ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 151

2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)

A Review of Anomaly Detection Techniques Based on Nearest Neighbor

Ming Zhao[1,2,*], Jingchao Chen[1] and Yang Li[2]

[1]School of Information Science & Technology, Donghua University, Shanghai, 201620, China

[2]School of Air Transportation, Shanghai University of Engineering Science, Shanghai, 201620, China

[*]Corresponding author

***Abstract*—The concept of nearest neighbor has been used in several anomaly techniques, which supposes normal data instances occur in dense neighbors and anomalies occur far from their closest neighbors. So the techniques require a distance or similarity measure defined between two data instances. By now, there are several variants of basic technique extended by researchers in three different ways. The first set is to modify the definition of the anomaly score. The second set is to select different distance or density measure for different data type. The third set is to reduce the computation complexity. In this paper we have attempted to provide an overview of the previous work, although it is limited.**

Keywords—outlier detection; anomaly detection; distance based outlier detection; k nearest neighbor

I. INTRODUCTION

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are also known as outliers, exceptions, aberrations, or contaminants in different application domains. Anomaly is distinct from *noise*. Anomaly can be of great interest to the analyst, but noise is unmeaning.

Anomalies can be classified into point anomalies, contextual anomalies and collective anomalies. A label or a score is assigned to denote whether that instance is normal or anomalous. Anomaly detection techniques can be categorized in several approached, such as depth-based, distance-based, density-based, clustering-based, statistical-based, and so on. A distance-based approach usually works by calculating a distance of a data point from their nearest neighbors.

Nearest neighbor anomaly detection techniques assume that normal data instances occur in dense neighborhoods while anomalies occur far from their closet neighbors, and require a distance or similarity measure defined between two data instances. Distance or similarity can be computed in different ways. The distance of a data instance to its nearest neighbor or the relative density of each data instance is defined as anomaly score in nearest neighbor-based anomaly detection techniques. In this paper, we primarily discuss nearest neighbor-based anomaly detection techniques using distance measure.

Several variants of the basic nearest neighbor anomaly detection technique have been proposed by researchers in three different ways. The first set focuses on the definition of the anomaly score. The second set uses the different distance measures for different data types. The third set focuses on reducing computation complexity for improving the algorithm efficiency.

II. PREVIOUS IMPROVEMENTS

A. Definition of Anomaly Score

An anomaly score represents the degree to which that instance is considered an anomaly. An analyst may select top n instances with the largest anomaly scores as the anomalies, or use a specific threshold to determine the most relevant anomalies. Notably, anomaly score is not unique output form of anomaly detection. The other form is label, which indicates the test instance is normal or anomalous. Comparing with anomaly score, scoring techniques provide a ranked list of anomalies.

One way is that the anomaly score of a data instance is defined as the sum of its k nearest neighbors, the details refer to [1][2][3][4]. A similar technique is to compute the distance of a data instance to its k-th nearest neighbor as the anomaly score [5][9].

A different way to compute the anomaly score is to count the number of nearest neighbors k that are not more than d distance. For example in a 2-D data set, the inverse of $k/\pi d^2$ is the anomaly score [6]. This method can be also viewed as density-based technique, since it involves counting the number of neighbors in a sphere of radius d. Liu and Pan apply the minimum distance d_{min} between an instance and samples in class l which the instance belongs to [7]. So the d_{min} is the anomaly score, if d_{min} is larger than a given threshold, the instance is anomaly.

So far most techniques have been proposed to handle continuous attributes. Other variants have been proposed for categorical attributes or a mixture of categorical and continuous attributes. Otey et al. defined the anomaly score as the inverse of the sum of the *link* strength between the instance and the other instance in data sets [8]. The associated *link* strength is equal to the number of attribute-value pairs shared between two instances.

A hybrid approach, which calculates the average distance of n objects as neighborhood distance, and calculates the average number of neighbors within the average distance. In the selection process, this average value is used as a threshold. If the number of nearest neighbors is less than the threshold, an outlier is detected [19].

B. Distance/Similarity Measure

Nearest neighbor anomaly detection techniques require a distance or similarity measure. For continuous attributes, Euclidean distance is a popular choice. For example, the authors [5] use the square of the Euclidean distance as the distance metric, since it involves fewer and less expensive computations. However, almost all real-world data sets contain a mixture of categorical and continuous attributes. Several variants have been proposed to handle other data types.

Otey et al. presented a tunable algorithm for distributed anomaly detection in mixed-attribute data sets [8]. They capture the *link* between the points in the mixed categorical and continuous attribute space. For categorical attributes, two data points are considered linked if they have at least one attribute-value pair in common. For continuous attributes, a covariance matrix is maintained to capture the dependencies between the continuous values.

For categorical attributes, a hypergraph-based technique measure distance between two data instance by analyzing the connectivity of the graph [11]. The algorithm named ORCA is based on using Hamming distance to compute distance between categorical observations [12][13]. ORCA takes two parameters, number of outliers m and number of nearest neighbors k in kNN. Palshikar et al. extend the notion of distance-based outliers to time series data, their implementation offer a choice of various distance measure to the user (e.g., Manhattan, Euclidean, etc.) [14].

Density-based anomaly detection techniques estimate the density of neighborhood of each data instance. An instance with sparse neighborhood is regarded as anomalous, while an instance with dense neighborhood is declared to be normal. Most density-based anomaly detection techniques use *Local outlier Factor* (LOF) as the anomaly score. The LOF score is equal to ratio of average local density of the k nearest neighbors of the instance and the local density of the instance itself. Firstly, k nearest neighbors is found in the radius of smallest hyper-sphere centered at the instance. Then the local density of the instance is the volume of this hyper-sphere divide by k.

A variant of LOF is *Connectivity-based Outlier Factor* (COF). The difference between LOF and COF is the manner of computing k nearest neighbors. In COF, the neighborhood is computed in an incremental mode [20]. A drawback of the LOF technique is the $O(N^2)$complexity required.

C. Computation Complexity

Several variants have of nearest neighbor anomaly detection techniques have been proposed to improve the efficiency.

Ramaswamy et al. develop a highly efficient partition-based algorithm, which first partitions the input data set into disjoint subsets, and then prunes entire partitions that cannot possibly contain the top k anomalies[5]. In contrast, the partition-based algorithm is faster than the other algorithms for higher dimensions.

To prune the search space for nearest neighbors, the attribute space is partitioned into a hypergrid in several works. Angiulli and Pizzuti linearize the search space through the Hibert space filling curve to deal with high dimensional data sets [2]. The algorithm executes at most $d+1$ scans of the data set with a low time complexity cost, where d is the number of dimensions of the data set.

A simple sampling algorithm to efficiently detect distance-based outliers is proposed by Wu and Jermaine in [15]. For each data point i, the algorithm compute the kth nearest neighbor in a randomly sampled set from given data set. The top γ points whose distance to its kth nearest neighbor is greatest are outliers. Compared with the state-of-the-art algorithm, the complexity of the proposed algorithm is reduced to $O(MN)$, Where M is the sample size chosen.

As dimensionality increases, a novel insight into the nearest neighbors in the unsupervised outlier detection is proposed in [16]. The authors define the notion of k-occurrences, hubs and antihubs. The experimental results demonstrate that k-NN is the fastest algorithm, although it doesn't make exact empirical comparisons of running time. Proposed approach in [19], the total time complexity of the algorithm is $O(N^2 \log N)$.

D. Comparison of Outlier Detection Techniques

A distance-based method works on complete data set, so the required computation time in a distance-based method is more as compared to cluster-based methods. But this method is best suitable with time series and multi-dimensional datasets [18].

The advantages of nearest neighbor-based anomaly detection techniques are, (i) it needn't to know the data distribution model, (ii) the possibility that an anomaly has a close neighborhood is very low, (iii) for different data type, it only need define an appropriate distance measure.

The disadvantages of nearest neighbor-based anomaly detection techniques are, (i) in testing phase, the computational complexity is challenge since it need compute the distance of each test instance to the training data, (ii) defining distance measure between instances is also challenge when the data is graphs, sequences, and so on.

III. CONCLUSION

In this paper we have discussed the improvements of nearest neighbor-based anomaly detection techniques using distance to kth or k nearest neighbors in three different ways. Only we focus on the distance measure with Manhattan or Euclidean, the similarity with density measure isn't be concerned here. We have attempted to provide an overview of the previous work, maybe it is limited by the literatures we have read.

REFERENCES

[1] Bolton R J, David J H. Unsupervised Profiling Methods for Fraud Detection[J]. Proc Credit Scoring & Credit Control VII, 2001:5--7.

[2] Angiulli F, Pizzuti C. Fast Outlier Detection in High Dimensional Spaces[C].European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, 2002:15-26.

[3] Zhang J, Wang H. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance[J]. Knowledge & Information Systems, 2006, 10(3):333-355.

[4] Zhao R, Du B, Zhang L. GSEAD: Graphical Score Estimation for Hyperspectral Anomaly Detection[C]//IEEE, Workshop on

Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. IEEE, 2016.

[5] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets[C]. ACM SIGMOD International Conference on Management of Data. ACM, 2000:427-438.

[6] Knorr, Edwin M, Ng, et al. Distance-based outliers: algorithms and applications[J]. Vldb Journal, 2000, 8(3-4):237-253.

[7] Liu D, Pang J, Xu B, et al. Satellite Telemetry Data Anomaly Detection with Hybrid Similarity Measures[C]. International Conference on Sensing, Diagnostics, Prognostics, and Control. 2017:591-596.

[8] Otey M E, Ghoting A, Parthasarathy S. Fast Distributed Outlier Detection in Mixed-Attribute Data Sets[J]. Data Mining & Knowledge Discovery, 2006, 12(2-3):203-228.

[9] V Chandola, D Cheboli and V Kumar. Detecting anomalies in a time series database[J]. Technical report, 2009, p.12.

[10] Abid A, Kachouri A, Guiloufi A B F, et al. Centralized KNN anomaly detector for WSN[C]. International Multi-Conference on Systems, Signals & Devices. IEEE, 2015:1-4.

[11] Wei L, Qian W, Zhou A, et al. HOT: Hypergraph-Based Outlier Test for Categorical Data[J]. 2003, 2637(2):399-410.

[12] Bay S D. Mining distance-based outliers in near linear time with randomization and a simple pruning rule[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003:29-38.

[13] Taha A, Hegazy O M. A proposed outliers identification algorithm for categorical data sets[C]. The, International Conference on Informatics and Systems. IEEE, 2010:1-5.

[14] Palshikar G K. Distance-Based Outliers in Sequences[M]. Distributed Computing and Internet Technology. Springer Berlin Heidelberg, 2005:547-552.

[15] Wu M, Jermaine C. Outlier detection by sampling with accuracy guarantees[C]. Twelfth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2006:767-772.

[16] Radovanović M, Nanopoulos A, Ivanović M. Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(5):1369-1382.

[17] Behera S, Rani R. Comparative analysis of density based outlier detection techniques on breast cancer data using hadoop and map reduce[C]. International Conference on Inventive Computation Technologies. IEEE, 2017:1-4.

[18] Dang T T, Ngan H Y T, Liu W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data[C]. IEEE International Conference on Digital Signal Processing. IEEE, 2015:507-510.

[19] Huang Y, Zhang Z, Liao M, et al. A hybrid distance-based outlier detection approach[D]. IEEE Computer Society, 2012.

[20] Chandola V, Banerjee A, Kumar V. Anomaly detection:A survey[J]. Acm Computing Surveys, 2009, 41(3):1-58.