International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)

ISSN: 0976-1353 Volume 7 Issue 1 –MARCH 2014.

OUTLIER DETECTION THROUGH UNSUPERVISED APPROACH WITH HOLOENTROPY

N.Ganesa Moorthy[1] S.Nandhini Devi [2]

[1]PG Scholar, Computer Science and Engineering, Srinivasan Engineering College
nganeshm1988@gmail.com

[2]Assistant Professor, Computer Science and Engineering, Srinivasan Engineering College
Nandhini@gmail.com

Abstract- Outlier detection can usually be considered as a pre-processing step for analyzing in a data set, each and every object that does not conform to well-defined meaning of expected operation. It is very important in data mining concept for discovering unusually happens, bad actions, exceptional events, etc. Investigating outlier detection for categorical data sets. This abnormal data set are challenging because of the difficulty of defining a meaningful similarity measure for categorical data. So the formal definition of outliers and an selection of a best model of outlier detection, these are all new concept of Holoentropy that takes both entropy and total correlation into consideration. Based on this concept, that we define a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently. There are two practical, parameter outlier detection methods, named Information Theory based SS(step by step) and SP(single Pass), are used to the analyzer that user defined parameters for deciding whether an object is an outlier. Users can only prefer the number of outliers they want to detect. Outcome results show that SS and SP algorithm are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets.

Index Terms -Entropy, Total Correlation, Holoentropy , Outlier Factor

I.INTRODUCTION

In real applications, a large portion or the entire data set is often presented in terms of categorical attributes. examples of such data sets include transactional data, financial records, and customer information in commercial banks, etc. the problem of outlier detection in this type of data set is more challenging since there is no inherent measurement of distance in between the objects. Existing unsupervised outlier detection approaches, e.g., lof, loci, and effective on data sets with

In general, outlier candidates can be estimated based on data distribution or on attribute correlation, which provides a global measure for outlier detection. They can also be estimated using a between-object similarity or local density, which provides a local measure for outlier detection. various outlier detection techniques such as rule based methods , proximity-based methods, and information-theoretic methods have been proposed

The common problem with the existing methods is the lack of a proper definition for the outlier detection problem. Without a proper definition, outlier detection is often designed as an ad hoc process. Several user defined parameters are required to define whether an object to be qualified as an outlier. the parameter based results are heavily dependent on suitable parameter settings only, which are very difficult to estimate without background knowledge about the data. many existing outlier and anomaly detection methods also suffer from low effectiveness and low efficiency due to high dimension and large size of the data set.

The supervised anomaly detection approach learns a classifier using labeled objects belonging to the normal classes and anomaly classes containing objects, and assigns appropriate labels to test objects. The supervised outlier detection approach has been studied extensively and many methods have been developed to detect the outlier .

The semi-supervised anomaly detection approach primarily learns a model representing normal behavior from a given training data set of normal objects in a class , and then calculates the likelihood of a test object’s being generated by the learned model. it is difficult to obtain a training data set which covers all possible abnormal behavior that can occur in the data.

The unsupervised anomaly detection approach detects anomalies in an unlabeled data set under the assumption that the majority of the objects in the data set are normal.. Moreover, this approach is applied to different kinds of outlier detection tasks and data sets, As this unsupervised approach does not require a labeled training data set and is suitable for different types of outlier detection tasks, it is the most widely used and applicable for high dimensional data set.

During the implementation of supervised and semi-supervised outlier detection approaches , one must first label the training data in a data set. However, when faced with a large high dimensional data set with millions of high-dimensional objects and a low anomaly data rate. In this data set picking the abnormal and normal objects to compose a good training data set assigned with labels are labor intensive and require more time . The unsupervised approach is more widely used method compared to the other approaches because it does not need any labeled information. If one wants to use a supervised approach or semi-supervised approach for outlier detection , an unsupervised method can be used as the first step to find a candidate set of outliers, which will help to build the training data set. The unsupervised outlier detection method is our research focus (to detecting the irrelavent data i.e outlier data)in this paper.

Propose a formal optimization based model to detect the outlier in categorical data set , through a new concept of weighted Holoentropy which captures the distribution and correlation information of a data set is proposed.

To solve the optimization problem, derive a new outlier factor function 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 also estimate an upper bound of outliers to reduce the search space.

Propose two effective and efficient algorithms, based on Information Theory. They are Step-by-Step and Single-Pass 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.

Information-Theory Based Step by Step(SS)

Information-Theory Based Single Pass(SP)

II RELATED WORK

Large Scale Categorical Data sets are defining a meaningful similarity using some methods. Methods for outlier detection are classified according to the availability of labels in the training data sets in given data set, there are three categories:

Supervised Approach

In a Supervised outlier detection approach a training set should be provided with labels for anomaly data sets and then assign with labels for normal objects.

Semi-supervised Approach

In semi supervised outlier detection approach The training set are created with normal object only labels are not required for anomaly data's in given data set .

Unsupervised Approach

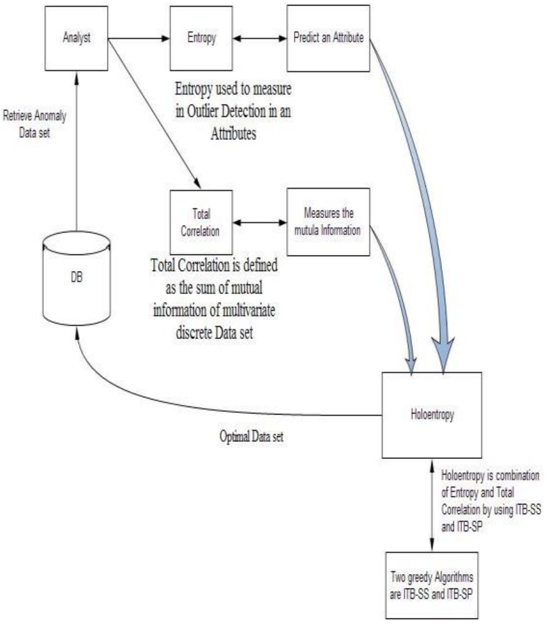
The unsupervised approach does not require any object label information. Investigating outlier detection for large scale categorical data sets. This problem difficult to defining a meaningful similarity measure for categorical data. Models within the supervised or the semi-supervised approaches all need to be trained before use.The current work focuses on finding single records that are anomalous. Sometimes in real world applications , more interested in detecting groups of unusual records that deviate from the norm, rather than detecting the records separately.

III. OUR SYSTEM AND ASSUMPTIONS

This paper is deals with Unsupervised approach with the lack of a formal definition of outliers and modeling of the outlier detection problem; second, aim to propose effective and efficient methods that can be used to solve the outlier detection problem in real applications. in this paper, these two goals are achieved by exploring the information theoretic approach.

First, approach, is adopt the deviation-based strategy which, according to, avoids the use of statistical tests and proximity-based measures to identify exceptional objects. Explore information theory to derive several new concepts. in particular and then combine entropy and total correlation with attribute weighting to define the concept of weighted Holoentropy, where the entropy measures the global disorder of a data set and the total correlation measures the attribute relationship. Based on this concept, build a formal model of outlier detection and propose a criterion for estimating the “goodness” of a subset of objects as potential outlier candidates.

The formulated outlier detection as an optimization problem and then proposed a two practical, unsupervised approach one parameter algorithms for detecting outliers in large categorical data sets. The effectiveness of our algorithms got 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 get from the outlier factor function derived from the Holoentropy. The outlier factor of an object is only 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, information theory based ss and sp, that provide practical solutions to the optimization problem for outlier detection in categorical data set. And also estimate an upper bound for the number of outliers and an anomaly candidate set. this upper bound, obtained under a very reasonable hypothesis on the number of possible outliers.


Analyst
Entropy
Predict an Attribute
Entropy used to measure in Outlier Detection in an Attributes
Retrieve Anomaly Data set
Total Correlation
Measures the mutula Information
DB
Total Correlation is defined as the sum of mutual information of multivariate discrete Data set
Holoentropy
Optimal Data set
Holoentropy is combination of Entropy and Total Correlation by using ITB-SS and ITB-SP
Two greedy Algorithms are ITB-SS and ITB-SP

Formal and proper definition of outliers and an optimization model of outlier detection in categorical data set using a new concept of Holoentropy that takes both entropy and total correlation into consideration.

Two practical parameter outlier detection methods, named Information Theory Based SS and SP.

1. Information-Theory Based Step by Step(SS)
2. Information-Theory Based Single Pass(SP)

By using these methods ,deals with both large and high-dimensional data sets. In an outlier detection , using the unsupervised approach method in a data sets. The unsupervised anomaly detection approach detects anomalies in an unlabeled data set under the assumption that the majority of the objects in the data set are normal. The unsupervised approach is more widely used than the other approaches because it does not need labeled information.

IV.SYSTEM PRELIMINARIES

Entropy

We will first introduce the concept of entropy, which is a measure of uncertainty of given random variable . Let X be a discrete random variable and probability mass function p(x) = Pr{X = x}, x E X. We denote the probability mass function by p(x) . Thus, p(x) and p(y) refer to two different random variables, and are in fact different probability mass functions, pX(x) and pY(y) respectively. The entropy H(X) of a discrete random variable X is defined by

$$H_x(y_m|y_{m-1},\dots,y_1) =$$

$$=\sum_{i=1}^{m} H_x(y_i\,|y_{i-1,\dots,}y_1)$$

$$= H_x(y_1) + H_x(y_2|y_1) + \cdots . + H_x(y_m|y_{m-1},\dots,y_1)$$

Total Correlation

The total correlation is defined as the sum of mutual information of multivariate discrete random vectors Y, denoted as $C_x(Y)$.Where r_1…….r_i are attribute numbers chosen from1 to m. $I_x(y_{r1};\dots;y_{ri})=I_x(y_{r1};\dots;y_{ri-1})- I_x(y_{r1};\dots;y_{ri-1}|y_{ri})$ is

$$C_x(Y) = \sum_{i=2}^{m} \sum_{\{r1,\dots ri\}C\{1,\dots,m\}} I_x(y_{r1},\dots;y_{rj}) =$$

$$= \sum_{(r_1,r_2)C\{1,\dots m\}} I_x(y_{r1};y_{r2}) + \cdots .. + I_x(y_{r1;\dots}y_{rm})$$

The multivariate mutual information of y_{r1}….y_{ri}, where $I_x(y_{r1};\dots;y_{ri-1}|y_{ri})$ $E(I_x(y_{r1};\dots;y_{ri-1})|y_{ri})$ is the conditional mutual information .The total correlation is a quantity that measure the mutual dependence or shared information of a data set.

Holoentropy

The Holoentropy $HL_x(Y)$ 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. The weighted Holoentropy denoted as

$$W_x(Y) = \textstyle\sum_{i=1}^{m} w_x(y_i)H_x(y_i)$$

Single Pass Algorithm

Step1:Predict attributes and requested to Outlier o
Step 2: Weighted Holoentropy execute the loop while log is always small then no of attributes
Step 3: Execute for loop up to no. of attributes in data set
Step 4: Calculate outlier factor to get anomaly candidate
Step 5: If object is less then upper bound on outlier
Step 6: Object and upper bound on outlier becomes equals
Step 7: Else object becomes greater then Outlier factor and check the data set using Heap sort

Step By Step Algorithm

Step1:Predict attributes and requested to Outlier o
Step 2: Weighted Holoentropy execute the loop while log is always small then no. of attributes
Step 3: Execute for loop up to no. of attributes in data set
Step 4: Calculate outlier factor to get anomaly candidate set
Step 5: If object is less then upper bound on outlier
Step 6: Object and upper bound on outlier becomes equals
Step 7: Loop executes up to object
Step 8: Search for the object with greatest outlier factor from anomaly candidate set
Step 9: Add attributes one by one up to outlier set
Step 10: Update all outlier factor attributes of anomaly candidate set

V.EXPERIMENTAL EVALUATION

Data Deployment

In this module, Deploy dataset from the database. And then to predict the number of attributes in a data set. And also predict the total correlation between the data set. In this module investigating outlier detection for categorical data sets.

Entropy Based Data Set

In this module, accessing the data set using Entropy. The entropy can be used as a global measure in to detect outlier in large categorical data set. 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, need to predict the correct value.

Total Correlation Based Data Set

In this module, data sets are accessed using Total Correlation. The total correlation is a quantity that measures the mutual dependence or shared information of a data set. It check the two different attributes, if one attribute is large, it means that the number of duplicate pairs of attribute values is small in these two attributes.

Holoentropy Based Data Set Using Information Theory Based SS And SP

In this module, the data sets are accessed using Holoentropy. Holoentropy is a combination of Entropy and Total correlation to detect outlier from the data set. Entropy alone is not a good enough measure for outlier detection and the contribution of the total correlation is necessary. Outlier detection from the data set by using Two Greedy Algorithms are Information Theory based SS and SP

VI.CONCLUSION

In this paper, formulated outlier detection as an optimization problem and proposed two practical, unsupervised approach based , one parameter algorithms for detecting outliers in large-scale categorical data sets. The effectiveness of our algorithms get 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 only 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, Information Theory based SS and SP, which provide practical solutions to the optimization problem for outlier detection. and 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.

The proposed algorithms have been evaluated on real and synthetic data sets, and compared with different mainstream algorithms. In a large-scale data set, those objects that conform to well-defined notions of expected behavior. To identify the outlier from the large data set, that means identify the vicious data set or abnormal data set from the large data set. The main work is to detecting the outlier from the large data set and updates the vicious data from the data set.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 009.

[2] V.J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Rev., vol. 22, no. 2, pp. 85-126, 2004.

[3] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, “Semi-Supervised Adapted HMMs for Unusual Event Detection,” Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR ’05),

[4] S.R. Gaddam, V.V. Phoha, and K.S. Balagani, “K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods,”

[5] T. Cover and J. Thomas, Elements of Information Theory. John Wiley & Sons, 1991.

[6] M. Filippone and G. Sanguinetti, “Information Theoretic Novelty Detection,” Pattern Recognition, vol. 43, pp. 805-814, 2010.

[7] 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.

[8] W. Lee and D. Xiang, “Information-Theoretic Measures for Anomaly Detection,” Proc. IEEE Symp. Security and Privacy, 2001.

[9] J. Han and M. Kamber, Data Mining—Concepts and Techniques.Elsevier, 2006.