Outlier Detection using Hybrid Approach for Mixed datasets

Anjali Barmade, Prof. Manjusha Deshmukh
Computer Department
Pillai's Institute of Information Technology
Mumbai, India
anjali.barmade08@gmail.com

Abstract - Several approaches of outlier detection are employed in many study areas amongst which distance based and density based outlier detection techniques have gathered most attention of researchers .So we are using hybrid of these two methods. The proposed model uses hybrid of distance and density outlier detection methods and weighted squeezer method for clustering. Advantages of both outlier methods will be combined giving higher result. The clustering algorithm which does not require to specify number of clusters as input which is drawback of many clustering algorithms. Most of the models deals with only single datatype datasets. Here the project deals with mixed datatype datasets. Here we will compare hybrid system with single method system. From performance measures it will be cleared how hybrid system gives better results as compared to single method.

Keywords—*outlier detection, weighted squeezer clustering, hybrid, distance, density based, k-means, mixed datasets.*

*****

I. Introduction

The method of summarizing data by analyzing it from different directions to get required useful information is data mining. Meaningful data is required for increasing profits in companies and reducing errors. We can define data mining as method of collecting the data required by us or required for our system and using it. Working on any number, facts, text by computer is nothing but data. Data mining can be said as one of the analytical tools for analyzing data.

Analysing whether a system works well or not depends mainly on data mining. Analyst will gain a lot if mapping between real world problems and outliers will be done. The process of finding patterns in data that do not follow routine ways is called outlier detection.

The fraud or fault detection is done in areas like credit card, insurance, tax fraud detection, intrusion detection and fault detection in safety critical systems, military surveillance.

The remaining part of the paper is organized as follows. Section II introduces to related work on proposed system. Section III describes existing system. Section IV describes proposed system. Section V describes stages of proposed system. Section VI provides experimental evidences for system and Section VII concludes the paper.

II. Related work

This section describes the work related to our hybrid system

A. Outliers

Data points from the datasets which are separated or far away or different from remaining data points are outliers. The patterns which is very different from present data sets patterns and which will generate some other mechanism is outlier. Example given a method based on some formula or some approach is followed by normal objects but in case of outliers are abnormal objects deviate from this generating mechanism.

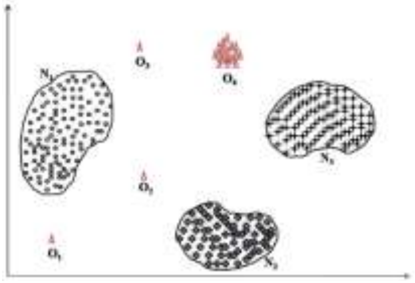


Fig.1. Outliers

B. Outlier Detection:

The method of analyzing huge datasets and extracting information required for our system is data mining. Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior.

C. Outlier detection methods

It is broadly divided into

1. Non transaction specific outlier detection methods
2. Transaction specific outlier detection methods

1. Non- Transaction specific outlier detection methods:

- Supervised, unsupervised and semi supervised.
- Distance, density, depth based method
- Statistical, classification
- Univariate, multivariate method

2 Transaction specific outlier detection methods [23]

Methods used to detect specifically abnormal transactions called outlier transaction from transactional databases.

- Association rule based outlier detection method.
- Frequent pattern based outlier detection method.

III. EXISTING SYSTEM

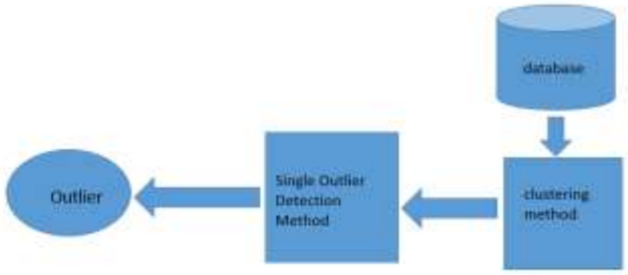


Figure 2. Existing system

Existing system uses distance method or density method for outlier detection. The clustering method used is k means method.

The distance approach uses distance between the centroid and point as detection criteria.

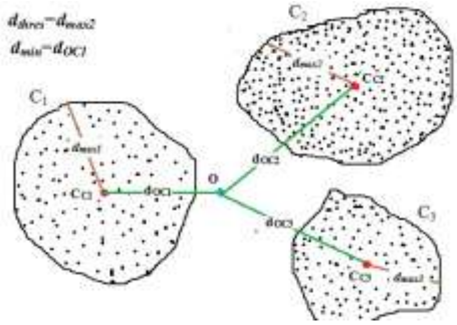


Figure 3. Distance based outlier detection.

Algorithm:

1) Identify threshold distance,

$d_{thres} = \max(d_{max1}, d_{max2} \ldots . d_{maxk})$

2) Identify minimum distance,

$d_{min} = \min(d_{OC1}, d_{OC2} \ldots d_{OCK})$

3) If $d_{min} > d_{thres}$ than object O is outlier, otherwise normal object.

In density based approach value of ε-neighborhood or we can say the number of points in its neighborhood i.e. density is used as criteria to detect outlier.

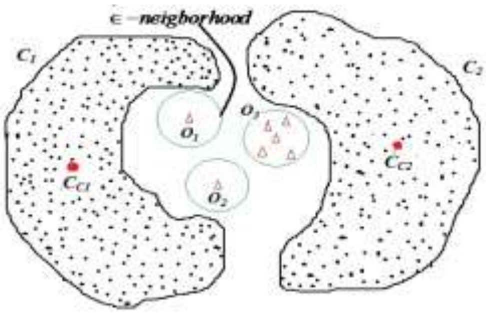


Figure 4. Density based outlier detection.

Algorithm:

1) For object O, find all points in ε –neighborhood of O.

2) Check whether these neighboring points are labelled, if not then object O is an outlier otherwise normal object.

A. *Limitations of existing model:*

In these model k-means method used for clustering works well only for single type attributes (integer or float).But the real world datasets are of mixed type. The performance of the existing outlier detection algorithms are dataset dependent.

In the existing system k-means clustering algorithm is used for clustering data points into datasets which requires number of clusters to be formed mentioned as input to algorithm which is not possible in real world databases with millions of tuples.

IV. PROPOSED SYSTEM

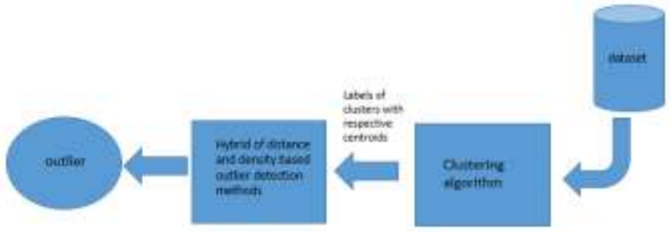

dataset
Clustering algorithm
Labels of clusters with respective centroids
Hybrid of distance and density based outlier detection methods
outlier

Figure 5. Proposed system.

The way to detect the data points that are much variant from others present i.e. outlier detection is done using hybrid approach. Two types of datasets:

- Single datatype attribute dataset ex: Iris.
- Mixed datatype attribute datasets ex: Abalone, Adult census, Housing.

In these proposed model outlier detection and analysis is done using hybrid approach and handle mixed datasets [4]. The clustering method takes the datasets from database and clusters the data using appropriate algorithm. It clusters data based on some similarity factor. Than the valid and invalid clusters are given to distance or density method for analysis based on shape or distribution of datasets.

V. STAGES IN SYSTEM

Following are the stages in the system:

A. Database
B. Clustering method
C. Distance based outlier detection method
D. Density based outlier detection method
E. Output

A. Database:

The database will contain all tables or datasets of mixed type. Example Abalone dataset and adult census. The database will contain cleaned data no repetitions or missing fields will be present. Following shows information about datasets used:

Dataset	Information
1]Fisher's Iris Dataset	Shows information about iris flower species. Datatype: Real. Attributes:4 Instances:150
2] Abalone Dataset	Shows values used for measuring abalone age. Datatype: Categorical, Real, Integer. Attributes:8 Instances:4177
2] Adult Census dataset	Shows details about adults living in Boston. Datatype: Categorical, Integer. Attributes:14 Instances:48842

Table 1: Dataset information.

B. Clustering method:

The process to group together the data points which have some similarities in them is clustering which has a vast variety of application areas. Making the classes such that same class has high similarity while huge dissimilarity with other classes. There are many clustering method e.g. k-means, k-means with farthest seeds, ROCK, CP etc.

The clustering method used is squeezer method [19]. These is clustering method for mixed datasets. One after another scanning is done in these method on dataset. Initial cluster are formed from first tuple arrival. The similarity factor will decide whether to go to new group or same old one.

The outlier are find with less efforts and with one scan or reading. Here the developer does not know how much classes are going to form and it is not required to mention also. Only similarity value is mentioned. Some of the basic concepts of weighted squeezer method [19]:

Definition 1: (Cluster) Cluster = {tid | tid belongs to TID} is subset of TID

Definition 2: The group of different attributes values of A_i w.r.t. C is, Defined as: VAL_i (C) = {tid.A_i |tid belongs to C} where 1<=i<=m

Definition 3: Let a_i belongs to D_t, the support of ai in C w.r.t. Ai is defined as (a_t) = |{tid| tid.Ai=ai}|

Definition 4: Summary for C is defined as:
Summary = {VSt| 1<=i<=m} where VSi = {(at, sup (at))| at belongs to VAL i(C)}

Definition 5: (Cluster Structure, CS) Cluster Structure (CS) for C is defined as: CS= {Cluster, Summary}

Definition 6: Given a Cluster C and tuple t with t_{id} belongs to TID the similarity between C and tid is defined as

Sim (C, tid)= $\sum_{i=1}^{m} wi\ (Sup(ai)) / \sum_{aj\ \in VAli(C)} Sup(aj)$

Where tid.Ai =ai and wi is the weight of attribute A.

Squeezer algorithm:

```
Algorithm Squeezer(D,s)
Begin
While (D has unread tuple){
Tuple =getCurrentTuple(D)
If(tuple.tid==1){
addNewClusterStructure(tuple.tid)}
else{
    for each existed cluster C
    simComputation(C,tuple)
Get the max value of similarity : sim_max
Get the corresponding cluster index: index
If  sim_max>=s
    addTupleToCluster(tuple,index)
Else
    addNewClusterStructure(tuple.tid)}
}
```

```
Sub_Function addNewClusterStructure(tid)
Cluster={tid}
For each attribute value ai on Ai
   $VS_i$ ={a1,1}
 Add $VS_i$ to Summary
CS={Cluster,Summary}
```

```
Sub_Function addTupleToCluster(tuple,index)
Cluster=Cluster U {tuple,tid}
For each attribute value ai on Ai
$VS_i$ =(ai,Sup(ai)+1)
Add $VS_i$  to Summary
CS={Cluster,Summary}
```

```
Sub_Function simComputation(C,tuple)
    1. Defin sim=0
    2. For eacj attribute value ai on Ai
    3.  sim =sim+ probability of ai on C
    4. Return sim.
```

The input to squeezer algorithm is n tuples and output is clusters.First the initial tuple is taken from database and a cluster structure (CS) is formed with C=1. In the same way other rows are read .We compute similarity function for all tuples with all present clusters and stored in CS.Then the max value of similarity is selected .It is compared with threshold value s, if it is larger than s than it is put into it.The CS also updated else a new cluster is created using the function.These id done till all the tuples are scanned.

C. Distance based outlier detection method:

Distance between the data points is used as measure of detection for outlier detection. Distance from its neighbors is used as criteria to judge the points. The point is outlier if the point is far away from centroid otherwise normal point. The distance between two data points or we can say two rows will be calculated using Euclidian distance formula:

$$\sqrt{values\ of\ one\ row^2 + values\ of\ other\ second\ row^2}\ .$$

Algorithm:

1) First identify threshold distance which is maximum of all maximum distance points in each cluster taken from there centroids.
2) The distance between two data points or we can say two rows will be calculated using Euclidian distance formula
$$\sqrt{values\ of\ one\ row^2 + values\ of\ other\ second\ row^2}$$
3) Identify the minimum distance which is smallest of all distances taken from all cluster centroids.
4) If minimum distance is greater than threshold than the point is outlier. Otherwise normal point.

D. Density based outlier detection method:

These approach depends on value of ε-neighborhood distance or we can say the number of points in its neighborhood i.e. density.

Algorithm:

1) For each point find ε-neighborhood density using formula: density = (Area / number of tuples).
2) In finding area maximum distance in cluster will be the diameter.
3) If the density of the point is different or deviates than its neighboring point it is outlier, otherwise normal point.

E. Output:

The output of the project will be in two forms:

1] Scatter graph.

2] Excel sheet form.

1] **Scatter graph:**

In these form the output i.e. the outliers detected will be shown in the scatter graph form. In these graph the formed clusters and outliers will be seen in dots format. Each cluster will be seen in different color so to identify easily. Centroids of cluster will be shown by red circle for all clusters. The X and Y axis will show the distances between each data points.

2] **Excel sheet form:**

In these form output will be shown on excel sheets. It will contain columns for all attributes in dataset with its values, actual output which is expected, the output based on method used for outlier detection i.e. distance based or density based or hybrid. The output will show difference between single method and hybrid method.

VI. EXPERIMENTAL RESULTS

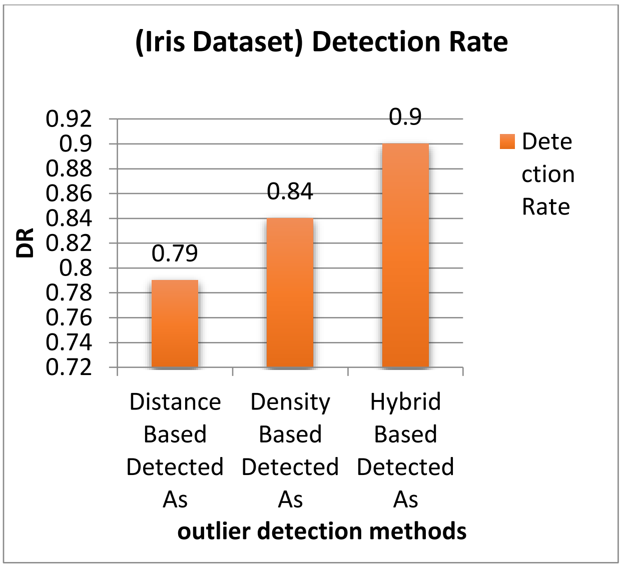
The performance of the system is decided based on one factors or we can say one measures:

1] Detection rate

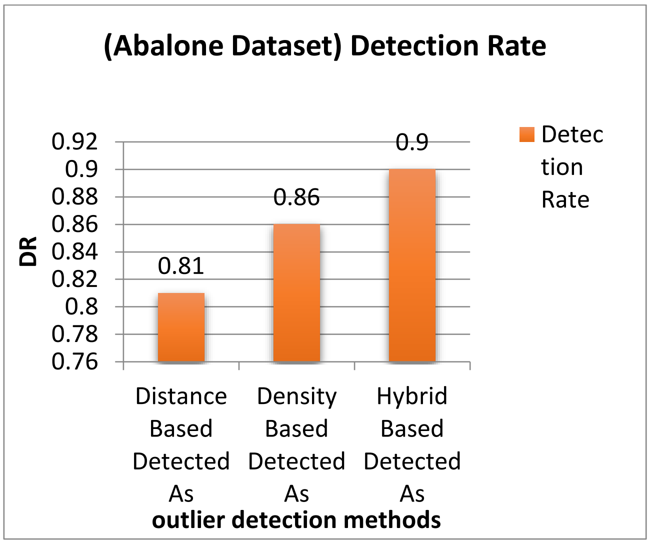
1] Detection rate:

The detection rate is nothing but percentage of outliers detected from actual present. The percentage of output which is same as actual is detected. It is count (actual value = output of system)/100.

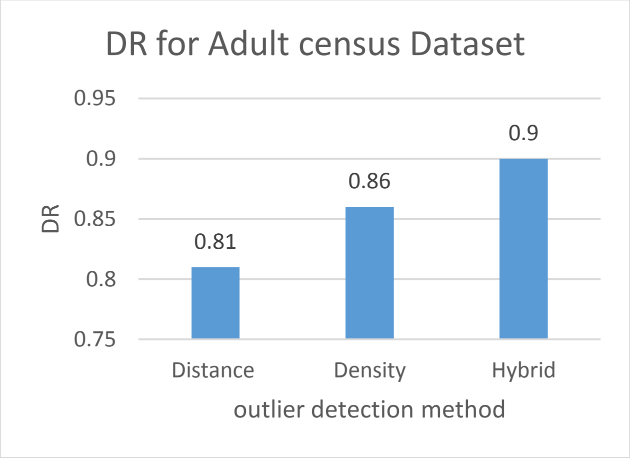
The detection rate of iris dataset for distance, density and hybrid is 0.79, 0.84, 0.9 respectively. These values shows the detection rate of hybrid system is higher than single method systems. Means hybrid mode detects outliers with more accuracy.



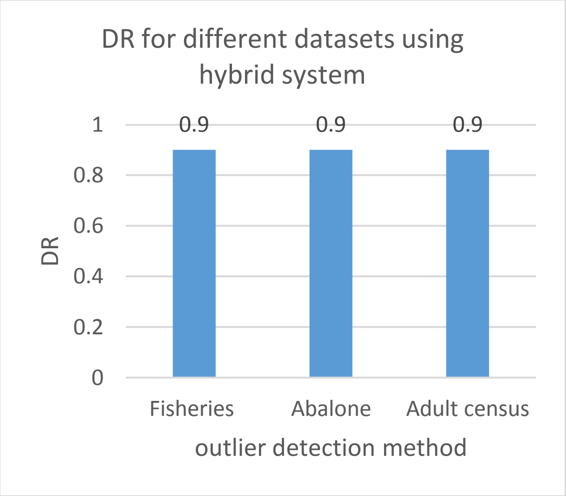
Graph 1: showing DR for iris dataset



Graph 2: showing DR for abalone dataset



Graph 3: showing DR for adult census dataset



Graph 4: DR comparison for different datasets using hybrid system.

VII. CONCLUSION

Large or huge study is done by writing algorithm for numerical datasets (pure integer or pure real) using hybrid method or single outlier detection methods like distance or density. But dataset can be of mixed data type also. So these hybrid model proposed in these project works for mixed datasets. The effectiveness of model results from combined effect of distance and density based method. Distance based method detects outliers from uniform data while density based from non-uniform datasets. Both these methods are popular methods and works well when combined.

The clustering algorithm used works well with mixed datasets. Time required for executing a squeezer algorithm is constant and not depended on number of cluster formed or dataset size which is the drawback of many clustering methods e.g. k means. Most of the real world datasets like medical, breast cancer, teaching ass. Evaluation, credit card fraud detection are mixed datasets. The cleaning of datasets and making it free from error and mistakes done by finding outliers.

Outlier finding can be very helpful for separating useful data and finding information. The k-means method requires number of clusters to form as input which is difficult for real life datasets which contains millions of attributes and rows. The number of groups in output is not required to be specified first so performance or working is not depended on it. From performance analysis it is seen that detection rate which is the measure used for performance analysis is highest for hybrid model as compared to single methods. It is 0.9 for different datasets.

The testing also shows how the DR is constant for different datasets and is higher for hybrid system. The future scope can be dealing with dynamic data or real life datasets like stock exchange dataset or credit card using hybrid model and the performance of the algorithm will be tested on datasets which are related to network errors or the intruders.

REFERENCES

[1] Cao, Lei , Yang, Di ,Wang, Qingyang , Yu, Yanwei ,Wang, Jiayuan , Elke , “Scalable distance-based outlier detection over high-volume data stream”, 2014.

[2] Madhu Nashipudimath, Anjali Barmade, “ Efficient Strategy to detect outlier Transactions”, *International Journal for Soft Computing and Engineering*, [Online].Available: http://www.ijsce.org/attachments/File/v3i6/F2037013614.pdf ISSN: 2231-2307, vol. 3, Issue-6, January 2014.

[3] Shengrui Wang, “Information-Theoretic Outlier Detection for Large-Scale Categorical Data”, 2013.

[4] S. Saharia , “RODHA: Robust Outlier Detection using Hybrid Approach A. Mira*, D.K. Bhattacharyya”, 2012.

[5] S.D.Pachgade, “Outlier Detection over datasets using cluster based and distance based Approach”, 2012.

[6] Jongwoo Lim, “A Framework for Clustering Mixed Attribute Type Datasets”, 2012.

[7] M. V. Jagannatha Reddy1 and B. Kavitha2 ,“Clustering the Mixed Numerical and Categorical

[8] Dataset using Similarity Weight and Filter Method”,2012.

[9] Prasad Pinisett, “Hybrid Algorithm for Clustering Mixed Data Sets”2012.

[10] Murugavel P, “Improved Hybrid clustering and distance based method for outlier removal”, 2011.

[11] Yi Shih*, Jar-Wen Jheng and Lien-Fu Lai, “A Two-Step Method for Clustering Mixed Categorical and Numeric Data Ming”, 2010.

[12] Hans-Peter Kriegel, Peer Kröger, Arthur Zimek, “Outlier Detection Techniques”, 2010.

[13] Anna Koufakou , “Scalable and Efficient Outlier Detection in Large Distributed Data Sets with Mixed-Type Attribute”, 2009.

[14] Yunzin Tao, “Unifying Density-Based Clustering and Outlier Detection”, 2009.

[15] Peng Yang, “A Modified Density Based Outlier Mining Algorithm for Large Dataset, 2008.

[16] Amir Ahmad a, Lipika Dey, “A k-mean clustering algorithm for mixed numeric and categorical data”, 2007.

[17] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, “ Lof: Identifying density-based local outliers” , In Proceedings of ACM SIGMOD on Management of Data, pp. 386-395, 2005.

[18] S. Bay,M. Schwabacher, “ Mining Distance-based outliers in near linear time with Randomization and a simple pruning rule”, pp. 29-38, 2003.

[19] Deng shengum, “Squeezer: An efficient algorithm for clustering mixed data”, 2002.

[20] Shenchung Deng, “Clustering mixed numerical and categorical data: A cluster ensemble approach”, 2002.

[21] E. M. Knorr, R. T. Ng, and V. Tucakov, “Distance-based outliers: Algorithms and applications”, VLDB Journal, 8, pp. 237-253, 2000.

[22] Zengyou He, Xiaofei Xu, Shenchun Deng, “Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches”, (unpublished).

[23] [Online]. Available: http://archive.ics.uci.edu/ml/datasets