ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

Distance-Based Outlier Detection in Large Datasets Using Hybrid Algorithm

Raja.K [1], Matheswaran.V [2]

1. P.G. Student, Dept. of MCA, VSB Engineering College, Karur, Tamilnadu, India
2. Asst.Professor, Dept. of MCA, VSB Engineering College, Karur, Tamilnadu, India

Abstract— **Outlier detection is the process of finding outlying pattern from a given dataset. Outlier detection became important subject in different knowledge domains. Data size is getting doubled every years there is a need to detect outliers in large datasets as early as possible. In high-dimensional data outlier detection presents various challenges because of curse of dimensionality. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the dataset. Proposed Method for outlier detection uses hybrid approach. Purpose of approach is first to apply clustering algorithm that is kmeans which partition the dataset into number of clusters and then find outliers from the each resulting clusters using distance based method. The principle of outliers finding depend on the threshold. Threshold is set by user. The main objective of the second stage is a finding out the objects, which are far away from their cluster centroids. In proposed approach, two techniques are combining to efficiently find the outlier from the data set. The experimental results using real dataset demonstrate that proposed method takes less computational cost and performs better than the distance based method. Proposed algorithm efficiently prunes of the safe cells (inliers) and save huge number of extra calculations.**

Keywords— Outlier, Cluster-based, Distance-based. high-dimensional data, distance concentration

I. INTRODUCTION

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. Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. In recent years, conventional database querying methods are inadequate to extract useful information, and hence researches nowadays are focused to develop new techniques to meet the raised requirements. It is to be noted that the increase in dimensionality of data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of attributes. Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the database. Medical application is a high dimensional domain hence determining outliers is found to be very tedious due to the Curse of dimensionality. There are various origins of outliers. With the growth of the medical dataset day by day, the process of determining outliers becomes more complex and tedious. Efficient detection of outliers reduces the risk of making poor decisions based on erroneous data, and aids in identifying, preventing, and repairing the effects of malicious or faulty behavior. Additionally, many data mining and machine learning algorithms and techniques for statistical analysis may not work well in the presence of outliers. Outliers may introduce skew or complexity into models of the data, making it

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

difficult, if not impossible, to fit an accurate model to the data in a computationally feasible manner. For example, statistical measures of the data may be skewed because of erroneous values, or the noise of the outliers may obscure the truly valuable information residing in the data set. Accurate and efficient removal of outliers may greatly enhance the performance of statistical and data mining algorithms and techniques [6]. Detecting and eliminating such outliers as a pre-processing step for other techniques is known as data cleaning. As can be seen, different domains have different reasons for discovering outliers: They may be noise that we want to remove. 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. Finding anomalous points among the data points is the basic idea to find out an outlier. Distance based techniques use the distance function for relating each pair of objects of the data set. Distance based definition (these definitions are computationally efficient) [7, 10] represent a useful tool for data analysis [8].

II. OBJECTIVES OF STUDY

Basic aims to reduce the number of pair wise distance calculations, to let user free to provide sensitive parameters .We are first testing with distance based approach; this approach applies to all data, then testing with hybrid approach. In that we are first partition the data in to number of clusters and then we apply distance based approach. The principle of outlier's detection depends on the threshold. This approach takes less computational time than distance based method.

III. RELATED WORK

Outlier detection (deviation detection, exception mining, novelty detection, etc.) is an important problem that has attracted wide interest and numerous solutions. These solutions can be broadly classified into several major ideas:

Model-Based [2]: An explicit model of the domain is built (i.e., a model of the heart, or of an oil refinery), and objects that do not fit the model are flagged.

Disadvantage: Model-based methods require the building of a model, which is often an expensive and difficult enterprise requiring the input of a domain expert

Connectedness [11]: In domains where objects are linked (social networks, biological networks), objects with few links are considered potential anomalies.

Disadvantage: Connectedness approaches are only defined for datasets with linkage information

Density-Based [3]: 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.

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

Distance-Based [1]: Given any distance measure, objects that have distances to their nearest neighbors that exceed a specific threshold are considered potential anomalies. In contrast to the above, distance-based methods are much more flexible and robust. They are defined for any data type for which we have a distance measure and do

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

not require a detailed understanding of the application domain.

Cluster based approach [4]: The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. The assumed behavior of outliers is that they either do not belong to any cluster, or belong to very small clusters, or are forced to belong to a cluster where they are very different from other members. Clustering based outlier detection techniques have been enveloped which make use of the fact that outliers do not belong to any cluster since they are very few and different from the normal instances.

K-Nearest Neighbor Based Approach [12]: 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 obvious strength of these techniques is that they can work in an unsupervised mode, i.e. they do not assume availability of class labels.

In this work, we are introducing clustering method that will reduce size of datasets, and groups the data having similar characteristic. Next we apply distance based to detect the outliers as per given threshold. Within a cluster get outliers, that are far from their cluster centroid.

IV. PROPOSED WORK

1) System architecture

Outlier (anomaly) detection refers to the task of identifying patterns that do not conform to established regular behavior [1]. Despite the lack of a rigid mathematical definition of outliers, their detection is a widely applied practice [2]. The interest in outliers is strong since they may constitute critical and actionable information in various domains, such as intrusion and fraud detection, and medical diagnosis. The task of detecting outliers can be categorized as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and/or regular instances. Among these categories, unsupervised methods are more widely applied [1], because the other categories require accurate and representative labels that are often prohibitively expensive to obtain. Unsupervised methods include distance-based methods [3], [4], [5] that mainly rely on a measure of distance or similarity in order to detect outliers. A commonly accepted opinion is that, due to the "curse of dimensionality," distance becomes meaningless [6], since distance measures concentrate, i.e., pairwise distances become indiscernible as dimensionality increases [7], [8]. The effect of distance concentration on unsupervised outlier detection was implied to be that every point in high-dimensional space becomes an almost equally good outlier [9]. This somewhat simplified view was recently challenged [10].

Our motivation is based on the following factors:

1) It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in [10] the actual challenges posed by the "curse of dimensionality" differ from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space [9]. We will present further evidence which challenges this view, motivating the (re)examination of methods.

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

2) Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points [11], [12],1 but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data warrant their reexamination for the outlier-detection task. In this light, we will revisit the ODIN method [11]. Christo Ananth et al. [14] proposed a system in which FASTRA downloads and data transfers can be carried over a high speed internet network. On enhancement of the algorithm, the new algorithm holds the key for many new frontiers to be explored in case of congestion control. The congestion control algorithm is currently running on Linux platform. The Windows platform is the widely used one. By proper Simulation applications, in Windows we can implement the same congestion control algorithm for Windows platform also. The Torrents application which we are currently using can achieve speeds similar to or better than —Rapid share (premium user) application.

An input of collection of large data set will be provided to the proposed system, as data is collected from standard data set repositories, data preprocessing will be applied before passing data to the next phase of the system. Further, this preprocessed input is being passed through to the partition module, where these datasets are been partitioned among many nodes from that one of the node is supervisor node and generate partition statistics and this statistical data is been visualized. After this, in outlier detection module,distributed algorithms is proposed on the preprocessed input data set for identifying outliers. These results will be evaluated for proposed algorithmic distributed approaches in the performance evaluation.

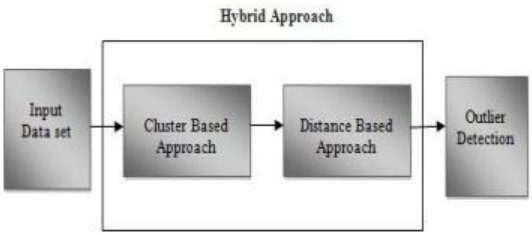

Figure 1: System Architecture

2) Proposed Algorithm

Generating cluster: *K*-means clustering is a partitioning method .Initially, cluster the entire dataset into k cluster using K-mean clustering and calculate centroid of each cluster.

Kmean Clustering: Given *k*, the *k-means* algorithm is implemented in four steps:

a) Select k observations from data matrix X at random

b) Calculate distance with each instances (with respect to randomly selected instances)

c) Assign each instance to the cluster with the nearest seed

d) Go back to Step b, stop when no instance to move group

Calculate Threshold % for each cluster

-- finding min max values from each clusters

-- finding maximum distance from centroid

-- take threshold from user

-- find threshold % value for each cluster

Calculate distance of each point of cluster from centroid of the cluster. If the distance is greater

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

than threshold then it will declare as ―outlier.

3) Modules Description

Data collection and data preprocessing :

In data collection the initial input data for this system will be collected from standard dataset portal i.e. UCI data set repository. As proposed in system, the standard dataset will be used for this system includes Cover type, IPS datasets. Collected datasets may be available in their original, uncompressed form therefore; it is required to preprocess such data before forwarding for future steps. To preprocess large dataset contents, techniques available is data mining such as data integration, data transformation, data cleaning, etc. will be used and cleaned, required data will be generated.

Data partitioning:

In this module, as stated earlier in system execution plan, the preprocessed data is divided into number of clients from central supervisor node i.e. server as per the data request made by desired number of clients. This partitioned data will be then processed by individual clients to identify outliers based on applied algorithm strategy.

Outlier detection:

Cluster Based Approach**:** Clustering is a popular technique used to group similar data points or objects in groups or clusters. Clustering is an important tool for outlier analysis. Cluster based approach is here act as data reduction. First, clustering technique is used to groups the data having similar characteristics. And calculate the centroids for each group.

Distance Based Approach**:** Distance based technique is used to calculate maximum distance value for each cluster. If this maximum distance is greater than some threshold then it will declare as ―outlier‖ otherwise as a real object or inliers. Threshold is given by user.

Outlier Detection**:** Outlier detection is an extremely important task in a wide variety of application domains. 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.

Performance Evaluation and Result Visualization :

In this module, the outlier detected by above approach will be evaluated on the basis of set evaluation parameters for their performance evaluation. The performance evaluation will also provide details about implemented system performance metrics, constraints and directions for future scope. With the help of proper visualization of results, the system execution will be made more understandable and explorative for its evaluators.

V. CONCLUSION

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 groups the data (having similar characteristics) in to number of clusters. Due to reduction in size of dataset, the computation time reduced considerably. Then we take threshold value from user and calculate outliers according to given threshold value for each cluster. We get outliers within a cluster. Hybrid approach takes less computation time. Approach is only deals with numerical data, so future work requires modifications that can make applicable for textual mining also. The approach needs to be implemented on more complex datasets. Future

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

work requires approach applicable for varying datasets.

REFERENCES

[1] F. Angiulli and F. Fassetti, "Detecting Distance-based Outliers in Streams of Data," In Proceedings of CIKM'07, Pages 811-820, November 6-10 2007.

[2] F. J. Anscombe and I. Guttman, "Rejection of Outliers," Techno metrics, vol. 2, Pages 123-147, May 1960.M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "OPTICS-OF: Identifying Local Outliers," In Proceedings of PKDD'99, Pages 262- 270, September 15-18 1999.

[4] Parneeta Dhaliwal, MPS Bhatia and Priti Bansal,‖ A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner) JOURNAL OF COMPUTING, VOLUME 2, ISSUE 2, FEBRUARY 2010, ISSN: 2151-9617.PAGES 74-80.

[5] Manzoor Elahi, KunLi, Wasif Nisar, Xinjie Lv, Hongan Wang, ‖Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream‖ In Proc .of Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD.2008),ISBN: 978-0-7695-3305-6/08, pages 298-304.

[6] Hadi A.S., A.H.M.R. Imon, and M. Werner, ―Detection of outliers,‖ *Computational Statistics*, vol. 1, 2009, 57-70.

[7] E. M. Knorr and R. T. Ng. ―Algorithms for mining distance based outliers in large datasets‖ In Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pages 392– 403, 1998.

[8] M. Knorr and R. T. Ng. ―Finding intentional knowledge of distance-based outliers‖ In VLDB '99: Proceedings of the 25thInternational Conference on Very Large Data Bases, pages 211–222, 1999.

[9] Rajendra Pamula,Jatindra kumar Deka,Sukumar Nandi.‖An Outlier Detection Method based on Clustering‖, Second International Conference on Emerging Applications of information Technology,2011. ISBN: 978-0-7695-4329-1/11, Pages 253-256 Ramaswamy, R. Rastogi, and K. Shim. ―Efficient algorithms for mining outliers from large data sets‖ pages 427–438, 2000.

[10] J. Tang, Z. Chen, A. W.-C. Fu and D. W.-L. Cheung, "Enhancing Effectiveness of Outlier Detections for Low Density Patterns," In Proceedings of PAKDD'02, Pages 535-548, May 6-8 2002.

[11] Peng Yang; Biao Huang;‖ KNN Based Outlier Detection Algorithm in Large Dataset‖ International Workshop on Education Technology and Training, ISBN: 978-0-7695-3563-0, Pages 611 – 613, 2008.

[12] http://archive.ics.uci.edu/ml

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Survey, vol. 41, no. 3, p. 15, 2009.

[14] Christo Ananth, A. Ramalakshmi, S. Velammal,B. Rajalakshmi Chmizh, M. Esakki Deepana, "FASTRA –SAFE AND SECURE", International Journal For Technological Research In Engineering (IJTRE), Volume 1, Issue 12, August-2014,pp: 1433-1438.

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 11, March 2016

[15] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," SIGMOD Rec., vol. 29, no. 2, pp. 427–438, 2000.

[16] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," in Proc. Conf. Appl. Data Mining Comput. Security, 2002, pp. 78–100.

[17] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237– 253, 2000.

[18] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217–235.

[19] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proc. 8th Int. Conf. Database Theory, 2001, pp. 420–434.

[20] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7, pp. 873–886, Jul. 2007.

[21] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in Proc. 27th ACM SIGMOD Int. Conf. Manage. Data, 2001, pp. 37–46.

[22] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 363–387, 2012.

[23] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in Proc 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 430–433.

[24] J. Lin, D. Etter, and D. DeBarr, "Exact approximate reverse nearest neighbor search for multimedia data," in Proc 8th SIAM Int. Conf. Data Mining, 2008, pp. 656–667.

All Rights Reserved © 2015 IJARMATE