ISSN 22773061

Outlier Analysis of Categorical Data Using Infrequency

M. Krishna Murthy, A. Govardhan, Lakshmi SreenivasaReddy D
Research scholar, ANU, Guntur, India
Krishna_mudimbi@yahoo.co.uk
Director of Evaluation, JNTUH, Hyderabad, India
govardhan_cse@jntuh.ac.in
Department of CSE, Rise Gandhi Groups of Institutions, Ongole, India.
urdlsreddy@yahoo.com.

ABSTRACT

Anomalies are those objects, which will act with different behavior and do not follow with the remaining records in the databases. Detecting anomalies is an important issue in many fields. Though many methods are available to detect anomalies in numerical datasets, only a few methods are available for categorical datasets. In this work, a new method has been proposed. This algorithm finds anomalies based on infrequent itemsets in each record. These outliers are generated by Apriori property on each record values in datasets. Previous methods may not distinguish different records with the same frequency. These give same score for each record. For each record a score is generated based on infrequent itemsets which is called MAD score in this paper. This algorithm utilizes the frequency of each value in the dataset. FPOF method is used the concept of frequent itemset and otey method used infrequent itemset. But these cannot distinguish records perfectly. The proposed algorithm has been applied on Nursery dataset and Bank dataset taken from "UCI Machine Learning Repository". Numerical attributes are excluded from Datasets for this analysis. The experimental results show that it is efficient for outlier detection in categorical dataset.

Indexing terms/Keywords

Data Mining, Outlier detection, FPOF score, FDOD Score, MAD score

Academic Discipline and Sub-Disciplines

Computer Science and Engineering

SUBJECT CLASSIFICATION

Data Mining

TYPE (METHOD/APPROACH)

Outlier Analysis.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 8, No 3
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

1. 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. The mechanism in the previous methods which are depends on frequent itemsets is that, these calculates frequency of each value in each records and checked with the threshold value, whether that value frequent or not. Then they formed all combinations of itemsets and find their frequency by scanning the dataset. All these combinations are the subsets of records. Based on infrequent itemsets their respective scores are generated. Top k-outliers are selected based on the least k-scores. The parameter s used in this method 'k', the number of outliers and a threshold value 'σ' to decide frequent item sets in each data object' [1].

2. TERMINOLOGY

Table.1. Terminology used in this paper.

Term	Description
K	Target number of outliers
N	Number of objects in Dataset
M	Number of Attributes in Dataset
x_i	i^{th} object in Dataset ranging from 1 to n
A_j	j^{th} Attribute ranging from 1 to m
$D(A_j)$	Domain of distinct values of j^{th} attribute
x_{ij}	cell value in i^{th} object which takes from domain d_j of j^{th} attribute A_j
D	Dataset
V	Set of all distinct values in Dataset D
I	Item set
F	Frequent Item set
IF	Infrequent item set.
$f(x_{ij})$	Frequency of x_{ij} value
$FS(x_i)$	Set of frequent Item sets of x_i object
$IFS(x_i)$	Set of infrequent Item sets of x_i object
Minsup	Minimum support of frequent item set
Support(I)	Support of Item set I

Some of the Existing Approaches for Categorical Datasets based on Item Frequency Frequent Pattern Outlier Factor (FPOF) algorithm:

This algorithm utilizes the Apriori algorithm as a first step to find all frequent Item sets. This method needs a human defined threshold value called" minimum support" 'σ' 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. The formula utilized in this algorithm is

$$FPOFScore = \sum_{F \subset xi \wedge F \in IF(x_i)} \frac{\text{support}(F)}{|FS(x_i)|} \qquad \textbf{(1)}$$

Where Dataset D= {A1, A2-------- Am},
Minimum support = 'σ',
Number of outliers = 'k',

Where F is the frequent item set which satisfies the minimum support,
FS (x_i) is the set of all frequent itemsets which are subsets of the record "x_i",

This model finds FPOF score for each record and selects k-outliers as least k-scores. If there is no frequent itemset at all in any record, identifying the score is a problem in this method.

a) Fast Distributed Outlier Detection(FDOD) (Otey) Algorithm

This algorithm also used the concept of frequent pattern method. The inverse Apriori is used in this model which say that "every super set of an infrequent itemset is again an infrequent set. So that, model reduces number of scannings of the dataset. This model first finds all combinations of subsets of each record and checks its support with threshold 'σ'.
It considers infrequent item sets and finds the FDOD score for each record by the below formula,

$$FDODScore = \sum_{IF \subset x_i \wedge IF \in IFS(x_i)} \frac{1}{|IFS(x_i)|} \quad \textbf{(2)}$$

Where IF is the infrequent item set which does not satisfy the minimum support,
IFS (x_i) is the set of all infrequent itemsets which are subsets of the record "x_i",

This model finds FDOD score for each record and selects k-outliers as top k-scores. If there is no infrequent itemset at all in any record, then identifying the score is another problem in this method

b) Attribute Value Frequency (AVF) algorithm

This algorithm is simple for finding scores of each object. 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 "x_i" is defined based on the AVF Score below:

$$AVFScore = \frac{1}{m}\sum_{i=1}^{m}\frac{f(x_{ij})}{n} \quad \textbf{(3)}$$

Where f (x_{ij}) is the frequency of each value involved in each record,
'm' is the number of attributes,
'n' is the dataset size,
"x_{ij}" is the cell value in i^{th} record and j^{th} Attribute.
This model finds AVF score for each record and selects k-outliers as least k-scores. When these above algorithms applied on a sample data below with threshold value =5, one problem is identified.

Table.2. Comparison of existing model scores on Sample Health data.

S.No	Age	Income level	Disease	FPOF	FDOD	AVF
1	Middle	Low	Cancer	0.35	1.33	0.53
2	Middle	High	Cancer	0.35	1.33	0.53
3	Young	Low	Cancer	0.27	1.83	0.53
4	Middle	High	Cancer	0.35	1.33	0.53
5	Young	Low	Sugar	0.17	2.83	0.47
6	Young	Low	Sugar	0.17	2.83	0.47
7	Middle	High	Cancer	0.35	1.33	0.53
8	Young	Low	Sugar	0.17	2.83	0.47
9	Middle	High	Cancer	0.35	1.33	0.53
10	Young	High	Sugar	0.17	2.83	0.47

From all the above scores it is observed that some of the scores are same for different combination of values. If it is needed to select 10% of outliers which record will be selected? Since 5th, 6th, 8th and 10th records scores are same, there exists some ambiguity. If it is compulsory, the 5th may be selected because 5th is the first one. By observing these four records 10th one is very different. But the scores are same. The proposed model can distinguish this 10th record which is different from the 5th, 6th and 8th records.

c) Proposed Model (MAD Score)

This proposed algorithm also used the infrequent itemsets which are generated by Appriori concept. This proposed model finds the score for each record. We call this score as MAD score.

$$MADscore = \frac{n}{1+\sum Freq(\text{Infrequentitemset})} = \frac{n}{1+\sum_{\forall IF \in IFS(x_i)} \sup(IF)} \quad \textbf{(4)}$$

Where 'n' is the dataset length,
IF is the infrequent itemset,
IFS is the Set of Infrequent itemsets,
$x_{i\ is}$ the i^{th} record in dataset.
Sup (IF) is Frequency of IF

From the below table it is revealed that 10th record is the most different outlier, and then the records 5th 6th 8th are in the next order. 5th 6th 8th and 10th records gave the same scores in FPOF, FDOD and AVF. But the proposed model distinguishes the 10th record separately. Not only the only the 10th record, for all different records are giving different scores by the proposed model. So the proposed model is reliable to find reliable outliers. The proposed model is applied on Bank Data set, Nursery Data set and Breast Cancer Dataset which are taken from UCI ML repository [10].The comparison of results with FPOF, FDOD and AVF on Trail data is given below.

Table.3. Comparison of MAD score with FPOF, FDOD and AVF.

S.No	Age	Income level	Disease	FPOF	FDOD	AVF	MAD
1	Middle	Low	Cancer	0.35	1.33	0.53	1.6666
2	Middle	High	Cancer	0.35	1.33	0.53	1
3	Young	High	Cancer	0.27	1.83	0.53	1.25
4	Middle	Low	Cancer	0.35	1.33	0.53	1.6666
5	Young	Low	Sugar	0.17	2.83	0.47	0.714
6	Young	Low	Sugar	0.17	2.83	0.47	0.714
7	Middle	High	Cancer	0.35	1.33	0.53	1
8	Young	Low	Sugar	0.17	2.83	0.47	0.714
9	Middle	High	Cancer	0.35	1.33	0.53	1
10	Young	High	Sugar	0.17	2.83	0.47	0.8333

3. Experimental Results

When the experiments are conducted on bank data with 45212 records by the proposed model, it has achieved the maximum classifier accuracy better than AVF. The experimental results are compared with AVF results because in previous research work AVF has given good results when compared with FPOF and FDOD. The bank data contains 7 variables and 46 values. The Bank Sample has partitioned into two parts, one is with "Yes" Class label (5299) and another is with "no" class label (39922 records) using Clementine tool. The "yes" label records are considered as outliers in this experiment. In this experiment 50% of outliers (2645 records) are selected randomly and mixed up with "no" class label. The mixed records (42567 records) considered for experiments. Both AVF and MAD models applied on the same built mixed records to delete top 100,200,300,400,500,600,700 and 800 outliers. After deleted these outliers different classifiers are tested. The tested results are given bellow.

ISSN 22773061

Table.4. Comparison of classifier accuracies for Bank data (1-in-2 sample).

Classifier	NN		LR		CHAID	
Number of outliers deleted	AVF Score	MAD Score	AVF Score	MAD Score	AVF Score	MAD Score
100	98.726%	98.796%	98.726%	98.796%	98.726%	98.796%
200	98.763%	98.865%	98.763%	98.865%	98.763%	98.865%
300	98.842%	98.924%	98.842%	98.924%	98.842%	98.924%
400	98.904%	98.991%	98.904%	98.991%	98.904%	98.991%
500	98.926%	99.041%	98.926%	99.041%	98.926%	99.041%
600	98.966%	99.117%	98.966%	99.117%	98.966%	99.117%
700	99.034%	99.198%	99.034%	99.198%	99.034%	99.198%
800	99.082%	99.208%	99.082%	99.208%	99.082%	99.208%

From the above results it is concluded that Classifiers built on the records when outliers are deleted by MAD Score algorithm gives better results when compared with AVF Score. Classifiers also applied on the original data without deleting outliers. These classifiers gave 88.302% only.

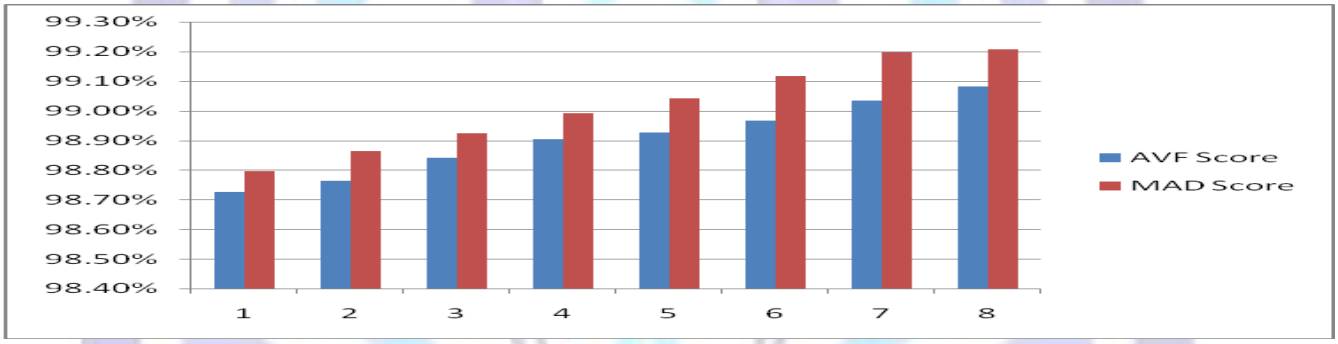

Fig.1.NN, LR and CHAID Classifiers accuracy when outliers deleted by AVF and MAD for below results

4. Conclusion and Future work

To sum up, this proposed method finds distinguished score for distinguishable records, where as the previous methods may not find different scores for different records. This model also gives reliable records as the classifiers get maximum accuracies when compared with old models. To form the combinations of item sets and scanning of the dataset for every itemset for frequency is a big problem in these models. Some possibility of solving this problem is parallel computing. When the attributes are increasing the complexity becomes more in these models.

5. References

[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 Outlier mining” Proc. of PAKDD, 2006.

[3] 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.

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

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

[6] 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

[7] 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

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

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

[10] 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.

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

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

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

[14] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases." Proceedings International Conference on Very Large Data Bases, 1994, pp. 487-499.

[15] LakshmiSreenivasaReddy.D, .B.RaveendraBabu and A.Govardhan, "A model for Improving Classifier Accuracy for Categorical data using Outlier Analysis", International Journal of Computers and Technology" vol 7, 2013.

THE AUTHORS

Mr. Mudimbi.Krishna Murthy did His M.C.A in first Class in 2003 from MKU Madurai. He Has 15 years of technical experience in Computer Science and Engineering at School of Information Technology (SIT), Jawaharlal Nehru Technological University Hyderabad, India. He has five research papers at international and national conferences. His area of research is Data Mining and Information Retrieval Systems.

Dr.A.Govardhan is presently a Professor of Computer Science & Engineering and Director of Evaluation, Jawaharlal Nehru Technological University Hyderabad (JNTUH), India. He did his B.E(CSE) from Osmania University College of Engineering, Hyderabad in 1992, M.Tech from Jawaharlal Nehru University(JNU), New Delhi in 1994 and Ph.D from Jawaharlal Nehru Technological University, Hyderabad in 2003. He is a recipient of several International and National Awards including A.P. State Best Teacher Award, Bharat Seva Ratna Puraskar, CSI Chapter Patron Award, Bharat Jyoti Award and Mother Teresa Award for Outstanding Services, Achievements, Contributions for Meritorious Services, Outstanding Performance and Remarkable Role in the field of Education and Service to the Nation. He is a Chairman and Member on several Boards of Studies of various Universities. He is the Chairman of CSI Hyderabad Chapter. He is a Member on the Editorial Boards for Eight International Journals. He is Member of several Advisory Boards and Committee Member for several International and National Conferences. He has guided 14 Ph.D theses and he has published 152 papers at International/National Journals/Conferences including *IEEE, ACM, Springer and Elsevier*. He has delivered more than 35 Keynote addresses and invited lectures. He served as Principal, Head of the Department and Students' Advisor. He is a member in several Professional and Service oriented bodies. His areas of research include Databases, Data Warehousing & Mining and Information Retrieval Systems.

Mr.Lakshmi Sreenivasareddy.D obtained his Masters degree from Jawaharlal Nehru Technological University Hyderabad (JNTU). He is pursuing 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. His area of interest is Data Warehousing & Mining.