Open access Journal International journal of Emerging Trends in Science and Technology

Outlier Detection and Analysis using Hybrid Approach for Mixed datasets

Authors

Anjali Barmade, Prof. Madhu Nashipudimath

Computer Department
Pillai's Institute of Information Technology
Mumbai, India
Email: *anjali.barmade08@gmail.com*

Abstract
There are several approaches of outlier detection 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 existing system uses distance based method for outlier detection and K-means as clustering method.But distance based method has limitation that it fails for non-uniform datasets.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.So we move to proposed model.The proposed model uses hybrid of distance and density outlier detection methods and weighted squuzer method for clustering.Most of the models deals with only single datasets.Here the project deals with mixed datasets.Future scope will be to handle dyanamic data.
Keywords—*outlierdetection,weighted squeezer clustering, hybrid, distance ,density based, k-means, mixed datasets.*

I. INTRODUCTION

Data mining is Extraction of interesting patterns or knowledge from huge amount of data.Outlier detection is a important data mining technique to detect rare events, deviant objects, and exceptions from data called outliers.Outlier detection has been a very important concept in the realm of data analysis. Recently, several application domains have realized the direct mapping between outliers in data and real world anomalies, that are of great interest to an analyst.Outlier detectionrefers to the problem of finding patterns in data that do notconform to expected normal behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains.

Outlier detection has been a widely researched problem and immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas.Outliers are non-conforming patterns in data; that is, they are patterns that do not exhibit normal behaviour. Outlier detection has several applications. For example, outlier detection can be employed as a pre-processing step to clean the data set from erroneous measurements and noisy data points. On the other hand, it can also be used to isolate suspicious or interesting patterns in the data.

There are several approaches of outlier detection 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 existing system uses distance based method for outlier detection and K-means as clustering method.But distance based method has limitation that it fails for non-uniform datasets.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.So we move to proposed model.

The proposed model called outlier detection and analysis uses hybrid of distance and density outlier detection methods and weighted squuzer method for clustering.Most of the models deals with only single datasets.Here the project deals with mixed datasets Takes the benefit of distance based, density based as well as information theoretic approach while identifying an outlier.Performance is independent of dimensionality and number of clusters.

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

Outliers are non-conforming patterns in data; that is, they are patterns that do not exhibit normal behaviour.Points that are sufficiently far away from the normal region (e.g., points O1,

O2, O3 and points in O4 regions) are outliers.[9] .Hawkins (Hawkins, 1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

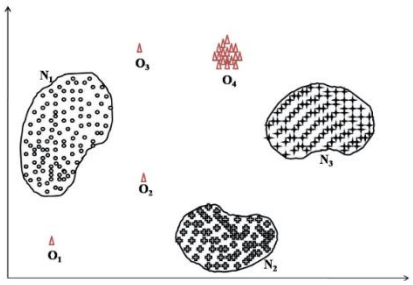

Fig.1. Outliers

B. Outlier Detection:

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,semisupervised
- Distance,density,depth based method
- Statistical, classification
- Univariant,multivariant 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

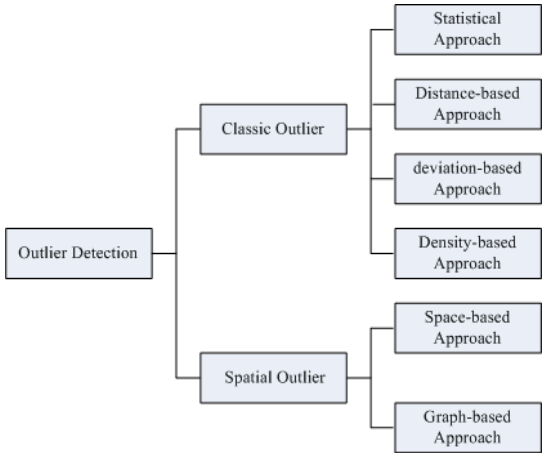

Figure 2. Non Transaction specific outlier detection method

III. EXISTING SYSTEM

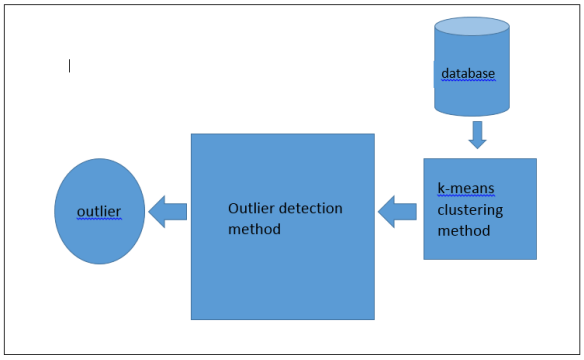

Figure 3. Existing system

The existing model contains a single method system for outlier detection.The k-means method is used for clustering.The formed clusters are than given to outlier detection method for outlier detection .

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 datapoints 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

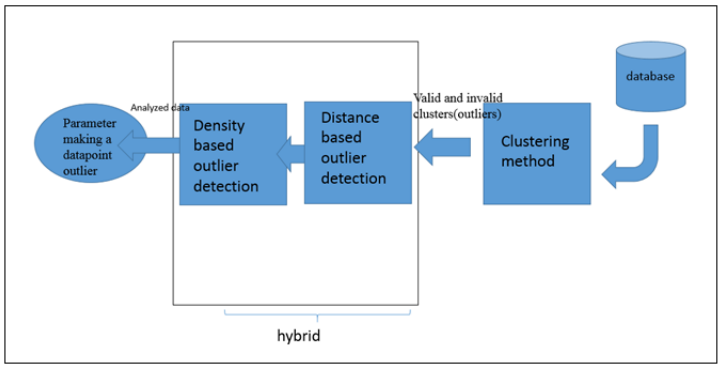

Figure 4. Proposed system.
Finding patterns in data that do not conform to expected normal behavior i.e. outlier detection is done using hybrid approach. Two types of datasets:

- Single datatype attribute dataset ex: synthetic, iris etc
- Mixed datatype attribute datasets ex: credit card fraud, breast cancer detection, teaching ass. Eval.

In these proposed model outlier detection and analysis is done using hybrid approach and handle mixed datasets. Clustering method based on some similarity factor.Than the valid and invalid clusters are given to hybrid of distance and density method for analysis .These methods finds the parameter which causes a datapoint to become outlier.So proper action can be taken.

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 credit card data,stock exchange , breast cancer detection ,teaching assistance evaluation etc.The database will contain cleaned data no repetations or missing fields will be present.

B. Clustering method:

Grouping the objects in datasets such a that objects in same group have high degree of similarity and high degree of dissimilarity with objects in other group.(cluster).

The clustering method used is weighted squeezer method[15].These is clustering method for mixed datasets. Squeezer repeatedly reads tuples from a dataset one by one. When the first tuple arrives, it forms a cluster alone. The consequent tuples are either put into existing clusters or rejected by all existing clusters to form a new cluster by given a similarity function defined between a tuple and a cluster.

The Squeezer algorithm only makes one scan over the dataset. Outliers can be handled efficiently and directly. the algorithm does not require the number of desired clusters as an input parameter. This is very important for the user who usually does not know this number in advance. The only parameter to be pre-specfied is the value of similarity between the tuple and the cluster.Some of the basic concepts of weighted squeezer method:

Defination 1: (cluster)cluster={tid|tid E TID }is subset of TID

Defination 2: Given a cluster C ,the set of diff. attributes values on Ai, wrt C is defined asVALi(C)={tid.Ai|tid E C}

Defination 3: Given a cluster C, thesupport of ai in C wrt Ai is defined as Sup(ai)={tid|tid.Ai=ai}

Defination 4: Given a cluster C ,summary for C is defined as, summary={VSi}, where VSi ={(ai,sup(Ai)|aiE VALi(C))}

Defination 5: Cluster structure for C is, CS={cluster,summary}

Defination 6: Similarity between C and tid is,

$$Sim(C, tid) = \sum_{i=1}^{m} w_i \left(\frac{Sup(a_i)}{\sum_{a_j \in VAL_i(C)} Sup(a_j)} \right)$$ where $tid.A_i = a_i$ and w_i is the *weight*

Weighted squeezer algorithm:

Algorithm *Squeezer* (*D*, *s*)
Begin
1. while (*D* has unread tuple){
2. tuple = getCurrentTuple (*D*)
3. if (tuple.tid = = 1){
4. *addNewClusterStructure (tuple.tid)* }
5. else{
6. for each existed cluster *C*
7. *simComputation(C,tuple)*
8. get the max value of similarity: *sim_max*
9. get the corresponding Cluster Index: *index*
10. if *sim_max* >= *s*
11. *addTupleToCluster(tuple, index)*
12. else
13. *addNewClusterStructure (tuple.tid)* }
14. }
15. *outputClusteringResult ()*

End

Sub_Function *addNewClusterStructure*(*tid*)
1. $Cluster = \{tid\}$
2. for each attribute value a_i on A_i
3. $VS_i = (a_i, 1)$
4. add VS_i to *Summary*
5. $CS = \{Cluster, Summary\}$

Sub-function *addNewClusterStructure*().

Sub_Function *addTupleToCluster*(*tuple, index*)
1. $Cluster = Cluster \cup \{tuple.tid\}$
2. for each attribute value a_i on A_i
3. $VS_i = (a_i, Sup(a_i) + 1)$
4. add VS_i to *Summary*
5. $CS = \{Cluster, Summary\}$

Sub-function *addTupleToCluster*().

Sub_Function *simComputayion*(*C, tuple*)
1. defin $sim = 0$
2. for each attribute value a_i on A_i
3. $sim = sim + probability\ of\ a_i on\ C$
4. return sim

Sub-function *simComputation*().

The squeezer algorithm has n tuples as input and produce clusters as final results.First tuple I read and CS constructed with c={1}.Then next tuples are read.For each tuple similarity function is computed with existing clusters using simComputation() methodand CS is updated.Squeezer algo makes only one scan of dataset.addNewClusterStructure() method forms new cluster.If sim_max larger than threshold s than addTupletoCluster method executed.

These subfunction 1 uses new tuple to initialize cluster and summary and than new CS is created. Subfunction 2 updates the specified CS with new tuple. Subfunction 3 makes use of information stored in CS to get statistics based similarity.

C. Distance based outlier detection method:

Here outliers are detected using distance between the datapoints as a measure for detection. An object O in a dataset T is a DB(p; D) outlier if at least fraction p of the objects in T lies greater than distance D from O. Distance based outlier detection algorithm are as follows: indexed based, nested-loop method, partition based. Here we are using basic distance based method algorithm.

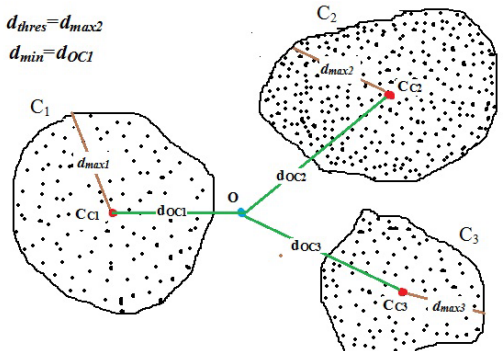

Fig. 5. Distance based outlier detection method

Let us consider the dataset D has the spatial distribution as shown in the figure 3, i.e. the clusters C1, C2,.....Ck in the data set are of convex nature being well-separated from one another. Let CC1, CC2,.....CCk be the respective centroids of the k clusters in the dataset D. Find the maximum distances dmax1, dmax2,.....dmaxk of the centroids to the cluster objects. Then identify the threshold distance dthres as the maximum of the values dmax1, dmax2,.....dmaxk. For a test object O, find dmin, the minimum of the distances dOC1, dOC2,.....dOCk of the object O to the centroid CC1, CC2,.....CCk. If dmin is greater than dthres, then the test object O is identified as an outlier, otherwise it is a normal object.

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

Limitation of distance based method:

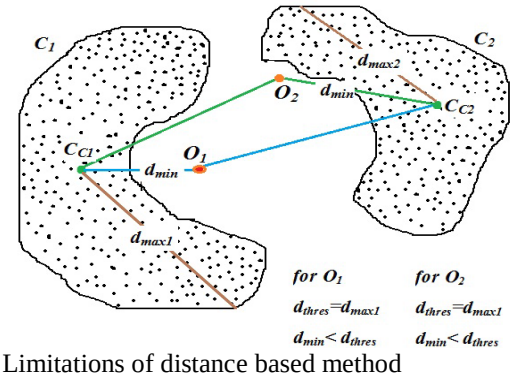

Fig. 6. Limitations of distance based method

This distance-based approach can detect outliers where the dataset is of convex in nature. But the approach fails for datasets of concave nature (as shown in figure

D. Density based outlier detection method:

These approach requires selection of parameter value ε which is distance to check availability of any training samples within ε –neigbourhood of test sample O.[7]

ε –neigbourhood: For object O , ε –neigbourhood finds all samples within distance of ε from object O. Compare the density around a point with the density around its local Neighbors. The relative density of a point compared to its neighbors is computed as an outlier score. O1,O2 are candidate outliers. Performance of approach sensitive to ε value.

Density based outlier detection algorithm: Local outlier correlation integral (LOCI),local outlier factor(LOF). Here basic algorithm used.[3]

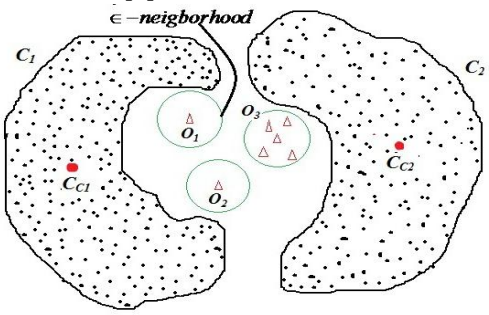

Fig. 7. Density based outlier detection method

Here the objects *O1* and *O2* are the candidate outliers. C_{c1}, C_{c2} are cluster centroids.Within the -neighbourhood of *O1* and *O2* there are some points, but none are labelled i.e. none are the points included within any of the clusters in the data set.

Algorithm:

1)For object O find all points in ε –neigbourhood of O
2)Check whether these neigbouring points are labelled, if not than object O is a outlier otherwise normal object.

Hybrid Algorithm:

Final 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 normal point.But it can be outlier cluster also.so go to step 4.
4)Check whether these neigbouring points are labelled,
if not than object O is a outlier otherwise normal object.

E. Output:

So here the final output is the parameter which causes the datapoint to become outlier and the points which are outlier is displayed.

VI EXPERIMENTAL RESULTS

In these section we will describe some examples and there results.We ran our system for real time dataset credit card transaction database which is of mixed type.

Example:

Table 1 Credit card transaction database

Name	Address	Profession	Card no.	Credit limit	Due date	Payment date	Overseas purchase	Spend on	spending	Avg. spending	PAN card no	Status
Ajay	pune	Engg	111	1 lac	10/3/13	9/4/13	No	Grocery	1000	1000	123	Available
Ajay	pune	Engg	111	1 lac	20/4/13	12/4/13	No	Grocery	2000	1500	123	available
Neha	Mumbai	Doctor	222	3 lac	10/3/13	6/4/13	Yes	Hotel	3000	3000	345	Available
Neha	Mumbai	Doctor	222	3 lac	20/4/13	11/4/13	Yes	Hotel	3000	3000	345	Available
Pooja	Delhi	Teacher	333	1 lac	10/3/13	8/4/13	No	Utility bills	2000	2000	567	Available
Pooja	Delhi	Teacher	333	1 lac	20/4/13	19/4/13	No	Utility bills	2000	2000	567	Available
girish	Surat	Architect	444	2 lac	10/3/13	9/4/13	No	Movie Tickets	5000	5000	678	Stolen
girish	surat	Architect	444	2 lac	20/4/13	18/4/13	no	Utility bills	2000	3500	678	Stolen

Table 2. current credit card transactions of 4 customer

number	city	payment	Spending on	Overseas purchase
111	pune	2000	Grocery	No
222	Mumbai	3000	Grocery	No
333	delhi	2000	Utility Bills	No
444	tirupati	80,000	diamonds	No

$$Sim(t\ \ C,\ tid) = \sum_{i=1}^{m} w_i \left(\frac{Sup(a_i)}{\sum_{a_j \in VAL_i(C)} Sup(a_j)}\right)$$ where $tid.A_i = a_i$ and w_i is the *weight* of att

Calculate similarity between tuple and clusters using below formula

- S1=Sim(1,tid1)=(6(2/2)+2(2/2)+1(1/2)+1(2/2)+4(2/2))-1=12.5
- S2=Sim(2,tid1)=(6(0/2)+2(0/2)+1(2/2)+1(0/2)+4(0/2))-1=0
- S3=Sim(3,tid1)=(6(0/2)+2(0/2)+1(2/2)+1(0/2)+4(2/2))-1=4
- S4=Sim(4,tid1)=(6(0/2)+2(0/2)+1(1/2)+1(0/2)+4(2/2))-1=3.5

Simmax=s1=12.5
Simmax>=s (if s=10)
Index=1
So tid1 similar to cluster 1.
Now,
Tuple tid4 to be added to cluster
S1=Sim(1,tid4)=(6(0/2)+2(0/2)+1(0/2)+1(0/2)+4(2/2))-1=3
S2=Sim(2,tid4)=(6(0/2)+2(0/2)+1(0/2)+1(0/2)+4(0/2))-1=0
S3=Sim(3,tid4)=(6(0/2)+2(0/2)+1(0/2)+1(0/2)+4(2/2))-1=3
S4=Sim(4,tid4)=(6(2/2)+2(0/2)+1(0/2)+1(0/2)+4(2/2))-1=9

Simmax=9 <=s (if s=10)
So create a new cluster will be created for tid4 using addnewclusterstructure(tid)

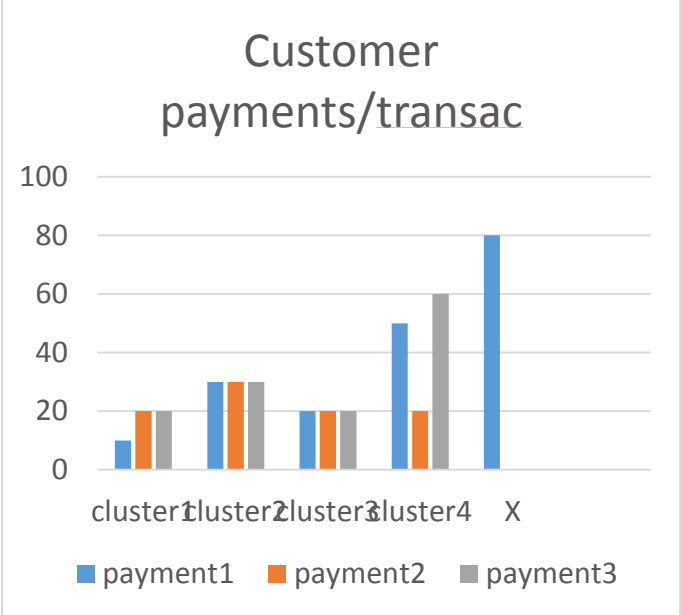

Figure 8. Chart showing average payments of each customer for recent three transactions.

Here ,
Avgofcluster=(16,30,20,44) (centroid of each cluster)
dthres=max((20,30,20,60)-(16,30,20,44))=max(4,0,0,16)=16
dmin=(80-min(16,30,20,44)) =min(64,50,60,36)=36
Here dmin>dthres so X is outlier according to distance based outlier method using payment as parameter.

Consider ε –neigbourhood Radius is 10 .In these radius no cluster point appear. So X is outlier according to density based method.
So analysis says payment parameter helps to detect outlier.
Same for city or some other parameter can be done Diagramatically shown a below , Graph showing use of distance based outlier detection method for uniform data or density based method for non- uniform cluster to detect which dimension make datapoint outlier.
Here in below fig. dmin>dthres . So outlier.

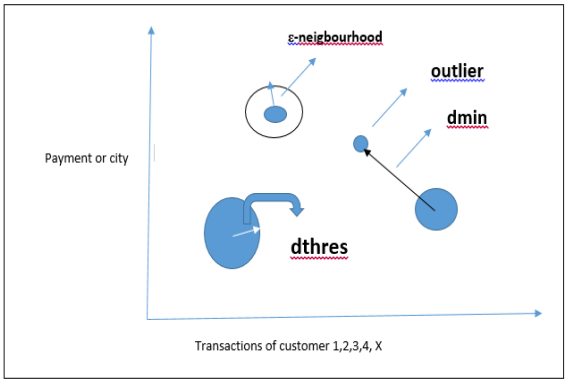

Figure 8. Graph showing use of distance/density method for analysis.

VII. CONCLUSION

Thus in these paper we have seen that there are many algorithm used for outlier detection for pure numerical or pure categorical datasets.But these hybrid method can be used for

mixed datasets present in real world.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.

Time complexity of squeezer algo is linear with size of dataset,number of attributes, number of clusters. Most of the real world datasets like medical,breast cancer,teaching ass. Evaluation,credit card fraud detection are mixed datasets. Most of the clustering algorithms for mixed datasets convert categorical data to numerical or vice versa which causes loss of data and accuracy. outlier detection can be employed as a pre-processing step to clean the data set from erroneous measurements and noisy data points.

On the other hand, it can also be used to isolate suspicious or interesting patterns in the data. 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 clustering used in proposed model does not require number of clusters to form as input. Performance is independent of dimensionality and number of clusters.In the future work we wil investigate to deal with outlier detection from dynamic data using hybrid approach specially for mixed datasets .

REFERENCES

1. Information-Theoretic Outlier Detection for Large-Scale Categorical Data Shu Wu ; Shengrui Wang ,2013
2. Scalable distance-based outlier detection over high-volume data streams Cao, Lei ; Yang, Di ; Wang, Qingyang ; Yu, Yanwei ;Wang, Jiayuan ; Rundensteiner, Elke A.,2014
3. RODHA: Robust Outlier Detection using Hybrid Approach A. Mira*, D.K. Bhattacharyya, S. Saharia,2012
4. Improved Hybrid clustering and distance based method for outlier removal, P. Murugavel ,2011
5. utlier Detection over datasets using cluster based and distance based approach, S.D.Pachgade ,2012
6. S. Bay and M. Schwabacher. Distance-based outliers in near linear time with randomization and a simple pruning rule. In Proceedings of the Ninth ACM SIGKDD, pages 29-38. Keleuven Press
7. 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, pages 386-395.
8. A Two-Step Method for Clustering Mixed Categroicaland Numeric DataMing-Yi Shih*, Jar-Wen Jheng and Lien-Fu Lai ,2010
9. Scalable and Efficient Outlier Detection in Large Distributed Data Sets with Mixed-Type Attributes by Anna Koufakou
10. Improving Categorical Data Clustering Algorithm byWeighting Uncommon Attribute Value MatchesZengyou He, Xiaofei Xu, Shenchun Deng
11. A Framework for Clustering Mixed Attribute Type Datasets ,,Jongwoo Lim .
12. E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. VLDB Journal, 8: 237-253, 2000.
13. http://archive.ics.uci.edu/ml/datasets
14. A k-mean clustering algorithm for mixed numeric and categorical data,Amir Ahmad a,*, Lipika Dey b
15. Squeezer: An efficient algorithm for clustering mixed data,Deng shengum
16. Clustering mixed numerical and categorical data:A cluster ensemble approach, shenchugn deng
17. A Framework for Clustering Mixed Attribute Type Datasets Jongwoo Lim1, Jongeun Jun2, Seon Ho Kim2 and Dennis McLeod1
18. Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method ,M. V. Jagannatha Reddy1 and B. Kavitha2
19. Hybrid Algorithm for Clustering Mixed Data Sets,V.N. Prasad Pinisetty
20. Outlier Detection Techniques,Hans-Peter Kriegel, Peer Kröger, Arthur Zimek
21. Unifying Density-Based Clustering and Outlier Detection, Yunzin Tao
22. A Modified Density Based Outlier Mining Algorithm for Large Dataset, Peng Yang
23. An Efficient Strategy to detect outlier transactions, http://www.ijsce.org/attachments/File/v3i6/F2037013614.pdf, International Journal for Soft Computing and Engg., Madhu Nashipudimath, Anjali Barmade.