Detection of Outlier Schema for Mixed Data using ITB-SP and HilOut Algorithms

G. Ashkan Jahanban, T. Sumit Singh*
Associate Professor, AIMT Grater Noida

Abstract-- *Outliers are the data or attributes that does not belong to any group or cluster. There are basically two types of data, categorical data and numerical data. The data set that contains both these types of data are called mixed data. There are different types of algorithms to form clusters and to detect the outliers for different data set. It is easy to detect the outliers if it is either purely categorical or numerical. But it is really challenging to detect the outliers in a mixed data. In this paper, we propose the new concept to detect the outliers. For the given data set we first partition the data set into categorical and numerical data set. For the partitioned categorical data set we apply the ITB-SP algorithm to detect the outlier set. This algorithm uses the concept of holoentropy which is the summation of both entropy and total correlation. Here the outlier is detected based on the outlier factor. Then we find the outlier candidate set for the numerical data using HilOut algorithm. This result in deriving two outlier set from the categorical and the numerical data set which will be clustered together to form perfect top n outliers.*

Keywords- Holoentropy, entropy, total correlation, outlier factor, candidate set

I. INTRODUCTION

Data mining is used in many domains including finance, engineering, biomedicine and cyber security. There are two categories of data mining methods – supervised and unsupervised. Supervised data mining techniques predict a hidden function using training data. The training data have pairs of input variables and output labels or classes. Unsupervised data mining is an attempt to identify hidden pattern from the given data without introducing training data. We live in a world where vast amount of data are collected daily. Analyzing such data is an important need. Data is produced at a phenomenal rate, our ability to store data has grown. Users also expect more sophisticated information. How to uncover hidden information is Data mining. There are a number of data mining functionalities. These include characterization and discrimination; the mining of frequent patterns, associations, and correlations; classification and regression; clustering analysis; and outlier analysis. Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize properties of the data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions.

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Clustering as a data mining tool has its roots in many application areas such as biology, security, business intelligence and web search.
Basic Clustering Techniques

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods

1.1 Cluster Analysis

It is the process of partitioning a set of data objects into subsets. Each subset is a cluster. Objects in a cluster are similar to one another yet dissimilar to objects in other cluster. Different clustering methods may generate different clustering on the same data set. Partitioning is performed by clustering algorithm. As a cluster is a collection of data objects that are similar to one another within the cluster and dissimilar to objects in other clusters, a cluster of data objects can be treated as an implicit class. In this sense, clustering is sometimes called automatic classification. Again a critical difference here is that clustering can automatically find the grouping. This is a distinct advantage of cluster analysis. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection where outliers may be more interesting than common cases. Clustering is known as unsupervised learning because the class label information is not present. For this reason, clustering is a form of learning by example.

1.2 Outliers

A data object that deviates significantly from the normal object as if it were generated by a different mechanism.

Example: unusual credit card purchase.
Types of outliers: There are three kinds of outliers. They are: global, contextual and collective.

Global : object is O_g if it significantly deviates from the rest of the data set.
Contextual : object is O_c if it deviates significantly based on a selected context.
Collective : a subset of data objects collectively deviates significantly from the whole data set, even if the individual data objects may not be outliers.

1.3 Machine Learning

Machine learning investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. Machine learning is a fast-growing discipline.

1.4 Classic Problems In Machine Learning

Supervised learning is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.

Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled. Typically, we may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits.

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. In one approach, labeled examples are used to learn class model and unlabeled examples are used to refine the boundaries between classes.

II. RELATED WORKS

Mainstream methods/algorithms designed for outlier detection for categorical data can be grouped into four categories.

Proximity-Based Methods

For categorical data sets, the proximity-based methods must confront the problems of how to choose the measurement of distance of density and how to avoid high time and space complexity in the distance computing process. Proximity-based methods also suffer from the curse of dimensionality when using distance or local density measures on the full dimensions. In general, these methods are time and space consuming and consequently are not appropriate for large data sets.

Rule-Based Methods

Rule-based methods borrow the concept of frequent items from association-rule mining. For each object, all support rates of associated frequent patterns are summed up as the outlier factor of this object. The objects with the 'o' smallest factors are considered as the outliers. Based on the infrequent items, the outlier factors of the objects are computed. The objects with the 'o' largest scores are treated as outliers. The time complexity of both algorithms is determined by the frequent-item or infrequent-item generating processes.

Information-Theoretic Methods

Several information-theoretic methods have been proposed in the literature. For anomaly detection in audit data sets, a series of information-theoretic measures, i.e., entropy, conditional entropy, relative conditional entropy, and information gain, to identify outliers in the univariate audit data set, where the attribute relationship does not need to be considered. In these methods, heuristic local search is used to minimize the objective function. In general, information-theoretic methods focus either on a single entropy-like measurement or on mutual information, and require expensive estimation of the joint probability distribution when the data set is shrunk following elimination of certain outliers.

Statistical/Model-Based Approaches

Statistical outlier detection was one of the earliest approaches. Statistical-based methods assume that a parametric model, which is usually univariate, describes the distribution of the data. Multivariate statistical approaches have been proposed, including use of robust estimates of the multidimensional distribution parameters. One inherent problem of statistical-based methods is finding the suitable model for each dataset and application. Also, as data increases in dimensionality, it becomes increasingly more challenging to estimate the multidimensional distribution.

Other Methods

In few high dimensional data set the proximity based definition is used to find the outliers. In most of the methods which are used to detect the outliers there occurs

some drawbacks in their methodologies, using an outlier factor to determine its relationship with its neighboring data is acceptable. There are approaches which deal with the kernel matrix which finds the outliers from single type and mixed attribute data sets. The Otey's algorithm is based on computing the anomaly scores that take into account both types of data similar to the concept of holoentropy. The Otey's algorithm calculates the covariance matrix for the continuous values in the particular data set. The data is considered to be an outlier if it contains infrequent categorical set or if its value differ from the covariance. Based on the idea of Entropy the outlier is first detected. The technique by Otey focuses on data set with mixed attributes. An outlier detection method for categorical data called Attribute Value Frequency (AVF) uses the frequency to calculate the anomaly score for individual data.

III. PROPOSED SYSTEM

The mixed data set is a data set that contains both categorical and numerical data. A categorical variable is also known as nominal variable which has two or more categories but there is no ordering to the categories. For example, hair color is a categorical variable which may be brown, black, red etc. but there is no perfect way to order them from highest to lowest. An ordinal variable is similar to a categorical variable but there is a class ordering of the variables. For example the economic status, has three categories such as low, medium and high.

In this paper we consider a mixed data set from UCI repository, we first split the data set to categorical and numerical data set. For the categorical data we apply the concept of holoentropy.Holoentropy is the summation of both entropy and total correlation. Holoentropy assigns equal importance to all the attributes, but in actual data set each attribute has different weightage. To weight the entropy of each attribute we use reverse sigmoid function which is defined as follows:

$$w\chi(yi) = 2(1 - \frac{1}{1+\exp(-H\chi(yi))})$$

The ITB-SP (Information-Theory Based Single Pass) algorithm is used to detect the outliers. In single pass the outlier factors are computed only once and the objects with the largest outlier factors are identified as outliers. The input to this algorithm will be the data set χ and the output is the outlier set OS. The outlier set is first initialized to null then the attribute weight is calculated for each object and is compared with the upper bound value of the outliers.

For the numerical data set we use HilOut algorithm which is designed to detect the outliers in large and high dimensional data set. For an integer k, the weight of the object or a point is the sum of the distances separating it from its k nearest neighbors. The objects or the points that have the highest value of weight are determined as outliers. Since the computation of the weight of the object is expensive to calculate we consider a set of points in which this set of points represents the point's candidate to be the true outliers. During each iteration of HilOut algorithm the number of candidate objects which belongs to the result set is reduced. Therefore a set of approximate outliers that are considered to be the true outliers are derived.Now, from both the categorical and numerical data set we derived two sets of outliers which can be combined together to give the required outlier set.

CONCLUSION

In this paper we present two algorithms ITB-SP and HilOut algorithm. The ITB-SP algorithm is used for detecting the outliers from a high dimensional categorical data set. This algorithm is proved to be very effective and is scalable. The HilOut algorithm efficiently detects the outliers from numerical data set. By combining both the algorithm we result in a better performance in detecting the outliers from a mixed data set.

REFERENCES

[1] Shu Wu, Member, IEEE, and Shengrui Wang, Member, IEEE, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data," vol.25 No.3, March 2013, (references)

[2] C.C. Aggarwal and P.S. Yu, "Outlier Detection for High Dimensional Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '01),2001.

[3] M. Breunig, H-P. Kriegel, R. Ng, and J. Sander, "LOF: IdentifyingDensity-Based Local Outliers," Proc. ACM SIGMOD Int'l Conf.Management of Data (SIGMOD '00), 2000.

[4] F. Angiulli and C. Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 2, pp. 203-215, Feb. 2005.

[5] Knorr, E.,Ng, R., and Tucakov,V. 2000. Distance-based outliers: Algorithms and applications.VLDB Journal.

[6] Guha, S., Rastogi, R., and Shim, K. 2000. ROCK: A robust clustering algorithm for categorical attribute.

