Ramalan Kani K *et al*, International Journal of Computer Science and Mobile Applications,
Vol.2 Issue. 3, March- 2014, pg. 47-54 **ISSN: 2321-8363**

MINING OF OUTLIER DETECTION IN LARGE CATEGORICAL DATASETS

Mrs. Ramalan Kani K[1], Ms. N.Radhika[2]

[1]M.TECH Student, Department of computer Science and Engineering, PRIST University, Trichy
[2]Asst.Professor, Department of Computer Science and Engineering, PRIST University, Trichy

Abstract

Outlier detection will typically be thought of as a pre-processing step for locating, throughout a data set, those objects that do not fits well-defined notions of expected behaviour. it is vital in process for locating novel or isolated events, anomalies, vicious actions, exceptional phenomena, etc. We have got an inclination to reinvestigating outlier detection for categorical data sets. This draw back is very hard owing to the matter of shaping a pregnant similarity live for categorical data. Throughout this paper, we have got an inclination to propose an accurate definition of outliers associated academic degree improvement model of outlier detection, via a replacement construct of holoentropy that takes each entropy and total correlation into thought. Supported this model, we have got an inclination to stipulate a perform for the outlier issue of associate object that's fully determined by the item itself and will be updated expeditiously. We have got AN inclination to propose 2 wise 1-parameter outlier detection ways in which, named ITB-SS and ITB-SP, that need no user-defined parameters for deciding whether or not or not or not associate object is associate outlier. Users would like solely supply the number of outliers they need to notice. Experimental results show that ITB-SS and ITB-SP square measure easier and economical than thought ways in which and will be accustomed agitate each vast and high-dimensional data sets wherever existing algorithms fail. Other ways like possibility, Hyper graph theory or agglomeration ways in which goes fail in outlier detection in categorical data. We have got an inclination to area unit measure the outlier detection pattern entropy and total correlation.

Keywords: Outlier detection, holoentropy, total correlation, outlier factor, attributes weighting

1. Introduction

".Outlier exposure, that is a lively analysis space refers to the matter of finding objects in an exceedingly information set that don\'t adapt to well-defined notions of expected behavior. The objects detected square measure referred to as outliers, conjointly spoken as anomalies, surprises, aberrant, etc. Outlier detection may be enforced as a preprocessing step before the applying of a sophisticated information analysis methodology. It can even be used as an efficient tool to get interest patterns like the expense behavior of a to-be bankrupt credit cardholder. Outlier detection is a vital step in exceedingly sort of sensible applications together with intrusion detection, health system observance, and criminal activity detection in E-commerce, and may even be employed in research project for information analysis and data discovery in biology, chemistry, astronomy, earth science, and different fields.

The existing ways for outlier detection square measure classified in line with the supply of labels within the coaching information sets, there square measure three broad categories: superintended, semi-supervised, and unsupervised approaches. in essence, models among the supervised or the semi-supervised approaches all got to be trained before use, whereas models adopting the unsupervised approach don\'t embrace the coaching part. Moreover, in an exceedingly supervised approach a coaching set ought to be given labels for anomalies in addition as labels of traditional objects, in distinction with the coaching set with traditional object labels alone needed by the semi-supervised approach. On the opposite hand, the unsupervised approach doesn\'t need any object label info. So the 3 approaches have completely different conditions and limitations, and that they match {different|totally completely different|completely different} sorts of information sets with different amounts of label info.

The three broad classes of outlier detection techniques square measure mentioned below. The supervised anomaly detection approach learns classifier victimisation tagged objects happiness to the traditional and anomaly categories, and assigns applicable labels to check objects. The supervised approach has been studied extensively and plenty of ways are developed. for example, the cluster of proximity-based ways includes the cluster-based "K-Means+ID3" algorithmic program that cascades K-Means clump associated an ID3 call tree for classifying abnormal and traditional objects. The work of Barbara´ et al is predicated on applied math testing associated an application of Transduction Confidence Machines, which needs k neighbors. Moreover, one-class SVMs are applied generally during this field as they are doing not ought to create likelihood density estimation. a spread of ways supported scientific theory have conjointly been projected. The work of Filippone and Sanguinetti proposes a technique to regulate the false positive rate within the novelty detection drawback.

The semi-supervised anomaly detection approach primarily learns a model representing traditional behavior from a given coaching information set of traditional objects, then calculates the chance of a check object's being generated by the learned model. Zhang et al propose associate custom-made hidden mathematician model for this approach to anomaly detection, whereas bureau set al propose a clustering-based algorithmic program that punishes deviation from familiar labels. ways that assume handiness of solely the outlier objects for coaching square measure rare, as a result of it\'s tough to get a coaching information set that covers all doable abnormal behavior which will occur within the information.

The unsupervised associateomaly detection approach detects anomalies in an unlabelled information set below the belief that the bulk of the objects within the information set square measure traditional. Angiulli et al propose a KNN distance-based methodology. clump is another wide enforced methodology, of that is associate example. Moreover, this approach is applied to completely different sorts of outlier detection tasks and information sets, e.g., conditional anomaly detection, context-aware outliers, and outliers in linguistics graphs. As this approach doesn't need a tagged coaching information set and is appropriate for various outlier detection tasks, it's the foremost wide applicable. To implement supervised and semi-supervised outlier detection ways, one should initial label the coaching information. However, once long-faced with an outsized information set with voluminous high-dimensional objects and a coffee abnormal rate, choosing the abnormal and traditional objects to compose a decent coaching information set is long and labour-intensive. The unsupervised approach is a lot of wide used than the opposite approaches as a result of it doesn\'t want tagged info. If one desires to use a supervised or semi-supervised approach, associate unsupervised methodology may be used because the opening to search out a candidate set of outliers, which can facilitate specialists to create the coaching information set. The unsupervised approach is our analysis focus during this paper.

2. Related Work

Thought ways algorithms designed for outlier detection from categorical information may be classified into four classes. A number of these algorithms are compared with the projected algorithms.

2.1 Proximity-Based ways

Being intuitively simple to grasp, proximity-based outlier detection, that measures the distance of objects in terms of distance, density, etc., is a vital technique adopted by several outlier detection ways. For numerical outlier detection, there are a spread of ways ,] during this class. for example, LOF is a good technique that utilizes an inspiration of native density to live however isolated associate object is w.r.t. the encompassing Minpts objects.

For categorical information sets, the proximity-based methods must confront the issues of a way to select the measure of distance or density and the way to avoid time and area complexness within the distance computing method. For example, dolphin uses the acting distance and CNB employs a common-neighbor-based distance to live the gap between categorical objects. The CNB algorithmic program consists of 2 steps, the neighbor-set generating step and also the outlier mining step. The neighbor-set of the k nearest neighbors with similarity threshold nine to all or any objects is computed within the neighbor-set generation step. Each k and a pair of are user-defined parameters. within the second step, associate outlier issue for every object is computed by summing its distance from its neighbors. The objects with the o (number of outliers) largest values are set to be outliers. The proximity-based approach has several requirement parameters, which require continual trial-and error to achieve the required result. Proximity-based ways conjointly suffer from the curse of spatiality once victimisation distance or native density measures on the complete dimensions. In general, these ways are time- and space-consuming and consequently aren't applicable for giant information sets.

2.2 Rule-Based ways

Rule-based ways borrow the thought of frequent things from association-rule mining. Such ways take into account the frequent or sporadic things the info set. for example, within the work of , objects with few frequent things or several sporadic things ar a lot of probably to be thought of as abnormal objects than others.

Frequent Pattern Outlier issue (called the FIB technique during this paper) and Otey's algorithmic program (called the OA method during this paper) are 2 well-known rule-based techniques. The procedure of the FIB algorithmic program includes associate initial computation of the set of frequent patterns, using a predefined minimum support rate. for every object, all support rates of associated frequent patterns are summed up because the outlier issue of this object. The objects with the o smallest factors are thought of because the outliers. Contrary to the FIB algorithmic program, OA begins by collection the sporadic things from the info set. supported the sporadic things, the outlier factors of the objects are computed. The objects with the o largest scores are treated as outliers. The time complexness of each algorithms is set by the frequent-item or infrequent-item generating processes. for example, the time complexness of the FIB technique is exponentially increasing with the quantity of attributes due to the Apriori algorithmic program. Therefore, this approach is restricted to low-dimensional information sets.

2.3 Information-Theoretic ways

Several information-theoretic ways are projected within the literature. For anomaly detection in audit information sets, Lee and Xiang [36] gift a series of information-theoretic measures, i.e., entropy, conditional entropy, relative conditional entropy, and data gain, to spot outliers within the univariate audit information set, wherever the attribute relationship doesn't got to be thought of. The work of He et al. employs entropy to live the disorder of an information set with the outliers removed. In these ways, heuristic native search is employed to reduce the target operate. The ways projected in and set a threshold of mutual info and acquire a collection of dependent attribute pairs. Based on this set, associate outlier issue for every individual object is outlined. In general, info-theoretic ways focus either on one entropy-like measure or on mutual information, and need dear estimation of the chance distribution once the info set is shrunken following elimination of sure outliers.

2.4 Different ways

Several different approaches victimisation the stochastic process, Hypergraph theory, or cluster ways are projected to handle the matter of outlier detection in categorical information. for example, supported hypergraph theory, HOT captures the distribution characteristics of associate object within the subspaces and these characteristics are then wont to determine outliers. within the random-walk-based technique [34], outliers are those objects with a coffee likelihood of jumping to neighbors. In different words, they need a high likelihood of staying in their states. In , the relationships among the neighbors are thought of and a mutual-reinforcement based native outlier issue is projected to spot outliers. This may even be viewed as a random-walk technique with a set variety of walk steps. {in a|during a|in associate exceedingly|in a very} cluster-based native outlier detection technique is projected to spot the physical significance of an object. The outlier think about this technique is measured by each the scale of the cluster the item belongs to and also the distance between the item and its nearest cluster. These ways aren't terribly economical for giant or high-dimensional information sets as a result of they contain some high-complexity procedures, e.g., frequent-item generating processes in HOT, similarity computation in the random-walk-based ways, and also the cluster method within the cluster-based technique.

Mainstream methods algorithms designed for outlier detection from categorical data can be grouped into four categories. Some of these algorithms are compared with the proposed algorithms.

3 Proposed System

The formal optimization-based model of categorical outlier detection, for which a new concept of weighted holoentropy which captures the distribution and correlation information of a data set is proposed. To solve the optimization problem a new outlier factor function is derived from the weighted holoentropy and show that computation/updating of the outlier factor can be performed without the need to estimate the joint probability distribution and estimate an upper bound of outliers to reduce the search space. This paper proposes two effective and efficient algorithms, named the Information-Theory-Based Step-by-Step (ITB-SS) and Single-Pass (ITB-SP) methods. These algorithms need only the number of outliers as an input parameter and completely dispense with the parameters for characterizing outliers usually required by existing algorithms.
The goal of this paper is twofold.
Deal with the lack of a formal definition of outliers and modeling of the outlier detection problem.
Aim to propose effective and efficient methods that can be used to solve the outlier detection problem in real applications.

4. Measurement for Outlier Detection

4.1 Entropy and Total Correlation

In this module entropy, mutual information, and total correlation computed on the data set X.
The entropy can be used as a global measure in outlier detection. In information theory, entropy means uncertainty relative to a random variable: if the value of an attribute is unknown, the entropy of this attribute indicates how much information needs to predict the correct value.

Here X is a dataset containing n objects $\{x_1, x_2, \ldots, x_n\}$, each x_i for $1 \leq i \leq n$ being a vector of categorical attributes $[y_1, y_2, \ldots, y_m]^T$, where m is the number of attributes, y_j has a value domain determined by $[y_{1,j}, y_{2,j}, \ldots, y_{nj,j}]$ $(1 \leq j \leq m)$ and n_j indicates the number of distinct values in attribute y_j. The total correlation is defined as the sum of mutual information of multivariate discrete random vectors Y.

Where $r_1 \ldots r_i$ are attribute numbers chosen from 1 to m. $I_X (y_{r1}, \ldots, y_{ri}) = I_X (y_{r1}, \ldots, y_{ri-1}) - I_X(y_{r1}, \ldots, y_{ri-1}|y_{ri})$ is the multivariate mutual information of $y_{r1} \ldots y_{ri}$, where $I_X (y_{r1}, \ldots, y_{ri-1}|y_{ri}) = E(I(y_{r1}, \ldots, y_{ri-1})|y_{ri})$ is the conditional mutual information. The total correlation is a quantity that measures the mutual dependence or shared information of a data set

4.2 Holo-entropy

The holoentropy is defined as the sum of the entropy and the total correlation of the random vector Y, and can be expressed by the sum of the entropies on all attributes. Given that the holoentropy is defined as the sum of entropies of individual attributes and outliers are detected by minimizing the holoentropy through the removal of outlier candidates

$$HL_{\mathcal{X}}(\mathcal{Y}) = H_{\mathcal{X}}(\mathcal{Y}) + C_{\mathcal{X}}(\mathcal{Y}) = \sum_{i=1}^{m} H_{\mathcal{X}}(y_i). \tag{3}$$

When the components of Y are independent or Y has only one component, $HL_X(Y) = H_X(Y)$, i.e., the holoentropy coincides with the entropy.

4.3. Attribute Weighing

In this module the attribute weight is computed. Given that the holoentropy is defined as the sum of entropies of individual attributes and outliers are detected by minimizing the holoentropy through the removal of outlier candidates, our strategy consists in weighting the entropy of each individual attribute in order to give more importance to those attributes with small entropy values. This increases the impact of removing an outlier candidate that is outstanding on those attributes. To weight the entropy of each attribute, propose to employ a reverse sigmoid function of the entropy, as follows:

$$w_{\mathcal{X}}(y_i) = 2\left(1 - \frac{1}{1 + \exp(-H_{\mathcal{X}}(y_i))}\right). \tag{4}$$

The weighted holoentropy $W_X(Y)$ is the sum of the weighted entropy on each attribute of the random vector Y

$$W_{\mathcal{X}}(\mathcal{Y}) = \sum_{i=1}^{m} w_{\mathcal{X}}(y_i) H_{\mathcal{X}}(y_i). \tag{5}$$

4.4. ITB-SP (Single Pass)

This algorithm is used to detect outliers. The outlier factors are computed only once, and the o objects with the largest OF(x) values are identified as outliers. The initialization of AS (anomaly candidate set) requires computation of the outlier factors of all the objects. In ITB-SP, the attribute weights, the OF(xi) of all the objects, initialization of AS and the heap sort search to find the top-o outlier candidates are computed.

The outlier factor of an object xo, denoted as $OF(x_o)$, is defined as

$$\begin{aligned} OF(x_o) &= \sum_{i=1}^{m} OF(x_{o,i}) \\ &= \sum_{i=1}^{m} \begin{cases} 0, & if\ n(x_{o,i}) = 1; \\ w_{\mathcal{X}}(y_i) \cdot \delta\big[n(x_{o,i})\big], & else. \end{cases} \end{aligned} \tag{6}$$

Where *OF(xo,i)* is defined as the outlier factor of xo on the ith attribute. *OF (xo)* can be considered as a measure of how likely it is that object xo is an outlier. An object xo with a large outlier factor value is more likely to be an outlier than an object with a small value.

Algorithm 1. ITB-SP single pass
1: **Input:** data set $\mathcal{X}$ and number of outliers requested o
2: **Output:** outlier set OS
3: Compute $w_{\mathcal{X}}(y_i)$ for $(1 \le i \le m)$ by (4)
4: Set $OS = \phi$
5: **for** $i = 1$ to n **do**
6: Compute $OF(x_i)$ and obtain AS by (13)
7: **end for**
8: **if** $o > UO$ **then**
9: $o = UO$
10: **else**
11: Build OS by searching for the o objects with greatest $OF(x_i)$ in AS using heapsort
12: **end if**

4.4. ITB-SS (Step by Step)

At each step of SS, the object with the largest OF(x) is identified as an outlier and is removed from the data set. Following this removal, the outlier factor OF(x) is updated for all the remaining objects. The process repeats until o objects have been removed. For ITB-SS, the attribute weights, initial outlier factors including initialization of AS, and the step-by-step top-o outlier selection procedure are computed. ITB-SS does benefit, however, from the reduced search space. In designing the two algorithms, assumed that the number of requested outliers o is always smaller than UO. That AS is

indeed large enough to include all the candidate objects that can reasonably be considered as outliers. Nevertheless, only minor modifications need to be made if a user wants to obtain more than UO "outliers."

Algorithm 2. ITB-SS Step-by-Step

Input: data set $\mathcal{X}$ and number of outliers requested o
Output: outlier set OS
Set $OS = \phi$
Compute $w_{\mathcal{X}}(y_i)$ for $(1 \leq i \leq m)$ by (4)
for $i = 1$ to n **do**
    Compute $OF(x_i)$ and obtain AS by (13)
end for
if $o > UO$ **then**
    $o = UO$
else
    for $i = 1$ to o **do**
        Search for the object with greatest $OF(x_o)$ from AS
        Add x_0 to OS and remove it from AS
        Update all the $OF(x)$ of AS
    end for
end if

5. Conclusion

The formulated outlier detection as an optimization problem and proposed two practical, unsupervised, 1-parameter algorithms for detecting outliers in large-scale categorical data sets. The effectiveness of our algorithms results from a new concept of weighted holoentropy that considers both the data distribution and attribute correlation to measure the likelihood of outlier candidates, while the efficiency of our algorithms results from the outlier factor function derived from the holoentropy. The outlier factor of an object is solely determined by the object and its updating does not require estimating the data distribution. Based on this property, apply the greedy approach to develop two efficient algorithms, ITB-SS and ITB-SP that provide practical solutions to the optimization problem for outlier detection. We also estimate an upper bound for the number of outliers and an anomaly candidate set. This bound, obtained under a very reasonable hypothesis on the number of possible outliers, allows us to further reduce the search cost. The proposed algorithms have been evaluated on real and synthetic data sets, and compared with different mainstream algorithms. First, our evaluations on a small real data set and a bundle of synthetic data sets show that the proposed algorithms do tend to optimize the selection of candidates as outliers. Moreover, our experiments on real and synthetic data sets in comparison with other algorithms confirm the effectiveness and efficiency of the proposed algorithms in practice. In particular, show that both of our algorithms can deal with data sets with a large number of objects and attributes.

References

[1] E.M. Knorr and R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. 24rd Int'l Conf. Very Large Data Bases (VLDB '98), 1998.

[2] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, Feb. 2006.

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

[4] K. Das, J. Schneider, and D.B. Neill, "Anomaly Pattern Detection in Categorical Data Sets," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), 2008.

[5] S. Ramaswamy, R. Rastogi, K Shim "Efficient Algorithms for Mining Outliers from Large Data Sets " in proceeding of the 2000 ACM SIGMOD international conference ,Volume 29 Issue 2, June 2000.

[6] S. Srinivasa, "A Review on Multivariate Mutual Information,"Univ. of Notre Dame, Notre Dame, Indiana, vol. 2, pp. 1-6, 2005.

[7] S. Watanabe, "Information Theoretical Analysis of Multivariate Correlation," IBM J. Research and Development, vol. 4, pp. 66-82,1960.

[8] L. Wei, W. Qian, A. Zhou, W. Jin, and J.X. Yu, "HOT:Hypergraph-Based Outlier Test for Categorical Data," Proc.Seventh Pacific-Asia Conf. Advances in Knowledge Discovery andData Mining (PAKDD '03), 2003.

[9] M. Breunig, H-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," Proc. ACM SIGMOD Int'l Conf.Management of Data (SIGMOD '00), 2000.

[10] P.K. Chan, M.V. Mahoney, and M.H. Arshad, "A Machine Learning Approach to Anomaly Detection," technical report,
Florida Inst. of Technology, 2003.

