ISSN 22773061

A Model for Improving Classifier Accuracy for Categorical Data Using Outlier Analysis

Lakshmi Sreenivasa Reddy.D, Dr B.Raveendrababu, Dr A.Govardhan
Department of CSE, Rise Gandhi Group of institutions, Ongole, India
VNR VJIET, Hyderabad, India
Director of Evaluation, JNTUH, Hyderabad, India
urdlsreddy@yahoo.com, rboghapathi@yahoo.com, govardhan_cse@yahoo.co

Abstract

Anomalies are those records, which have different behavior and do not comply with the remaining records in the dataset. Outlier analysis is the concept to find anomalies in Datasets. Detecting outliers efficiently is an important issue in many fields of science, medicine and technology. Many methods are available to detect anomalies in numerical datasets but a limited number of methods available for categorical datasets. In this work, a novel method to detect outliers in categorical data based on entropy is proposed. This algorithm finds anomalies based on each record score and has great intuitive appeal. These scores called BAD scores. This algorithm utilizes the frequency of each value in the dataset. Greedy method needs k- scans of dataset to find 'k' outliers where as the proposed method needs only one scan of dataset and it calculates BAD score of each record directly. It avoids the problem of giving 'k' as an input and can find any number of outliers based on our data set directly.AVF method has less time complexity when compared with the other methods like Greedy, FPOF and FDOD. Greedy has good accuracy when compared with other methods like AVF and FPOF, FDOD (which are based on frequency patterns of all combinations of values in each record). Our algorithm shows better results in accuracy than AVF algorithm and Greedy. But this method has reached nearest to AVF in time complexity. This algorithm has been applied on Nursery dataset and Bank dataset taken from "UCI Machine Learning Repository". In this work, it is proposed to extend Normal distribution [11], and Fuzzy concept [12] to BAD score [13] that is NAVF combined with Fuzzy AVF is applied to BAD Score. Numerical attributes are excluded from Datasets for our analysis. The experimental results show that it is efficient for outlier detection in categorical dataset.

Keywords: Data Mining, Outlier detection, BAD Score, Normal AVF, Fuzzy AVF

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol7, No 1
editor@cirworld.com
www.cirworld.com, member.cirworld.com

Introduction

Outlier analysis is an important research field in many fields like networks, medicine and Business decisions. This analysis concentrates on detecting infrequent data records in dataset. Most of the existing systems concentrate on numerical attributes or ordinal attributes and sometimes, categorical attribute values can be converted into ordinal values there to categorical values. This process is not always preferable. This paper presents a novel method for finding anomalies in categorical data. AVF method is one of the efficient methods to detect outliers in categorical data in time complexity and greedy in accuracy. The mechanism in this AVF method is that, it calculates frequency of each value in each data attribute and finds their probability, and then it finds the attribute value frequency for each record by averaging probabilities and selects top k- outliers based on the least AVF score. The parameter used in this method is only “k”, the number of outliers. FPOF is based on frequent patterns of each record which are adopted from Appriori algorithm [1]. This calculates frequent item sets from each object. From these frequencies it calculates FPOF score and finds the top k-outliers as the least k- FPOF scores. Time Complexity is more for FPOF algorithm when compared to AVF algorithm. The parameters used in FPOF and FDOD are σ, a threshold value to decide frequent item sets in each data object and ‘k’, the number of outliers. The next method is based on Entropy score.

I. Existing Approaches for numerical datasets

a) Statistical based Methods

Statistical Methods adopt a parametric model that describes the distribution of the data and the data is mostly univariate [3, 4]. There are many drawbacks in this method like difficulty of finding a correct model for different datasets and the efficiency of these models decreases as the number of dimensions increases [4]. The remedy for this problem is applying the Principle Component Analysis. Another method to handle high dimensional datasets is attribute relevance analysis .These ideas are not useful for more dimensions in any Dataset.

b) Distance-Based Methods

These methods are not based on any assumptions about the distribution of the data records because they should compute the distances between all records. But these methods make a high complexity. So these methods are impractical for large datasets with more records. Knorr’s et al. [5], achieved some improvements in the distance-based algorithms. They have explained that a part of dataset records belong to each outlier must be less than some threshold value. Still it is an exponential on the number of nearest neighbors.

c) Density Based Methods

Density based methods adopt on finding the density of the data records and identifying outliers as those lying in areas with low density. Breunig et al. have described a local outlier factor (LOF) to identify local outliers whether an object contains sufficient neighbor around it or not [6]. LOF decided a record as an outlier when the record LOF is less than the user defined threshold. Papadimitriou et al. described a similar method called Local Correlation Integral (LCI). This method selects the minimum points (min pts) in LOF through statistical methods [7]. These density based methods have some advantages that they can detect outliers those are left by techniques with single, global criterion methods.

d) Deviation based Methods

These methods find characteristics of objects instead of finding distances, densities and statistical parameters. The objects that deviate from the given description are treated as outliers. The complexity is Linear with the dataset size. The terminology used in our paper is given below

II. TERMINOLOGY

Term	Description
K	Target number of outliers
N	Number of objects in Dataset
M	Number of Attributes in Dataset
xi	ith object in Dataset ranging from 1 to n
Aj	jth Attribute ranging from 1 to m
D(Aj)	Domain of distinct values of jth attribute
xij	cell value in ith object which takes from domain dj of jth attribute Aj
D	Dataset

V	Set of all distinct values in Dataset D
I	Item set
F	Frequent Item set
f(xij)	Frequency of xij value
FS(xi)	Set of frequent Item sets of xi object
IFS(xi)	Set of infrequent Item sets of xi object
minsup	Minimum support of frequent item set
Support(I)	Support of Item set I

III. Existing Approaches for categorical datasets

a) Greedy algorithm

If any dataset contain outliers then it deviates from its original behavior and the dataset gives wrong results in data analysis. Greedy algorithm proposed the idea of finding a small subset of records; these contribute to eliminate the uncertainty of the dataset. This disturbance is also called entropy or disturbance. We can define it formally as ‘let us take a dataset D with ‘m’ attributes A1, A2----- Am and d(Aj) is the domain of distinct values in the variable Aj, then the entropy of single attribute Aj is

$$E(A_j) = \sum_{x \in D(A_j)}^{M} p(x) \log_2 (p(x)) \qquad (1)$$

Since all attributes are independent to each other, Entropy of the entire dataset D= { A1, A2-------- Am} is equal to the sum of the entropies of each one of the ‘m’ attributes, and it is defined as follows

$$E\left(A1, A2 - - Am\right) = E\left(A1\right) + E\left(A2\right) + - - - E\left(Am\right) \qquad (2)$$

If we want to find entropy the Greedy algorithm takes k outliers as input [2]. All objects in the dataset are initially designated as non-outliers. Initially all attribute value’s frequencies are computed and using these frequencies the initial entropy of the dataset is calculated. Then, Greedy algorithm scans k times over the data to determine the top k outliers keeping aside one non-outlier each time. While scanning each time every single non-outlier is temporarily removed from the dataset once and the total entropy is recalculated for the remaining dataset. For any non-outlier point that results in the maximum decrease for the entropy of the remaining dataset is the outlier data-point removed by the algorithm. The Greedy algorithm complexity is O(k *n*m*d), where k is the required number of outliers, n is the number of objects in the dataset D, m is the number of attributes in D, and d is the number of distinct attribute values, per attribute. Pseudo code for the Greedy Algorithm is as follows

Algorithm: Greedy

```
Input: Dataset – D
Target number of outliers – k
Output: k outliers detected
label all data points x1,x2,---xn as non-outliers
Calculate initial frequency of each attribute value and update hash table in each iteration
calculate initial entropy
counter = 0
while ( counter != k ) do
        counter++
        while ( not end of database ) do
```

```
                         read next record ‘xi ’ labeled non-outlier
                        label ‘xi’ as outlier
                        calculate decrease in entropy
                        if ( maximal decrease achieved by record ‘o’ )
                            update hash tables using ‘o’
                            add xi to set of outliers
                          end if
                  end while
end while
```

However entropy needs k as input and need to find number of outliers more times to get optimal accuracy of any classification model.

b) Attribute Value Frequency (AVF) algorithm

The algorithm discussed above is linear with respect to data size and it needs k-scans each time. The other models also exist which are based on frequent item set mining (FIM) which need to create a large space to store item sets, and then search for these sets in each and every data point .These techniques can become very slow when a low threshold value is chosen, to find frequent item sets from dataset. Another simpler and faster approach to detect outliers that minimizes the scans over the data and does not need to create more space and more search for combinations of attribute values or item sets is Attribute Value Frequency (AVF) algorithm. An outlier point xi is defined based on the AVF Score below:

$$\text{AVF Score } (xi) = \frac{1}{M}\sum_{j=1}^{M} f(x_{ij}) \qquad (3)$$

In this approach [1] again we need to find k-outliers many times to get optimal accuracy of any classification model.

Pseudo code for the AVF Algorithm is as follows

```
Input : Database D (n points _ m attributes), Target number of outliers - k
Output: k detected outliers
 Label all data points as non-outliers;
for each point xi, i = 1 to n do
      for each attribute j, j = 1 to m do
       Count frequency f(xij)of attribute value   xij;
      end
end
for each point xi, i = 1 to n do
      for each attribute j, j = 1 to m do
         AVF Score(xi) += f(xij);
      end
     AVF Score(xi) /= m;
end
Return k outliers with mini (AVF Score)
```

The AVF algorithm time complexity is lesser in comparison with Greedy algorithm. Since AVF needs only one scan to detect outliers, the time complexity is less. The complexity of AVF is O (n * m). AVF needs 'k' value as input to find 'k'-outliers.

In FPOF [8], frequent pattern based outlier detection is discussed and in this too, k-value and another parameter 'σ 'is required as threshold. Also it is discussed about frequent pattern based method to find infrequent object, in which also k-value, and another parameter 'σ' as input are required. In the proposed model(BAD), an optimal number of outliers in a single instance is considered to get an optimal precision classification model with good precision and low recall value. This method calculates 'k' value itself based on the frequency. Let us take the data set 'D' with 'm' attributes A1, A2-----Am and d (Ai) is the domain of distinct values in the variable Ai. k is the number of outliers which are normally distributed. To get 'k' this model uses Gaussian theory (NAVF) and fuzzy theory. If the frequency is less than "mean-3 S.D" then this model uses fuzzy logic. This method uses AVF score formula, but no k-value is required

c) Frequent Pattern Outlier Factor FPOF algorithm

This algorithm utilizes the Appriori algorithm as first step to find all frequent Item sets. This method needs a human defined threshold value called" minimum support" as input to find frequent item sets. By taking this threshold value, it makes all combinations of values of each record and compares the frequency of each combination with threshold value and finds each combination whether it is frequent or not. To find frequency of each combination, it needs one scan of the dataset. Even for one record it scans dataset, so many times. This is the major disadvantage of this algorithm. FPOF algorithm takes more memory and more time to generate combinations and their frequency.

FPOF algorithm:

```
Input: Dataset D= { A1,  A2-------- Am}, minsup, number of outliers-'k'.
Output: k detected outliers
Step 1: Mine frequent Item sets in D for each record xi in D by using Appriori
Step 2: Calculate all combinations and store them in vector V
Step 3: For each frequent pattern F in FS (xi)
                 if  'xi' contains F
                     Update V-Decrement number of combinations for the size of F
                end if
        end
Step 4:For each combination  in vector V
          FPOF score (xi) + = support (F)/size of vector V
        end
end
return 'k' outliers.
```

.

BAD score Algorithm

The algorithms discussed above, need many scans of dataset for each data object, but this proposed algorithm needs only one scan of dataset for all records and it finds frequency of each value in dataset .This algorithm declares the records as outliers if any record value having its frequency as one . This algorithm finds the disturbance of each record in data set and finds k-records as those highest BAD scores [13]. This algorithm is applied on Breast cancer data taken from UCI Machine Learning repository [10] and in this model it is defined that

1) Dataset as D= {A1, A2-------- Am},

2) D (Aj) = Domain of all distinct values in attribute 'j',

3) V=Set of all distinct values in dataset 'D'= D (A1)$\cup$ D (A2) $\cup$D (A3)….D (Am) = {V1j, V2j V3j V4j ….Vkj}

Where 1$\leq$k$\leq$n and 1$\leq$j$\leq$m for each record, then our approach to find BAD score for each record as below

$$Score1 = -\sum_{j=1}^{m}\left[\sum_{\forall V_{kj}\in D(A_j)\cap X_{ij}=V_{kj}} \frac{(f(V_{kj})-1)}{(n-1)}\log_{10}\left(\frac{(f(V_{kj})-1)}{(n-1)}\right)\right] \quad (4)$$

$$Score2 = -\sum_{j=1}^{m}\left[\sum_{\forall V_{kj}\in D(A_j)\cap X_{ij}\neq V_{kj}} \frac{(f(V_{kj}))}{(n-1)}\log_{10}\left(\frac{(f(V_{kj}))}{(n-1)}\right)\right] \quad (5)$$

$$BADScore = \frac{1}{Score1 + score2} \quad (6)$$

“BAD Score” Algorithm:

```
Input: Dataset D= {A1, A2-------- Am}, number of outliers-‘k’,
                           number of outliers-‘k’
Output: k detected outliers
Step 1: Find frequencies of all values in dataset D and store them in V= {V1, V2, V3,…..Vk}
Step 2: For each record ’xi’ in D
              if Xij=Vkj and f(X ij)=1 for any ‘j’
                 go to step 6
                 else if Xij=Vkj and VkjεD(Aj)
                         find Score1
      else if  Xij≠Vkj and VkjεD(Aj)
                                  find Score2
                         end else
                  end else
                endif
             end
Step -3: Find the Sum of Score1 and score2
Step 4: Then Find Reciprocal of the Sum
Step5: Find top k-Scores
Step6: return k-outliers
```

d) Our Approach 1 (Normally Distributed BAD Score)

An approach is derived based on mixing of normal distribution with BAD score (NBAD) like NAVF [11], and experiments conducted with this approach. This approach calculates N-seed values ‘a’ and ‘b’ as given below

b=mean (f(xi)), (7)

a=b-3*std (xi), if max (f(xi)) > 3*std (f(xi)) (8)

Comparing each record with these seed values this approach gives us outliers. With this model experiments are conducted on bank data taken from “UCI Machine Learning Repository” [10]. Experimental results are given below

e) Our Approach2 (Fuzzy based BAD Score)

Another approach is derived based on fuzzy logic applied on BADscore (FBAD) like FAVF [12], and experiments are conducted with this approach .This model calculates Fuzzy seed values ‘a’ ‘b’ and ‘c’ as given below

$$b = mean\ (fi) \qquad (9)$$

$$a = \begin{cases} b - 3 * STD(fi) & if \quad \max(fi) > 3 * STD(fi) \\ b - 2 * STD(fi) & if \quad \max(fi) > 2 * STD(fi) \\ b - STD(fi) & if \quad otherwise \end{cases} \qquad (10)$$

$$c = \begin{cases} b + 3 * STD(fi) & if \quad \max(fi) > 3 * STD(fi) \\ b + 2 * STD(fi) & if \quad \max(fi) > 2 * STD(fi) \\ b & if \quad otherwise \end{cases} \qquad (11)$$

Here, the outliers are derived based on fuzzy score based on S-fuzzy function .The S-fuzzy function is

$$s = \begin{cases} 0 & if \quad fi < a \\ 2\left\{\frac{fi-a}{c-a}\right\}^2 & if \quad a \le fi \le b \\ 1-2\left\{\frac{fi-a}{c-a}\right\}^2 & if \quad b \le fi \le c \\ 1 & if \quad fi > c \end{cases} \qquad (12)$$

Where ‘fi ‘ is the BAD score of ‘ith’ object in dataset ‘D’. This method has been applied on bank data taken from “UCI Machine Learning Repository” [10]. Experimental results are given below

IV. Experimental results

This model has been used on Bank Data from UCI Machine repository [10]. This method has implemented the approach of using MATLAB tool. Bank data consists of 45211 records and ten attributes “contact”, “default”, “education”,” housing”, “job”, “loan”, “marital status”, “month”, “p outcome” and a class label attribute ‘Y’. This dataset contain two types of classes. One is yes, other one is no. This data is divided into two parts based on class attribute, first part contains 39922 records with “no” class, and second part contain 5299 records with “yes” label which are used as outliers in our experiment. This separation is done by Clementine 11.1 tool. In first iteration 2645 sample records are selected randomly using Clementine tool; and then from each two records one is selected. These 2645 records are mixed up with part one which totals 42567 records and NBAD, NAVF, FBAD and FAVF are all applied to get outliers. Class attribute contains two values, all the remaining attributes “contact”, “default”, “education”,” housing”, “job”, “loan”, “marital status”, “month”, “p outcome” contain 3, 2, 4, 2, 12, 3,2, 12 and 4 values respectively, and the found outliers are given in Tables. Similarly 1058 records are selected randomly as one record from each five records and mixed up with first part and applied the same process .The results are given in the below Tables. Similarly one record is selected from each eight records and ten records and repeated the same process. Then these algorithms are applied on selected samples. Results are given in Tables below. This method has been implemented on Bank data which is taken from UCI Machine learning repository [10]. Comparison of results is given in Table II. Comparison graphs are given in the subsequent Figures.

TABLE I. COMPARISON OF NUMBER OF OUTLIERS FOUND IN BANK

From Table II, the results reveal that NBAD gives good number of outliers when compared with all other methods in any sample method. FBAD also gave good results in comparison with NAVF and FAVF except at 1-in-5 sample. Graph is given in Fig. I.

The same process has been applied on Nursery data with 6236 records [10]. The results show that NABD has found good number of outliers in this too. Graph is given in Fig. II

omparison of number of outliers found in nursery data

Sample Method	NBAD	FBAD	NAVF	FAVF
1-in-2	83	66	44	56
1-in-5	382	35	133	162
1-in-8	238	98	238	238
1-in-10	190	190	190	190

Different classifiers are tested on Bank data after deleting the outliers from NBAD, FBAD, NAVF, and FAVF.

Neural Network (NN), Logistic Regression (LR), CHAID, Decision Logic (DL) classifiers are applied by using Clementine 11.1 tool. The Accuracies of these models are given for the sample 1-in ten approaches in Table IV. Graph is given in Fig. III

Sample Method	NBAD	FBAD	NAVF	FAVF
1-in-2	933	363	274	279
1-in-5	543	264	366	199
1-in-8	396	231	152	154
1-in-10	336	187	126	126

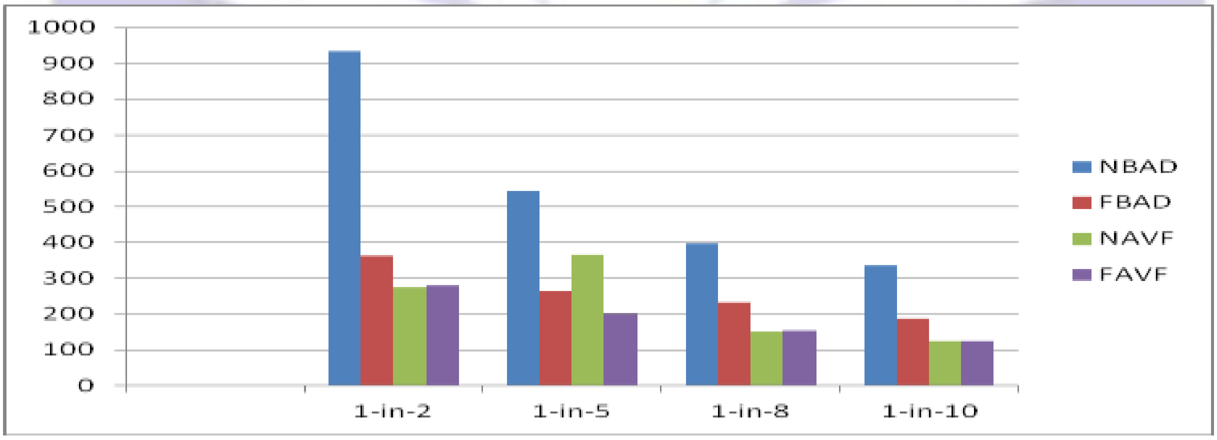

Fig I . Number of outliers found for Bank Data

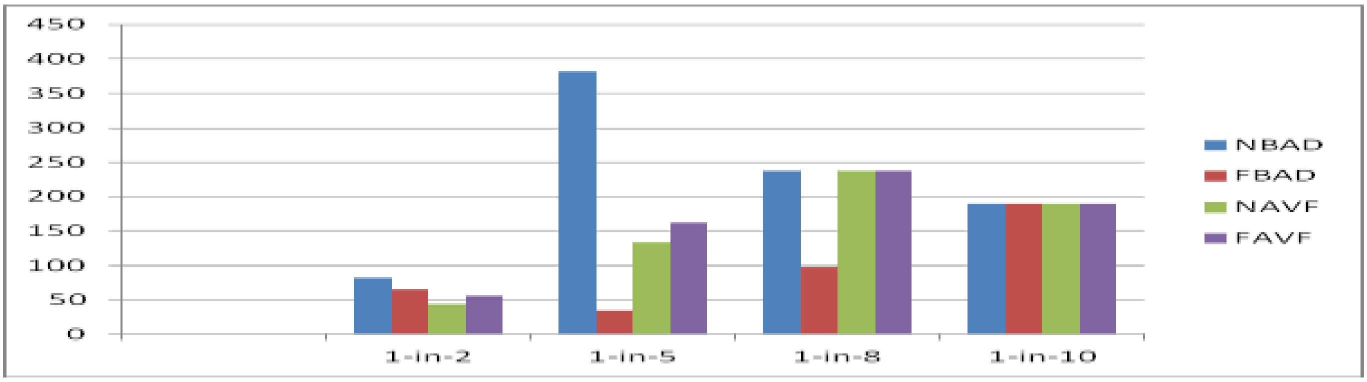

Fig II . Number of outliers found for Nursery Data

Comparison of accuracies of classifiers on bank data(1-in-10 sample)

Classifier	NBAD	FBAD	NAVF	FAVF	Direct (with outliers)
NN	99.503	99.147	98.998	98.998	89.586
LR	99.503	99.147	98.998	98.998	89.223
CHAID	99.503	99.147	98.998	98.998	89.386

From Table IV all classifiers gave good results after deleting outliers by NBAD and FBAD has performed next. Accuracy graph is given in Fig III.

V. Conclusion and Future work

To sum up, this proposed method gives more number of outliers when compared with existing models. Our model is good for categorical datasets to delete precise outliers. The combination of Normal distribution with our BAD score algorithm finds more outliers and train the classifiers with good accuracy. In future, different classifiers separately on mixed type of Datasets can be modeled.

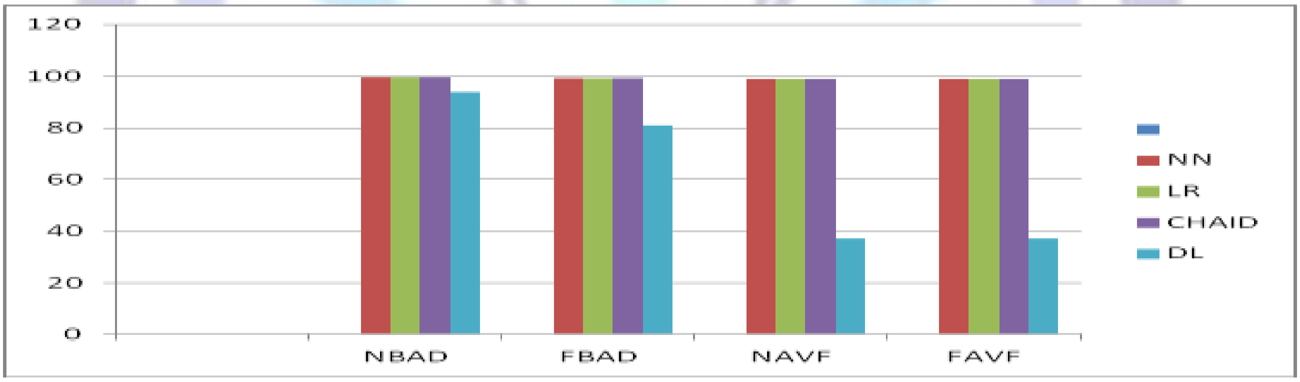

Fig III . Calssifiers accuracies for Bank Data

VI. References

M. E. Otey, A. Ghoting, and and A. Parthasarathy, "Fast Distributed Outlier Detection in Mixed-Attribute Data Sets," Data Mining and Knowledge Discovery

He, Z., Deng, S., Xu, X., "A Fast Greedy algorithm for

outlier mining", Proc. of PAKDD, 2006.

I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds.

New York: Academic, 1963, pp. 271–350.

P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining: Pearson Addison-Wesley, 2005

E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB Journal, 2000.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density based local outliers," presented at ACM SIGMOD International Conference on Management of Data, 2000

S. Papadimitriou, H. Kitawaga, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," presented at International Conference on Data Engineering, 2003

Z. He, X. Xu, J. Huang, and S. Deng, "FP-Outlier: Frequent Pattern Based Outlier Detection", Computer Science and Information System (ComSIS'05)," 2005

Shu Wu and Shengrui Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data, IEEE Transactions on Knowledge Engineering and Data Engineering,2011

Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

LakshmiSreenivasaReddy.D, .B.RaveendraBabu and A.Govardhan, "Outlier Analysis of Categorical Data using NAVF"', Informatica Economica vol 17, Cloud computing issue 1, 2013.

LakshmiSreenivasaReddy.D, B.RaveendraBabu "Outlier Analysis of Categorical Data using FuzzyAVF", presented at IEEE international conference ICCPCT-2013,pp 1259-1263.

LakshmiSreenivasaReddy.D, B.RaveendraBabu and A.Govardhan, 'A Novel Approach to Find Outliers in Categorical Dataset "presented at Elsevier AEMDS-2013

THE AUTHORS

Mr.Lakshmi Sreenivasareddy.D obtained his Masters degree from JNTU University, Hyderabad. He is doing his Ph.D in Computer Science and Engineering in JNTUH Hyderabad. He is currently heading the Department of Computer Science & Engineering, RISE Gandhi Groups of Institutions Ongole. He has 10 years of teaching experience

Dr B. Raveendra Babu, obtained his Masters in Computer Science and Engineering from Anna University, Chennai. He received his Ph.D. in Applied Mathematics at S.V University, Tirupati. He is currently working in VNR &VJIET Hyderabad,. He has 27 years of teaching experience. He has several publications to his credit. His research areas of interest include VLDB, Image Processing, Pattern analysis and Wavelets.

Dr.A.Govardhan did his BE in Computer Science and Engineering from Osmania University College of Engineering, Hyderabad in 1992, M.Tech from Jawaharlal Nehru University(JNU), Delhi in 1994 and he earned his Ph.D from Jawaharlal Nehru Technological University, Hyderabad (JNTUH) in 2003. He has several International publications to his credit. He is the Director of Evaluation for the same University now.