International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Information Theoretic Outlier Detection for Categorical Data

Nilam S. Khairnar[1], Prof D.B.Kshirsagar[2]
[1]*Department of Computer, PG Student,SRES COE,Kopargaon,423601,India*
[2]*Department of Computer Engg, Faculty of SRES COE,Kopargaon,423601,India*

Abstract-OUTLIER detection means the problem of finding objects in a data set that do not satisfy well-defined criteria of expected behavior. It can be implemented as a preprocessing step earlier to the application of an advanced data analysis method. It can also be used as an effective tool to discover interest patterns such as the expense behavior of to be bankrupt credit cardholder. Outlier detection is a crucial step in a variety of practical applications including intrusion detection, health system monitoring, and criminal activity detection in E-commerce, and can also be used in scientific research for data analysis and knowledge discovery in biology, chemistry, astronomy, oceanography, and other fields.Outlier detection can be done for Numerical as well as Categorical data. Most existing methods are designed for Numerical data but some or many of them encounter problem with real life applications that contain categorical data. Discovering rare events from categorical data is very vital in data mining because of the difficulty of defining a meaningful similarity measure. Also the common difficulty with the existing methods is the lack of a formal definition for the outlier detection problem. Without a formal definition, it is often designed as an ad-hoc process. In general, several user-defined parameters are often needed to define whether an object possesses sufficiently different properties, which separates them from others to be qualified as an outlier. The algorithms or methods which are taking input parameters from users are heavily dependent on suitable parameter settings for their results, which are very difficult to estimate without background knowledge about the data. Also many existing methods suffer from low effectiveness and low efficiency due to high dimensionality and large size of the data set, more complex statistical tests, or inefficient proximity-based measures. Here we are giving formal definition of outliers and an optimization model of outlier detection, via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model, we have defined a function for the outlier factor of an object which is solely determined by the object itself and is updated efficiently. We are implementing 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. Users need to only provide the number of outliers they want to detect. Experimental results show that ITB-SS and ITB-SP are more effective and efficient algorithms for large scale categorical data.

Keywords: Holoentropy, Entropy, Data Mining, Categorical Data, Outlier Factor.

I. INTRODUCTION

Anomalies might be induced in the data for a variety of reasons, such as malicious activity, for example, credit card fraud, terrorist activity, cyber-intrusion, or breakdown of a system, but all of the reasons have the common characteristic that they are interesting to the analyst. The real life relevance of anomalies is a key feature of anomaly detection. Traditional Data Mining task aims to find the general pattern applicable to the majority of data. Outlier Detection can be considered beneficial in both the ways like It targets the finding of rare data whose behavior is very exceptional when compared with rest large amount of data, so studying extra ordinary behavior of them helps

uncovering the valuable knowledge hidden behind them and can be used for number of practical applications like credit card fraud detection, Marketing and Weather prediction.
The main objectives of this paper are
1) To build a formal model of outlier detection. For this a new concept of weighted holoentropy is used. It captures the distribution as well as correlation information of a data set.
2) To implement ITB-SS and ITB-SP algorithms, which effectively bypass probability estimation.
3) To solve the optimization problem. For this we 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. We also estimate an upper bound of outliers to reduce the search space.

II. LITERATURE SURVEY

V. Chandola, A. Banerjee, and V. Kumar, discussed different ways in which the problem of anomaly detection has been formulated in the literature of [1] and they provided an overview of the huge literature on various techniques. For each category of anomaly detection techniques, they indented a unique assumption regarding the notion of normal and anomalous data. These assumptions can be used as guidelines to assess the effectiveness of the technique in that domain.
There is no single universally applicable or generic outlier detection approach. But from the previous descriptions, authors V.J. Hodge and J. Austin have applied a wide variety of techniques covering the full gamut of statistical, neural and machine learning techniques. They tried to provide a broad sample of current techniques and a feel of the diversity and multiplicity of techniques available in their survey [2].
The work of K. Das and J. Schneider [5] focuses on finding single records that are anomalous. S. Srinivasan, supported the concept that measure Multivariate Mutual Information can be used very effectively and systematically in analyzing discrete experimental data, while the traditional techniques of the analysis of variance cannot be employed because no numerical values are associated with the events in [6] also stated that contrary to the bivariate case, multivariate information can be negative but is much stronger when used for error analysis. In order to study all possible source interactions, total correlation was shown to be an important, yet a tractable measure.
In [7], W. Lee and D. Xiang, proposed to use some information theoretic measures for anomaly detection. Entropy can be used to measure the regularity of an audit dataset of unordered records. Conditional entropy can be used to measure the regularity on sequential dependencies of an audit dataset of ordered records. Relative (conditional) entropy can be used to measure the similarity between the regularity measures of two datasets. Information gain of a feature describes its power in classifying data items. Information cost measures the computational cost of processing audit data by an anomaly detection model. They discussed that these measures can be used to guide the model building process and to explain the performance of the model. In the case studies on send mail system call data, they showed that we can use conditional entropy to determine the appropriate sequence length for accuracy only or for the trade-o between accuracy and cost, a problem that has been posed but not solved by the community. They also showed that when relative conditional entropy is low, the detection performance on the testing dataset is comparable to that on the training dataset. In the case study on network data, they showed that entropy can be used to direct the partitioning of a dataset and build better models. They also showed evidence that conditional entropy can be used to guide the constructionof temporal and statistical features.

III. IMPLEMENTATION DETAILS

3.1 Entropy and Holoentropy

Real or synthetic data set is taken as an input, The Algorithms ask user for no of outliers they want to find out from that dataset. Both the algorithms calculate Entropy for every object in a dataset by using formula (1), But in some cases, entropy alone as a measure does not give the accurate results. So the idea is to combine or sum up entropy with total correlation which gives new measure Holoentropy. The holoentropy HLx(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.

$$HX(Y) = -\sum_{x\epsilon X} P(Y) log2P(Y) \quad (1)$$

$$HLX(Y) = HX(Y) + CX(Y) = \sum_{i=1}^{m} HX(yi) \quad (2)$$

3.2 Weighted Holoentropy

Objects with lower HLX values are good candidate to be an outlier. In Holoentropy, we just sum up entropies of all attributes considering all attributes with equal weight as 1. But in real applications some attributes often contribute differently to form the overall structure of dataset. So we assign weight to the attributes using formula (3) and calculate weighted holoentropy using formula (4).

$$wx(yi) = 2\left(1 - \frac{1}{1 + \exp(-Hx(yi))}\right) \quad (3)$$

$$Wx(Y) = \sum_{i=1}^{m} wx(yi)Hx(yi) \quad (4)$$

We can show that weighted holoentropy helps us a lot to give more accurate anamoly candidates.

3.3 Differential Holoentropy

In both the algorithms, user will give the dataset and no of outliers he wants to remove from that dataset. In general suppose dataset contains 8 objects and user wants 3 outliers from them, then possible candidate sets or combinations for this are calculated by using nCr = n!/r!(n-r)! So here, possible ways are 8! / 3! (5!) = 56 which is very large no.

So strategy is to calculate differential holoentropy. We can calculate differential holoentropy by using formula given below.

$$h\hat{}X(x0) = \sum_{i=1}^{m} wx(yi)[HX(yi) - HX\{x0\}(yi)] \quad (5)$$

The objects with non positive h^(xi) are defined as elements of the normal object set (NS), And with positive h^(xi) are defined as elements of the anamoly set (AS)

$$NS = \{xi \mid h\hat{}(xi) <= 0|\}.$$

$$AS = \{xi, |h(xi) > 0|\} \quad (6)$$

The number of objects in AS is defined as UO which is the upper bound on outliers.

$$UO = N(AS) = \sum_{i=1}^{n} (h^{xi} > 0). \quad (7)$$

AS will be used as the outlier candidate set; i.e., only the UO objects from AS will be examined by our algorithms.

3.4 Outlier Factor

For every object in the Anamoly set, OF Value is calculated using formula

$$OF(x0) = \sum_{i=1}^{m} \begin{cases} 0, & if n(x0, i) = 1; \\ wx(yi).\delta[n(x0, i)], & else \end{cases} \quad (8)$$

Where, $(x0, i)$ means value appear in the ith attribute of the $x0$ object.

$n(x0,i)$means the times $(x0,i)$ appears in the ith attribute and

$$\delta[n(x0,i)]= [n(x0,i)1]* \log [n(x0,i)1] - [n(x0,i)\log [n(x0,i)] \quad (9)$$

3.5 Updating the Outlier Factor

Here, we discuss the issue of updating the outlier factor within a constant time in a step-by-step process. According to Definition of Outlier factor of an object and the definition of attribute weight, to update OF (xo), we should first update the entropy of each attribute. Since the attribute entropy is always changing when outliers are detected and removed from the data set, the direct computation of *HX\ {xo}(yi)* is very time consuming. But the unweighted differential holoentropy*HLX(Y) - HLX\ {xo} (Y)* can be deduced as follows:

$$\begin{aligned} &HLX(Y) - HLX\{x0\}(Y) \\ &= m\left[\left(\frac{a}{b} - a\right) log\ a - (b+1)logb\right] - bHLX(Y) \\ &+a \sum_{i=1}^{m} \begin{cases} 0, & if n(x0,i) = 1; \\ \delta[n(x0,i)], & else \end{cases} \end{aligned} \quad (10)$$

Based on this expression, we can obtain the simple updated form of the holoentropy*HL\ {xo}(Y)* as

$$\begin{aligned} HLX\backslash\{x0\}(Y) &= (1+b)HLX(Y) \\ &- m\left[\left(\frac{a}{b} - a\right) log\ a - (b+1)logb\right] \\ -a \sum_{i=1}^{m} &\begin{cases} 0, & if n(x0,i) = 1; \\ \delta[n(x0,i)], & else \end{cases} \end{aligned} \quad (11)$$

From this, the formula for each individual attribute entropy *HX\{xo}(yi)* is obtained

$$\begin{aligned} HX\backslash\{x0\}(yi) &= (1+b)HX(yi) \\ &- \left[\left(\frac{a}{b} - a\right) log\ a - (b+1)logb\right] \quad - a \begin{cases} 0, & else \\ \delta[n(x0,i)], & if\ n(x0,i) = 1; \end{cases} \end{aligned} \quad (12)$$

This can be efficiently implemented in a step-by-step process. After calculating the entropy by (12), we can easily compute the updated attribute weight using (3). Finally, using Definition of OF (8), the outlier factor can be efficiently updated.

IV. ITB-SP AND ITB-SS ALGORITHM

At each step of ITB-SS, the object with the largest *OF (xo)* 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. In SP, the outlier factors are computed only once, and the o objects with the largest *OF(x)* values are identified as outliers. In both algorithms, search is conducted only within the anomaly candidate set AS, although this does not make any difference for the algorithm ITB-SP since the initialization of AS requires computation of the outlier factors of all the objects. ITB-SS does benefit, however, from the reduced search space. In designing the two algorithms, we assumed that the number of requested outliers 'o' is always smaller than UO. Because, AS is indeed large enough to include all the candidate objects that can reasonably be considered as outliers.

Algorithm 1. 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 = Φ
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

Algorithm 2. ITB-SS Step-by-Step

1: Input: data set X and number of outliers requested o
2: Output: outlier set OS
3: Set OS = Φ
4: Compute wx(yi) for $(1 \leq i \leq m)$
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: for i = 1 to o do
12: Search for the object with greatest OF(xo) from AS
13: Add x0 to OS and remove it from AS
14: Update all the OF(x) of AS
15: end for
16: end if

V. RESULTS

The input data set for the system is a synthetic data set or a real dataset from UCI repository which is available at https://archive.ics.uci.edu/ml/datasets.html. We tested our results with some datasets like Zoo, Hepatitis, Diabetes, Heart, Glass and many more. We enlisted some of them below with their no of instances and attributes.

Table 1.1: Real Datasets with their instances and attributes

SR NO	Dataset	No of Objects	No of Attributes	Actual Outliers	Detected Outliers	Detection Rate	AUC
1	Zoo	101	17	64	56	87.50	0.866
2	Hepatitis	155	19	72	65	90.27	0.8691
3	Heart-h	294	14	132	120	90.90	0.8719
4	Credit-a	691	16	350	279	79.71	0.756
5	Diabetes	768	9	340	283	83.23	0.8738
6	Glass	215	10	83	77	92.77	0.8708
7	Heart-s	270	14	128	115	89.84	0.836

For ITB-SP and ITB-SS algorithms, the system calculates OF for every object. After calculation of OF, system derive Differential Holoentropy value for each object. The objects with positive differential holoentropy value are then added to Anamoly set. The count of objects in AS is nothing

but an upper bound on outliers.i.e. These are the total no of outliers or anamolies detected by our system. By comparing these objects with actual outliers we calculated detection rate of our system which can be seen from Detected outliers from Table 1.1

We are measuring performance of our system by using AUC value. The value of AUC lies between 0 to 1, and it shows that how accurate the given algorithm is. The greater the value, the outlier detection is more precise. For our example datasets we got AUC values, which are given in AUC column of table 1.1.

The Graphical representation for comparison of our results with Actual no of outliers in the respective datasets is given by following graph in fig 1.3

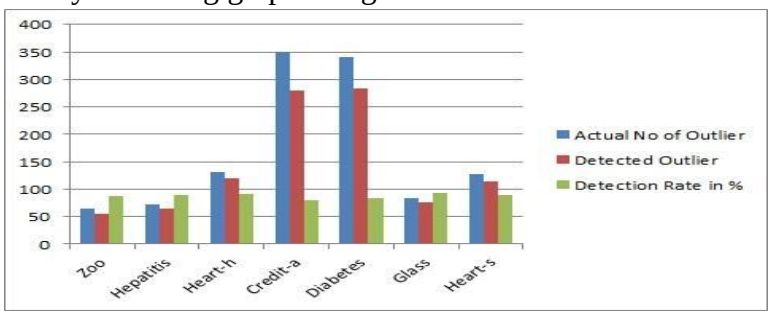

Figure 1.3: Graphical representation of Detection rate

VI. CONCLUSION

We have formulated outlier detection as an optimization problem and implemented 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. The efficiency of our algorithms results from the outlier factor function derived from the holoentropy. And is solely determined by the object and its updating does not require estimating the data distribution. Based on this property, we have developed two efficient algorithms. We have also estimated an upper bound for the number of outliers and an anomaly candidate set. This bound reduce the search cost. The algorithms are evaluated on real and synthetic data sets. As AUC values for datasets we have tested lies between 0.7 and 0.9 we can say that our system performs better for all these tested datasets.

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," Ar-ticial Intelligence Rev., vol. 22, no. 2, pp. 85- 126, 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] Z. He, X. Xu, and S. Deng, "An Optimization Model for Outlier Detection in Categorical Data," Proc. Int'l Conf. Advances in Intelligent Computing (ICIC '05), 2005.

[5] SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), 2007.

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

[7] Shu Wu, Shengrui Wang "Information-Theoretic Outlier Detection for Large-Scale Categorical Data", IEEE Transactions on Knowledge and Data Engineer-ing, vol. 25, No.3, March 2013.

