International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Detection of Outliers Using Hybrid Algorithm on Categorical Datasets

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

[1]YCCE Nagpur, Department of Information Technology, Hingna Road, Nagpur, India

[2]Professor, YCCE Nagpur, Department of Information Technology, Hingna Road, Nagpur, India

Abstract: ***The outlier is an observation that is different from the other remaining values in a data set. Real life contains large number of categorical data. There is some outlier detection algorithms have been designed for categorical data. There are two main problems of outlier detection for categorical data, which are the time complexity and accuracy for detection of outliers in categorical dataset. Categorical dataset have some limited approaches as compared to numeric dataset. This paper describes about some existing algorithms for outlier detection in categorical dataset. The novel Hybrid method which overcomes limitations of previous methods (NAVF and ROAD) has been implemented. The algorithm is implemented and tested on different types of networking datasets, in which detected outliers are virus or intrusion whose behavior is different than behavior in normal networking data.***

Keywords: outliers, categorical data, hybrid approach, networking dataset, ranking and NAVF algorithm.

1. Introduction

1.1 Overview

Outlier analysis is very important research field in many applications like credit card fraud, intrusion detection in networks, medical field. This analysis concentrates on detecting infrequent data records in dataset. Preprocessing and classification technique has applied on dataset. In data transformation, the data are transformed or consolidated into binary forms appropriate for mining. Normalization, where the attribute data are scaled so as to fall within a small specified range, such as 1:0 to 1:0, or 0:0 to 1:0. In scaling it will read transformed file and convert it to max min value range from (0-1). i.e. if you are having number form 0 -10 i.e. 5,3,9 then 5 will become 0 and 3 will become 0 and 9 will become 1.

Frequent attacks on computer systems may result in systems being disabled, even completely collapsing. The identification of such intrusions could find out malicious programs in computer operating system and also detect unauthorized access with malicious intentions to computer network systems and so effectively keep out hackers. In this project we will consider the Networking Dataset in which we will detect outlier is virus or intrusion whose behavior will be different than in normal data.

2. Literature Review

2.1 Existing Algorithm for Categorical Datasets

In many data mining applications, the data objects are described using qualitative (categorical) attributes. The acceptable values of such a qualitative attribute are represented by various categories. The information on the occurrence frequencies of various categories of a categorical attribute in a given data set is very useful for many data-dependent tasks such as outlier detection.

2.1.1 NAVF (Normally distributed attribute value frequency)

NAVF has been defined as an optimal number of outliers in a single instance to get optimal precision in any classification model with good precision and low recall value.While taking the number of outliers sometimes the original data may be missed. If any classifier modeled using this data, wrong classifiers may be modeled.

This method calculates ‘k’ value itself based on the frequency. Let us take the data set ‘D’ with ‘m’ attributes A1, A2----- Am and d (Ai) is the domain of distinct values in the variable Ai. kN is the number of outliers which are normally distributed. To get ‘kN’ this model used Gaussian theory. If any object frequency is less than “mean-3 S.D” then this model treats those objects as outliers. This method uses AVF score formula to find AVF score but no k-value is required. Let D be the Categorical dataset, contains ‘n’ data points, xi, where i= 1…n. If each data point has ‘m’ attributes, we can write xi = [xi1, …..xil,…..xim], where xil is the value of the lth attribute of xi.

Algorithm

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.[1]

2.1.2 ROAD (Ranking-based Outlier Analysis and Detection)

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].

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).

3. Work Done

The existing algorithms have their own limitations such as time complexity, lower accuracy rate of detection outliers. The proposed system typically focuses on achieving higher accuracy rate of detection outliers so that it can overcome these limitations.

In proposed system, we try to overcome the problem of lower accuracy rate of detection outliers in categorical datasets. For that purpose system uses the Hybrid algorithm approach.
Therefore we have divided the work into following modules as:
3.1 Dataset
3.2 Preprocessing
3.3 Implementation of Hybrid Algorithm

3.1 Dataset

Proposed system needs categorical dataset for detection of outliers. We need dataset having networking information on the features containing protocol type, attack, flag and service having different number of attributes. Therefore proposed system uses categorical dataset.

3.2 Preprocessing

Preprocessing routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Normalization, where the attribute data are scaled so as to fall within a small specified range, such as -1:0 to 1:0, or 0:0 to 1:0.Min-max normalization performs a linear transformation on the original data. Suppose that minA and maxA are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v, of A to v0 in the range [new_minA,new_maxA] by computing
v0 =[(v-minA)/(maxA-minA)]*(new _maxA-new_minA)+new minA.

3.3 Implementation of Hybrid Algorithm

The proposed model has been developed by using NAVF and ROAD algorithm. NAVF algorithm calculates TP score of outliers, but it takes more time for computation. The ROAD algorithm gives maximum accuracy for detection of outliers in categorical datasets by using kmode algorithm, but it does not calculate TP score of outliers.

Hybrid algorithm uses feature of TP score calculation from NAVF algorithm and finds accurate outliers by using kmeans algorithm.

Given a set of data points (local group or global set)
- Outliers are points that do not fit to the general characteristics of that set, i.e., the variance of the set is minimized when removing the outliers. Outliers are the outermost points of the data set Given a smoothing factor SF(I) that computes for each I ? DB how much the variance of DB is decreased when I is removed from DB
- With equal decrease in variance, a smaller exception set is better
- The outliers are the elements of the exception set E ? DB for which the following holds:

SF(E) = SF(I) for all I ? DB

Similar idea like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution
- Naïve solution is in O(2n) for n data objects
- Applicable to any data type (depends on the definition of SF)
- Originally designed as a global method
- Outputs a labeling

We identify the points which are not outliers using clustering and distance functions, and prune out those points. Next, we

calculate a distance-based measure for all remaining points, which is used as a parameter to identify a point to be an outlier or not. We assume that there are n outliers in data set, and top n points will be reported as outliers by our method.

We use K-means algorithm to cluster the data set. Once clusters are formed, we calculate radius of each cluster. Prune the points whose distance from the centroid is less than the radius of the respective clusters. After that for each unpruned points in every cluster we calculate the ldof(local deviation outlier) .We report the top-n points with high ldof value as outliers.

The main idea underlying the new algorithm is to first cluster the data set into clusters, and then prune the points in different clusters if determined that they cannot be outliers. Since n (number of outliers) will typically be very small, this additional preprocessing step helps to eliminate a significant number of points which are not outliers.

Steps for Hybrid Algorithm

Step 1: Generating clusters:
Initially, we cluster the entire dataset into c clusters using K-means clustering algorithm and calculate radius of each cluster.

Step 2: Clusters having less number of points:
If a cluster contains less number of points than the required number of outliers, the radius pruning is avoided for that cluster.

Step 3: Pruning points inside each cluster:
Calculate deviation of each point of a cluster from the centroid of the cluster. If the distance of a point is less than the radius of a cluster, the point is pruned.

Step 4: Computing outlier points:
Calculate ldof for all the points that are left unpruned in all the clusters. Then n points with high ldof values are reported as outliers

4. Results and Discussions

4.1 Results

The proposed system is tested on various networking datasets. The same algorithm can be used for different datasets having different number of records. Hence we have derived necessary graphs, which show the comparison between NAVF, ROAD and Hybrid algorithm. We used the dataset of networking data. Proposed system has been designed for detection of outliers in categorical dataset. This approach can be used for different applications and corresponding datasets.

4.1.1 Comparison between NAVF and Hybrid Algorithm:

We have tested the proposed system on various datasets of networking data containing different number of records. The comparison is based on different three parameters, which are accuracy obtained by class, total accuracy obtained and time required for detection of outliers.

Detailed accuracy by class

This graph shows comparison between existing NAVF algorithm and Hybrid algorithm. It shows detailed accuracy obtained by class according to their TP rate. From this graph we can conclude that hybrid algorithm gives higher TP rate than NAVF algorithm according to their class.

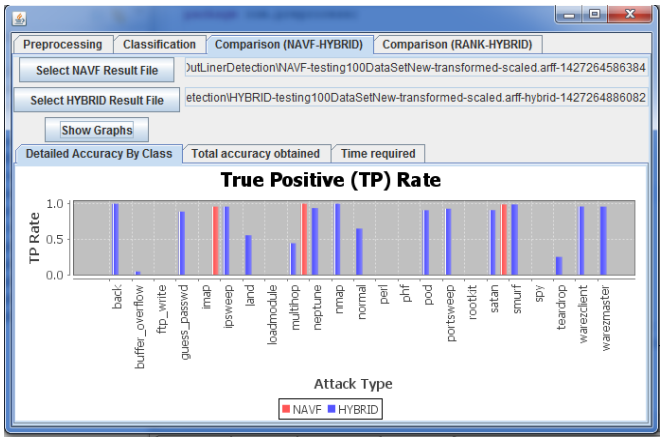

Figure 4.1: Detailed accuracy by class (Comparison between NAVF and Hybrid Algorithm)

4.1.2 Comparison between RANK and Hybrid Algorithm

Next, ROAD algorithm is compared with Hybrid algorithm and it gives following results:

Detailed accuracy by class

This graph shows comparison between existing ROAD algorithm and Hybrid algorithm. It shows detailed accuracy obtained by class according to their TP rate, where ROAD algorithm does not have feature of TP rate. So, only Hybrid algorithm calculates detailed accuracy by class. From this graph we can conclude that hybrid algorithm gives higher TP rate than ROAD algorithm according to their class.

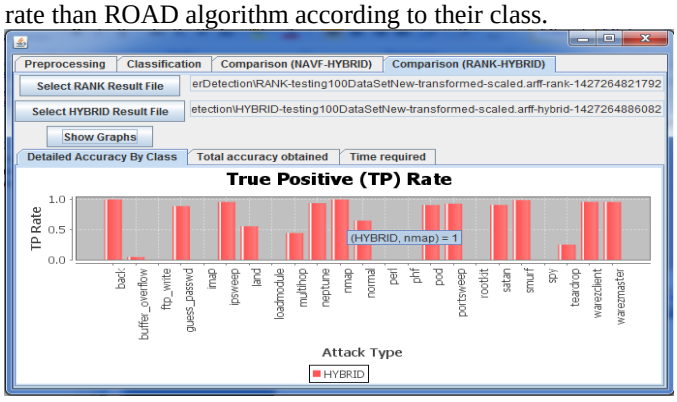

Figure 4.4: Detailed accuracy by class (Comparison between ROAD and Hybrid Algorithm)

Comparison with existing system

For our proposed system, we have compared following two algorithms shown in table 4.1.

Table 4.1: Comparison with existing system

Algorithm / Datasets		Hybrid	ROAD	NAVF
dataset1	time	1295ms	735ms	42979ms
	accuracy	88%	83%	25%
dataset2	time	1968ms	1159ms	43569ms

	accuracy	86%	81%	23%
dataset3	time	2397ms	2459ms	50653ms
	accuracy	85%	80%	20%

The above table shows comparison between Hybrid, ROAD and NAVF algorithm based on accuracy and time parameter. The time required for Hybrid algorithm is less as compared with NAVF algorithm. ROAD algorithm does not calculate TP score. So, time required for ROAD algorithm is less as compared with Hybrid algorithm. Second parameter is total accuracy obtained by algorithm. It gives the result of total accuracy obtained by all three algorithms. Hence, Hybrid algorithm has higher accuracy than NAVF and ROAD algorithm.

5. Conclusion and Future Scope

The novel method which overcomes limitations of previous methods. In particular, we can say our algorithms can deal with data sets with a large number of objects and attributes. Hence, derived hybrid algorithm and implemented it on different types of networking datasets, in which detected outliers are virus or intrusion whose behavior is different than behavior in normal networking data. Compared algorithms based on different parameters like time and accuracy. Hybrid algorithm can be implemented with different datasets with different number of records for the detection of outliers in categorical datasets. There are many key areas for future work including, the same method can also be applied on mixed type of dataset.

References

[1] D. Lakshmi Sreenivasa Reddy, B. Raveendra Babu, A. Govardhan, “Outlier Analysis Of Categorical Data Using Navf”, IEE 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 Isolation 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.