Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Outlier Detection for Large Scale Categorical Data

Mahesh Shinde, Chandrashekhar Shitole, Vishal Tidke, Ravindra Trimukhe, Asst. Rupali Shishupal
C.E., Pune University, Pune, Maharashtra,
India

Abstract: Outlier detection is an important issue occur within various research and applications domains in today. It aims to detect the object that are considerably distinct, exceptional and inconsistent the majority data in input data sets. Many outlier detection terms have been specifically developed for certain application domains. To identify abnormal data which forms non-conforming pattern is referred to as outlier, anomaly detection. This leads to knowledge and discovery. In this paper, we propose a formal definition of outliers that decide the data set is within outlier or not .Many outlier detection method have been proposed based on classification clustering, classification, statistics and frequent patterns. Among them information theory have some different perspective while its computation is based on statistical approach only. The outlier detection from unsupervised data sets in more challenging since there is no inherent measurement of distance between these objects. We propose two methods 1-parameter outlier detection methods, named ITB-SS and ITB-SP, which finds data can be in inner layer or outlier. Users need only to set number of outliers they want to detect in different data set.

Keywords:-Outlier detection, total correlation, outlier factor, holoentropy, attribute weighting, greedy algorithms

I. INTRODUCTION

Outlier detection is term of data mining, An outlier is a data point which is significantly different from the remaining data the concept of outlier is observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. Outlier detection in many times known as anomaly detection in advanced technology for a high range of real time applications like security, medical, industrial, e-commerce and engineering purpose. Outlier arises due to faults in systems, changes in the system, human errors, behavioral and instrumental errors. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or in a set of manner. When the generating process behaves in an unusual way, it results in the creation of outliers. This paper mainly discusses about outlier detection approach from data mining perspective. The inherent idea is to research and compare to special database.

The existing system for outlier detection measure classified in line with the supply of labels within the information sets, there three broad categories of square measure are : superintended, semi-supervised, and unsupervised approaches. in substance , models among the supervised or the semi-supervised got to be trained before use, whereas models adopting the unsupervised approach don't embrace the coaching part.

The supervised anomaly detection approach learn classifier scam on tagged objects happiness to the traditional and anomaly categories, and assigns applicable labels to checking 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 semi-supervised anomaly detection approach primarily learns a model representing traditional behavior from training information set of traditional objects, then calculates the chance of a check object's being generated by the learned model. In outlier propose associate custom-made hidden mathematician model for this approach to anomaly detection, propose a clustering-based algorithmic program that punishes deviation from familiar labels.

The unsupervised associate anomaly detection approach can detects anomalies in an-unsorted unlabelled information set below the belief that the set of the objects within the information set square measure traditional. The unsupervised approach is a lot of wide used than the opposite approaches as a result of tagged information. If one desires to use a supervised or semi-supervised approach, associate unsupervised methodology may be used for because the opening to search out a candidate set of outliers, which can provide specialists to create the trained information set.

To implement supervised and semi-supervised outlier detection methods approach first label the training data. However, when faced with a large data set with millions of high-dimensional objects and a low data rate, picking the normal and abnormal objects to compose a good coaching data set is time-consuming and labor-intensive. As compare to other approaches unsupervised approach is used more other than approaches because it does not need labeled information. If one wants to employ a supervised or semi-supervised approach, this unsupervised approach can be used as the first step to find a candidate set of outliers, which will help experts to build the training data set. The unsupervised approach is our research focus in this paper.

II. PREVIOUS WORK

The classic definition of an outlier is due to Hawkins who defines "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".Thought ways for outlier detection approach algorithms designed from categorical information, these are classified into four classes, 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 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 Min pts objects.

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 .example, within the work of, objects with few frequent things or several sporadic things are a lot of probably to be thought of as abnormal objects than others.

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 audit information set, wherever the attribute relationship.

2.4 Different ways

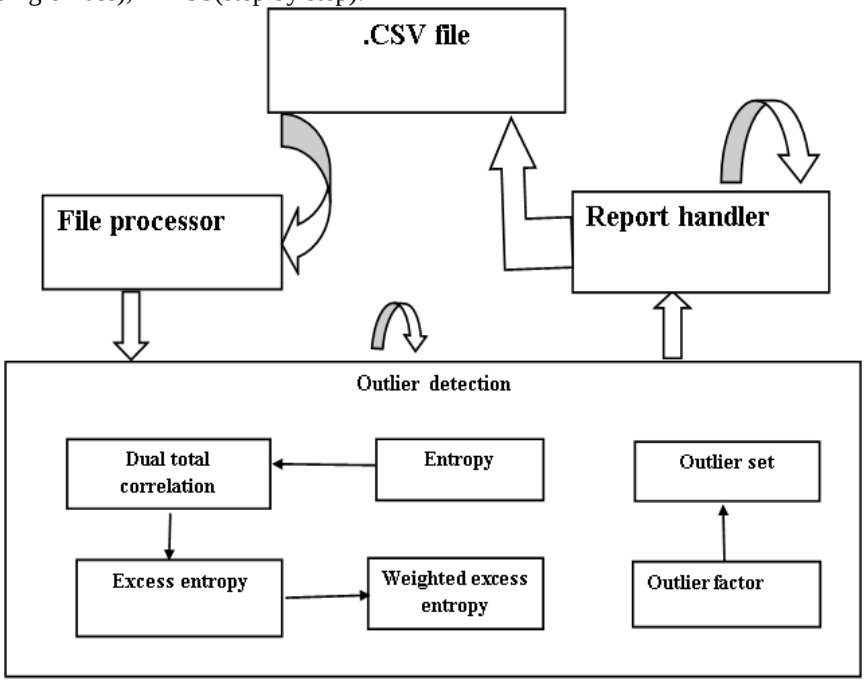
Several different approaches victimization the stochastic process, Hyper graph theory, or cluster ways are projected to handle the matter of outlier detection in categorical information. for example, supported hyper graph theory, HOT captures the distribution characteristics of associate object within the subspaces and these characteristics are then wont to determine outliers.

III. PROPOSED SYSTEM

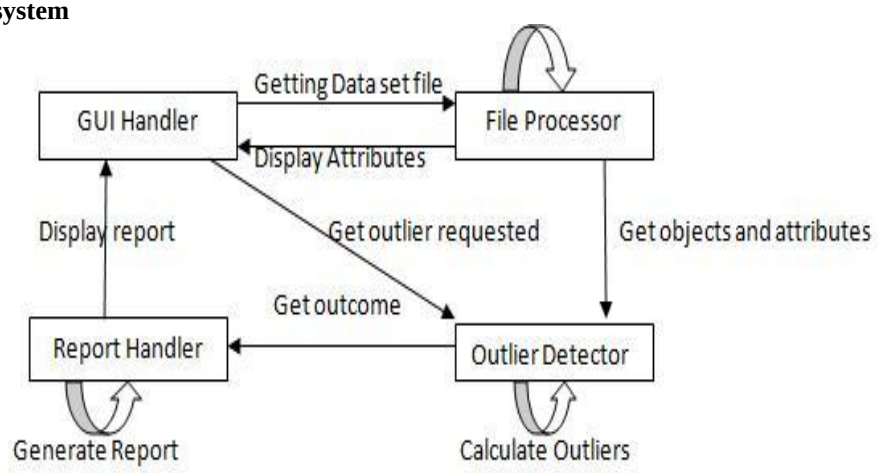
In the formal optimization-based model of categorical outlier detection, for which deriving a new method of weighted holoentropy, that work for detecting the distribution and correlation information of a data set is proposed. For the solving this problem a new outlier factor function is derived from 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. In this paper proposes two effective and efficient algorithms, named the ITB-SS (Information-Theory-Based Step-by-Step) and ITB-SP (Single-Pass) methods. these algorithm only needs number of outliers and input parameter for detecting outliers required by existing system. Aim to propose effective and efficient methods that can be used to solve outlier detection problem in real applications.

IV. SYSTEM ARCHITECTURE

In the system architecture of outlier detection use the different stages for calculate terms entropy, Holoentropy, Attribute Weighing, ITB-SP(Single-Pass),ITB-SS(step by step).



4.1 Overview of system



V. MEASUREMENT FOR OUTLIE DETECTION

5.1 Entropy and Total Correlation

The entropy can be used for global measure in outlier detection. In information theory, entropy means uncertainty relative to a random variable. In a dataset ,If the value of attributes is unknown then entropy of this attribute indicate how much information is needed to predict the correct one value.

$$Hx(y) = Hx(y1, y2 \ldots, ym) = \sum_{i=1}^{m} Hx\left(\frac{yi}{yi-1}, \ldots, y1\right)$$

$$Hx(y1) + Hx\left(\frac{y2}{y1}\right) + \cdots + Hx\left(\frac{ym}{ym-1}, \ldots y1\right)$$

$$\text{Where} Hx\left(\frac{ym}{ym-1}, \ldots, y1\right) = -\sum_{ym,ym-1,\ldots y1} p(ym, ym-1, \ldots, y1) \log p\left(\frac{ym}{ym-1}, \ldots y1\right).$$

5.2 Holo-entropy

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

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

5.3. Attribute Weighing

In this outlier detection strategy consists in weighting the entropy of each individual attribute in order to give more importance to those attributes with small entropy values. This use to increases the impact of removing an outlier candidate on those attributes. To weight the entropy of each attribute, propose to employ a reverse sigmoid function of the entropy, as follows:

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

5.4. ITB-SP (Single Pass)

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.

1: Input: data set X and number of outliers requested o
2: Output: outlier set OS
3: Compute Wx(yi) for($1 \leq i \leq m$)
4: Set OS =0
5: for i= 1 to n do
6: Compute OF(xi) and obtain AS
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(xi) in AS using heapsort
12. end if

VI. RECENT ADVANCED IN OUTLIER DETECTION

Using the fast development of data mining technique, identification of outliers in large dataset has received more and more attention. In that some new method are developed for special background.

6.1 High Dimension-based Approach:

In outlier detection High dimension space is a difficult problem. a new method designed for high dimension proposed ODHDP based on the concept of projection is proposed in that paper, deal with the sparsity of high dimensional points.

6.2 SVM-based Approach:

A SVM-based method based on outlier detection approach. That used several models for varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This has the advantage that the decision does not depend on the quality of a single model. Other outlier detection Support Vector approach, using Replicator Neural Networks (RNNs), and using a relative degree of density with respect only to a few fixed reference points

VII. CONCLUSION

This paper mainly discusses about outlier detection approaches from position of data mining term. In that we reviews related work in outlier detection. Based on this paper we apply the greedy approach to develop two efficient algorithms, ITB-SS, and ITB-SP, that provide practical solutions on optimization problem for outlier detection. In that we have formulated outlier detection as an optimization problem and proposed a practical for parameter less algorithm for detecting outliers in large-scale categorical data sets. In particular, we can say our algorithms can deal with data sets with a large number of objects and attributes.

ACKNOWLEDGEMENT

We would like to Thanks all authors in reference section. Their methods, formulas, algorithms, conceptual techniques are very helpful for our research to publishing this paper . All papers in the reference section are very useful for our proposed system.

REFRENCESS

[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] Mrs. Ramalan Kani K and Ms. N.Radhika "MINING OF OUTLIER DETECTION IN LARGE CATEGORICAL DATASETS" IJCSI ,Vol.2 Issue. 3, March- 2014

[4] Wenbin Zhou and Su Yang "Outlier Detection on Large-Scale Collective Behaviors" Fourth International Joint Conference on Computational Sciences and Optimization 2011

[5] Karanjit Singh and Dr.Shuchita Upadhyaya "Outlier Detection: Applications and Techniques" IJCSI International Journal of Computer Science Issues, Vol. 9

