Distance based Parallel Outlier Detection by subset sequence: An SVM Regression Approach

Gujja Sathish Kumar
M.Tech in Software Engineering
Aurora's Technological & Research Institute,
Parvathapur, Uppal, Hyderabad-500039

B. Malathi
Sr. Asst. Professor in CSE DEPARTMENT
Aurora's Technological & Research Institute,
Parvathapur, Uppal, Hyderabad-500039

Abstract***:*** **Outlier recognition happens to be very active area of research in data set mining community. Finding outliers in a collection of designs is a really well-known problem in the data-mining area. An outlier is a design which can be dissimilar regarding the rest of the designs in the dataset. Hybrid approach is used by proposed Method for outlier detection. Intent behind method is first to utilize clustering algorithm that's kmeans which partition the dataset in to number of clusters and then find outliers from the each resulting clusters using range based technique. The theory of outliers finding rely on the threshold. Limit is defined by individual. The main objective of the second stage is a learning the objects, which are a long way away from their cluster centroids. In planned approach, two practices are combining to efficiently discover the outlier in the data set. Proposed formula effortlessly prunes of the cells (inliers) and save yourself huge number of additional measurements.**

Keywords— *Outlier, Cluster-based, Distance-based.*

I. INTRODUCTION

Data mining is a procedure of extracting hidden and useful information in the data and the knowledge discovered by data mining is probably useful, previously unknown, and valid and of high quality. Finding outliers is an essential job in data mining. Outlier diagnosis being a branch of data mining has many important applications and deserves more attention from data mining community. In recent years, typical database querying methods are inadequate to extract useful information, and hence researches nowadays are focused to produce new processes to meet the requirements. It is to be noted that the increase in dimensionality of data gives rise to a number of new computational challenges not merely due to the increase in number of data objects but also due to the increase in number of attributes. Outlier recognition is an essential research problem that aims to get items that are dramatically dissimilar, extraordinary and contradictory in the database. Medical software is just a high-dimensional area thus determining outliers is available to be very boring due to the Curse of dimensionality. There are various origins of outliers. With all the development of the medical dataset day by day, the procedure of deciding outliers becomes more technical and tedious. Successful detection of outliers reduces the chance of making poor decisions centered on erroneous data, and helps with preventing, determining, and correcting the results of harmful or bad behavior. Also, many data-mining and machine learning techniques and methods for statistical analysis might not work well in the presence of outliers. As an example, statistical measures of the data may be skewed because of incorrect values, or the noise of the outliers may obscure the truly valuable information residing in the data set. Precise and effective removal of outliers might significantly enhance the performance of statistical and data mining algorithms and practices [6]. Detecting and eliminating such outliers as a pre-processing step for other practices is known as data cleaning. Different areas have different reasons for discovering outliers: They may be noise that people desire to remove, as is visible.

In this work, we are introducing clustering method that will reduce size of datasets, and groups the info having similar attribute. Next we apply distance based to find the outliers depending on given threshold. Inside a cluster get outliers, which might be removed from their cluster centroid.

Finding outliers has important applications in data cleaning together with in the exploration of irregular points for fraud detection, stock-market analysis, intrusion detection, advertising, system devices. Finding anomalous points one of the data points could be the fundamental idea to discover an outlier. Distance based techniques utilize the distance function for relating each pair of objects of the data set. Distance based definition (these definitions are computationally effective) [7, 10] represent a good instrument for data analysis [8].

II. OBJECTIVES OF STUDY

Basic aims to lessen the amount of pair wise distance measurements, to let consumer free to give painful and sensitive parameters. We are first testing with distance based approach; this approach pertains to all information, then testing with hybrid approach. In that we're first partition the information in to variety of clusters and then we apply range based approach. The theory of outlier's recognition depends on the threshold. This process requires less computational time than distance based method.

III. RELATED WORK

Outlier detection (deviation detection, exception mining, novelty detection, etc.) is an important issue that has attracted broad interest and numerous options. These remedies can be broadly classified into several important 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 items which don't match the model are flagged.

Downside: Model based techniques need the construction of a model, which is typically an expensive

And tough venture requiring the input of a domain expert

Connectedness [11]: In domains where objects are connected (social networks, biological networks), objects with few hyperlinks are considered possible anomalies.

Disadvantage: Connectedness tactics are simply defined for datasets with linkage info **DensityBased [3]:** Items in low-density regions of space are flagged.

Disadvantage: Density based models require the options of many parameters.

It demands quadratic time complexity.

It could exclude outliers close to some non - outliers designs that has low density.

Space-Based [1]: Given any distance measure, items that have distances to their nearest neighbors that surpass a particular threshold are believed to be potential anomalies. In contrast to the above, space-based techniques are way more flexible and robust.

Cluster based approach [4]: The clustering based methods involve a step which partitions the information into groups which include similar items. Clustering based outlier detection methods are enveloped which make practical use of the reality that outliers don't belong to any bunch since they can be very few and distinct from the standard examples.

K-Nearest Neighbor Based Approach [12]: The fundamental idea behind such schemes is that the outlier is going to have neighborhood though a regular item will have a neighborhood where all its neighbors will soon be exactly like it, where it's going to stand out. The apparent strength of these techniques is the fact that they're able to work in an unsupervised manner, I.e. they don't assume availability of category labels.

IV. PROPOSED WORK

1. System architecture

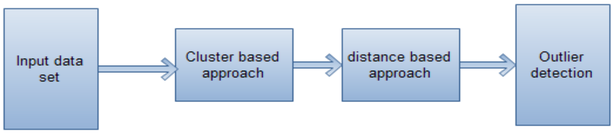

Figure 1: System Architecture

Input Data Set: Collecting dataset from UCI Machine wisdom repository [13].

Cluster Based Approach: Clustering is a popular technique used to group related data points or things in groups or clusters. Cluster based strategy will be here act as data reduction. First, clustering technique is used to groups the data having similar characteristics. And compute the centroids for every single team. Distance Based Method: Distance based technique is employed to compute maximum distance value for each bunch. If this maximum distance is more than some threshold then it'll declare as "outlief' otherwise being a real object or inliers. Threshold is given by user.

Outlier Detection: Outlier detection is a highly important job in a broad selection of application domains. Outlier detection is a job that finds items that are dissimilar or inconsistent regarding the remaining information or which are far away from their cluster centroids.

2. Distance based Algorithm

This approach is extremely determined by parameter supplied by the users and computationally costly when applied unbounded data set. With the development of information technologies, the number of their measurements, in addition to databases and complexity grow fast. With distance is calculated by high dimensional dataset with each examples will increase the computational cost. We're comparing space based method with proposed approach.

Pairwise distance computes the Euclidean distance among pairs of objects in n-by-p data matrix X. Rows of X communicate to observations; columns communicate to variables. y is a row vector of length n(n-1)/2, equivalent to pairs of observations in X. The distances are approved in the order (2,1), (3,1), ..., (n,1), (3,2), ..., (n,2), ..., (n,n-1)). y is commonly used as a variation matrix in clustering or multidimensional scaling.

Euclidean distance

$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)'$$

Where,

$$\bar{X}_r = \frac{1}{n}\sum_j x_{rj} \quad \text{and}$$

$$\bar{X}_s = \frac{1}{n}\sum_j x_{sj}$$

1) Calculate pairwise detachment that is computing the Euclidean distance among pairs of object.
2) Take square detachment. Calculate maximum values from square detachment values
4) Take threshold since user.
5) If distance > threshold assessment that will be the outliers.

3. Proposed Clustering and Distance-Based Algorithm

Generating bunch: K-means bunching is really a partitioning strategy. Initially, cluster the entire dataset into k bunch using K-mean clustering and calculate centroid of each bunch. Kmean Clustering: Given k, the k-means algorithm is implemented in four steps:

a) Pick k observations from data matrix X at random

b) Compute space with each instances (with respect to randomly selected examples)

C) Assign each instance to the cluster with the seed

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

2) Compute Threshold to look for each bunch

-- finding min-max values from each clusters -- finding maximum distance from centroid -- consider threshold from user -- locate threshold quality value for every bunch

3) Calculate distance of each point of clustering from centroid of the clustering. In the event the distance is greater than threshold then it is going to declare as "outlier".

V. EXPERIMENTAL RESULTS

MATLAB tools are used by us for applying our algorithms. Experiments were conducted in Matlab 7.8.0 (R2009a) on several different data sets. Data is gathered from UCI machine learning repository that provided various sorts of datasets. This dataset can be utilized for regression, classification and clustering. Dataset has several aspect and examples. A repository of databases, domain theories and data generators are used by the machine learning community for the empirical evaluation of machine. This data file will be taken to find the outlier.

Medical Diagnosis Data Set: In real world data repositories, it is tough to discover a data set for evaluating outlier detection algorithms, because only for very few real-world data sets it is exactly known which items are actually acting differently. In this experiment, we use a health data set, WDBC (Diagnosis), that is used for nuclear feature extraction for breast tumor analysis. The data set includes 428 medical diagnosing records (things), each with 32 attributes (ID, diagnosing, 30 realvalued input features).The diagnosis is binary: Benign and Malignant. There are 2 types of datasets so we're splitting dataset into 2 numbers of clusters.

Number of Data Points in each cluster for reuters dataset	
1 Cluster	71
2 Cluster	357

Table 1: Shows number of data points for reuters Dataset

Here, we obtain groups of 71 instances in first cluster and 357 instances in second cluster among 428 instances. Group of 71 fit in to malignant record and group of 357 belongs to caring record. From first cluster 10 numbers of outliers notice at 75 percent of threshold. From second cluster 21 outliers notice at 75 percent of threshold.

Threshold %		***75***	***80***	***85***	***90***	***95***
No. of	1 Cluster	10	8	7	6	6
Outliers	2 Cluster	21	19	17	11	9

Table 2: Test on different threshold value for reuters dataset and getting variations in number of outliers

Elapsed Time	
Distance Based Approach	Proposed Approach
0.29246s	0.0954145s

Table 3: CPU Time in Second for Reuters Dataset

2) Bupa liver disorder datasets: which refers to the first 5 variables are all blood tests that are believed to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the bupa Data file constitutes the document of the single male person. It appears that drinks > 5 is some sort of the selector on this particular database. Selector field used to split data into two sets. The Dataset includes 7 attributes and 345 instances. Within this dataset only few numbers of outliers are detected at 75%. From first cluster simply one outlier is detected and in cluster two outliers detected.

Number of Data Points in each cluster for Liver Disorder		**Number of outliers at 75 %**
1 Cluster	37	1
2 Cluster	308	2

Table 4: Shows number of data points and outliers for liver Disorder Dataset

VI. DISCUSSION

Finding outliers is an essential task in data mining. Outlier detection for a branch of data mining has many significant applications and deserves more focus from data mining community. Comparison between Distance based approach and suggested approach are as follows:

Distance-Based Method

- Operate on whole data. Cannot offer number of clusters.
- Calculation time will increases
- Give only one value as the majority expected outlier

Clustering and Distance-Based

- Can group the information in to number of clusters
- Reduce the size of database that determination reduces computation time
- To each cluster user can give positive radius to find outliers.

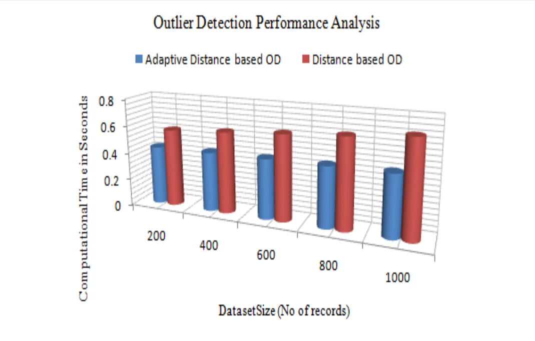

Figure 1: Computational Time Analysis of Distance based and Adaptive distance Based Outlier Detection Models

VII. CONCLUSION AND FUTURE WORK

This papers aims to find outliers may be the task that finds objects that are dissimilar or inconsistent with respect to remaining data. We first groups the data (having similar characteristics) in to amount of clusters. Due to reduction in size of dataset, the computation time reduced drastically. Then we take threshold value from user and compute outliers according to specified threshold value for every cluster. Hybrid approach takes less computation time.

Approach is only deals with numerical data, so future work needs changes that may make applicable for textual mining also. The approach requires to be put into place on more complex datasets. Future work demands strategy applicable for changing 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.

[3] 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: 978-07695-3305-6/08, pages 298-304.

[6] Hadi A.S., A.H.M.R. Imon, and M. Werner, "Detection of outliers," *Computational Statistics,* vol. 1, 2009, 5770.

[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 392403, 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.

[10] Ramaswamy, R. Rastogi, and K. Shim. "Efficient algorithms for mining outliers from large data sets" pages 427-438, 2000.

[11] 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.

[12] 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.

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