ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

Differentiate Clustering Approaches for Outlier Detection

Ms. Neeraj Bansal[1], Mr.Amit Chugh[2]

Student, Departments of Computer Science, Lingaya's University, Faridabad, India[1]

Assistant professor, Departments of Computer Science, Lingaya's University,Faridabad, India[2]

Abstract: Data mining is a process of extracting hidden and useful information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, and valid and of high quality. There are several techniques exist for data extraction. Clustering is one of the techniques amongst them. In clustering technique, we form the group of similar objects (similarity in terms of distance or there may be any other factor). Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. Therefore, it is important to detect outlier from the extracted data. There are so many techniques existing to detect outlier but Clustering is one of the efficient techniques. In this paper, I have compared the result of different Clustering techniques in terms of time complexity and proposed a new solution by adding fuzziness to already existing Clustering techniques.

Keywords: Clustering, Data Mining, Outlier Detection, Data Mining

I. INTRODUCTION

Data mining is the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Now, data mining is becoming an important technique to convert the data into valuable information. It is commonly used in a wide series of profiling practices, such as marketing, fraud detection and scientific discovery. Data mining is the method of extracting patterns from data. The mining process will be ineffective if the samples are not good representation of the larger body of the data. Therefore, it is important to detect whether the extracted is either useful to us or not. Outliers are the set of objects that are considerably dissimilar from the remainder of the data. Outlier detection is an extremely important problem with a direct application in a wide variety of application domains, including fraud detection, identifying computer network intrusions and bottlenecks, criminal activities in e-commerce and detecting suspicious activities. Different approaches have been proposed to detect outliers.

Outlier Detection: Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data or which are far away from their cluster centroids. A failure to detect outliers or their ineffective handling can have serious impact on the strength of the inferences drained from the exercise. There are large number of techniques are available to perform this task, since there is no standard algorithm exist for detecting it. Below are the some algorithms used for detecting the outlier.

Existing Algorithms for detecting outlier:

II. DISTANCE BASED OUTLIER DETECTION

In the distance-based approach, outliers are detected as follows. Given a distance measure on a space, a point k in a data set is an outlier with respect to the parameters M and d, if there are less than M points within the distance d from q, where the values of M and d are decided by the user. The problem with this approach is that it is difficult to determine the values of M and d.

III. DISTRIBUTION BASED OUTLIER DETECTION

Develop statistical models (typically for the normal behavior) from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects that have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications.

IV. DENSITY BASED OUTLIER DETECTION

Objects in low-density regions of space are flagged.

Disadvantage: Density based models require the careful settings of several parameters.
It requires quadratic time complexity.

V. CLUSTERING BASED OUTLIER DETECTION

Consider clusters of small sizes as clustered outliers. In these approaches, small clusters (i.e.clusters containing significantly less points than other clusters) are considered outliers. The advantage of the clustering-based approaches is that they do not have to be supervised. Moreover, clustering-based techniques are capable of being used in an incremental mode.

There are various kinds of Clustering based outlier detection approach have been proposed which are following:

A. PAM (Partition around Medoid):

PAM uses a k-medoid method for clustering. It is very robust when compared with k-means in the presence of noise and outliers. Mainly it contains two phases Build phase and Swap phase.

Build phase: This step is sequentially select k objects which is centrally located. This k objects to be used as k-medoids. Swap phase: Calculates the total cost for each pair of selected and non-selected object.

PAM Procedure:

- Input the dataset D
- Randomly select k objects from the dataset G
- Calculate the Total cost T for each pair of selected Si and non selected object Sh
- For each pair if T si < 0, then it is replaced Sh
- Then find similar medoid for each non-selected object
- Repeat steps 2, 3 and 4, until find the medoids.

B. CLARA (Clustering Large Applications):

CLARA is introduced to overcome the problem of PAM. This works in larger data set than PAM. This method takes only a sample of data from the data set instead of taking full data set. It randomly selects the data and chooses the medoid using PAM algorithm [1].

CLARA Procedure

1. For i=1 to 5, repeat Steps 2 to 5.
2. Draw a sample of 40 + 2k objects randomly from the entire data set and call PAM algorithm to find k medoids of the sample.
3. For each object O in the entire data set, determine k-medoids which is most similar to O.
4. Calculate average dissimilarity of the clusters obtained from Step 3. If this value is less than current minimum, use the new value as current minimum and retain the k medoids found in Step 2 as the best set of medoids obtained so far.
5. Return to Step 1 to start the next iteration.

C. CLARANS (Clustering Large Applications Based on Randomized Search):

This method is similar to PAM and CLARA. It starts with the selection of medoids randomly. It draws the neighbour dynamically. It checks “max neighbour” for swapping. If the pair is negative then it chooses another medoid set. Otherwise it chooses current selection of medoids as local optimum and restarts with the new selection of medoids randomly. It stops the process until returns the best.

CLARANS Procedure:

- Input parameters numlocal and max neighbour.
- Select k objects from the database object D randomly.
- Mark these K objects as selected Si and all other as non- selected Sh.
- Calculate the cost T for selected Si
- If T is negative update medoid set. Otherwise selected medoid chosen as local optimum.
- Restart the selection of another set of medoid and find another local optimum.

- CLARANS stops until returns the best.

D. ENHANCED CLARANS (ECLARANS):

This method is different from PAM, CLARA AND CLARANS. Thus method is produced to improve the accuracy of outliers. ECLARANS is a new partitioning algorithm which is an improvement of CLARANS to form clusters with selecting proper arbitrary nodes instead of selecting as random searching operations. The algorithm is similar to CLARANS but these selected arbitrary nodes reduce the number of iterations of CLARANS.

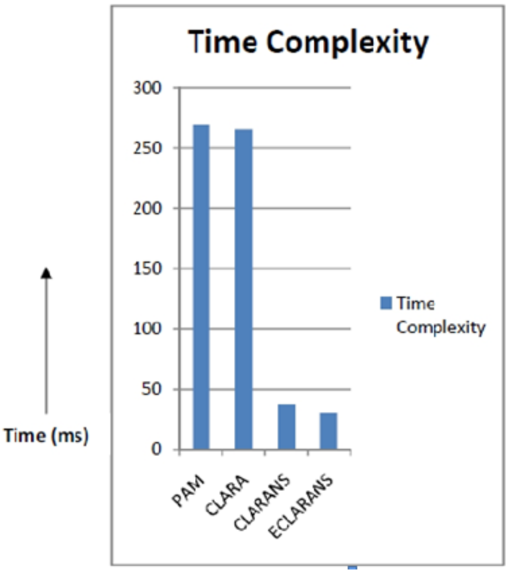
ECLARANS Procedure:

1. Input parameters numlocal and maxneighbour. Initialize i to 1, and mincost to a large number.
2. Calculating distance between each data points
3. Choose n maximum distance data points
4. Set current to an arbitrary node in n: k
5. Set j to 1.
6. Consider a random neighbour S of current, and based on 6, calculate the cost differential of the two nodes.
7. If S has a lower cost, set current to S, and go to Step
8. Otherwise, increment j by 1. If j max neighbour, goto Step6
9. Otherwise, when j > maxneighbour, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set best node to current.Increment i by 1. If i > numlocal, output best node and halt. Otherwise, go to Step 4

Comparison between above approaches:

Above algorithms are implemented in java (using Java Eclipse IDE). Then, their time complexity is compared i.e. time taken by them to detect the outlier using above approaches. I have taken the data set and applied all above algorithms which produce different result. Below is the table and graph which is showing the comparison of time complexity of all above stated algorithms.

Name of an algorithm	PAM	CLARA	CLARANS	ECLARANS
Time in millisecond	289.6	274.4	33.2	28.4



Graph 1- Comparison graph between different algorithms

VI. CONCLUSION

Data mining is the process of extracting the data from data set. Thus, outlier detection becomes important process in data mining since if we proceed with outlier, it can create problem in further analysis. There are different algorithms exist for detecting outlier. As we have seen, ECLARANS is the best technique amongst them. It takes lesser amount of time to detect the outlier.

As future lies, further advancement is going on in outlier detection methods. More work is being done on the basis of fuzzy approach in clustering techniques. It helps in detection of outlier for imprecise and incomplete data set.

REFERENCES

1. S. Vijayarani and S. Nithya, " An Efficient Clustering Algorithm for Outlier Detection", Computer Software and Applications Conference Workshops.
2. Al-Zoubi, M. (2009) An Effective Clustering-Based Approach for Outlier Detection, European Journal of Scientific Research.
3. Jiang, S. And An, Q. (2008) Clustering Based Outlier Detection Method, Fifth International Conference on Fuzzy Systems and Knowledge Discovery.
4. John Peter. S., Department of computer science and research center St.Xavier‟s College, Palayamkottai, An Efficient Algorithm for Local Outlier Detection Using Minimum Spanning Tree, International Journal of Research and Reviews in Computer Science (IJRRCS), March 2011.
5. Loureiro, A., Torgo, L. And Soares, C. (2004) Outlier Detection using Clustering Methods: A Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-Based Systems.
6. Knorr, E. and Ng, R. (1997). A unified approach for mining outliers. In Proc. KDD, pp. 219–222.
7. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, "LOF: identifying density-based local outliers", Jörg Sander, 2000 ACM SIGMOD international conference on Management of data, pp. 93-104, ACM, New York, NY, USA
8. Ian H. Witten and Eibe Frank, Morgan Kaufmann, "Data Mining: Practical machine learning tools with Java implementations", San Francisco 2000
9. Perarson R. K., "Outliers in process modeling and identification," IEEE Transactions on Control Systems Technology, pp.10, 55-63, 2002.
10. Ramaswamy S., Rastogi R., Shim K., "Efficient algorithms for mining outliers from large data sets," In Proceedings of the ACM SIGMOD International Conference on Management of Data, Dalas, TX, 2000.
11. Hadi A.S., A.H.M.R. Imon, and M. Werner, ―Detection of outliers,‖ *Computational Statistics*, vol. 1, 2009, pp. 57-70