# A Model to Find Outliers in Mixed-Attribute Datasets using Mixed Attribute Outlier Factor

**M. Krishna Murthy, A. Govardhan, Lakshmi SreenivasaReddy D**

**[1] Research scholar, ANU, Guntur, India**

**[2] Professor of CSE, JNTUH, Hyderabad, India**

**[3] Department of CSE, Rise Gandhi Groups of Institutions, Ongole, India.**

## Abstract

Outliers are records in real datasets which have abnormal behavior comparing with other records in datasets. Finding outliers in numerical dataset is easy. Many methods are available for numerical datasets. Number of methods is also available for categorical datasets. But very less number of methods is available for mixed attribute Datasets. In all the available methods, the concept of frequent pattern mining is used. Finding different frequent patterns from datasets for the categorical attributes is a cumbersome process. In proposed model, Mixed Attribute Outlier Factor (MAOF) is presented. Which is a simple technique and it requires only one scan of dataset. MAOF is derived based on Attribute Value frequency (AVF) for Categorical part of dataset and cosine factor for continuous part which is derived from the mean record to the remaining numerical data points in the dataset. Average of these two factors will give the MAOF score. This model has been applied on Bank dataset which is a real dataset taken from UCI ML repository [10]. This method shows the good results.

Keywords: ***Data mining, Outlier detection, Oteyscores, ODMADscore, MAOF score.***

.

## 1. Introduction

Most of the real datasets contain mixed type of attributes. Finding outliers of this type of datasets is very useful to model the data. Outlier analysis is an important task in many fields like medicine, bank, and networks. Existing systems concentrated on finding frequent patterns. Outlier factors are found by the frequent patterns gives us a very high complexity. There are so many methods derived from frequent mining concept. Some methodologies are derived based on Appriori property to reduce the complexity [14]. By this approach forming the number of subsets for each record and scanning the dataset for all these subsets for frequency is a problem. Even utilization of Appriori concept does not in prove efficiency. The proposed model solves all these problems. This proposed method utilized the concept of attribute value frequency for categorical part of dataset and cosine vector product concept for numerical part of records in datasets.

## 2. Terminology

Different terminologies are required for existing and proposed model about frequency, support and input number of outliers etc is given in Table.1 below.

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 |
| 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 |
| P | Set of all combinations of distinct attribute values, where each attribute occurs only once in any combination |
| I | Item set |

| | |
|---|---|
| F | Frequent Item set |
| IF | Infrequent 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 |

## 3. Existing approaches for mixed attribute datasets based on Item frequency

### 3.1 Otey Score Algorithm

In this approach anomaly score is computed by partitioning the entire mixed dataset in to two parts. First part contains categorical subspace and the second part contains numerical subspace. Outlier factor of categorical part is denoted as Score1 (xi) and outlier score of numerical part is denoted by Score2 (xi). This approach is described based on links between attributes.

This approach is derived as below:

Let V= set of all distinct categorical values included in Dataset

C= Set of all combinations (itemsets I) of distinct attribute values

i.e $P = FS \cup IFS$ (1)

Where FS = Set of Frequent Itemsets such that sup (Itemset) $\geq$ user defined threshold value.

IFS= Set of infrequent Itemsets

Sup (I) = support of itemset I

Now the outlier score of the categorical part is defined as below:

$$Score1(C(xi)) = \sum_{I \in FS(x_i)} \frac{1}{|I|} \tag{2}$$

Let CI is the covariance of the itemset I,

CIij is the covariance of I in 'i' and 'j' attributes from numerical part

CxiIij is the covariance of I in 'i' and 'j' attributes from numerical part for the object xi

$$C_{ij}^{xiI} = (x_i - \mu_i^I)\, x\, (x_i - \mu_i^I) \tag{3}$$

The violation score of an object xi is defined as below:

$$V_\Gamma(x_i) = \sum_i \sum_j v_\Gamma(x_{ij}) \tag{4}$$

$$v_\Gamma(x_{ij}) = \begin{cases} 1 & if \left|C_{ij}^{xiI} - C_{ij}^{I}\right| \leq \Gamma \\ 0 & otherwise \end{cases} \tag{5}$$

$$\sigma_{C_{ij}^I}^2 = \frac{1}{\sup(I)-1} \sum_{\{xi/I \in xi\}}^{n} (C_{ij}^{xiI} - C_{ij}^{I})^2 \tag{6}$$

Where $\dfrac{C_{ij}^{xiI} - C_{ij}^{I}}{\sigma_{C_{ij}^I}}$ follows the normal distribution

Now the outlier score of xi in a dataset

$$Score2(x_i) = \sum_{I \subseteq x_i} \left(\frac{1}{|I|} / (C_1 \vee C_2) isTrue\right. \tag{7}$$

Where C1: sup (I) $\leq$ threshold value C2: sup (I)> $\delta$, where $\delta$ is maximum violate score.

Based on these two scores we can find outlier factor of an object in mixed attribute dataset. Another approach of finding outlier factor for every object in mixed type of Dataset is defined by Koufakou et al in [3].

### 3.2 ODMAD Score

This algorithm also depends on two parts of mixed data. The first part is categorical subspace; second part is numerical subspace of the dataset.

$$Score1(xi) = \sum_{|IF(xi)|=1}^{MAXLEN} \frac{1}{\sup(IF(xi)) * |IF(xi)|} \tag{8}$$

MAXLEN = user entered maximum length of in frequent itemset,

Sup (IF (xi) = support of infrequent itemset in object xi,

$|IF(xi)|$ =length of infrequent itemset in object xi,

$$Score2((xi) = \frac{1}{\left|a \in xi^C\right|} \sum_{\forall a \in xi^C} COS(xi^N, \mu_a) \tag{9}$$

Where

$$\mathrm{COS}(xi^N, \mu_a) = \sum_{j=1}^{m_N} \frac{1}{\left\|x_{ij}^{\ N}\right\|} * \frac{\mu_{aj}}{\mu_a} \tag{10}$$

Here 'a' is a categorical value included in the object xi.

Based on the above scores the outlier factor is found in ODMAD.In both approaches finding frequent itemsets is a big problem. So we approached the below way.

### 3.3 Proposed Method

In this approach outlier factor is found with forming any frequent patterns in an object. Instead of this the attribute value frequency has been proposed. From the above two approaches number of scans of a dataset is required. Proposed method needs only on scan of the dataset. This proposed method finds again two scores, one is for

categorical part of dataset and other is for numerical part of the dataset. Score1 is defined like below:
Let there are 'm' categorical attributes and 'n' numerical attributes in a dataset.

$$Score_1(C(x_i)) = \sum_{j=1}^{m} \frac{\sup(x_{ij})}{|D|} \quad (11)$$

$$Score_2(N(x_i)) = COS(N(x_i)) = \frac{\langle \mu_{xiN}, x_N \rangle}{\|\mu_{xiN}\| * \|x_N\|} \quad (12)$$

Here $\mu_{xiN}$ is a vector of means of all attributes in Numerical part of Dataset.

$x_{iN}$ is the vector of all attribute values in the numerical part of the ith object.
MAOF factor can be defined as

$$MAOFscore(x_i) = \frac{Score_1(C(x_i)) + Score_2(N(x_i))}{2} \quad (13)$$

## 4. Experimental Results

Experiments are conducted on 1342, 1298, 1279, 1200 records respectively from 1-in-2, 1-in-5, 1-in-8, 1-in-10 samples are taken from Bank dataset which is taken from UCI machine Repository [10]. Different sample are selected from Bank date with two class values. These samples are selected like that one sample from each two records, one sample from each 5 records, one sample from each 8 records, and one sample from each 10 record from "yes" class records. These are mixed-up with normal records. Then 45 records are created randomly with much variation and mixed up with the above said samples. All these operations are conducted by Clementine 11.1 tool. Our model found these created records from each sample as given in the below Table.1.1

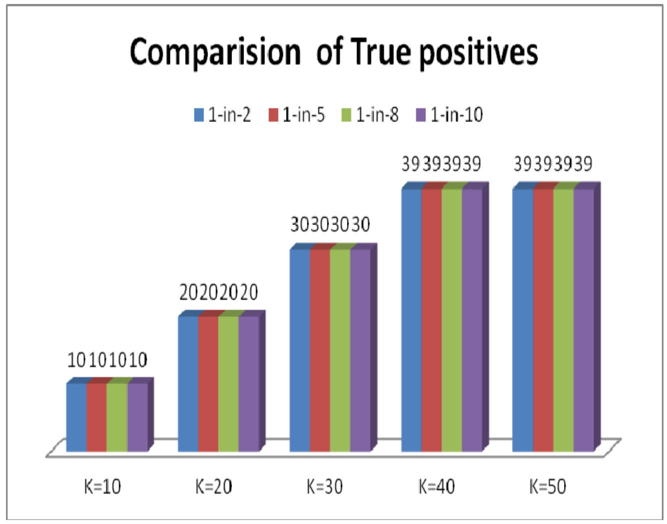


Fig.1 Comparison of true positives found in different samples for different inputs

Table.1 Comparison of the number of outliers found for different input 'K'

| Sample | Number of true and false outliers are found for different K values given as input | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | K=10 | | K=20 | | K=30 | | K=40 | | K=50 | |
| | True | False | True | False | True | False | True | False | True | False |
| 1-in-2 | 10 | 0 | 20 | 0 | 30 | 0 | 39 | 1 | 39 | 11 |
| 1-in-5 | 10 | 0 | 20 | 0 | 30 | 0 | 39 | 1 | 39 | 11 |
| 1-in-8 | 10 | 0 | 20 | 0 | 30 | 0 | 39 | 1 | 39 | 11 |
| 1-in-10 | 10 | 0 | 20 | 0 | 30 | 0 | 39 | 1 | 39 | 11 |

From the Table 1 it is shows that for inputs k=10, k=20, k=30 this model found exact number of true positives and for k=40 and for k=50 it has found 39 true positive out of 45 outliers which are included in Bank Dataset. The first score computes how much an object deviates from the others in categorical part and the score computes cosine product which calculates the similarity between mean for numerical part and an object in the Numerical part of each object. These results are not compared with others because the approach is entirely different with the existing ones.

## 5. Conclusion and Future Work

This model has been developed on Attribute Value Frequencies for categorical data and cosine dot product for numerical dataset. These two scores give the factors in the range of 0 to 1 and its average again gives the value in the range of 0 to 1. In future we investigate the numerical score by correlation factor between mean of the attribute values in numerical part and the numerical part in each object.

## References

[1] Anna Koufakou, Michael Georgiopoulos, “A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes” Data mining and Knowledge Discovery , Volume 20, 2010, pp.259-289.

[2] M.E. Otey, A. Ghoting, and S. Parthasarathy. “Fast Distributed Outlier Detection in Mixed-Attribute Data Sets." Data Mining and Knowledge Discovery, 12, 2006, 203-228.

[3] A. Koufakou, E.G. Ortiz, M. Georgiopoulos, G.C. Anagnostopoulos, and K.M.Reynolds. “A Scalable and Efficient Outlier Detection Strategy for Categorical Data.” IEEE International Conference on Tools with Artificial Intelligence ICTAI, 2007, pp. 210-217.

[4] Z. He, X. Xu, S. Deng, D. Calvanese, G. De Giacomo, and M. Lenzerini. “A Fast Greedy Algorithm for Outlier Mining.” Proceedings of 10th Pacific-Asia Conference on Knowledge and Data Discovery, 2006, pp. 567-576.

[5] Z. He, X. Xu, J.Z. Huang, and SC Deng.“FP-Outlier: Frequent Pattern Based Outlier Detection.” Computer Science and Information System, valume2, 2005, pp.103-118.

[6] Z. He, X. Xu, J.Z. Huang, and SC Deng. “FP-Outlier: Frequent Pattern Based Outlier Detection.” Computer Science and Information System, Volume2, No.1, 2005, pp103-118.

[7] J. Han, J. Pei, Y. Yin, and R. Mao. “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach.” Data Mining and Knowledge Discovery, Volume8, No1, 2004, pp53-87.

[8] S.D. Bay and M. Schwabacher. “Mining distance-based outliers in near linear time with randomization and a simple pruning rule.” Proceedings of the ACMSIGKDD International conference on Knowledge Discovery and Data Mining, 2003, pp. 29-38.

[9] C. Borgelt. “Efficient Implementations of Apriori and Eclat.” Workshop of Frequent Item Set Mining Implementations FIMI, 2003.

[10] L. Wei, W. Qian, A. Zhou, W. Jin, and J.X. Yu. “HOT: Hyper graph-Based Outlier Test for Categorical Data.” Proceedings Pacific-Asia Conference PAKDD, 2003, pp. 399.

[11] T. Calders and B. Goethals. “Mining All Non-Derivable Frequent Itemsets.” Proceedings PKDD International Conference Principles of Data Mining and Knowledge Discovery, 2002, pp.74-85.

[12] F. Pan, G. Cong, A.K.H. Tung, J. Yang, and M.J. Zaki. “CARPENTER: Finding closed patterns in long biological datasets.” Proceedings ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2003, pp. 637-642.

[13] He, Z., Deng, S., Xu, X., “A Fast Greedy algorithm for Outlier mining” Proc. of PAKDD, 2006.

[14] 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.

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

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

[17] 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

[18] 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

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

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

[21] 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.

[22] LakshmiSreenivasaReddy.D, .B.RaveendraBabu and A.Govardhan, “Outlier Analysis of Categorical Data using NAVF”’, Informatica Economica vol 17, Cloud computing issue 1, 2013.

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

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

[25] 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.

[26] 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" Volume 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, 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 15 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.LakshmiSreenivasareddy.D obtained his Masters degree from Jawaharlal Nehru Technological University Hyderabad (JNTU). He is pursuing his Ph.D in Computer Science and Engineering from 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. He has six research papers at international conferences and journals including IEEE and Elsevier. His area of interest is Data Warehousing & Mining and Information Retrieval Systems.