EXCESS ENTROPY BASED OUTLIER DETECTION IN CATEGORICAL DATA SET

ASHWINI G. SAGADE, RITEESH THAKUR

1Student, IOKCOE, Pune. 2H.O.D. (Computer), IOKCOE, Pune.
Email: ashwinisagade25@gmail.com

Abstract—Many outlier detection methods have been proposed because of need of finding meaningful information by removal of unwanted data based on classification, clustering, frequent patterns and statistics. Among them information theory has some different perspective while its computation is based on statistical approach. The outlier detection from unsupervised data sets is more challenging since there is no inherent measurement of distance between objects. We proposed a novel framework based on information theoretic measures for outlier detection in unsupervised data with the help of Excess Entropy. In which we are using different information theoretic measures such as entropy and dual correlation. Based on this model we proposed EEB-SP outlier detection algorithm which do not require any user defined parameter except input data set. We also used the formal definition of outliers which depends upon the weighted entropy. This algorithm detects outliers in large scale unsupervised datasets expertly than other existing methods.

Keywords— Attribute weightings, Entropy, Excess entropy, Information Theory, Mutual information, Outlier detection.

I. INTRODUCTION

Outlier detection is an important problem that has been researched within various research areas and application domains. It aims to detect objects that are considerably distinct, exceptional and inconsistent with respect to the majority data in input data sets. Many outlier detection techniques have been specifically developed for certain application domains while others are more generic. Outlier detection many times known as anomaly detection in many literature is an advance technology for wide range of real time applications in medical, industry, e-commerce, security and engineering. It can also utilize in scientific research work for analyzing data and knowledge discovery in astronomy, oceanography, chemistry, biology and other applications.

Outlier arises due to faults in mechanical systems, Changes in system’s behavior, mankind errors, fraudulent behavior, and instrument errors or simply through natural deviations in populations. Detection of these outliers helps in identification of system faults and frauds before they affect intensively with outcomes. The specific techniques or algorithms used in different outlier detection methods are varied notably which are mainly dependent upon the characteristics of data sets to be worked with. According to if we classify existing methods of outlier detection based on availability of labels (associated with data instances denote whether that instance is normal or abnormal) to the training data sets. There are three broad categories of data sets supervised, semi-supervised and unsupervised. Methods which are in supervised mode assumes that the training data has labeled instances for both normal and abnormal classes while semi-supervised training data set has labeled instances only for normal class.

Different from these two modes, an unsupervised method does not require any training data set.

These three approaches have different prerequisites, limitations and they use different kinds of data sets with different amounts of label information. All these modes are discussed in detail below.

Supervised outlier detection approach uses labeled objects belonging to the normal and outlier classes to learn the classifier and assign appropriate labels to test objects.

Semi-supervised outlier detection approach firstly learns a model denoting normal behavior from given training data set of normal objects and further calculates the likelihood of test objects.

Unsupervised outlier detection approach detects outliers in unlabeled data set. Considering that the most of the objects in data set are normal. This approach is applied to various kinds of outlier detection methods and data sets. The unsupervised approach is our focus area in this paper. To use supervised and semi-supervised approach one must first label the training data sets. But if we consider large data sets or high dimensional data then labeling will be tedious and time consuming task.

This paper is organized as follows: section II describes the some challenges and objective of work. Section III consists of literature survey of existing work with comparisons. Some background details and formulation is given in section III. Section IV describes proposed system with mathematical models. Results are shown in section V. Section VI describes conclusion and finally future work.

A. Challenges and Objectives:

1) At an abstract level outliers can be a pattern 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) The data sets like transaction data, financial records in commercial banks, demographic data are present in non-numerical attributes known as categorical attributes. 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) In categorical data, attributes can be examined by data distribution, attribute correlation, between-object-similarity or local density/distance. We studied information theoretic measures to derive some new methods. We have combined entropy and dual correlation with attribute weighting resulting into weighted excess entropy where entropy computes correlation with attribute weighting resulting into mutual information or attribute relation.

6) Number of user defined intrinsic and decision parameters are required for outlier detection problem. Thus results are based on correct values of these parameters.

II. LITERATURE SURVEY

Methods designed for unsupervised outlier detection in categorical data can be grouped into four categories as follows.

B. Proximity based methods:
To understand this concept easily, it is the method which measures compactness of objects in terms of distance and density. ORCA and CNB are algorithms for outlier detection in categorical data. ORCA uses hamming distance and CNB uses common neighbor set. These methods are not useful for high dimensional data because of problem of choosing the measurement of distance or density and high time and space complexity.

C. Ruled based methods:
Ruled based methods use the concept of frequent items from association rule mining. Frequent pattern outlier factor and Otey’s algorithm are two well known ruled based techniques. FPOF algorithm initially computes set of frequent patterns with minimum support rate for all objects. The outlier factor is calculated as sum of rates of supports associated frequent patterns and we can get outliers with smallest factors. While Otey’s algorithm, begins with computation of infrequent items from data set. Based on this, outlier factor is calculated. Objects with largest scores are treated as outliers.

D. Other methods:
Some methods are implemented using several approaches like Random walk, Hyper-graph theory. In random walk method objects those have low probability of combining with neighbors are outliers. That means they remain in their state. In method 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>CNB</th>
<th>ORCA</th>
<th>FPOF</th>
<th>ITB-SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>Proximity based</td>
<td>Proximity based</td>
<td>Rule based</td>
<td>Information theoretic based</td>
</tr>
<tr>
<td>Method</td>
<td>Distance</td>
<td>Distance</td>
<td>Item set frequency</td>
<td>Weighted holoentropy</td>
</tr>
<tr>
<td>Input Data Set</td>
<td>Low dimensional categorical data</td>
<td>high dimensional data in random</td>
<td>Low dimensional Numeric data</td>
<td>High Dimensional Categorical Data</td>
</tr>
<tr>
<td>Required parameters</td>
<td>M, ( \sum_y )</td>
<td>( k, M )</td>
<td>Mindig, manifold, ( M )</td>
<td>Number of outliers</td>
</tr>
<tr>
<td>Output Data Set</td>
<td>outliers</td>
<td>O-outliers</td>
<td>Value of FPOF, FFP-outliers</td>
<td>0-outlier set</td>
</tr>
<tr>
<td>Complexity</td>
<td>( \delta(\pi^2(k + (\delta(y) + 2)) + nx + M) )</td>
<td>( \delta(\pi^2(k)) )</td>
<td>( \delta(\pi^2(\pi^2)) )</td>
<td>( \delta(\pi(n)) )</td>
</tr>
</tbody>
</table>

In literature several methods have been proposed for outlier detection using information theoretic measures.

1) 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. Where, regularity is characterized but not the attribute relation.

2) Information theoretic outlier detection in large scale categorical data, this paper computes holoentropy –sum of entropy and total correlation. It gives optimal solution to outlier detection by using ITB-SP algorithm. This method computes holoentropy as follows:

\[
H_{\alpha}(y) = H_k(y) + C_k(y)
\]

After computing holoentropy, weighted holoentropy and differential entropy is calculated using (2) and (3) to get anomaly set \( AS \) from data set.

\[
W_X(y) = \sum_{i=1}^m w_X(y_i)H_X(y_i)
\]

\[
h_X(x_\alpha) = W_X(y) - W_{X(\alpha)}(y)
\]
In Table 1, we compare different outlier detection methods using parameters like approach, type, input data set, output data set, complexity and user defined parameters.

E. Need of advance outlier detection:
Existing systems are more dependable on user defined parameters and very few methods are dealing with unsupervised categorical data. So there should be appropriate method which will able to deal large scale data without requirement of any user defined parameter. There is requirement of method which will perform outlier detection using joint correlation between attributes.

III. BACKGROUND AND PROBLEM FORMULATION

In this section we first look at how entropy and dual total correlation can be used to capture likelihood of outlier candidates. We are proposing Weighted Excess Entropy and formulate the outlier detection problem.

F. Entropy: Entropy is measure of information and uncertainty of a random variable.
Let X be the set of n objects \( \{x_1, x_2, x_3, \ldots, x_n \} \), each \( x_i \) for \( 1 \leq i \leq n \) being a vector of categorical attributes \( [y_1, y_2, y_3, \ldots, y_m] \) where \( m \) number of is attributes. Now based on chain rule of entropy [4], Entropy of \( Y \) denoted as \( H_y(y) \) can be written as follows.

\[
H_y(y) = H_y(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H_y(y_i | y_{i-1}, \ldots, y_1)
\]

Where,

\[
H_y(y_i | y_{i-1}, \ldots, y_1) = H_y(y_i) + H_y(y_2^{i-1} | y_i) + \ldots + H_y(y_m | y_{m-1}, \ldots, y_i)
\]

(4)

If significant decrease in entropy of dataset occurs due to removal of object then that is good outlier candidate.

G. Total Correlation:
It is defined as sum of mutual information of multivariate discrete random vector \( y \), denoted as \( C_y(y) \) [4]. It is based on Watanabe’s proof that total correlation can be expressed as

\[
C_y(y) = \sum_{i=1}^{n} H_y(y_i) - H_y(y)
\]

(5)

H. Dual total correlation:
The dual total correlation calculates the amount of entropy present in \( Y \) beyond the sum of the entropies for each variable conditioned upon all other variables [13]. The dual total correlation has also been referred to as the excess entropy and the binding information. In this paper we describe dual total correlation as \( E_x(Y) \) and expressed as

\[
E_x(y) = \left( \sum_{X \subseteq Y} H_{X|Y}(y) \right) - (n-1)H_x(y)
\]

Where n is number of attributes. (6)

I. Weighted Excess Entropy:
We will begin this with showing that entropy alone is not sufficient to measure for outlier detection and correlation is necessary. Looking at the example in Table 2, where 14 objects with four attributes are illustrated, we represent data set by \( X \). \( X \) includes two objects \( x_5 \) and \( x_6 \) which can be identified as the most likely outliers with comparison with the other 12 objects. Moreover, \( x_5 \) is clearly more exceptional than \( x_6 \) since it shares none of its attributes with the rest of objects. By using equation (4), (5) and (6) we calculated entropy, total correlation and excess entropy as shown in Table 3. We combine entropy with dual total correlation to get excess entropy. As we look at the calculations in Table 3, we obtain \( x_5 \) to be distinguished as a more likely outlier than \( x_6 \).

While if we combine dual total correlation and entropy we get more precise results. As entropy assigns equal importance to all attributes, whereas in practical applications, different attributes often contribute differently to form overall structure of data set.
To weight the entropy of each attribute, we are using a reverse function of the entropy, as follows:

\[ w_\alpha(y_\alpha) = 2 \left( \frac{1}{1 + \exp(-H(Y))} \right) \]  

(7)

The weighted Excess entropy is defined as follows:

**Definition 1**: The weighted excess entropy \( EW_X(Y) \) is the sum of weighted entropy on each attribute of the random vector \( Y \).

\[ EW_X(Y) = \sum_{i=1}^{m} w_\alpha(y_i) H_X(y_i) \]  

(8)

Outliers are detected by minimizing the excess entropy through the removal of outlier candidates.

**Proposed strategy**: Have weighting the entropy of each individual attribute in order to give more importance to those attributes with small entropy values.

**J. Formal definition of outlier detection**: We are using formal definition of outliers based on the weighted excess entropy, assuming that set of outlier candidates is the best if its exclusion from the original data set \( X \) causes the greatest decrease in the weighted excess entropy value.

**Definition 2**: Given data set \( X \) with \( n \) objects, a subset \( Out(o) \) is defined as the set of outliers if it minimizes the weighted excess entropy of \( X \) with \( o \) objects removed.

**K. Differential excess entropy**: **Definition 3**: Given an object \( x_o \) of \( X \), the difference of weighted excess entropy \( e_X(x_o) \) between the data set \( X \) and the data set \( X \setminus \{x_o\} \) is defined as the differential excess entropy of the object \( x_o \).

\[ e_X(x_o) = W_X(y) - W_{X\setminus\{x_o\}}(y) \]  

(9)

**L. Outlier factor**: It can be considered as a measure of how likely it is that object \( x_o \) is an outlier. An object \( x_o \) with a large outlier factor value is more likely to be an outlier than an object with a small value. The outlier factor of an object \( x_o \), denoted as \( OF(x_o) \) is defined as

\[ OF(x_o) = \sum_{i=1}^{m} OF(x_o_i) \]  

(10)

**IV. PROPOSED APPROACH**

Proposed approach is based on weighted entropy and differential entropy which can be calculated using equation (8) and (9). System will take data set file of format .CSV and gives output file with outliers removed.

**M. System Architecture**: 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 Fig. 1

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**: It will handle following tasks:
   - Generate Report
   - Generate comparison model using graphs

**N. Mathematical model**: The proposed concept is constructed on the assumption that elimination of outliers will improve the purity of data set and reduces \( EW_X(y) \). When a normal object is removed from the data set, the value
of EW(y) should increase. Thus, the objects with positive \( e(x_i) \) are defined as the outlier candidate set (OS). The objects with non-positive \( e(x_i) \) are defined as elements of normal object set (NS).

\[
\text{OS} = \{x_i | e(x_i) > 0\} \quad \text{And} \quad \text{NS} = \{x_i | e(x_i) \leq 0\}
\]

**TABLE IV:**

| Outlier Factors of different Methods on Synthetic Data Set |
|---------------|-----------------|-----------------|-----------------|
|               | EEB-SP          | ITB-SP          | CNB             |
| \( y_1 \)     | 0.44            | 0.44            | 0.44            |
| \( y_2 \)     | 0.44            | 0.44            | 0.44            |
| \( y_3 \)     | 0.44            | 0.44            | 0.44            |
| \( y_4 \)     | 0.44            | 0.44            | 0.44            |
| \( y_5 \)     | 0.44            | 0.44            | 0.44            |
| \( y_6 \)     | 0.44            | 0.44            | 0.44            |

**Algorithm: EEB-Single Pass**

1. **Input:** data set \( X \)
2. **Output:** Outlier set \( S \)
3. Compute \( \omega(y_i) \) for \( 1 \leq i \leq n \) by (7)
4. Set \( \text{OS}=\text{null} \)
5. for \( i=1 \) to \( n \) do
6. Compute \( \text{OF}(x_i) \) and obtain \( \text{OS} \) by (10)
7. **end for**
8. Build \( S \) by searching in \( \text{OS} \)

The above algorithm is greedy approach to find out outliers from input data set. It firstly computes weighted entropy for each attribute. 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, then calculates outlier factor for each attribute and get the largest OF set which will convert to OS (Outlier set). After that set \( S \) will be built. Complexity of the algorithm is \( O(nm) \), as we are not using any searching algorithm.

**O. EEB-SP Algorithm for outlier detection:**

In this section, we have derived Excess entropy based single pass greedy algorithm for outlier detection. In this algorithm outlier factors are computed only once, and the objects with largest values are identified as outliers. This algorithm is parameter-less as we do not need to provide any user defined parameters.

**Algorithm:** EEB-Single Pass
1. **Input:** data set \( X \)
2. **Output:** Outlier set \( S \)
3. Compute \( \omega(y_i) \) for \( 1 \leq i \leq n \) by (7)
4. Set \( \text{OS}=\text{null} \)
5. for \( i=1 \) to \( n \) do
6. Compute \( \text{OF}(x_i) \) and obtain \( \text{OS} \) by (10)
7. **end for**
8. Build \( S \) by searching in \( \text{OS} \)

As shown in above Table 4, OF for attributes are varied from method to method. Specifically, the column CNB shows that all objects obtain the same outlier factor value. So for CNB, all the objects are equally likely to be outliers. FIB and OA make a similar distinction between objects 5-9 and the rest of the objects. They improve on the assessment of CNB by assigning a greater likelihood of being outliers to objects 1-4 and 10-13. While ITB-SP differentiate attributes with different OF values.

It is EEB-SP that provides the most precise assessment. It indicates that object 7 in the middle of the data set is less likely to be an outlier than objects 5, 6, 8, and 9, which are similar to each other but have a common similar object 7. Moreover, objects 1, 2, 3, and 13 are less likely to be outliers than objects 8-12 and 5-10, each of which is similar to only two other objects. These differences are important indices used by EEB-SP to accurately identify the most likely outlier candidate. As outlier factor of EEB-SP gives precise value for each attribute, it can remove outliers exactly than others.

We can also deal with real data sets like Zoo, Soybean data. Labor data efficiently using our algorithm. Input data set files for our systems are in the “.CSV” format. EEB-SP algorithm detects outliers from given data set file and remove them to provide result data set.

As shown in figure 2. Firstly, all the attributes will be listed from input set file and then weight and excess entropy for each attribute is calculated. Outliers and normal attributes are shown in following Figure 3. Comparison of above results is displayed using graph as shown in figure 4.

**Fig. 2. Attribute list in Soyabean.csv file**

**V. RESULTS**

This section helps to understand why EEB-SP is effective in solving the outlier detection problem. Experiments on different synthetic data in this section can be used as evidence to illustrate the effectiveness and stability of the proposed methods for large-scale data sets. Outlier factors of different methods are compared to gain a better understanding of the advantage of the proposed methods. We are taking input data set from [15]. In above Table 4, we are dealing with synthetic data set by \( y_1 \ldots y_6 \). The 13 objects are different from each other. We are computing OF (Outlier factor) for given synthetic data set. Previous existing methods like CNB, FPOF and ITB-SP are used to compute outlier factor for above data set and compare values of OF with proposed EEB-SP.
CONCLUSION

In this paper, many outlier detection methods based on information theory are discussed. We are proposing a novel method which will overcome limitations of previous methods. We have formulated outlier detection as an optimization problem and proposed a practical, unsupervised, parameter less algorithm for detecting outliers in large-scale categorical data sets. The effectiveness of our algorithms results from a new concept of weighted excess entropy that considers both the data distribution and joint attribute correlation to measure the likelihood of outlier candidates, while the efficiency of our algorithms results from the outlier factor function derived from the differential entropy. In particular, we can say our algorithms can deal with data sets with a large number of objects and attributes. We can extend our work in data stream mining and outlier detection in continuous data flow. Therefore, excess entropy based outlier detection in data stream will be another promising research direction.

ACKNOWLEDGEMENT

I am Ashwini Sagade and would like to thank the publishers, researchers for making their resources material available. I am greatly thankful to Prof. Ritesh Thakur for their guidance.

We also thank the college authorities Prof. Sachin Malave: PG coordinator and Principal Dr. R.S. Jahugirdar for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES


