig.3, Fig.4 and Fig.5 shows the results only for 5250 records, same as weighted k-means are also applied to 3000, 7500 records.

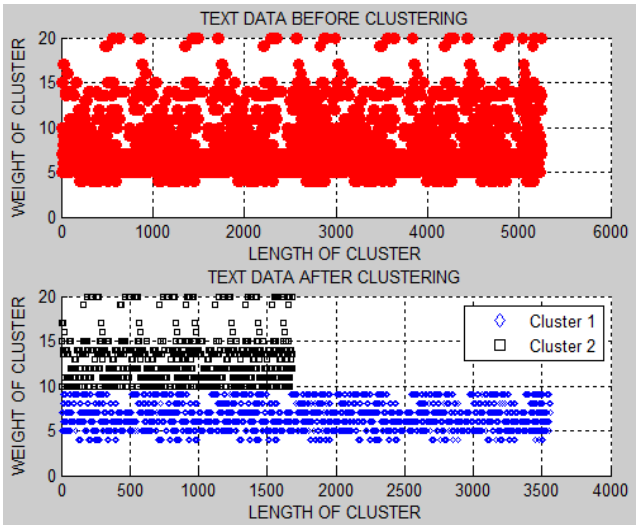

Fig. 4. Weighted k-means results for text data having 5250 records

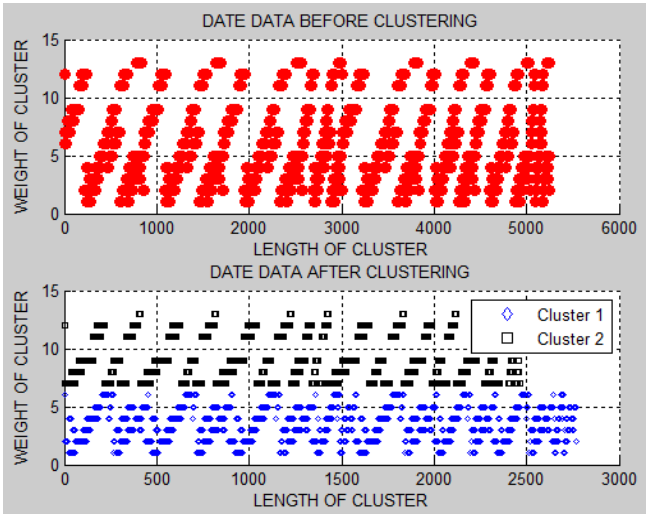

Fig. 5. Weighted k-means results for date data having 5250 records

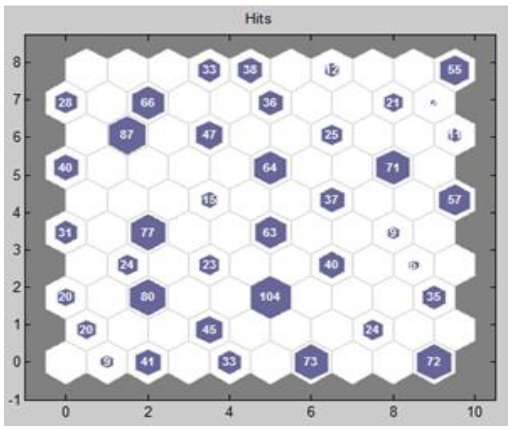

Fig. 6. SOM Hits

Fig.6. shows one of the neural output i.e. SOM hits (numbers of hits on neurons) for date data having 3000 records. Same as above neural result for date data containing 3000 record, the neural also generates the results for numeric and text data.

Table 1. Outlier Rate obtained from three different types of data

	Numeric data	Text data	Date/Time Data
3000 records	1.5941 %	3.8576 %	0.708 %
5250 records	2.8235 %	2.7209 %	0.771 %
7500 records	0. 3789 %	2.8132 %	0.697 %

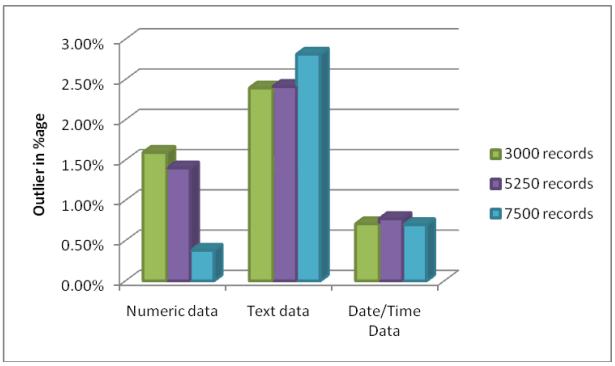

Fig. 7. Outlier Result

Fig.7. shows graphically the outlier rate that obtained from numeric data, text data and date data in percentage that mention in Table.1.

Table 2. Elapsed time to formulate the clusters

	Numeric data	Text data	Date/Time Data
3000 records	1.8643 s	0.86384 s	2.0295 s
5250 records	2.8054 s	1.1613 s	1.6391 s
7500 records	2.5551 s	1.4185s	7.8038 s

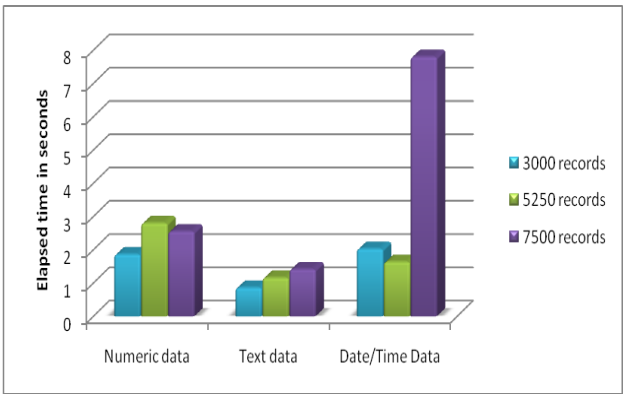

Fig. 8. Elapsed time

Fig.8. shows graphically the cluster formulation time for numeric data, text data and date data in seconds that mention in Table.2.

Table 3. Describe accuracy of proposed algorithm

	Numeric data	Text data	Date/Time Data
3000 records	93.1677 %	93.36 %	95.29 %
5250 records	95.97 %	96.36 %	94.76 %
7500 records	97.15 %	97.42 %	97.40 %

Fig. 9. Accuracy Result

Fig.9. shows graphically how accurately the proposed algorithm work for numeric data, text data and date data in percentage that mention in Table.3.

Table 4. Mean Square Error

	Numeric data	Text data	Date/Time Data
3000 records	9.64	19.35	22.36
5250 records	10.90	18.23	20.96
7500 records	12.16	18.05	20.93

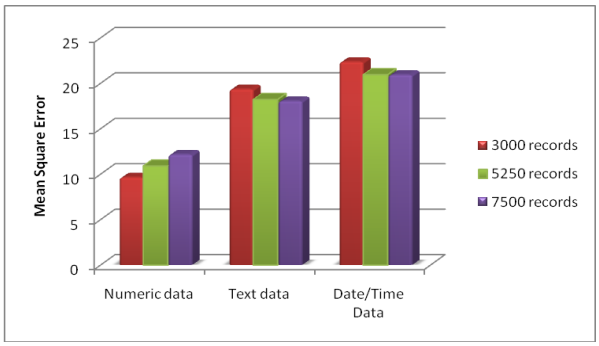

Fig. 10. Mean Square Error Result

Fig.10. shows graphically Mean Square Error (the average of the squares of the "errors") that mention in Table.4.

Table 5. Compare the cluster formulation time of existing and proposed method

	3000 records	5250 records	7500 records
Genetic K-means	579.7287 s	1006.7793 s	1592.22 s
Proposed Method	1.8643 s	2.8054 s	2.5551 s

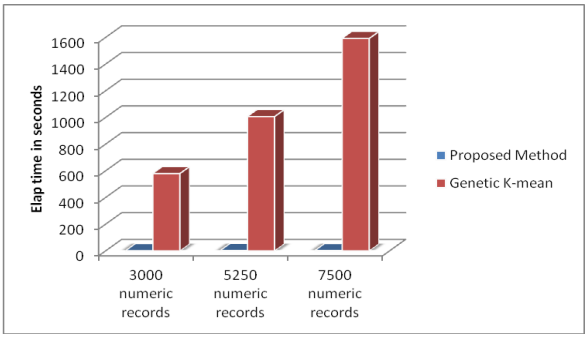

Fig. 11. Elapsed time Result comparison

Fig.11. shows comparison graphically of the total time taken to formulate the clusters by existing approach with proposed approach that mention in Table.4.

Table 6. Compare the Mean square error result of existing and proposed method

	3000 records	5250 records	7500 records
Genetic K-mean	14.77	26.37	800.69
Proposed Method	9.64	10.90	12.16

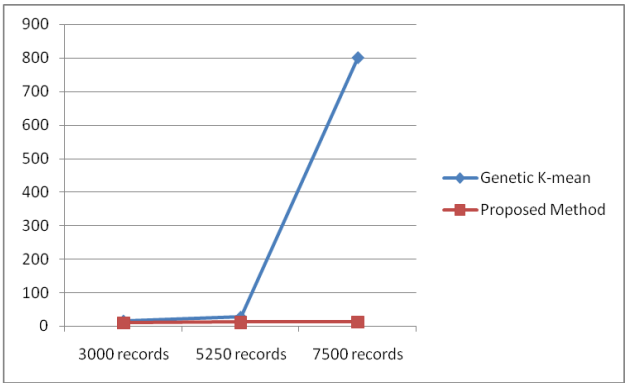

Fig. 12. Mean Square Error Result comparison

Fig.12. shows comparison of the average of the squares of the errors obtained by existing approach with proposed approach graphically that mention in Table.6.

Table 7. Compare Outlier Results of numeric data

	3000 records	5250 records	7500 records
Genetic K-means	2.5279 %	5.6183 %	1.0766 %
Proposed Method	1.5941 %	2.8235 %	0.3789 %

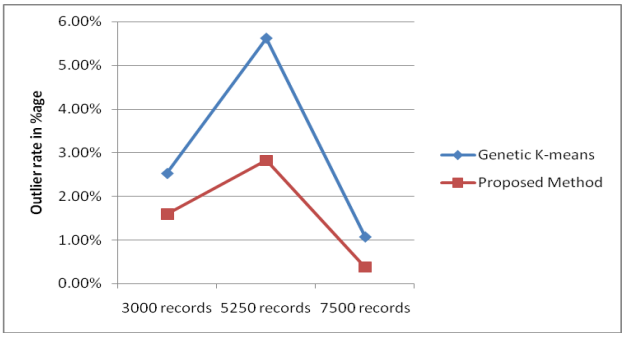

Fig. 13. Outlier rate comparison

Fig.13 shows graphically the outlier rate for existing method and proposed method that mention in Table.7. The graph shows that the outlier (unknown data) detected very less in the proposed work.

VII. Conclusion And Future Work

There are lots of methods for detecting the outlier in data mining. Every one mostly focuses on trying different-different techniques to detect the outlier for better results. The proposed method focuses on the reduction of outlier. The proposed work uses the hybrid Approach to formulate the clusters very effectively so that the number of outliers becomes reduced. The simulation results showed that the proposed algorithm performs better than that of genetic k-means. This proposed method deals with text and date dataset that has not been implemented before using genetic k-means on the text and date dataset but rather performed on numeric dataset. The proposed method takes very less time as compare with the existing method. The outlier rate in proposed method is also reducing as compare with the existing method. The mean square error is also reducing as comparing to that of existing method.

The future work requires modifications that can make applicable for dataset that contains multiple symbols as well as text and date. The approach needs to be implemented on more complex dataset (dataset contains at least 20,000 records) and also focus on reduction of mean square error.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine* Volume 17 Number 3 (1996).

[2] Dr. Sankar Rajagopal , " CUSTOMER DATA CLUSTERING USING DATA MINING TECHNIQUE", *International Journal of Database Management Systems (IJDMS)* Vol.3, No.4, November 2011.

[3] Samir Kumar Sarangi, Dr. Vivek Jaglan, Yajnaseni Dash, "A Review of Clustering and Classification Techniques in Data Mining", *In ternational Journal of Engineering, Business and Enterprise Applications (IJEBEA)* pp 140-145, 2013.

[4] Yabing Jiao, "Research of an Improved Apriori Algorithm in Data Mining Association Rule", *International Journal of Computer and Communication Engineering*, Vol. 2, No. 1, January 2013.

[5] V. Uma, M. Kalaivany, G. Aghila, "Survey of Sequential Pattern Mining Algorithms and an Extension to Time Interval Based Mining Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 12, December 2013.

[6] Ms. S. D. Pachgade, Ms. S. S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 6, June 2012.

[7] De Amorim, R.C. , "Constrained clustering with Minkowski Weighted K-meanss " *Computational Intelligence and Informatics (CINTI), 2012 IEEE* 13th International Symposium on 20-22 Nov. 2012.

[8] K. Amarendra, K.V. Lakshmi & K.V. Ramani , "Research on Data Mining Using Neural Networks" , *Special Issue of International Journal of Computer Science & Informatics (IJCSI)*, ISSN (PRINT) : 2231–5292, Vol.- II, Issue-1, 2.

[9] M. O. Mansur, Mohd. Noor Md. Sap, "Outlier Detection Technique in Data Mining: A Research Perspective", Proceedings *of the Postgraduate Annual Research Seminar* 2005.

[10] Jason J. Jung, Geun-Sik J, Semantic Outlier Analysis for Sessionizing Web Logs.

[11] Nancy Lekhi, Manish Mahajan, "Improving Cluster Formulation to Reduce Outliers in Data Mining",

International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 6, June 2014.

[12] Dharmendra S. Modha and W. Scott Spangler, "FeatureWeighting in k-meanss Clustering", *2002 Kluwer Academic Publishers. Printed in the Netherlands, Machine Learning*, Vol. 47, 2002.

[13] Reddy M. V. Jagannatha and B. Kavitha, " Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method", *International Journal of Database Theory and Application* Vol. 5, No. 1, March, 2012.

[14] Anand M. Baswade, Kalpana D. Joshi , Prakash S. Nalwade, "A Comparative Study Of K-meanss And Weighted K-meanss For Clustering", *International Journal of Engineering Research & Technology (IJERT)* Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.

[15] Pranjali Kasture, Jayant Gadge, "Cluster based Outlier Detection", *International Journal of Computer Applications (0975 – 8887)* Volume 58– No.10, November 2012.

[16] Garima Singh, Vijay Kumar, "An Efficient Clustering and Distance Based Approach for Outlier Detection", *International Journal of Computer Trends and Technology (IJCTT)* – volume 4 Issue 7–July 2013.

[17] Surekha V Peshatwar & Snehlata Dongre, "Outlier Detection Over Data Stream Using Cluster Based Approach And Distance Based Approach", *International Conference on Electrical Engineering and Computer Science (ICEECS-2012),*Trivendum May 12th, 2012.

[18] Ms. S. D. Pachgade, Ms. S. S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 6, June 2012.

[19] M. H. Marghny , Ahmed I. Taloba, " Outlier Detection using Improved Genetic K-meanss", *IEEE TRANSACTIONS ON COMMUNICATIONS*, VOL. 38, NO. 11, July 2013.

Authors' Profiles

Nancy Lekhi: Btech Information Technology from Baba Banda Singh Bahadur engineering college, Fatehgarh Sahib in 2012. Currently pursuing M.Tech (IT) from Chandigarh Engineering College,Landran interested in Data Mining and Neural Network.

Manish Mahajan: working as an Associate professor in the Department of Information Technology, CGC Landran, Punjab India. He has completed his M-Tech in CSE from PTU and now pursuing Ph.D. from PTU, Punjab. His research interests includes Digital Image Processing, Data Mining, Steganography, Data Security.