Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.873 – 881

RESEARCH ARTICLE

INFORMATION-THEORETIC OUTLIER DETECTION FOR LARGE-SCALE CATEGORICAL DATA

Srijoni Saha Pradip[1], Jesica Fernandes Robert[2], Jasmine Faujdar Hamza[3]

[1]JSPM's BHIVRABAI SAWANT INSTITUTE of TECHNOLOGY & RESEARCH WAGHOLI, PUNE - 412207

[2]DEPARTMENT of COMPUTER ENGINEERING

[1] srijonisaha@gmail.com; [2] jesica.r.fernades@gmail.com; [3] jasminefaujdar74@gmail.com

Abstract*— Outlier detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specially developed for certain application domains. The outlier detection techniques have been proposed for the removal of unwanted data in order to get meaningful information based on classification, clustering, frequent patterns and statistics. The detection of outliers from unsupervised datasets is more complicated work since there is no inherent measure between the objects. We have proposed a novel method with effectively and efficiently detect outliers in unsupervised datasets using holoentropy. We have proposed ITB-SP outlier detection algorithm which do not require any user-defined parameter. This algorithm just takes the dataset as input and provides the dataset with outliers removed as the output. We have also used the concept of weighted holoentropy. The proposed method removes outliers more effectively and efficiently than other existing methods.

***Keywords*— Outlier detection, entropy, total correlation, holoentropy, outlier factor**

I. INTRODUCTION

Outlier detection is the problem of finding objects in data that do not conform to expected behaviour. Such objects are called outliers. There are number of outlier detection techniques that have been developed specifically for some application areas [2]. Outliers arise due to faults in mechanical systems, system's behavioural changes, mankind errors, fraudulent nature, and instrumental errors [2]. Such outliers detection are very necessary as these objects may carry information which may be useful in various real time applications like medical, industrial, e-commerce security, public safety etc. [1]. According to the availability of labels in training of datasets, the following methods are used in outlier detection:

1. Supervised
2. Semi-supervised, and
3. Unsupervised approaches.

In supervised outlier detection approach labelling of datasets is required. This method requires labelling of anomalies as well as label for normal objects of the datasets. Semi-supervised approach also requires labelling of datasets. In this approach training datasets have labels for only normal objects is required. On the other hand, unsupervised approach do not require any labelling for the datasets. The models within the supervised or the semi-supervised approaches all need to be trained before use, while models adopting the unsupervised approach do not include the training phase [2].

Supervised outlier detection approach learns the classifier and assign appropriate labels to test objects using labelled objects belonging to the normal and outlier classes. The semi-supervised anomaly detection approach primarily learns normal behaviour from a given training data set of normal objects, and then calculates the likelihood of a test objects. Since this approach do not require labels for the anomaly class, they are more widely applicable than supervised approach.

In unsupervised approach for outlier detection labelling for the dataset is not required, and thus are most widely applicable approach. It works under the assumption that the majority of the objects in the data set are normal. To make use of supervised and semi-supervised approach labelling of the dataset is first required which could to be a tedious and time consuming for large or high dimensional datasets. So, we will mainly concentrate on the unsupervised approach in this paper.

This paper is organized as follows: section II describes the objectives and challenges. Section III deals with the literature survey of the various existing methods with limitations. This section also involves the comparisons among these existing methods. Section IV describes the proposed system. Section V gives conclusion and Section VI consists of references.

II. CHALLENGES AND OBJECTIVES

1. At an abstract level outliers are patterns that does not conform to expected normal behavior .Therefore we define region representing normal behavior and declare any observation which does not belong to normal region as an anomaly but several factors make this simple approach very challenging.
2. To define boundary between normal and anomalous behavior is often not precise, Adaption in anomalous observations to appear like normal, evolution in normal behavior, difference in exact notion of outliers for different application domains. This makes outlier detection problem more complex.
3. Existing unsupervised method are applicable on numerical data sets, however they cannot be adapted to deal with categorical data.
4. Using formal definition of outlier our aim is to develop effective and efficient method that can be used to detect outliers in large scale categorical unsupervised data sets for real applications.
5. Our main aim is to develop an efficient and effective method that do not require any user-defined parameters for outlier detection can detect outliers in large categorical datasets.
6. We have combined entropy and dual correlation with attribute weighting resulting into weighted holoentropy where entropy computes uncertainty and dual correlation measures mutual information or attribute relation.

III.LITERATURE SURVEY

Unsupervised outlier detection does not require labelling information about the datasets. This method detects anomalies in an unlabelled dataset with an assumption that the majority of the objects in the data set are normal. This method of outlier detection does not require labelling of objects, and hence, is widely used.

Methods that are designed for unsupervised outlier detection in categorical data can be grouped up into four categories as follows:

1. Proximity based method
2. Rule based method
3. Information Theoretic based method
4. Other methods

1. Proximity Based Methods

Proximity based method measures the nearness of the objects with respect to distance, density. An object is an outlier if the nearest neighbour of the object are far away i.e. the proximity of the object deviates from the other objects in the same datasets.ORCA and CNB are algorithms for outlier detection in categorical data that are proximity based. These two algorithms measures the distance between the objects inorder to calom the given calculate an outlier from the given datasets.

Disadvantages:

1. Do suitable for high dimensional datasets.
2. User parameters are required.
3. High time and space complexity.

2. Rule Based Methods

Rule based methods use the concept of frequent items from association rule mining. If an object has infrequent pattern from that of the other data object in the given datasets, then such object are considered as an outlier. Frequent pattern based outlier factor(FPBO) and Otey's algorithm are two that uses ruled based techniques. FPBO generalizes many concepts from distribution-based approach and enjoys better computational complexity. In this algorithm, outlier factor of each point is computed by summing the distance from its k nearest neighbours. The data object that is having the largest value is considered to be an outlier.

Disadvantages:

1. Limited for low-dimensional datasets.
2. User parameters are required.
3. Pattern is needed to be assumed.

3. Other methods

Several other approaches using the Random Walk, Hypergraph theory, or clustering methods have been proposed to deal with the problem of outlier detection in categorical data. In the random-walk-based method, outliers are those objects with a low probability of jumping to neighbours. There are various clustering algorithms that are used in the outlier detection. These algorithms differentiate the isolated objects and the clusters of the data objects. In these methods, we first find clusters and then the data objects that do not belong to that clusters are considered to be an outlier.

Disadvantages:
1. Inefficient for large datasets.
2. Clustering process is costly since first clustering is to be carried out.

4. Information Theoretic based methods

Several methods have been proposed for outlier detection using information theoretic measures. Anomaly detection in audit data sets presents information theoretic measures like entropy, conditional entropy, relative entropy & information gain to identify outliers in the univariate audit data set. ITB-SP algorithm is an information theoretic based algorithm that can be used for outlier detection in large scale categorical data. This algorithm computes holoentropy which is sum of entropy and total correlation. The outlier factor is calculated which gives the final result in the form to two sets- anomalous set and normal datasets. This method gives the optimal solution in the measurement of the outliers by using ITB-SP algorithm.

Advantages:
1. Suitable for high dimensional datasets.
2. No user parameters are required.
3. Effective and efficient.

Parameter	CNB	ORCA	FPOF	ITB-SP
Approach	Proximity Based	Proximity Based	Ruled Based	Information-theoretic based
Method	Distance	Distance	Item set frequency	High dimensional categorical data
Input Data Set	Low dimensional categorical data	High dimensional categorical data	Low dimensional categorical data	High dimensional categorical data
Required Parameters	M, sim, k	K, M	Minfreq, maxlen, M	Number of outliers o
Output Data Set	Outliers	O- outliers	Value of FPOF, FP-outliers	OS- outlier set
Complexity	O(n^2(k+S(θ)+q) + n(k + M))	O(n^2q)	O(n(2^{T-f}))	O (nm)

Table 1: Comparison of systems

The Table 1, shows the comparison between the various outlier detection methods with the help of parameters like approach, method used, input and output data set, required parameters and complexity.

IV. PROPOSED SYSTEM

In this section we are proposing a new concept of weighted holoentropy which captures the distribution and correlation information of a data set. We are also going to estimate an Upper bound of outliers to reduce the search space.

Measurement of outlier detection

1. Entropy
2. Total correlation
3. Holoentropy
4. Weighted holoentropy
5. Outlier factor
6. Upper Bound on outliers

Entropy:

Entropy is a measure of information and uncertainty of a random variable; if the value of an attribute is not known, then the entropy of this attribute indicates how much information is needed to predict the correct value in a particular dataset. The entropy of the dataset decreases with the removal of the outlier objects from the dataset.
Let X be a dataset containing n objects {x1,x2,……,xn}.
Each xi where $1 \leq i \leq n$ is a vector for categorical attributes {y1,y2,……,ym} where m is the number of attributes of the dataset X.
The random vector [y1,y2,….,ym] is represented by Y.
The Entropy that is to be computed on the dataset X is represented as $Hx()$.
Using the chain rule entropy, the entropy of Y which is denoted as $\mathrm{Hx(Y)}$ can be calculated as follows:

$$Hx(Y) = \sum_{i=1}^{m} Hx(yi|yi-1, \dots\dots\dots, y1)$$

$$= Hx(y1) + Hx(y2|y1) + \cdots\dots\dots\dots\dots, Hx(ym|ym-1, \dots.y1)$$

Where,

$$Hx(ym|ym-1, \dots\dots\dots, y1) = \sum p(ym, ym-1, \dots\dots., y1)\log p(ym|ym-1, \dots\dots, y1)$$

Total Correlation:

Total Correlation is defined as the summation of mutual information of multivariate discrete random vector Y. It is denoted by $Cx(Y)$. It calculates the mutual dependence or shared information of the dataset. Larger value of $Cx(Y)$ implies small number of objects share common attribute values.
Total Correlation is based on Watanabe's proof and can be

$$Cx(Y) = \sum_{i=1}^{m} Hx(yi) - Hx(Y)$$

Holoentropy:
Entropy measurement for detecting outliers is not sufficient. The total correlation is also necessary to get better outlier candidates. Likewise, contribution of holoentropy along with entropy and total correlation helps in giving appropriate results for outlier detection.
The holoentropy is defined as the sum of entropy and total correlation of random vector Y.
It is denoted as $HLx(Y)$ and is expressed as follows:

$$HLx(Y) = Hx(Y) + Cx(Y)$$

Attribute Weighting:
Different attributes contributes differently to the overall structure of the datasets. A reverse sigmoid function of the entropy is used to weight the entropy of each attribute and is given as follows:

$$wx(yi) = 2\left(1 - \frac{1}{1 + \exp(-Hx(yi))}\right)$$

Weighted Holoentropy:
The Weighted holoentropy is the sum of the weighted entropy of each attribute of the random vector Y.
It is denoted by $Wx(Y)$ and is expressed as follows:

$$Wx(Y) = \sum_{i=1}^{m} wx(yi)Hx(yi)$$

Formal Definition of Outlier Detection:
Given a dataset X having n objects and the number of outliers desired are o, a subset $Out(o)$ is defined as the set of outliers if it minimizes, the weighted holoentropy of the dataset X with the o objects removed $Jx(Y, o)$

$$Jx(Y, o) = Wx\backslash Set(o)(Y)$$

Where $Set(o)$ is any subset of o objects from X.

Differential holoentropy:
Given an object xo of X, the difference of weighted holoentropy $hx(xo)$ between the X and the dataset X/{xo} is defined as the differential holoentropy of the object xo.

$$hx(xo) = Wx(Y) - Wx\backslash\{xo\}(Y)$$

$$= \sum_{i=1}^{m}[wx(yi)Hx(yi) - wx\backslash\{xo\}(yi)Hx\backslash\{xo\}(yi)]$$

Outlier Factor:
It measures the likeliness of xo to be an outlier. The outlier factor of the object xo is denoted by $OF(xo)$.

$$OF(xo) = \sum_{i=1}^{m} OF(xo,i)$$

Upper Bound on Outliers:
The dataset can be divided into normal set and anomaly dataset.

The objects with the negative $\hat{h}(xi)$ are defined as the objects of normal set.

$$NS = \{xi|h(xi) \leq 0|\}$$

The objects with then positive $\hat{h}(xi)$ are defined as the objects of normal set.

$$AS = \{xi|h(xi) > 0|\}$$

The number of objects in AS is defined as UO

$$UO = N(AS) = \sum_{i=1}^{m} (h(xi) > 0)$$

Proposed Approach
Our proposed approach is based on the weighted holoentropy and differential holoentropy for outlier detection from large categorical dataset. The proposed system takes .csv file as the input data set file and will return a file with the outliers removed as an output.
System Architecture
A proposed system architecture with holoentropy is considered as shown in Fig.1

1. GUI Handler
It provides the following functionality:
- File selector (CSV File)
- Display for Attributes
- Display for Outliers

2. File Processor
It support following tasks:
- Separate objects and attributes.
- Saving outlier results.

3. Outlier Detector
It calculates the following:
- Entropy
- Total Correlation
- Weighted Holoentropy
- Outlier factor
- Outlier set
- Getting the result as a data set file with removal of attributes

4. Report generator
- Generate resultant report
- Generate graph

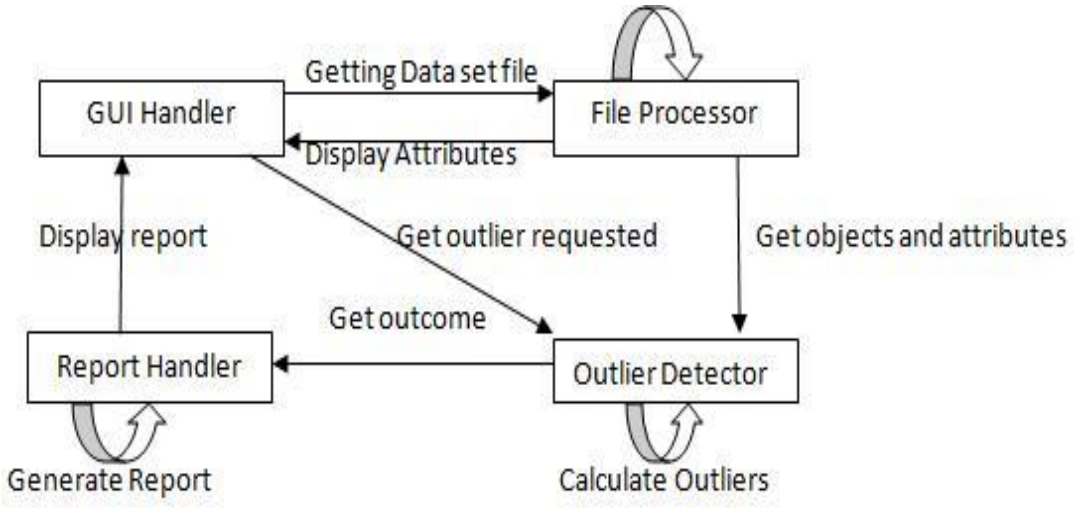

Fig 1: System Architecture

ITB - SP ALGORITHM

In this algorithm, the outlier factor for each object is computed only once. The object with the highest outlier factor value is identified as an outlier.

Algorithm: ITB-SP single pass

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 heap
12. end if

In ITB-SP, the attribute weights wx(yi)($1 \leq i \leq m$) , the outlier Factor OF(xi) for all objects, initialization of AS and the heap sort search to find outlier candidates are to be calculated. The time complexity of ITB-SP O (nm). The upper bound on outliers (UO) is to evaluate an upper limit on the number of outliers in a data set.

V. CONCLUSIONS

This paper formulates outlier detection problem as an optimization problem and proposed a practical, unsupervised, parameter less algorithm for the detection of outliers in large categorical datasets. The proposed method discussed in this paper will overcome limitations

of the previous methods used for outlier detection. Effectiveness of our approach results from a new concept of weighted attributes and holoentropy that considers both the data distribution and attribute correlation to measure the similarity of outlier candidates in data sets and the efficiency results from the outlier factor function derived from the holoentropy. An upper bound for the number of outliers is also estimated in this paper, which allows us to reduce the search cost.

REFERENCES

[1] Shu Wu, Member IEEE, and Shengrui Wang,Member IEEE, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data," IEEE transactions on knowledge and data engineering, vol. 25, no. 3, march 2013

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009

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

[4] Ramalan Kani K, Ms.N.Radhika "Mining of outlier detection in large categorical datasets" Ramalan Kani K et al, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 3, March- 2014

[5] M. O. Mansur, Mohd. Noor Md. Sap, Faculty of Computer Science and Information Systems Universiti Teknologi Malaysia "Outlier Detection Technique in Data Mining: A Research Perspective"

[6] Ashwini .G. Sagade, Student IOKCOE, Pune and Ritesh Thakur, H.O.D. (Computer), IOKCOE, Pune "Excess Entropy Based Outlier Detection In Categorical Data Set" Proceedings of 9th IRF International Conference, Pune, India, 18th May. 2014

