ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Enhanced K-Means with Greedy Algorithm for Outlier Detection

C.Sumithiradevi*
Research Scholar in
Computer science , Bharathiar University
Coimbatore, India
sumithradevic@yahoo.co.in

Dr.M.Punithavalli
Director –MCA
Sri Ramakrishna Engineering College
Coimbatore, India
mpunitha_srcw@yahoo.co.in

Abstract: Due to significant development in information technology, larger and huge volumes of data are accumulated in databases. In order to make the most out of this huge collection, well-organized and effective analysis techniques are essential that can obtain non-trivial, valid, and constructive information. Organizing data into valid groupings is one of the most basic ways of understanding and learning. Cluster analysis is the technique of grouping or clustering objects based on the measured or perceived fundamental features or similarity. The main objective of clustering is to discover structure in data and hence it is exploratory in nature. But the major risk for clustering approaches is to handle the outliers. Outliers occur because of the mechanical faults, any transformation in system behavior, fraudulent behavior, human fault, instrument mistake or any form of natural deviations. Outlier detection is a fundamental part of data mining and has huge attention from the research community recently. In this paper, the standard K-Means technique is enhanced using the Greedy algorithm for effective detection and removal of outliers (EKMOD). Experiments on iris dataset revealed that EKMOD automatically detect and remove outliers, and thus help in increasing the clustering accuracy. Moreover, the Means Squared Error and execution time is very less for the proposed EKMOD.

Keywords: K-Means, Greedy Algorithm, Outlier Detection.

I. INTRODUCTION

Data mining deals with the detection of nontrivial, unseen and interesting information from several types of data. Due to the continuous growth of information technologies, there is huge increase in the number of databases, in addition to their dimension and difficulty. An automated technique is essential to analyze this huge amount of information [1]. The analysis results can be used for making a decision by a human or program.

Data clustering has been extensively utilized for the following three major purposes [2].

a. Underlying structure: To expand insight into data, produce hypotheses, identify anomalies and recognize salient features.
b. Natural classification: To recognize the degree of similarity among forms or organisms.
c. Compression: As a technique for organizing the data and summarizing it through cluster prototypes.

One of the fundamental difficulties in data mining is the outlier detection. Clustering is a significant tool for outlier analysis [3-5]. Outliers are the collection of objects that are significantly unrelated from the remainder of the data [6]. Outlier detection is a very important difficulty with a direct application in an extensive variety of application domains, together with fraud detection [7], recognizing computer network intrusions and bottlenecks [8], illegal activities in e-commerce and detecting mistrustful activities [9, [10].

Many data-mining approaches discover outliers as a side-product of clustering techniques. On the other hand these approaches characterize outliers as points, which do not fit inside the clusters. As a result, the techniques unconditionally characterize outliers as the background noise in which the clusters are surrounded. Another class of techniques characterized outliers as points, which are neither a division of a cluster nor a division of the background noise; relatively they are specifically points which behave in a different way from the standard.

The difficulty in outlier detection in some cases is comparable to the classification problem. For instance, the major concern of clustering-dependent outlier detection approaches is to discover clusters and outliers, which are typically considered as noise that should be eradicated with the purpose of making more consistent clustering [11]. Few noisy points possibly will be distant from the data points, while the others might be nearer.

The distant noisy points would influence the result more considerably since they are more dissimilar from the data points. It is necessary to recognize and eliminate the outliers, which are distant from all the other points in cluster. Therefore, in order to enhance the clustering accuracy, a perfect clustering approach is necessary that should detect and remove these outliers.

II. RELATED WORKS

Outlier detection is used widely in various fields. The theme about the outlier factor of an object is unlimited to the case of cluster. Based on this factor of the cluster, a clustering-based outlier detection method, which is named as CBOD, is projected by Sheng-yizJiang and Qing-bo An [12]. This technique constitutes of two levels, the first level is cluster dataset by one-pass clustering algorithm and second level find out outlier cluster by outlier factor. The time difficulty of CBOD is almost linear with the amount of dataset and the number of attributes that ends in good scalability and become accustomed to huge datasets.

Eliminating the objects that are noisy is one of the major goal of data cleaning as noise delays most type of data analysis. Mostly used data cleaning techniques focuses on eliminating noise that is the product of low-level data errors that results from an imperfect data collection method, but data objects which are not related or only weakly related can also considerably hold back on data analysis. Therefore, if the goal is to improve the data analysis to the extent that is possible, these objects must also be considered as noise, at least with respect to the underlying analysis. As a result, there is a need for data cleaning techniques that eliminate both types of noise. Since data sets can include huge amount of noise, these methods also need to be able to remove a potentially large fraction of the data. Xiong et al., [13] discovered four methods projected for noise removal to improve data analysis in the occurrence of high noise levels. Three of the methods are based on usual outlier detection techniques: distance-based, clustering-based, and an approach based on the local outlier factor (LOF) of an object.

The other technique, that is a new method that is projected, is a hyperclique-based data cleaner (HCleaner). These techniques are examined based on the terms of their contact on the subsequent data analysis, specially, clustering and association analysis.

The idea about outlier factor of an object is extended to the case of cluster. Outlier factor of cluster determine the difference degree of a cluster from the entire dataset and two outlier factor definitions are projected by Sheng-Yi Jiang and Ai-Min Yang [14]. A framework of clustering-based outlier detection, called as FCBOD, is suggested. This framework contains two stages, the initial stage cluster dataset and the next stage determine outlier cluster by outlier factor. The time difficulty of FCBOD is almost similar with respect to both size of dataset and the number of attributes.

III. METHODOLOGY

Clustering approaches can be largely segmented into two divisions: hierarchical and partitional. Hierarchical clustering approaches recursively discover nested clusters either in agglomerative mode (initializing with each data point in its individual cluster and integrating the most similar pair of clusters consecutively to generate a cluster hierarchy) or in divisive (topdown) mode (initializing with all the data points in one cluster and recursively dividing each cluster into smaller clusters). Dissimilar to hierarchical clustering approaches, partitional clustering approaches discover all the clusters at the same time as a partition of the data and do not enforce a hierarchical structure [2].

The most recognized hierarchical approaches are single-link and complete-link; the extensively used and the simplest partitional approach is K-Means. Because partitional approaches are widely used in pattern recognition owing to its nature of available data. K-Means has a wealthy and diverse history as it was separately discovered in several scientific fields.

A. K-Means Algorithm:

Consider $X = \{x_i\}, i = 1, \ldots, n$ is a set of n d-dimensional points to be clustered into a set of K clusters, $C = \{c_k, k = 1, \ldots, K\}$. K-Means algorithm discovers a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is reduced. Consider μ_k be the mean of cluster c_k. The squared error between μ_k and the points in cluster c_k is given as

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

The main objective of K-Means is to reduce the sum of the squared error over all K clusters,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

Reducing this objective function is recognized to be an NP-hard problem (even for K = 2) [15]. As a result K-Means, which is a greedy algorithm, can only converge to a local minimum, although current study has shown with a large probability K-Means could converge to the global optimum when clusters are well separated [16]. K-Means begins with an initial partition with K clusters and allocate patterns to clusters in an attempt to lessen the squared error. While the squared error constantly decrease with an increase in the number of clusters K (with $J(C) = 0$ when $K = n$), it can be reduced only for a constant number of clusters. The major steps of K-Means algorithm are as follows.

Choose an initial partition with K clusters; reiterate steps 2 and 3 until cluster membership becomes constant.

Produce a new partition by assigning each pattern to its closest cluster center.

Generate new cluster centers.

Features of the data streams include their huge volume and potentially unrestrained size, sequential access and dynamically evolving nature. This enforces further necessities to conventional clustering approaches to quickly process and sum up the enormous amount of constantly arriving data. It also necessitates the capability of adapting to changes in the data distribution, the capability of detecting emerging clusters and differentiate them from outliers in the data and the capability of incorporating old clusters or remove expired ones. All of these necessities make data stream clustering a considerable challenge. Hence in order to detect and remove the outliers, K-Means is enhanced by integrating it with Greedy Algorithm and proposed EKMOD.

B. Enhanced K-Means with Greedy Algorithm for Outlier Detection (EKMOD):

In this section, proposed Enhanced K-Means with Greedy Algorithm for Outlier Detection (EKMOD), which is efficient and has the potential to identify and remove the outliers.

This approach takes the number of preferred outliers (consider it to be k) as input and selects points as outliers in a greedy approach. At first, the set of outliers (represented by OS) is specified to be empty and all points are represented as non-outlier. Subsequently, k scans are necessary over the dataset to choose k points as outliers. In every scan, for each point labeled as non-outlier, it is temporally removed from the dataset as outlier and the entropy object is re-evaluated. A point that accomplishes maximal entropy impact, i.e., the maximal reduction in entropy experienced by removing this point, is taken as outlier in current scan and accumulated in OS. The algorithm ends when the size of OS reaches k.

Figure 1 shows the greedy algorithm of EKMOD. The collection of records is stored in a file on the disk and each record t is read in sequence.

During the initialization phase of the greedy algorithm, each record is represented as non-outlier and hash tables for attributes are also built and updated.

In the greedy procedure, the dataset is scanned for k times to discover exact k outliers, that is, one outlier is found and removed in each pass. In every scan over dataset, read each record t that is represented as non-outlier, its label is changed to outlier and the changed entropy value is calculated. A record that accomplishes maximal entropy impact is chosen as outlier in current scan and accumulated to the set of outliers.

In this approach, the vital step is computing the transformed value of entropy. In the following theorem, the decreased entropy value is only reliant on the attribute values of the record to be temporally eliminated.

```
Algorithm outlier detection Algorithm
Input: D // the categorical database
k // the number of desired outliers
Output: k identified outliers
/* Phase 1-initialization */
Begin
for each record t in D
update hash tables using t
label t as a non-outlier with flag "0"
/* Phase 2-Greedy Procedure */
counter = 0
Repeat
counter++
while not end of the database do
read next record t which is labeled "0" //non-outlier
compute the decrease on entropy value by labeling t as outlier
if maximal decrease on entropy is achieved by record b then
update hash tables using b
label b as a outlier with flag "1"
Until counter = k
End
```

Figure 1: Algorithm for outlier detection

IV. EXPERIMENTAL RESULTS

To evaluate the Enhanced K-Means with Greedy Algorithm for Outlier Detection (EKMOD), experiments were carried out using University of California, Irvine (UCI) Machine Learning Repository [17]. For the purpose of evaluating the proposed technique iris dataset [18] is used and the results are compared with standard K-Means and Semi-Supervised K-Means Clustering for Outlier Detection (SKOD).

Clustering results are generated using Standard K-Means, SKOD and the proposed EKMOD for the iris dataset. The performance of the proposed EKMOD scheme is evaluated against the Standard K-Means, SKOD based on the following parameters.

a. Clustering Accuracy,
b. Mean Squared Error and
c. Execution Time

A. Clustering Accuracy:

Since the outliers are detected and removed using the proposed EKMOD, clustering accuracy is drastically increased. Table I shows the comparison of the accuracy of clustering accuracy for the proposed method with the standard K-Means and SKOD method.

Table 1: Comparison of Clustering Accuracy in Iris Dataset

Clustering Technique	Clustering Accuracy (%)
Standard K-Means	89.80
SKOD	93.25
EKMOD	97.23

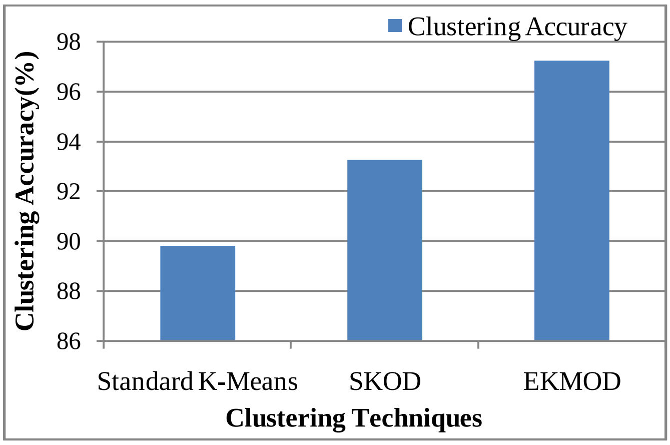

Figure 2: Comparison of Clustering Accuracy in Iris Dataset

From the Figure 2, it can be observed that the accuracy of clustering result using standard K-Means and SKOD method is 89.80 % and 93.25% respectively and that of the proposed EKMOD is 97.23% for iris dataset.

B. Mean Squared Error (MSE):

As mentioned above the formula for MSE is

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \left\| x_i - \mu_k \right\|^2$$

MSE of the iris dataset for the two cluster centers of the three methods are provided in table II.

Table 2: Comparison of Mean Squared Error in Iris Dataset

Cluster	Standard K-Means	SKOD	EKMOD
Cluster 1	0.6923	0.6012	0.4325
Cluster 2	0.5256	0.4706	0.3029

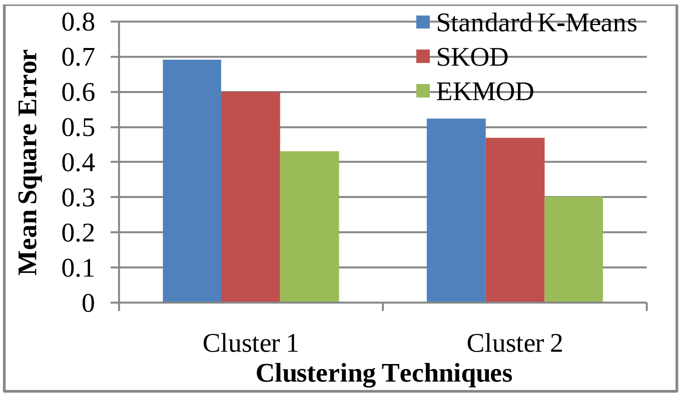

Figure 3: Comparison of Mean Squared Error in Iris Dataset

From figure 3, it is observed that the proposed EKMOD gives very low MSE values for both the clusters (0.4325 and 0.3029) than the Standard K-Means (0.6923 and 0.5256) and SKOD (0.6012 and 0.4706).

C. Execution Time:

The execution time is calculated based on the machine time (i.e., the time taken by the machine to run the proposed algorithm).

Table 3: Comparison of Execution Time in Iris Dataset

Clustering Technique	Execution Time (Sec)
Standard K-Means	2.31
SKOD	1.45
EKMOD	0.92

Table 4 shows the execution time taken by the Standard K-Means, SKOD and the proposed EKMOD in iris dataset. It can be observed that the time required for execution using the proposed EKMOD scheme for iris dataset is 0.91 seconds, whereas more time is needed by other two clustering techniques for execution.

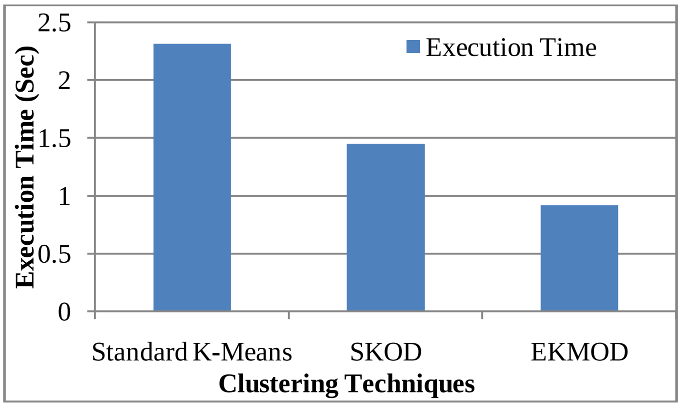

Figure 4: Comparison of Execution Time in Iris Dataset

From figure 4.3, it is observed that the proposed EKMOD takes very low execution time when compared with the Standard K-Means and SKOD which takes 2.31 and 1.45 seconds respectively in iris dataset.

V. CONCLUSION

K-Means is one of the standard clustering approaches which is widely used in several applications. The major concern in this clustering approach is that detection and removal of outliers. Outlier detection is an essential subject in data mining, particularly it has been extensively utilized to identify and eliminate anomalous or irrelevant objects from data cluster. In this paper, proposed an Enhanced K-Means with Greedy Algorithm for Outlier Detection (EKMOD) which uses greedy algorithm to identify and remove the outliers in the clusters. The effectiveness of the proposed approach is tested using the iris dataset based on the clustering accuracy, MSE and execution time. From the results, it is revealed that the proposed EKMOD provides the very accurate cluster results with low MSE. Moreover the execution time of this approach is very low when compared to other two clustering approaches.

VI. REFERENCES

[1] M. H. Marghny and Ahmed I. Taloba, “Outlier Detection using Improved Genetic K-means”, International Journal of Computer Applications, Vol. 28, No. 11, Pp. 33-36, 2011.

[2] Anil K. Jain, “Data Clustering: 50 Years Beyond K-Means”, Pattern Recognition Letters, 2009.

[3] Jain, A. and R. Dubes, “Algorithms for Clustering Data”, Prentice-Hall, 1988.

[4] Loureiro, A., L. Torgo and C. Soares, “Outlier Detection using Clustering Methods: a Data Cleaning Application”, Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany, 2004.

[5] Niu, K., C. Huang, S. Zhang, and J. Chen, “ODDC: Outlier Detection Using Distance Distribution Clustering”, T. Washio et al. (Eds.): PAKDD 2007 Workshops, Lecture Notes in Artificial Intelligence (LNAI) 4819, Pp. 332–343, Springer-Verlag, 2007.

[6] Han, J. and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2nd ed, 2006.

[7] Bolton, R. and D. J. Hand, “Statistical Fraud Detection: A Review”, Statistical Science, Vol. 17, No. 3, Pp. 235-255, 2002.

[8] Lane, T. and C. E. Brodley, “Temporal Sequence Learning and Data Reduction for Anomaly Detection”, ACM Transactions on Information and System Security, Vil. 2, No. 3, Pp. 295-331, 1999.

[9] Chiu, A. and A. Fu, “Enhancement on Local Outlier Detection”, 7th International Database Engineering and Application Symposium (IDEAS03), Pp. 298-307, 2003.

[10] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufman Publishers, 2000.

[11] Z. He, X. Xu and S. Deng, “Discovering cluster-based Local Outliers”, Pattern Recognition Letters, Vol. 24, No. 9-10,
Pp. 1641–1650, 2003

[12] Sheng-yizJiang and Qing-bo An, “Clustering-Based Outlier Detection Method”, Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '08), Vol. 2, Pp. 429–433, 2008.

[13] Xiong, H.; Gaurav Pandey; Steinbach, M.; Vipin Kumar; “Enhancing data analysis with noise removal”, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3, Pp. 304–319, 2006.

[14] Sheng-Yi Jiang and Ai-Min Yang; “Framework of Clustering-Based Outlier Detection”, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 1, Pp. 475–479, 2009.

[15] Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V, “Clustering large graphs via the singular value decomposition”, Machine Learning, Vol. 56, No. 1-3, Pp. 9–33, 1999.

[16] Meila, Marina. “The uniqueness of a good optimum for k-means”, Proceedings of the 23rd International Conference on Machine Learning, Pp. 625–632, 2006.

[17] http://www.uci.edu/

[18] http://archive.ics.uci.edu/ml/datasets/Iris