www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 6 June 2016, Page No. 17084-17087

Detection of Outlier in Uncategorical Dataset using Hybrid Algorithm

Ms.Kanchan D.Shastrakar[1], *Prof.Pravin G.Kulurkar*[2]

[1]Department of Computer Science &Engineering
VIT, RTMNU
Nagpur, India
Kanchan.Shastrakar7@gmail.com

[2]Department of Computer Science &Engineering
VIT, RTMNU
Nagpur, India
Pravinkulurkar@gmail.com

Abstract: ***Outlier mining is an important task of discovering the data records which have an exceptional behavior comparing with other records in the remaining dataset. Outliers do not follow with other data objects in the dataset. There are many effective approaches to detect outliers in numerical data. Most of the earliest work on outlier detection was performed by the statistics community on numeric data. But for categorical dataset there are limited approaches By using NAVF (Normally distributed attribute value frequency) and ROAD (Ranking-based Outlier Analysis and Detection algorithm) and new hybrid approach for outlier detection in categorical dataset will be formed.***

Keywords: NAVF, ROAD, Outliers, Categorical

1. Introduction

Outlier detection is the process of detecting instances with unusual behavior that occurs in a system. Effective detection of outliers can lead to the discovery of valuable information in the data. Over the years, mining for outliers has received significant attention due to its wide applicability in areas such as detecting fraudulent usage of credit cards, unauthorized access in computer networks, weather prediction and environmental monitoring.

A number of existing methods are designed for detecting outliers in continuous data. Most of these methods use distances between data points to detect outliers. In the case of data with categorical attributes, attempts are often made to map categorical features to numerical values. Such mappings impose arbitrary ordering of categorical values and may cause unreliable result.

Outliers are also referred to as irregularities, discordants, deviants, or differences in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an uncommon way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights.

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.

Though there exist a number of methods [1], [2], [6] for outlier detection in numerical data, the problem of outlier detection in categorical data is still evolving. The fundamental challenge in solving this problem is the difficulty in defining a suitable similarity measure over the categorical values. This is due to the fact that the various values that a categorical variable can assume are not inherently ordered [7]. As a result, many data mining related tasks such as determining the nearest neighbor of a categorical object turn out to be non-trivial. Some research efforts [8] in this direction are indicative of the importance of this issue.

Some examples are as follows:

Medical condition monitoring (such as heart rate monitors) or Medical Treatments (unusual responses to various drugs).

The analysis of outlier data is denoted to as outlier mining. Most of the existing systems are concentrated on numerical attributes or ordinal attributes. By using NAVF (Normally distributed attribute value frequency) and ROAD (Ranking-based Outlier Analysis and Detection algorithm) new hybrid approach for outlier detection in categorical dataset will be formed.

Interesting Sensor Events: In many real applications Sensors are used to track various environmental and location parameters. The sudden changes in the underlying patterns may represent events of interest. In the field of sensor network event detection is one of the primary motivating applications.

Satellite image analysis: identification of novel features or misclassified features.

Detecting unexpected entries in databases (in data mining application, to the aim of detecting errors, frauds or valid but unexpected entries).Detecting mislabeled data in a training data set.

2. Literature Survey

2.1 Existing Methodology

Outlier detection algorithms attempt to find data points that are different from the rest of the data points in a given data set. The problem is of considerable importance, arising frequently in many real-world applications, for data mining researchers. Many practical applications concerning outlier detection occur in different domains such as fraud detection, cyber-intrusion detection, medical anomaly detection, image processing and textual anomaly detection [1].

Statistics-based approaches (see [2, 3]) were first used for outlier detection based on an assumption that the distributions of datasets are known. A data point was defined as an outlier if it deviates from the existing distribution. With sufficient knowledge about the dataset, statistics-based methods work effectively. But in real-world, unfortunately, distributions of datasets are unknown, signify all points that belong to clusters. The effectiveness of this approach depends on the clustering algorithm. Knorr and Ng [6] propose to detect an outlier based on its distances from neighboring data points, many other variations of distance-based approaches have been discussed in the literature [7{9].

2.1.1. Greedy

The Greedy algorithm proposed the idea of finding a small subset of the data records that contribute to eliminate the disturbance of the dataset. This disturbance is also called entropy or uncertainty.

The Greedy algorithm complexity is O(k *n*m*d), where k is the required number of outliers, n is the number of objects in the dataset D, m is the number of attributes in D, and d is the number of distinct attribute values, per attribute.

2.1.2. AVF(Attribute Value Frequency)

The AVF algorithm complexity is lesser than Greedy algorithm since AVF needs only one scan to detect outliers. The complexity is O (n * m). It needs ‘k’ value as input.

2.1.3. NAVF(Normally Distributed Attribute Value Frequency)

This proposed model (NAVF) has been de-fined as an optimal number of outliers in single instance to get optimal precision in any classification model with good precision and low recall value. This method calculates ‘k’ value itself based on the frequency.

2.1.4. ROAD(Ranking-based Outlier Analysis and Detection Algorithm)

The computational complexity of the proposed algorithm turns out to be O(nm+nlog(n)).It is important to note that the computational complexity of this algorithm is not affected by the number of outliers to be detected.

3. Work Done

The existing algorithms have their own limitations such as lower accuracy rate, time complexity of detection outliers. The proposed system normally 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 Uncategorical dataset by merging NAVF (Normally distributed attribute value frequency) and Ranking algorithm. 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 networking dataset to exactly evaluate the behavior of our new method for different dimensionalities and database sizes, we generated multiple data sets having 25, 50 and 100 dimensions. As database sizes (dB size) we selected 500, 1,000, 5,000 and 10,000 data objects.

3.2 Preprocessing

Preprocessing routines work to filter the raw 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 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
- Heuristics like random sampling or best first search are applied
- 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 outlinear) .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.

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

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

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.

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. Result Analysis and Discussion

4.1 Results

In various networking datasets proposed system is tested and for different datasets having different number of records same algorithm can be used. Hence we derived some graphs which show the comparison between NAVF, ROAD and Hybrid algorithm. We used networking dataset and this approach can be used for different applications.

4.1.1 Comparison of NAVF with New Algorithm Hybrid

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.

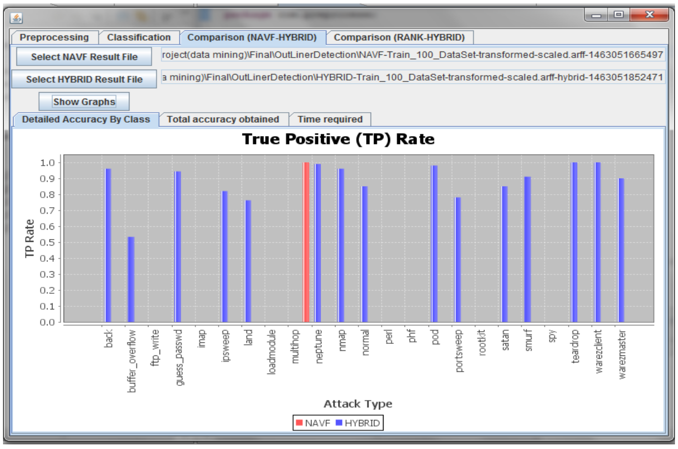
Figure 1: Comparison of NAVF with Our New Algorithm (Hybrid)

4.1.2 Comparison of RANK with New Algorithm Hybrid

This Graph show the comparison of RANK with our new algorithm (Hybrid)

Detailed accuracy by class

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.

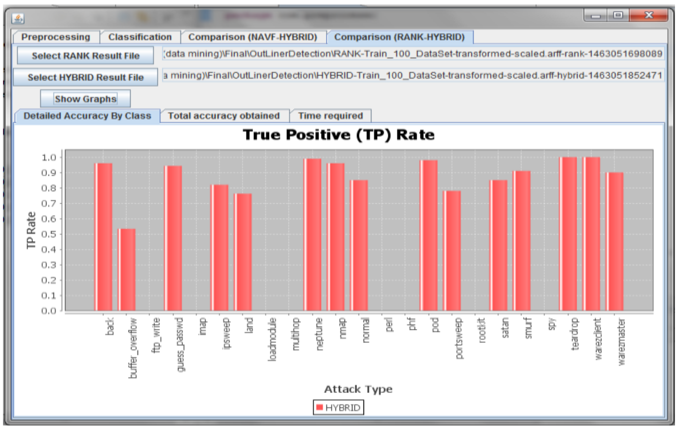


roject(data mining)\Final\OutLinerDetection\NAVF-Train_100_DataSet-transformed-scaled.arff-1463051665497

a mining)\Final\OutLinerDetection\HYBRID-Train_100_DataSet-transformed-scaled.arff-hybrid-1463051852471

Figure 1: Comparison of RANK with Our New Algorithm (Hybrid)

In Figure 2 a novel algorithm for mining categorical outliers through ranking data set consists of labeled instances belonging to two different classes. As per the standard practice in this field, objects with missing attribute values have been removed and objects belonging to small sized class in every data set are considered as outliers. Though these designated outliers are not outliers in real sense, they are considered so for validating the proposed method. In order to impose imbalance in the number of objects belonging to normal and outlier categories, only a selected sub-set of the objects belonging to outlier class have been considered, by taking every fifth object of the designated outlier class. Summarizes the description of various benchmark categorical data sets considered in this experimentation. As the proposed algorithm works in unsupervised learning mode, it doesn't require labeled data. However, class labels are used to measure its performance in detecting outliers.



(data mining)\Final\OutLinerDetection\RANK-Train_100_DataSet-transformed-scaled.arff-rank-1463051698089

a mining)\Final\OutLinerDetection\HYBRID-Train_100_DataSet-transformed-scaled.arff-hybrid-1463051852471

DOI: 10.18535/ijecs/v5i6.50

5. Conclusion & Future Scope

Outlier detection is an important task for data mining applications. Existing algorithms are effective and have been successfully applied in many real-world applications. But these algorithms, especially density-based algorithms, have low efficiency in datasets with different densities or when datasets consist of clusters with special shapes. In this paper, we introduce a two algorithm i.e. NAVF RANK to measure an object's outlierness. Sum of two of an object is naturally meaningful to measure the degree of isolation of an object. Based on this idea, we propose the Hybrid Algorithm which is combination of both NAVF and RANK that is effective to solve the problems mentioned above for many situations.

In this project, we introduced a novel, parameter-free approach to outlier detection based on the variance of angles between pairs of data points. This idea alleviates the effects of the curse of dimensionality" on mining high-dimensional data where distance-based approaches often fail to offer high quality results. In addition to the basic approach NAVF, we proposed two variants: RANK as acceleration suitable for low-dimensional but big data sets, and Hybrid, alter-refinement approach as acceleration suitable also for high-dimensional data. In a thorough evaluation, we demonstrate the ability of our new approach to rank the best candidates for being an outlier with high precision and recall. Furthermore, the evaluation discusses efficiency issues and explains the influence of the sample size to the runtime of the introduced methods.

There are two directions for future work. The first one is to improve the performance of Hybrid

in datasets consisting of clusters with special shapes such as lines or circles. Currently, Hybrid

doesn't perform as good as COF(connectivity-based outlier factor) for this kind of datasets. The second is to further improve the effectiveness of ranking.

References

[1] M. E. Otey, A. Ghoting, and and A. Parthasarathy, "Fast Distributed Outlier Detection in Mixed-Attribute Data Sets," Data Mining and Knowledge Discovery

He, Z., Deng, S., Xu, X., “A Fast Greedy algorithm for outlier mining”, Proc. of PAKDD, 2006.

[2] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining: Pearson Addison-Wesley, 2005

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

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density based local outliers," presented at ACM SIGMOD International Conference on Management of Data, 2000

[5] 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

[6] Z. He, X. Xu, J. Huang, and S. Deng, "FP-Outlier: Frequent Pattern Based Outlier Detection”, Computer Science and Information System (ComSIS'05)," 2005

[7] S. Wu and S. Wang, “Information-Theoretic Outlier Detection for Large-Scale Categorical Data, IEEE Transactions on Knowledge Engineering and Data Engineering,2011

[8] A. Frank, & A. Asuncion, (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] E. Muller, I. Assent, U. Steinhausen, and T. Seidl, “Outrank: ranking outliers in high dimensional data,” in IEEE ICDE Workshop, Cancun, Mexico, 2008, pp. 600–603.

[10] K. Das and J. Schneider, “Detecting anomalous records in categorical datasets,” in ACM KDD, San Jose, California, 2007, pp. 220–229.

[11] Z. He, X. Xu, and S. Deng, “A fast greedy algorithm for outlier mining,” in PAKDD, Singapore, 2006, pp. 567–576.

[12] A. Koufakou, E. Ortiz, and M. Georgiopoulos, “A scalable and efficient outlier detection strategy for categorical data,” in IEEE ICTAI, Patras, Greece, 2007, pp. 210–217.

[13] S. Guha, R. Rastogi, and S. Kyuseok, “ROCK: A robust clustering algorithm for categorical attributes,” in ICDE, Sydney, Australia, 1999, pp. 512–521.

[14] Z. Huang, “A fast clustering algorithm to cluster very large categorical data sets in data mining,” in SIGMOD DMKD Workshop, 1997, pp. 1–8.

[15] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognition Letters, vol. 31, pp. 651–666, 2010.

[16] 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.

[17] A. Asuncion and D. J. Newman. (2007) UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml

[18] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In Proc. VLDB, 1998.

[19] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In Proc. VLDB, 1999.

[20] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchthold. Efficient biased sampling for approximate clustering and outlier detection in large datasets. IEEE TKDE, 15(5):1170{1187, 2003.

[21] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998.

[22] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In Proc. ICDE, 2003.

[23] P. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41:212{223, 1999.