Outlier Analysis of Categorical Data using NAVF

Lakshmi Sreenivasa Reddy.D † and B.Raveendrababu††and A.Govardhan†††

Rise Gandhi Group of institutions, India VNR VJIET, India JNTUH, India

Summary
Outlier mining is an important task to discover the data records which have an exceptional behavior comparing with other records in the remaining dataset. Outliers do not follow with other data objects in the dataset. There are many effective approaches to detect outliers in numerical data. But for categorical dataset there are limited approaches. We propose an algorithm NAVF (Normally distributed attribute value frequency) to detect outliers in categorical data. This algorithm utilizes the frequent pattern data mining method. It avoids problem of giving k-outliers to get optimal accuracy in any classification models in previous work like Greedy, AVF, FPOF, and FDOD while finding outliers. The algorithm is applied on UCI ML Repository datasets like Nursery, Breast cancer mushroom and bank dataset by excluding numerical attributes. The experimental results show that it is efficient for outlier detection in categorical dataset.

Key words:
Outliers, Categorical, AVF, NAVF, FPOF, FDOD

1. Introduction

Outlier analysis is an important research field in many applications like credit card fraud, intrusion detection in networks, medical field .This analysis concentrate on detecting infrequent data records in dataset.

Most of the existing systems are concentrated on numerical attributes or ordinal attributes .Sometimes categorical attribute values can be converted into numerical values. This process is not always preferable. In this paper we discuss a simple method for categorical data is presented.

AVF method is one of the efficient methods to detect outliers in categorical data. The mechanism in this 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 no. of outliers. FPOF is based on frequent patterns which are adopted from Appriori algorithm [1]. This calculates frequent patterns item sets from each object. From these frequencies it calculates FPOF score and finds the least k- outliers as the least FPOF scores. This method takes more time to detect outliers comparing with AVF. The parameters used in it are σ, a threshold value to decide frequent sub sets in each data object. The next method is based on Entropy score. Greedy [2] is another method to detect outliers from categorical data. The previous approaches used to detect outliers were Statistical based

This method adopted a parametric model that describes the distribution of the data and the data was mostly univariate [3, 4]. The main drawbacks of this method are difficulty of finding a correct model for different datasets and their efficiency decreases as the no. of dimensions increases [4]. To rectify this problem the Principle component method can be used. Another method to handle high dimensional datasets is to convert the data records in layers however; these ideas are not practical for more than or equal to three dimensions.

1.1 Distance-Based

Distance based methods do not make any assumptions about the distribution of the data records because they must compute the distances between records. But these make a high complexity. So these methods are not useful for large datasets. There are some improvements exist in the distance-based algorithms, such as Knorr's et al. [5], they have explained that apart 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.

1.2 Density Based

These methods are based on finding the density of the data and identifying outliers as those lying in regions with low density. Breunig et al. have calculated a local outlier factor (LOF) to identify whether an object contains sufficient neighbor around it or not[6]. They have decided a record as an outlier when the record LOF which is a user defined threshold. Papadimitriou et al. presented a similar technique called Local Correlation Integral, which deals of selecting the minimum points (min pts) in LOF through statistical methods in [7]. The density based methods have some advantages that they can detect outliers that are missed by techniques with single, global criterion methods. The terminology used in this paper is given below

1.3 Statistical based

This method adopted a parametric model that describes the distribution of the data and the data was mostly univariate [3, 4]. The main drawbacks of this method are difficulty of

Manuscript received January 5, 2014
Manuscript revised January 20, 2014

finding a correct model for different datasets and their efficiency decreases as the no. of dimensions increases [4]. To rectify this problem the Principle component method can be used. Another method to handle high dimensional datasets is to convert the data records in layers however; these ideas are not practical for more than or equal to three dimensions.

TABLE1. TERMINOLOGY

Term	Description
k	Target number of outliers
n	Number of objects in Dataset
m	Number of Attributes in Dataset
d	Domain of distinct values per attribute
xi	i th object in Dataset ranging from 1 to n
Aj	jth Attribute ranging from 1 to m
xij	A value in xi th object which takes from domain dj of j th attribute Aj
D	Dataset
I	Item set
F	Frequent Item set
Fi	AVF score(xi)
P(xij)	Frequency of xij value
FS	Set of frequent Item sets
IFSi	Set of infrequent Itemsets of ithobject
Minsup	Minimum support of frequent itemset
Support(I)	Support of Itemset I

2. Algorithms

2.1 Greedy algorithm

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

$$E(Aj) = -\sum_{x \varepsilon d(Aj)} p(x) \log_2(p(x)) \quad (1)$$

Because of 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 is defined as follows

$$E(A1, A2\text{--------} Am) = E(A1) + E(A2) + \text{-------} E(Am) \quad (2)$$

When we want to find entropy the Greedy algorithm takes k outliers as input [2]. All records in the set 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 in put and need to find number of outliers more times to get optimal accuracy of any classification model.

2.2 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) 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 we

select low threshold value 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:

$$\text{AVF Score (xi)} = Fi = \frac{1}{m}\sum_{j=1}^{m} f(xi\,j) \qquad (3)$$

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 complexity is lesser than Greedy algorithm since AVF needs only one scan to detect outliers. The complexity is O (n * m). It needs 'k' value as input. In FPOF [8] this has discussed frequent pattern based outlier detection, in this too k-value and another parameter 'σ 'are required as threshold. This also discussed about frequent pattern based method to find infrequent object, in this too it requires k-value, and another parameter 'σ' as input.

2.3 N AVF algorithm

This proposed model (NAVF) has been defined as an optimal number of outliers in a single instance to get optimal precision in any 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. kN is the number of outliers which are normally distributed. To get 'kN' this model used Gaussian theory. If any object frequency is less than "mean-3 S.D" then this model treats those objects as outliers. This method uses AVF score formula to find AVF score but no k-value is required. Let D be the Categorical dataset, contains 'n' data points, xi, where i= 1…n. If each datapoint has 'm' attributes, we can write xi = [xi1, …..xil,…..xim], where xil is the value of the lth attribute of xi.

Algorithm

Input: Dataset – D,
Output: K detected outliers.

Step 1: Read data set D
Step 2: Label all the Data points as non-outliers
Step 3: calculate normalized frequency
of each attribute value for each point xi
Step 4: calculate the frequency score of
each record xi as, Attribute Value Frequency of xi is

$$\text{AVF Score (xi)} = Fi = \frac{1}{m}\sum_{j=1}^{m} f(xi\,j)$$

Step 5: compute N-seed values a and b as
b=mean (xi), a=b-3*std (xi), if max (Fi) > 3*std (Fi)
Step6: If Fi< a, then declare xi as outlier
Step 7: return KN detected outliers.

TABLE 2. NURSERY

3. Experimental results

In this paper this model has been applied on Breast Cancer, Nursery data and Bank marketing data from UCI Machine repository. This method has implemented the approach of using MATLAB tool. We ran our experiments on a workstation with a Pentium(R) D, 2.80 GHz Processor and 1 .24 GB of RAM.
Nursery data consists of nine attributes and 6236 records. This data divided into two parts based on parent attribute, first part contains 4320 records with usual parent type, and second part contain 1916 records with pretentious parent type which is used as outliers in our experiment. In first iteration 956 sample records are selected randomly using Clementine tool; from each two records one is selected.

Sample method	Actual outliers	Total records	NAVF	
			True positives	False negatives
1-in-2	956	5276	44	1
1-in-5	382	4702	132	1
1-in-8	238	4558	238	0
1-in-10	190	4570	190	0

These 956 records are mixed up with part one and applied normally distributed AVF to get outliers. The found outliers are given in table2. Similarly in the next iteration

382 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 table2. Similarly one record is selected from each eight records and ten records and repeated the same process. This method has been implemented on Nursery dataset, Breast cancer and Bank dataset which are taken from UCI Machine learning repository. This method compared with different number of outliers from each sample. Comparison graph is given in Figure 1.

In the first sample from nursery the NAVF model found out only 4.60% of outliers from 956 outliers which are mixed up with 4320 records which totals to 5276 records.. In the next sample of 382 records, 34.8% of correct outliers are found by NAVF. For the sample of 238 records NAVF found 239 outliers in which 238 are correct, which means that NAVF model found 100% outliers correctly. Similarly NAVF model found 100% outliers in the sample of 190 records (as outliers) mixed up with 4320 records in part one.

TABLE3. BREAST CANCER

Sample method	Actual outliers	Total records	NAVF	
			True positives	False negatives
1-in-2	119	577	35	0
1-in-5	48	506	9	0
1-in-8	29	487	9	0
1-in-10	23	481	14	1

In case of breast cancer dataset, correct outliers found by NAVF model did not touch 100%. In breast cancer data 119, 48, 29, 23 outliers are selected respectively using 1-in-2, 1 -in-5, 1-in-8, 1-in-10 sampling from benign breast cancer.. NAVF found 35, 9, 9, and 14 correct and 0, 0, 0, 1 wrong from 119, 48, 29, 23 outliers. The results are given in table 3

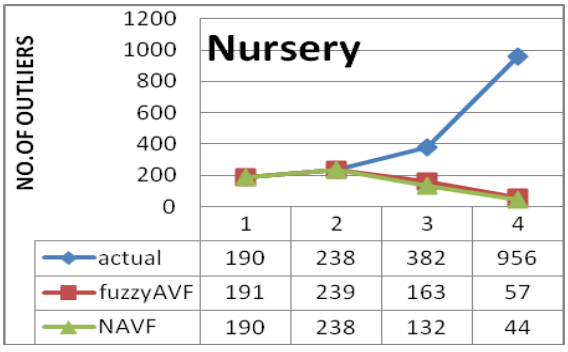

Fig4. NURSERY-DATA

In Bank marketing data, only categorical attributes are selected and 2644, 1027, 661, 528 outliers are selected respectively using 1-in-2, 1 -in-5, 1-in-8, 1-in-10 sampling

TABLE 4. BANK DATA

Sample method	Actual outliers	Total records	NAVF	
			True positives	False negatives
1-in-2	2644	39922	274	100
1-in-5	1027	39922	198	168
1-in-8	661	39922	152	202
1-in-10	528	39922	126	213

method from the attribute Y=”yes” and applied the same process as above. In this data NAVF found 274, 198, 152, 126 correct outliers and 100, 168, 202, 213 wrong outliers from the random sample of 2644, 1027, 661, 528 outliers taken by 1-in-2, 1 -in-5, 1-in-8, 1-in-10 sampling method.

Similarly we applied the same process for all samples in banking data. The results are summarized in the table 4 and its graph is given in Fig 3. Different classification models are tested for accuracy of the bank dataset after deleting the outliers. Different classification models are tested on 1-in-5 sample data which contain 39922 original records mixed up with 1027 outliers. The NAVF model has found 366 records as outliers in which 198 are correct outliers and 168 are wrong outlier (original records).When Neural network,C5,CRT,QUEST,CHAID Linear Regression ,Decision Logic Classifiers are applied on the above sample data ,the classifiers have given the accuracies as given in the table 5

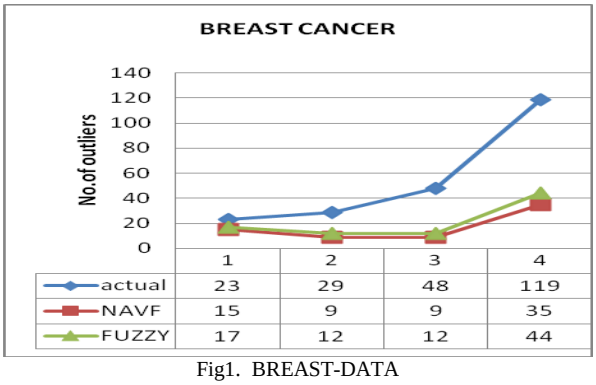

Fig1. BREAST-DATA

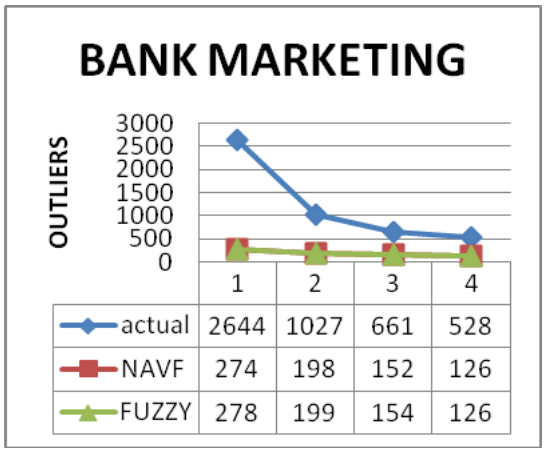

Fig. 1. BANK DATA

TABLE 5. CLASSIFIERS RESULTS ON BANK DATA

Classifier	NAVF	K=100	K=354	K=370	K=500	K=700
NN	98.735	97.581	97.878	97.885	97.994	98.135
C5	98.735	97.581	97.881	97.885	97.994	98.138
CRT	98.735	97.581	97.881	97.885	97.994	98.138
QUEST	98.735	97.581	97.881	97.885	97.994	98.138
DL	37.058	95.445	92.933	92.94	92.339	69.148
CHAID	98.735	97.581	97.881	97.885	97.994	98.138
LR	98.735	97.581	97.881	97.882	97.994	98.138

Different classifiers are applied on the remaining dataset after deleted the outliers by NAVF model. All classifiers have given good results for NAVF. Only the decision logic classifier gave very less accuracy (37.058) by NAVF.

4. Conclusion and Future work

To sum up, this proposed method gives the optimal number of outliers 'KN '.In existing models it is mandatory to give the number of outliers to find them .While taking the number of outliers sometimes the original data may be missed. If any classifier modeled using this data, wrong classifiers may be modeled. In future there is a possibility of checking the precision and recall values of each model with the existing models. The same method can also be applied on mixed type of dataset.

Reference

[1] M. E. Otey, A. Ghoting, and and A. Parthasarathy, "Fast Distributed Outlier Detection in Mixed-Attribute Data Sets," Data Mining and Knowledge Discovery
[2] He, Z., Deng, S., Xu, X., "A Fast Greedy algorithm for
[3] outlier mining", Proc. of PAKDD, 2006.
[4] 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.
[5] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining: Pearson Addison-Wesley, 2005
[6] E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: Algorithms andapplications," VLDB Journal, 2000.
[7] 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
[8] 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
[9] Z. He, X. Xu, J. Huang, and S. Deng, "FP-Outlier: Frequent Pattern Based Outlier Detection", Computer Science and Information System (ComSIS'05)," 2005
[10] http://archive.ics.uci.edu/ml/