Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Outlier Detection Using a New Hybrid Approach Based on Group Weighted K-Mean and Greedy Method on Mixed Dataset

Navneet Kaur
Computer Science, SGGSWU, Fatehgarh Sahib,
Punjab, India

Abstract*— Outlier detection is currently very active area of research in data mining. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in dataset. Proposed method for outlier detection uses hybrid approach. Purpose of approach is first to apply the clustering algorithm that is group weighted k-means (GWK-Mean) which partition the dataset into number of groups and second using greedy algorithm for detect outliers. The principal of outliers finding depend on the threshold. Threshold is set by user. Entropy is also used to measure of disorder present in a system. The hybrid approach, two techniques are combined to improve efficiently find the outlier from the dataset.

Keywords*— Outlier, threshold, entropy, execution time, greedy algorithm, and group weighted k-mean.

I. INTRODUCTION

Data mining is a process of extracting hidden and useful information from the data. Finding outlier is an important task in data mining. An outlier is data object that is different from the remaining dataset. According to Hawkins (1980), "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". An outlier definition (Barnett and Lewis, 1994) is:"An observation which appears to be inconsistent with the remainder of that set of data ".

John (1995) states that an outlier may also be "surprising veridical data", a point belonging to class A but actually situated inside class B so the true (veridical) classification of the point is surprising to the observer.

Aggarwal and Yu [1] notes that "outliers may be considered as noise points lying outside a set of defined clusters or alternatively outliers may be defined as the points that lie outside of the set of clusters but are also separated from the noise." These outliers behave differently from the norm.

Detecting outliers has important applications in data cleaning as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing, network sensors. In this paper, we have proposed a novel hybrid approach to solve the problem mixed attributes. First the original mixed dataset is divided into sub groups, the integer, text, custom. Second phase is detecting the outliers.

II. OBJECTIVES OF STUDY

Basic aims to detect the outliers, to let user free to provide sensitive parameters. In that we are first partition the data into groups with weights. Second step is using greedy algorithm to detect the outlier. The principle of outlier's detection depends on the threshold. This approach takes less computational time.

III. RELATED WORK

Outlier detection has been very interesting topic for research community. Several clustering approaches were used and implemented for detection of outliers from a data set. These approaches can be broadly classified into several major ideas:

Distribution-based approach [6,7]develop statistical models from given data and then apply statistical test to determine if an object belongs to this model or not.

Disadvantage**:** Distribution-based approaches cannot be applied in multidimensional sceneries.

Distance-based approach [8,9] this approach has been originally proposed by Knorr and Ng. Given any distance measure, objects that have distances to their nearest neighbors that exceed a specific threshold are considered potential anomalies.

Density-based approach (Breunig, 2000; Papadimitriou, 2003) compute the density of regions in the data and declare the objects in low dense regions as outliers. [8,9]

Disadvantage: Density based models require the careful settings of several parameters.

It requires quadratic time complexity.

It may rule out outliers close to some non-outliers patterns that has low density.

Cluster based approach (Loureiro, 2004; Gath and Geva, 1989; Cutsem and Gath, 1993; Jiang, 2001; Acuna and Rodriguez, 2004), consider clusters of small sizes as clustered outliers.[9,10]

K-Nearest neighbor based approach [9,10]: K-nearest neighbor based schemes analyses each object with respect to its local neighborhood. The basic idea behind such schemes is that an outlier will have a neighborhood where it will stand out, while a normal object will have a neighborhood where all its neighbors will be exactly like it.
The k-means clustering algorithm is used. As mentioned, the k-means is sensitive to outliers, and hence may not give accurate results.

IV. PROPOSED CLUSTERING ALGORITHM

The Hybrid Approach a new method for outlier detection. The proposed method is using the advantages of existing outlier detection algorithms that are group weighted K-Mean and greedy algorithms. Purpose of new hybrid approach is first to apply the clustering algorithm that is GWK-Mean which partition the dataset into number of groups and second using greedy algorithm for detect outliers. The principal of outliers finding depend on the threshold. Threshold is set by user. Entropy is also used to measure of disorder present in a system. The problem of this research work is to find out the outliers using a new hybrid approach on mixed type datasets and to verify the performance of existing clustering algorithms.

4.1Following procedure steps are used in Hybrid Approach

```
Input –D dataset.
Divide data into groups
calculate weights of each groups and update hash table
calculate threshold
counter= 0
while ( counter!= k) do
counter++
      while ( not end of database ) do
      read next record x labeled non-outlier label x as outlier
if ( current mode>threshold than data is classified and flag with 0)
otherwise (current<threshold than data is outlier and flag with 1)
end if
```

Figure: 4.1 Procedure Steps Are Used In Hybrid Approach

4.2 System Architecture OF Hybrid Approach:

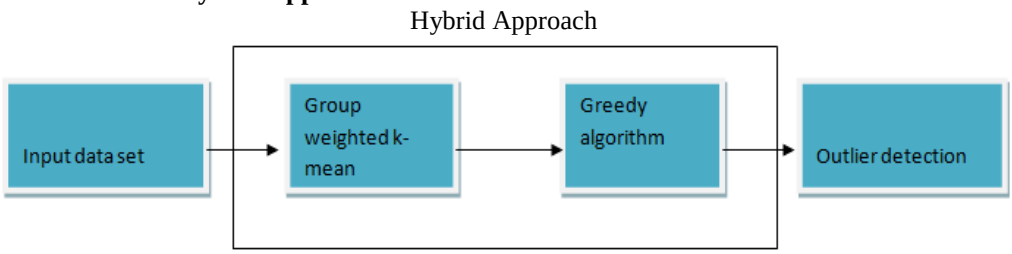

Figure 4.1 System Architecture OF Hybrid Approach

Input Data Set: Collecting dataset from UCI Machine learning repository.
Group Weighted K-Mean: In this algorithm, the variables of the high dimensional data can be divided into several variable groups and a group weight is assigned to each variable group to identify the importance of the variable group.[14]
Greedy algorithm: In the greedy procedure, the dataset is scanned for k times to discover exact outliers, that is, one outlier is found and removed in each pass. In every scan over dataset, read each record 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.[16]
Outlier Detection: Outlier detection is a fundamental part of data mining and has huge attention from the research community recently. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data.[17]

V. EXPERIMENTAL RESULTS

To evaluate the Enhanced Group Weighted K- Means with Greedy Algorithm for Outlier Detection, experiments were carried out using University of California, Irvine (UCI) Machine Learning Repository .For the purpose of evaluating the

proposed technique iris dataset is used and the results are compared with standard K-Means and Semi-Supervised K-Means Clustering for Outlier Detection (SKOD), Enhanced K-Means with Greedy Algorithm for Outlier Detection (EKMOD). We use MATLAB tools for implementing our algorithms. We conducted all experiments on a Windows 7 Home Premium with Intel® Core™ i3 CPU M380 @ 2.53 GHz with 6.00 GB RAM. Experiments were conducted in Mat lab 7.8.0 (R2009a) on various data sets. The performance of clustering algorithm is presented in this section.

5.1 Outlier Accuracy

Outlier detection accuracy is calculated, in order to find out more number of outliers detected by the existing clustering algorithms. Table shows the comparison of the accuracy of clustering accuracy for the proposed method with other algorithms. From the table, it can be observed that the accuracy of clustering result using standard K-Means and SKOD, EKMOD method is 90% and 93%, 97% respectively and that of the proposed hybrid approach is 99% for iris dataset.

Table: 5.1 Comparative Analysis of Hybrid approach with Existing Algorithms in Iris Dataset

Algorithms	Accuracy
K-Mean	90%
Semi K-Mean For Outlier Detection(SKOD)	93%
EKMOD	97%
Hybrid Approach	99%

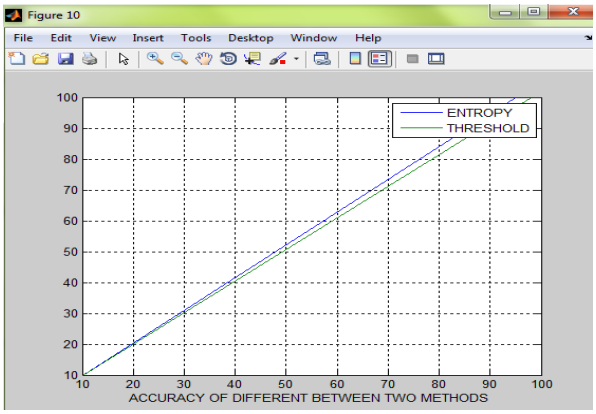

Figure: 5.1 Graphical Comparative Analysis of Hybrid approach with Existing Algorithms In Iris Dataset

5.2 Mean Squared Error

The mean square error is one way to evaluate the difference between an estimator and the true of the quantity being estimated. MSE measures average of the square of the "error," with the error being the amount by which the estimator differs from the quantity to be estimated. Table 4.5 shows the comparison of Existing Algorithms with Hybrid Approach on the basis of MSE in Iris Dataset.

Table: 5.2 Comparison of Existing Algorithms with Hybrid Approach on the basis of MSE in Iris Dataset

Algorithm	Mean squared error
K-Mean	0.6923
Semi K-Mean For Outlier Detection(SKOD)	0.4706
EKMOD	0.3029
Hybrid Approach	0.2890

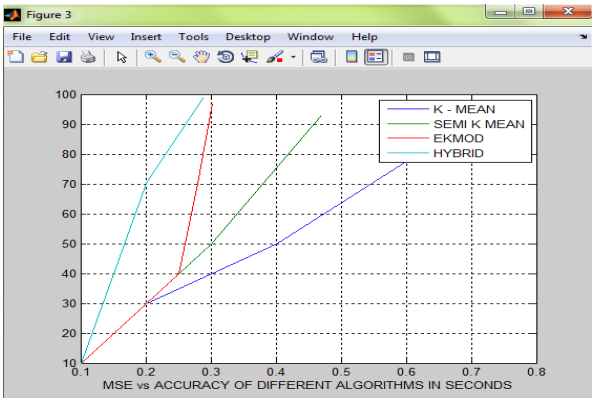

Figure: 5.2 Graphical Comparison of Existing Algorithms with Hybrid Approach on the basis on MSE in Iris Dataset

5.3 Execution Time

The execution time is calculated based on the machine time. Table shows the execution time taken by the Standard K-Means, SKOD and EKMOD the proposed hybrid approach in iris dataset. It can be observed that the time required for execution using the proposed hybrid approach scheme for iris dataset is 1.08 seconds, whereas more time is needed by other two clustering techniques for execution.

TABLE: 5.3 COMPARISON OF EXISTING ALGORITHMS WITH HYBRID APPROACH

Algorithm	Mean squared error
K-Mean	0.6923
Semi K-Mean For Outlier Detection(SKOD)	0.4706
EKMOD	0.3029
Hybrid Approach	0.2890

BASED ON EXECUTION TIME IN IRIS DATASET

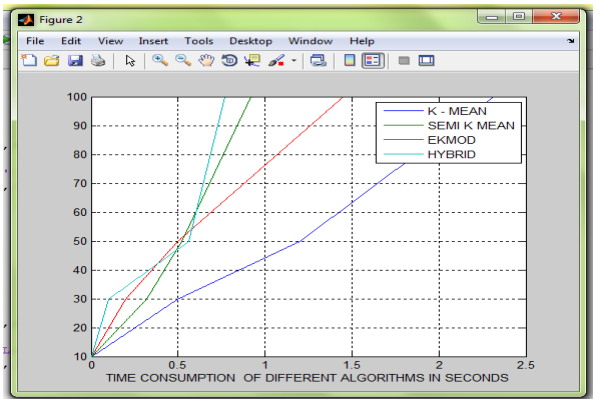

FIGURE: 5.3 GRAPHICAL COMPARISON OF EXISTING ALGORITHMS WITH HYBRID APPROACH BASED ON EXECUTION TIME IN IRIS DATASET

VI. CONCLUSION

K-Means is one of the standard clustering approaches which are widely used in several applications. The major concern in this clustering approach is that detection and removal of outliers. This papers aims to detect outliers is the task that finds objects that are dissimilar or inconsistent with respect to remaining data. We proposed an efficient outlier detection method. We first divide data into groups. Then we take threshold value from user and calculate outliers according to given threshold value for each group. 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 Hybrid Approach provides the very accurate cluster results with low MSE. Hybrid approach takes less computation time. Approach is only deals with mixed data, so future work requires modifications that can make applicable for image mining also.

REFERENCES

[1] Han, J. and Kamber, M., "Data Mining: Concepts and Techniques", Morgan Kaufman Publishers, 2006.
[2] Neelama Padhy and Pragnyaban Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
[3] S.P.Deshpande , "Data Mining System and Application: A Review**",** International Journal of Distributed and Parallel System, Vol.1, September 2010.
[4] Prabdeep and Shubha Singh, "A Survey Of Clustering Techniques", International Journal of Computer Science, Vol.7, October 2010.
[5] Varun Chandola and Banerjee and Kumar," Outlier Detection: A Survey".
[6] R. R. Rathod and Dr. B. F. Momin," Performance evaluation of Outlier Detection with Normalized Data Set", Department of Information Technology Walchand College of Enginee- ing Sangli, Maharashtra State, India.
[7] H. Desai, "Comparative Study of K-means Type Algorithms", UNIASCIT, Vol. 2, 2011.
[8] A.Mira and S.Saharia," A Robust Outlier Detection Using Hybrid Approach", Aamerican Journal of Intelligent System 2012.
[9] S.Vijayarni and S.Nithya,"An Efficient Clustering Algorithm for Outlier Detection", (IJCS) Vol.32, October 2011.
[10] Mohd - Al-Zoubi," New Outlier Detection Method Based On Fuzzy Clustering", (IJAR) Vol.4, October 2010.
[11] Jae-Gil, "Trajectory Outlier Detection: A Partition-and-Detect Framework", Department Of Computer Science, University of Illinois at Urbana-Champaign Urbana, IL 61801, USA.
[12] Deevi Radha Rani and Naya Dhulipala , "Outlier Detection For Dynamic Data Streams Using Weighted K-Means" IJEST, Vol.3, October 2011.

[13] Yogita and Durga Toshniwal," Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering", World Academy of Science, Engineering and Technology 2012.
[14] He. Xu, and S.Deng,"A Fast Greedy Algorithm for Outlier Mining, "PAKDD Conference, Singapore, 2006.
[15] Ms. S. D. Pachgade and Ms. S. S. Dhande," Outlier detection Over Data Set Using Cluster Based and Distance-Based Approach", (IJARCSSE), Volume 2, Issue6, June 2012.
[16] C.Sumithiradevi and Punithavalli,"Enhanced K-Means with Greedy Algorithm For Outlier Detection", IJARCS, Vol. 3, No.3, May-June 2012.
[17] S. John Peter," Hybrid Algorithm for Noise-free High Density Clusters with Detection Of Best Number of Clusters", (IJHIT) Vol. 4, No. 2, April, 2011.
[18] Neeraj Bansal," Differentiate Clustering Approaches for Outlier Detection", (IJIRCS), Vol. 1, Issue 2, April 2013.
[19] H.S.Behera," New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining Vol. 1, Issue 2, April 2013.
[20] H. Desai, "Comparative Study of K-means Type Algorithms", UNIASCIT, Vol. 2, 2011.
[21] E. M. Knorr ," Algorithms For Mining Distance Based Outliers In Large Datasets", VLDB, pages 392–403, 1998.
[22] http://www.uci.edu/.
[23] http://archive.ics.uci.edu/ml/datasets/Iris.
[24] http://archive.ics.uci.edu/ml/datasets/pima.
[25] http://archive.ics.uci.edu/ml/datasets/.

