A Proposed Outliers Identification Algorithm for Categorical Data Sets

Ayman Taha[#1], Osman M. Hegazy[#2]
[#] *Faculty of Computers and Information*
Cairo University - Egypt.
[1] iyman.taha@gmail.com
[2]osman.hegazy@gmail.com

***Abstract*— Outliers are a minority of observations that are inconsistent with the pattern suggested by the majority of observations. Outliers identification algorithms for categorical data sets face many limitation because measuring distance is not common in categorical data. In this paper, we propose a new unsupervised outliers identification method in categorical data sets. In contrast to other outliers identification methods, the proposed method considers number of categories inside categorical variables. Experimental results show that the proposed method has a comparable performance results with respect to other outliers identification methods in performance.**

***Keywords*— Outliers Detection, Categorical Data, Data Mining .**

I. INTRODUCTION

Outlier are observations that are highly inconsistent with other observations and arouse suspicion that they were generated by a different mechanism [7]. There are two main viewpoints to outliers identification process; as a major pre-processing process for data mining and as a data mining technique. The first viewpoint considers that outliers, if they exist, can affect measurements on other points. Consequently, this approach defines outliers identification process as an important process before data mining applications. However, the second viewpoint classifies outliers identification as one of data mining processes where outliers are the most informative observations in the data sets.

Outliers Identification has several important applications such as identifying errors and unexpected entries in databases, identifying new topic in text mining applications [11], identifying abnormal locations in spatial domain [18], identifying abnormal or catastrophic events in time series data, identifying fraudulent credit cards [15] and identifying unauthorized access or intrusion in computer networks [12].

Outliers detection approaches can be classified into three main approaches: supervised, unsupervised and semi supervised approach. In the supervised outliers detection approach labeled samples are used in training phase to learn the behavior of normal and abnormal points and then other points are tested. Observations having behavior similar to abnormal points are labeled as outliers and other observations are labeled as normal observations. While unsupervised outliers detection approach processes the data as a static distribution, finds the most remote points and highlights them as outliers. Unsupervised approach does not require prior knowledge but it requires all data to be available before processing. However semi-supervised outliers detection approach learns only the behavior of normal observations to define a boundary of inliers and then it labels observations outside this boundary as outliers [11]. In this paper, we focus on unsupervised learning outliers detection algorithms because they are more practical than supervised learning; especially in real situations where labeled examples may be unknown or a new fraud patterns that did not appear in training phase may appear in testing phase.

Variables can be classified into two types: continuous variables and categorical variables. Continuous variables have an infinite domain of values such as length, height and depth, while categorical variables have a finite domain of values such as colors, nationalities and types. This work concerns the problem of outlier identification in categorical data sets, where all variables are categorical variables.

The distance among categorical values is not regular term, which leads to disappearance of categorical data sets in data mining algorithms. Several distance functions have been proposed in the literature to compute the the distance between categorical observations. These distance functions make use of the following categorical data sets characteristics [4] and [5]:

- n: Number of observations.
- q: Number of categorical variables.
- c_i: Number of categories in the i^{th} categorical variable.
- $f_k(x)$: The frequency of category x in the k^{th}

categorical variable.

- $\hat{p}_k(x)$: The sample probability of the category x in the k^{th} categorical variable in D. It is calculated as follows:

$$\hat{p}_k(x) = \frac{f_k(x)}{n}.$$

The rest of this paper is organized as follows; section II presents related work in outlier identification in categorical domain, while the proposed method is presented in section III . Section IV shows experimental results. Finally, section VI provides a conclusion and future works.

II. Outliers Identification in Categorical Data sets

Since most of existing outliers identification techniques are based on measuring distance, outliers identification techniques are more prevalent in continuous domain such as [1], [3] and [12] rather categorical data sets [16], [6] and [13]. Outliers identification algorithms in categorical domain make use of frequency, number of occurrence of categories, instead of distance. Although there are few outliers identification algorithms in categorical domain, there is no agreement on the definition of outliers. We will classify outliers identification methods according to their definitions of outliers in categorical data sets to the following types:

A. Converting Categorical Variables to Continuous Variables

Outliers identification algorithms in this type do not have a certain definition for outliers in categorical data sets. They transform categorical variables to continuous variables then they use classical outliers identification method in continuous domain. An example of this type presented in [19] which presents a categorical outliers identification method that transforms a categorical variable to $c - 1$ binary variables, where c is number of categories in that categorical variable. The resultant binary data is treated as continuous data and outliers identification method for continuous data is applied. Converting categorical values to continuous values algorithms have two problems; the first problem is time complexity because they take long preprocessing time for converting categorical values to binary values. Moreover, they greatly increases number of dimensions of the tested data sets.

B. Observations with Small Marginal Frequency

Outliers identification algorithms in this type define outliers are observations with infrequent categories. They assume independence of categorical variables. Examples of of such methods are ORCA, name of software presented in [2] and Attribute Value Frequency (AVF) [14].

ORCA presents an distance- based outliers identification technique in categorical and mixed data sets called ORCA [2] . It is based on using hamming distance to compute distance between categorical observations. ORCA takes two parameters; *M*, number of outliers and *k*, number of K nearest neighbors in *k*NN.

AVF is a fast and scalable outliers identification algorithm for categorical data sets. It can efficiently deal with large and dynamic categorical data sets [14]. AVF defines an outlier in a categorical data set as those observations which are infrequent in the data set. AVF defines the ideal outliers, the most outlying observations, in categorical data sets as observations whose all categories are infrequent. It presents an outlying score for each observation called AVF, which is based on the marginal frequencies of categories. The lower AVF an observation has, the higher probability of being outlier. AVF score is calculated as follows:

$$\text{AVf}(x_i) = \frac{1}{q}\sum_{k=1}^{q} f(x_{ik}),$$

where x_i is i^{th} observation, q is number of categorical variables, $f(x_{ik})$ is the relative frequency of the category of i^{th} observation of k^{th} variable .

C. Observations with Small Item Set Frequency

These algorithms consider not only the frequency of marginal categories, but also the frequency of item sets, combinations of categories. They start with building a set of frequent item sets, then declare observations with infrequent item sets as outliers. Finding frequent item set process usually takes long time especially when number of categorical data is large, so these algorithms have a combinatorial time complexity. These algorithms require two additional parameters for finding frequent item set; minimum frequency threshold and maximum length of item set. Their results are sensitive to the choices of parameters values. Examples of such methods are LOADED (Link-based Outlier and Anomaly Detection in Evolving data sets) [17] and FPOF (Frequent Pattern Outlier Factor) [10].

LOADED, an outliers identification algorithm in categorical and mixed data sets, represents every observation as node in space then it links among observations that have common categories or item sets. Moreover, each link has a weight, number of common categories, that captures the degree of similarity between observations. LOADED defines outlier in categorical data set as " a data point that has either (a) very few links to other points or (b) links with very small weight to other points".

LOADED computes an anomaly score , $\varphi(x_i)$ for each observation x_i, as follows:

$$\varphi(x_i) = \sum_{d \subset x_i} \frac{1}{|d|} * \mid f(d) \leq minfreq \mid,$$

where d is the item set and $minfreq$ is the minimum frequency threshold. Observations that have high anomaly score will have high probability to be outliers.

FPOF, Frequent Pattern Outliers Factor, is an outliers identification algorithm in categorical domain. It defines outliers as " observations that contain less frequent patterns in their item sets". FPOF uses the *minfreq* parameter to define a list of frequent item sets, *S*. For each observation x_i, a measure of outlying called FPOF(x_i) is defined as follows:

$$\text{FPOF}(x_i) = \sum_{d \subseteq x_i} f(d) * \mid d \in S \mid$$

The lower FPOF score an observation has, the higher probability for being outlier.

D. Observations with Small Joint Frequency

Algorithms in this type are considering the frequency of an observation as a whole neither the frequency of categories nor item set. They calculate the full joint frequency of each record, then they identify infrequent records as outliers. An example of such methods is Greedy algorithm [8] and [9].

The idea behind the greedy algorithm is based on the direct relationship between numbers of outliers and amount of noise in data sets [9]. As the number of outliers increases in a data set as the disorder level increases. The greedy algorithm uses Entropy as a measure of disorder in a data set. Consequently, it transforms the outliers identification problem to an optimization problem: finding *M* observations, number of expected outliers, where the *e*ntropy, $E(n \setminus M)$ observations is minimum.

$$E(X) = -\sum_{x \in X} p(x) * log(p(x)),$$

where X is set of $(n \setminus M)$ observations, x is a distinct observation in X and $p(x)$ is the number of occurrence of observation x within X.

III. PROPOSED METHOD

Most of outliers identification methods in categorical domains have large time complexity. For example, Greedy algorithm needs M scans over the data set to find M outliers. On the other hand, methods based on item sets (FPOF and LOADED) need combinatorial time for building set of frequent item sets. In addition to large computation times they create a potentially large space for storing item sets and then searching for these frequent item sets for each observation. These techniques can become extremely slow for low values of the *minfreq* threshold and large number of categorical variables.

In this section, we propose a fast outliers identification method for categorical data sets named SCF (Squares of the Complement of the Frequency). The proposed method aims at finding outliers, observations with small marginal frequencies. For each observation, it calculates frequency score named SCF(x_i):

$$SCF(x_i) = \sum_{j=1}^{q} \frac{[1 - f(x_{ij})]^2}{c_j},$$

where c_l is number of categories in the l^{th} categorical variable.

SCF uses the sum of squares of the complement of the marginal frequency instead sum of the marginal frequency to emphasize the difference between frequent and infrequent categories. In contrast to other outliers identification methods in categorical data sets, it considers number of categories in the categorical variables.

Algorithm 1 The proposed Algorithm Pseudocode

Input:
D: a categorical data set and
M: number of expected outliers
Output:
O: a complete set of outliers
Processing:

1: For each categorical variable X_j
   1.1 Compute number of categories, c_j.
   1.2 For each category, v_{jk} in a categorical variable c_j
   1.2.1 Compute relative frequency $fr(v_{jk})$.
2: Let $O = \Phi$.
3: For each observation $x_i \epsilon D$, Calculate SCF scores
   3.1 For each category v_{ij}
   3.1.1 Get its relative frequency $f(v_{ij})$.
   3.1.2 Compute the square of its complement, $(1 - fr(v_{ij}))^2$.
   3.2 Compute the weighted sum of squares of the complement $\sum_{j=1}^{q}(1 - f(v_{ij}))^2$.
   3.3 Divide the weighted sum by number of categories $SCF(x_i) = \sum_{j=1}^{q} \frac{(1-f(v_{ij}))^2}{c_j}$
4: Sort observations based on SCF scores in a descending order.
5: Add top M observations to outlier set O.

TABLE I
RESULTS ON LYMPHOGRAPHY DATA SET

M	FPOF	LOADED	Greedy Algorithm	AVF	SCF
6	4	4	5	4	5
9	5	5	6	5	6

TABLE II
RESULTS ON BREAST CANCER DATA SET

M	FPOF	LOADED	Greedy Algorithm	AVF	SCF
4	3	3	4	4	4
24	21	21	22	21	22
32	27	28	27	28	28
39	30	32	33	31	32
56	39	39	39	39	39

IV. EXPERIMENTAL RESULTS

We compare among our proposed method against Greedy algorithm, FPOF and AVF using real data sets available at the UCI Machine Learning Repository and Intelligent systems and synthesis data sets. All experiments were conducted on a Pentium4- 1.68 G core 2 duo 2M cache machine with 1 G of RAM and running Windows XP.

Lymphography:
This data set contains 148 observations and 18 categorical variables and class variable. There are 4 different classes where classes 1 and 4 contain 6 observations. These rare observations are considered as the outliers.

Wisconsin Breast Cancer:
This data set has 483 observations and 9 categorical variables. Each record is labeled as either benign or malignant. Number of malignant observations is 39 and number of benign 444 observations. As other methods, we consider malignant observation as outliers.

Here, we are comparing our proposed method with algorithms that are based on marginal frequency such as AVF,joint frequency such as Greed Algorithm and item sets such as LOADED and FPOF. Since, number of expected outliers, M, is given as an input to all methods, we are interested in number of true detected outliers (detection rate).

Table I and II show number of outliers detected by each algorithm using lymphography and breast cancer data sets. As depicted, the proposed method is slightly better than LOADED, FPOF and AVF and very close to Greedy that has large computation time. For example, in Table II, the proposed algorithm and greedy algorithm detect 22 true outliers from 24 expected outliers. While FPOF, LOADED and AVF detect 21 true outliers.

Since outliers identification techniques are targeting huge data sets, time complexity is an important issue in comparing among these techniques. Item set based outliers identification techniques such as FPOF take long time in finding frequent item sets, because the size of item sets grows combinatorially especially with large number of categorical variables. Time complexity of FPOF is: O $(n * (q2 + |S|) + logn)$, where S is the set of frequent item set.

On the other hand, the Greedy algorithm scans data set M times. Consequently, It takes long time especially with large number of outliers and huge data sets. Time complexity for Greedy Algorithm is O$(n * q * M)$.

Time complexity of the proposed algorithm and AVF is O$(n * (q + logn))$. It linearly increases with number of observations and number of categorical variables.

V. CONCLUSION AND FUTURE WORK

In this paper, We surveyed a representative list of available techniques for outlier detection in categorical data sets. These techniques did not agree on a unique definition of outliers. We classified them according to their definition of outliers. We have proposed an effective outlier detection technique. The proposed algorithm, SCF, makes use of the square of the complement of marginal frequency and number of categories within categorical variables.

Experimental results using real data sets show that the proposed method (SCF) and AVF require a short time with respect to other algorithms that depends on item set frequency or joint frequency. The performance of SCF is slightly better than other methods except Greedy algorithms which takes large time complexity.

There are some limitations in the proposed method that we are aiming to handle them in the future such as ignoring dependency among categorical variables, handling data streams and mixed data sets. Moreover, identifying number of outliers in advance is an impractical issue in the proposed algorithm and other outliers identification techniques. Defining a critical value instead of specifying the number of outliers in advance is another direction in future work.

REFERENCES

[1] P. Angiulli and S. Basta. Distance-based detection and prediction of outliers. *IEEE Transactions on Knowledge and Data Engineering*, 18:145–160, 2006.

[2] S. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38, 2003.

[3] N. Billor, A. Hadi, and P. Velleman. Blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34:279–298, 2000.

[4] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *In Proceedings of 2008 SIAM Data Mining Conference*, pages 243–254, Atlanta, GA., April 2008.

[5] V. Chandola, S. Boriah, and V. Kumar. A framework for exploring categorical data. In *In Proceedings of 2009 SIAM Data Mining Conference*, Sparks, NV, 1–13 October 2009.

[6] K. Das, J. Schneider, and D. B. Neill. Anomaly pattern detection in categorical datasets. In *KDD'08*, pages 169–176, 2008.

[7] D.M. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[8] Z. He, X. Xu, and S. Deng. An optimization model for outlier detection in categorical data. In *International Conference on Intelligent Computing*, pages 400–409, 2005.

[9] Z. He, X. Xu, and S. Deng. A fast greedy algorithm for outlier mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 567–576, Seoul-Korea, 2006.

[10] Z. He, X. Xu, J. Z. Huang, and S. Deng. Fp-outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems (COMSIS)*, 2:726–732, 2005.

[11] V.J Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

[12] S. Y. Jiang, X. Song, H. Wang, J. J. Han, and Q. H. Li. A clustering-based method for unsupervised intrusion detections. *Pattern Recognition Letters*, 27:802–810, 2006.

[13] A. Koufakou, M. Georgiopoulos, and G. Anagnostopoulos. Detecting outliers in high-dimensional datasets with mixed attributes. In *International Conference on Data Mining (DMIN 2008)*, L.Vegas,NV, 14-17 2008.

[14] A. Koufakou, E. Ortiz, M. Georgiopoulos, G. Anagnostopoulos, and K. Reynolds. A scalable and efficient outlier detection strategy for categorical data. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Patras-Peloponnese-Greece, 29–31 October 2007.

[15] S. Lin and D. E. Brown. An outlier-based data association method for linking criminal incidents. *Decision Support Systems*, 41:604âĂŞ–615, March 2006.

[16] K. Narita and H. Kitagawa. Detecting outliers in categorical record databases based on attribute associations. *Progress in WWW Research and Development*, 4976:111–123, 2008.

[17] M. E. Otey, A. Ghoting, and S. Parthasarathy. Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery*, 12(2-3):203 – 228, May 2006.

[18] S. Shekhar, C. T. Lu, and Pusheng Zhang. Detecting graph-based spatial outliers: algorithms and applications. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 371–376, San Francisco, California, 2001.

[19] M. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S. Chen, L. W. Chang, and T. Goldring. Handling nominal features in anomaly intrusion detection problems. In *International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications*, pages 55 – 62, 2005.