International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015
ISSN 2091-2730

Analysis of outlier detection in categorical dataset

Miss. Rachana P. Jakkulwar[1], Prof.R.A.Fadnavis [2]

Department of Information Technology,YCCE,Nagpur,rachana.j8@gmail.com,9420139656

Abstract— Outlier mining is one of the most important tasks of discovering the data records which has an exceptional i.e. different behaviour comparing with rest of remaining records in the dataset. Outlier contains different behaviour from other data objects in the dataset. There are various kinds of effective approaches to detect outliers in numerical dataset. But for categorical dataset there are some few limited approaches. This paper describes about different classification and clustering algorithms. The time complexity defines the amount of time taken by an algorithm to perform the given operation on a dataset. Hybrid approach can be developed for outlier detection analysis for Categorical dataset by using NAVF (Normally distributed attribute value frequency) and Ranking algorithm. In this paper we have considered the Networking Dataset in which we will detect outlier as virus or intrusion which will be different than behaviour in normal data object.

Keywords— Outlier , categorical dataset, NAVF, ROAD, Hybrid Algorithm, Networking Dataset, intrusion.

I. INTRODUCTION

A generalization of the binary variable in that it can take on more than two states is called as categorical variable. For example, map color is a categorical variable that may have, say, three states: red, green, and blue [12].

Outlier detection is one of the most important processes of detecting instances with unusual behavior that occurs in a given system. The discovery of valuable information in the data can be made by doing effective detection of outliers. From many years, mining for outliers has received significant attention because of its wide applications in various areas such as detecting fraudulent usage of credit cards in banking sector, unauthorized access in computer networks, medical field, weather prediction and environmental monitoring [11].

Most of the existing methods are designed for detecting outliers in continuous i.e. numeric data, but the problem of outlier detection in categorical data is still evolving. The basic difficulty is defining a suitable similarity measure over the categorical values. Because most of the values that a categorical variable can assume are not inherently ordered [1].

This paper describes about various clustering and classification algorithm applied to categorical data for finding out outliers. This paper is organized as follows; section 2 gives an overview of different categorical classification clustering algorithms and its methodologies and time complexity of various categorical clustering and classification algorithms. Finally in section 3, conclusions and future work are provided.

II. EXISTING CATEGORICAL ALGORITHM

There are two main learning approaches for detection of Outliers in a categorical dataset, which are supervised and unsupervised learning approaches.

Type 1- Supervised learning determines the outliers with background knowledge of the data. This approach is analogous to supervised classification and requires pre-label data, tagged as normal or abnormal.
Type 2 – Unsupervised learning determine the outliers with no background knowledge of the data. This is essentially a learning approach analogous to unsupervised clustering [11].

A. Algorithms for classification method:

The computational complexity of classification based techniques depends on how the classification algorithm being used.

Training Phase- The complexity of training classifiers has been discussed in [14].

Testing Phase- The testing phase of classification techniques is usually very fast as compare to training phase since it uses a learnt model for classification method.

Advantages - Use of Powerful Algorithms and Fast testing Phase are two advantages of classification technique.

Disadvantages- Non-Availability of Accurate Labels for Various Normal classes and Assigning Label to Each Test Instance are two disadvantages of classification technique. [13]

1. AVF algorithm

Statistic and density based algorithms are linear with respect to data size and requires k-scans each time. When we select low threshold value to find frequent item sets from dataset then these techniques can be-come very slow [11].

Attribute Value Frequency (AVF) algorithm is simple and faster approach to detect outliers in categorical dataset which minimizes the number of scans over the data. It does not create more space and more search for combinations of attribute values or item sets. An outlier point xi is de-fined based on the AVF Score in [1].

1.1. Methodology for AVF Algorithm

Input: Database D (n points and m attributes),

Target numbers of outliers are k

Output: detected outliers k

Initially label all data points as non outliers;

Step1: Count frequency f(xij)of attribute value xij for each point xi, i = 1..n and attribute j, j = 1..m

Step2: Calculate AVF Score (xi) += f(xij) and

AVF Score (xi) /= m for each point xi and attribute j.

Step3: Return k detected outliers with mini (AVF Score); [1].

AVF algorithm requires only one scan to detect outliers. The complexity is O (n * m). It needs 'k' value as input [1].

2. Methodology used for N AVF algorithm

Normally distributed attribute value frequency algorithm (NAVF) is a advance of AVF algorithm. It gives good precision and low recall value. This method calculates 'k' value itself based on the frequency. This method uses AVF score formula to find AVF score but no k-value is required [1].

Input: Dataset – D,

Output: detected outliers are k

Step 1: Read data set D

Step 2: Initially label all the Data points as non-outliers

Step 3: calculate normalized frequency of each attribute value for each point xi

Step 4: calculate the frequency score of each record xi as, Attribute Value Frequency of xi is discussed in[1]

Step 5: compute the N-seed values a and b as b=mean (xi), a=b-3*std (xi), if max (Fi) > 3*std (Fi)

Step 6: If Fi< a, then declare xi as outlier and return KN detected outliers.

B. Algorithms for clustering method:

Clustering is also called as data segmentation in few applications because clustering partitions large data sets into many groups by considering their similarity measure [12].

1. ROAD Algorithm

ROAD Algorithm is a two-phase algorithm for unsupervised detection of outliers. The object density computation and exploration of a clustering of the given data set is done by first-phase of this algorithm. The set of big clusters is identified in order to determine the

distance between various data objects and their corresponding nearest big clusters by using the resulting clustering structure. The frequency-based ranks as well as the clustering-based rank of each data object are determined by the second-phase. A unified set of the most similar outliers is constructed by using these two individual rankings. So, name of the method as Ranking-based Outlier Analysis and Detection (ROAD) algorithm.

The computational complexity of the proposed algorithm is basically contributed by the initial three steps. The first step requires O(nms) computations, where the maximum number of unique values of an attribute is s. Generally, s is called as small quantity compared to n. Second step requires O(nmk2) computations, as discussed in [15]. The third step contains the k-modes algorithm, which needs O(nmkt) computations, where t is said to be the number of iterations.

The ranking phase requires O(nlog(n)) iterations. Thus, the computational complexity of the proposed algorithm becomes to be O(nm + nlog(n)).The number of outliers to be detected dose not affected by computational complexity of this algorithm[2].

1.1. Methodology used for ROAD Algorithm

Input: Data set D with n data objects which is m-dimensional and values need for the parameters k and α.

Output: Set of likely outliers identified.

Phase (1): Computational phase

Step1: Compute density (Xi) of each data objects using (Equation 3) described in [2].

Step 2: Determine the initial set of k cluster representatives, using the method described in [15].

Step 3: Perform the k-modes clustering [16] on D using the distance measure given in Equation 2 and Determine the set of big clusters BC (Equation 4).

Step 4: Determine its cluster distance For each data object Xi, (as defined in Equation 5)[2].

Phase (2): Ranking phase

Step 5: Determine the frequency-based rank and the clustering-based rank of each data object as described in (Definition 6 and 7 respectively)[2].

Step 6: Using the two ranked sequences, for a given p value constructs the likely set LS (Definition 9).

2. ROCK

ROCK is a **RO**bust **C**lustering using lin**K**s [17]. It uses an Agglomerative hierarchy clustering. Links are used to measure similarity between different data point. Firstly all tuples are assigned as a separate cluster. Merging of Clusters is based on the smaller distance between clusters. It is convenient for Boolean and categorical datasets. The sample size decides scalability of the algorithm referred the criterion function and goodness measure[13].

2.1. Methodology used for ROCK algorithm

1. Draw a random sample

2. Compute the similarity of Link

3. Make Cluster with the link

4. Label it on the disk

3. Hierarchical clustering on feature selection for categorical data of biomedical application

. The author [18] concentrated on the feature association mining rule. Based on the contingency table, the distance (closeness) between different cluster features is calculated. Then hierarchical agglomerative clustering is used. The clustered results helps the domain expects to identify the feature association of their own interest. It works only for categorical data, which is the drawback of this system.

III. CONCLUSION AND FUTURE WORK

The paper describes a review on different classification and clustering methodologies associated with the categorical data for finding out outliers. Its advantage and limitation are discussed. Time complexities of various categorical classification and clustering algorithms are discussed. In future, there is possibility of developing a hybrid approach by using NAVF and ROAD algorithm.

REFERENCES:

[1] D. Lakshmi Sreenivasa Reddy, B. Raveendra Babu, A. Govardhan, "Outlier Analysis Of Categorical Data Using Navf",Ieee Conference ,2013.
[2] N N R Ranga Suri, M Narasimha Murty, G Athithan, "An Algorithm for Mining Outliers in Categorical Data through Ranking", IEEE CONFERENCE,2012.
[3] S. Wu and S. Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data, IEEE TRANSACTION on Knowledge Engineering and Data Engineering,2011.
[4] Z. He, X. Xu, J. Huang, and S. Deng, "FP-Outlier: Frequent Pattern Based Outlier Detection", IEEE CONFERENCE, 2003.
[5] E. Muller, I. Assent, U. Steinhausen, and T. Seidl, "Outrank: ranking outliers in high dimensional data," in IEEE ICDE Workshop, Cancun, Mexico, 2008
[6] C. Li, G. Biswas, "Unsupervised learning with mixed numeric and nominal data". IEEE TRANSACTION on Knowledge and Data Engineering, 2002
[7] S. Papadimitriou, H. Kitawaga, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," presented at International Conference on Data Engineering, 2003.
[8] V. Cheng, C. H. Li, J. T. Kwok, C-K. Li. "Dissimilarity learning for nominal data pattern Recognition", 2004
[9] S-G. Lee, D-K. Yun. "Clustering Categorical and Numerical Data : A New Procedure Using Multi-dimensional Scal-ing". International Journal of Information Technology and Decision Making",2003
[10] . Yu, M. Song, L. Wang "Local Isola-tion Coefficient-Based Outlier Mining Algorithm", International Conference on Information Technology and Computer Science 2009.
[11] Victoria J. Hodge and Jim Austin Dept. of Computer Science, University of York," A Survey of Outlier Detection Methodologies", Hodge+Austin_OutlierDetection_AIRE381.tex; 19/01/2004.
[12] Jiawei Han and Micheline Kamber ,University of Illinois at Urbana-Champaign"Data Mining: Concepts and Techniques Second Edition".
[13] Dr. Shuchita Upadhyaya, Karanjit Singh ,"Classification Based Outlier Detection Techniques", International Journal of Computer Trends and Technology- volume3Issue2- 2012.
[14] Kearns M. J," Computational Complexity of Machine Learning" MIT Press, Cambridge, MA, USA 1990.
[15] F. Cao, J. Liang, and L. Bai, "A new initialization method for categorical data clustering," Expert Systems with Applications, vol. 36, pp. 10 223–10 228, 2009.
[16] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in SIGMOD DMKD Workshop, 1997, pp. 1–8.
[17] S. Guha, R. Rastogi and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", Information Systems, Vol. 25, No. 5, pp. 345 – 366, 2000.
[18] Y. Lu and L. R. Liang, "Hierarchical Clustering of Features on Categorical Data for Biomedical Applications", Proceedings of the ISCA 21st International conference on Computer Applications in Industry and Engineering, pp. 26 - 31, 2008.
[19] A. Asuncion and D. J. Newman,"UCI machine learning repository"[Online]Available: http://archive.ics.uci.edu/ml, (2007)
[20] E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB Journal, 2000.
[21] Tan, P.-N., Steinbach, M., and Kumar," Introduction to Data Mining Addison-Wesley" , 2005.
[22] D. K. Roy and L. K. Sharma, "Genetic K means Clustering Algorithm for Mixed Numerical and Categorical data set", International journal of Artificial Intelligence & Applications, Vol. 1, No. 2, pp. 23 – 28, 2010.
[23] T. Palpanas, D. Papadopoulos, V. Kalogeraki, D. Gunopulos.,"Distributed deviation detection in sensor networks",SIGMOD Record 32(4): 77-82, 2003.
[24] Brause, R., Langsdorf T and Hepp m," Neural data mining for credit card fraud detection",In Proceedings of IEEE International Conference on Tools with Artificial Intelligence. 103 – 106, 1999.
[25] J. Zhang, Q. Gao, H. Wang, Q. Liu, K. Xu,"Detecting Projected Outliers in High-Dimensional Data Streams", DEXA 2009: 629-644, 2009.