www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 11 November, 2014 Page No. 9178-9182

Information-Theoretic Outlier Detection For Large_Scale Categorical Data

Jesica Fernandes[1], Srijoni Saha[2], Jasmine Faujdar[3], Prof. Nitin Shivale[4]
1. B.E, Department of computer. JSPM's BSIOTR, Pune (India)
2. B.E, Department of computer. JSPM's BSIOTR, Pune (India)
3. B.E, Department of computer. JSPM's BSIOTR, Pune (India)
4. B.E, Department of computer. JSPM's BSIOTR, Pune (India)

jesica.r.fernandes@gmail.com
srijonisaha@gmail.com
Jasmineaujdar74@gmail.com
Nitinrajini3@gmail.com

ABSTRACT

Outlier detection is an important problem that has been reached within various research and applications domains in today's world. It aims to detect the object that are considerably distinct, exceptional and inconsistent the majority data in input data sets. Many outlier detection techniques 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. Many outlier detection methods 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 practical 1-parameter outlier detection methods, named ITB-SS and ITB-SP, which require no user-defined parameters for deciding whether an object is an outlier or not. Users need only provide the number of outliers they want to detect in different data set. Experimental results show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail to work .Outlier detection in many times known as anomaly detection in advanced technology for a wide range of real time applications like medical, industrial, e-commerce ,security and engineering purpose. Outlier arises due to faults in systems, changes in the system, human errors, behavioral and instrumental errors. Detection of these outliers helps in identification of frauds and faults before they arises and affect our system intensively with outcomes. The data sets like transaction data, financial records in commercial bank, demographic data are present in non-numerical attributes known as categorical data. Existing unsupervised method are applicable on numerical data sets. However they do not work with categorical type data.

.

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

1. INTRODUCTION

OUTLIER detection, which is an active research area [1],[2], refers to the problem of finding objects in a data set that do not conform to well-defined patterns of expected behaviour in a data set. These objects detected are called outliers, also referred to as, aberrant anomalies, surprises, and so on. Outlier detection can be implemented as a pre-processing step prior to the application of an advanced data analysis method. Data mining involves extraction of valid, previously unknown information from large datasets and is used for organizational decision making. There are several problems associated with the data mining for large datasets such as data redundancy, incomplete data, and outliers etc .Outliers are the objects in the datasets that are exceptional observations than those of others in the datasets. The detection of such outliers are quite 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. These outliers are also referred as anomalies, surprises, aberrant etc. The outlier detection is helpful in identification of frauds and faults that may affect our system Outlier may occur due to faults in system, changes in the system, human errors, behavioural and instrumental errors. This paper discusses the various approaches used for outlier detection. This paper proposes an idea for using information theoretic-based way for large scale categorical datasets using ITB-SP.

Depending upon the labels in the datasets, methods for outlier detection can be categorized as supervised approach, semi-supervised approach and unsupervised approach. Supervised outlier detection requires labelled datasets. The models that are adopting this approach deals with trained datasets. This method includes labelling of anomalies as well as label for normal objects. Semi-supervised approach also works on labelled datasets. In this approach training datasets have labels for only normal objects.

To implement supervised and semi-supervised methods first labelling of datasets is to be done. Since, a supervised and semi-supervised approach requires labelling of datasets which is not suitable for large-scale datasets. Detecting of normal and abnormal objects from large datasets consisting of high-dimensional objects and low anomalous instances would be quite time-consuming and tedious work. Due to these limitations, unsupervised outlier detection is preferred.

2. Challenges and Objectives

Outliers at an abstract level can be a pattern that does not conform to expected normal behaviour. Thus we define region representing normal with computation of infrequent items from data set. Based on this, outlier factor is calculated. Objects with largest scores are treated as outliers.

Based on this concept, we build a formal model of outlier detection and propose a criterion for estimating the “goodness” of a subset of objects as potential outlier candidates. After that outlier detection is formulated as an optimization problem involving searching for the optimal subset in terms of “goodness” and number of outlier candidates. Lastly to solve the optimization problem, we will carry out a deep investigation of the analytical and statistical properties of the proposed criterion and propose two greedy algorithms that effectively bypass probability estimation and the high complexity of exploring the whole outlier candidate space.

3. Literature survey for unsupervised outlier detection

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. Since this approach does not require labelling of objects, it is widely applicable. There are several existing methods of unsupervised outlier detection methods such as LOF, LOCI that are suitable for numerical datasets and adapting such methods for categorical data is not easy.

Outlier detection can be classified by the way outliers are measured with respect to other objects. There are various algorithms that have been designed for large set of categorical data which can be grouped into the following categories :

3.1 Proximity-Based Methods:

This method carries out outlier detection by measuring the nearness between the objects with respect to distance, density etc. For example, LOF [3] [5] [6] is a technique for numerical outlier detection which assigns a degree called local outlier factor of an object. This degree assigned depends on how isolated the object is with respect to the neighbouring Minpts objects. If the LOF value is high then, outliers are data objects with high LOF values whereas data objects with low LOF values are normal objects with respect to their neighbourhood. High LOF indicates of low-density neighbourhood and hence high potential of being outlier. Outlier detection in the LOF algorithm depends on the value of Minimum points. If the value of Minpts is large then LOF has to be computed for each and every object before the outliers are detected. This is not a desirable since outliers only contribute a small fraction of the entire dataset. Data in high-dimensional spaces are sparse thus making this method unsuitable. This method is time and space consuming and is thus not appropriate for large scale datasets.

3.2 Rule-based methods:

Rule-based method brought into consideration the concept of frequent items. In such method, objects with few frequent items or many frequent items are considered as outliers than others in the datasets. Frequent Pattern Outlier Factor (FPOF) [8] and Otey’s Algorithm are the famous rule-based techniques. FPOF algorithm involves initially the computation of frequent patterns by using the predefined minimum support rate. All support rates of associated frequent patterns for each object is summed up as the outlier factor of this object. The object then with the smallest outlier factors are considered as outliers.

Unlike FPOF, Otey’s Algorithm considers the infrequent items from the datasets. The object with largest outlier factor is referred to as an outlier. In this approach, time complexity depends on the generation of frequent and infrequent items. As in FPOF, the time complexity increases with the number of attributes. This approach is convenient for low-dimensional datasets.

3.3 Information-theoretic methods:

Generally, information-theoretic method focuses either on the single entropy or on mutual information. It requires expensive estimation of the joint probability distribution when the data set is shrunk for elimination of certain outliers. [9]The two algorithms for outlier detection of Information-Theory- Based approach are namely- ITB-SS for Information-Theory- Based Step-by-Step (or SS for short) and the other ITB-SP for Information-Theory-Based Single-Pass (or SP for short). Both these algorithms detect outliers one by one.

4 4. Proposed System:

This paper proposes an idea for using information theoretic-based way for large scale categorical datasets using ITB-SP. To overcome and address the problem discussed in need of effective outlier detection in unsupervised data set. A proposed methodology with excess entropy and to deal with large scale categorical data is considered as shown in Figure. 1 given below.

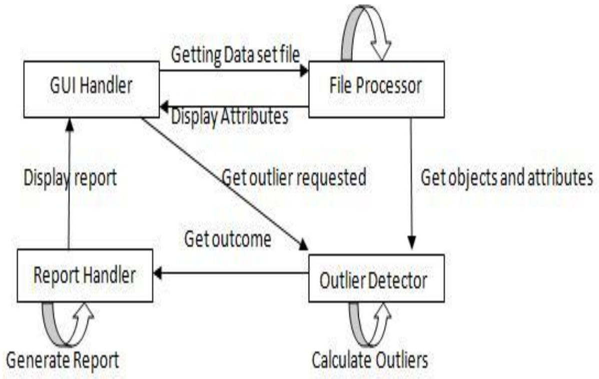


Fig.1: Proposed system

1) GUI Handler:

It provides following functionality:
File selector (CSV File)
Display for Attributes
Display for Outliers (Outcome)

2) File Processor:

It will handle following tasks:
Separate objects and attributes.
Saving outlier results.

3) Outlier Detector:

It will handle following tasks:
Calculate Entropy
Calculate Dual total Correlation
Calculate weighted excess entropy
Calculate Outlier factor
Getting outlier set
Getting data set file with removal of attributes

4) Report generator:

Generate Report
Generate comparison model using graphs

4.1 ITB-SS:

In ITB-SS, outlier factor is computed for each object. At each step of SS, the object with the largest OF(x) is referred as an outlier and is thus removed from the data set. With this removal, the outlier factor OF(x) is updated for all the remaining objects. The process continues to repeat until o objects have been removed. For ITB-SS consisting of the attribute weighting, initialised outlier factors including AS initialization, and the step-by-step top-o outlier selection procedure are calculated. The search for the outliers is conducted within anomalous set for the datasets. The time complexity of ITB-SS is comparatively higher than that of ITB-SP.

4.2 ITB-SP:

ITB-SP, the outlier factor for each object is computed only once. The object with the highest outlier factor value is identified as an outlier.
In ITB-SP, 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≤i≤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
```

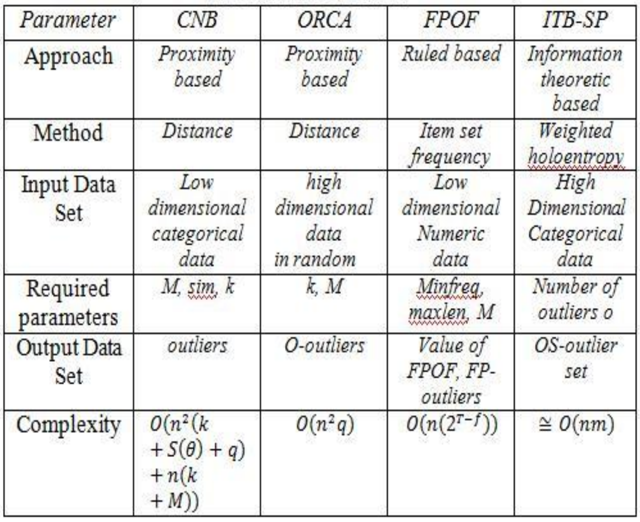
In ITB-SP, the attribute weights wx (yi)(1≤ i≤m) , the outlier Factor OF(xi) of all the objects, initialization of AS and the heap sort search to find the top-o outlier candidates are to be calculated. The time complexity of ITB-SP O (nm). The upper bound on outliers (UO) is to evaluate an upper limit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

5. Other methods

Some methods are implemented using several approaches like Random walk, Hyper-graph theory. [4].In random walk method[10]objects those have low probability of combining with neighbors are outliers. That means they remain in their state. In method [11] relationships are considered and mutual dependence based local outlier factor is proposed to detect outliers. There are many other methods cluster based local outlier detection method, classification based method. In literature several methods have been proposed for outlier detection using information theoretic measures.

1. Anomaly detection in audit data sets [12], [13] presents information theoretic measures like entropy, conditional entropy, relative entropy & information gain to identify outliers in the univariate audit data set. Where, regularity is characterized but not the attribute relation.
2. Information theoretic outlier detection in large scale categorical data, [4] this paper computes holoentropy –sum of entropy and total correlation. It gives optimal solution to outlier detection by using ITB-SP algorithm

TABLE I
COMPARISON OF SYSTEMS



Parameter	CNB	ORCA	FPOF	ITB-SP
Approach	Proximity based	Proximity based	Ruled based	Information theoretic based
Method	Distance	Distance	Item set frequency	Weighted holoentropy
Input Data Set	Low dimensional categorical data	high dimensional data in random	Low dimensional Numeric 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	$O(n^2(k + S(\theta) + q) + n(k + M))$	$O(n^2 q)$	$O(n(2^{T-f}))$	$\cong O(nm)$

6. Conclusion

The formulated outlier detection is an optimization problem proposed 2 practical, unsupervised, algorithms for detecting outliers in mild categorical data sets. Our algorithms can effectively result from a new concept of weighted attributes and holoentrophy that considers both the data distribution and attribute correlation to measure the similarity of outlier candidates in data sets. The efficiency of our algorithms results from the outlier factor function derived from the holoentropy. The outlier factor of an object is determined by the object and updating it does not require estimating the data distribution in the sets. Based on this property, apply the greedy approach to develop 2 efficient algorithms, ITB-SS and ITB-SP which provide practical solutions to the optimization problem in outlier detection. We estimate an upper bound for the number of outliers and an anomaly set of candidate set. In this bound can be obtained under a very reasonable hypothesis on the number of possible outliers, which 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.. Moreover our implementations 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.

7. References

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

[2] V.J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Rev., vol. 22, no. 2, pp. 85126, 2004

[3] 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.

[4] 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

[5] 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

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

[2] V.J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Rev., vol. 22, no. 2, pp. 85126, 2004

[3] 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.

[4] 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

[5] 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

[6] 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.

[7] S.D. Bay and M. Schwabacher, "Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), 2003

[8] 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.

[9] Z. He, X. Xu, Z.J. Huang, and S. Deng, "FP-Outlier: Frequent Pattern Based Outlier Detection," Computer Science and Information Systems, vol. 2, pp. 103-118, 2005.

[10] 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.

[11] H.D.K. Moonesignhe and P. Tan, "Outlier Detection Using Random Walks," Proc. IEEE 18th Int'l Conf. Tools with Artificial Intelligence (ICTAI '06), 2006.

[12] J.X. Yu, W. Qian, H. Lu, and A. Zhou, "Finding Centric Local Outliers in Categorical/Numerical Spaces, Knowledge and Information Systems, vol. 9, no. 3, pp 309-338, 2006.

[13] Nicholas Timme,_ Wesley Alford, Benjamin Flecker, and John M. Beggs," Multivariate information measures: an

experimentalist’s perspective”, 28 Nov 2011.