Application of clustering analysis in Intrusion Detection

Yifan Zhang, Xingshan Li[a], Min Xu

Luo he medical college, Luohe, 462000, China
[a]Email:604141388@qq.com

Keywords: Cluster, Analysis Techniques, Boundary Detection, Intrusion Detection

Abstract: Clustering technology and boundary point detection technology and its application in intrusion detection system are introduced in this paper from three aspects, which are the application of clustering analysis, boundary detection and clustering analysis in Intrusion Detection System. The data processing and the requirement of clustering algorithm for intrusion detection system are introduced in detail. Analyzed the result of the experiment environment and experiment, further validation of this project is based on the improved NPRIM algorithm applied to intrusion detection is effective and feasible.

1. Boundary point detection technology

The cluster boundary point is a point that has two or more clustering characteristics between the cluster and the cluster. The study of clustering boundary points is an important branch of clustering analysis, which plays an important role in disease prevention, biology, image retrieval, virtual reality, and improving clustering accuracy. Since Chenyi Xia first proposed the boundary point detection algorithm (BORDER) in 2006, researchers have proposed some boundary detection algorithms. In order to describe these algorithms, the algorithm is divided into four categories: density based boundary detection algorithm, grid based boundary detection algorithm and angle based boundary detection algorithm.

1.1. Boundary point detection algorithm based on density

Based on the density of the boundary point detection algorithm is the use of clustering near the boundary of the uneven distribution of data objects to extract the characteristics of the clustering boundary point. On the noisy data set, the algorithm can separate the boundary point from the noise region, especially the uniform data set. BRIM is a typical boundary detection algorithm in this algorithm.

In order to solve the existing problems of BORDER algorithm, BRIM is a density based boundary point detection algorithm, which can effectively detect the boundary of clustering in noisy data sets. The algorithm first according to the data object plus or minus the number of data points difference within half a neighborhood to calculating the boundary of points, and then the boundary is greater than the boundary degrees threshold δ marked point boundary point. Because of the

cluster boundary point forward neighborhood and negative half neighborhood point number differs greatly, two minutes boundary is high, and the clustering internal points/noise isolation of plus or minus half the number of neighborhood points within the difference is not big, small border degree, so the algorithm can contain data sets general boundary point and noise point/distinguish isolated point [1].

The BRIM algorithm can effectively detect the boundary points on noisy data sets, and the efficiency is high. But it is still in some deficiencies: For example, input parameters and parameters are difficult to determine, it is not possible to accurately identify the boundary points of the low density regions in the multi density dataset .etc.

1.2. Boundary point detection algorithm based on grid

Grid based boundary detection algorithm makes full use of the advantages of grid technology, which is the most efficient algorithm in the four kinds of algorithms. The algorithms in the clusters with arbitrary shape, different density of noise data sets quickly and efficiently detect the boundary point, its typical representative is GRIDEN, EDGE.

GRIDE combines the grid technology with the information entropy and the joint entropy technique. Entropy is an important concept in the information theory to describe the uncertainty of random distribution. The GRIDEN algorithm uses the entropy to detect the boundary points, because the density of the grid at the boundary points is usually uneven, and entropy of the uneven distribution of grid points in the grid will be bigger, otherwise will be smaller [2]. The algorithm in the implementation process, the data space is divided into grids, and be identified according to the characteristics of neighbor grid density boundary grid and uneven distribution, and then calculate the measured data object density variation degree of entropy in the boundary grid range, finally determine the boundary points according to entropy.

GRIDEN algorithm in clusters with arbitrary shape, different size of noise data sets, can eliminate the interference of noise points, quickly and efficiently detect the boundary point. At the same time run much faster than BORDER, BRIM algorithm, etc. But too many parameters (Three parameters: the number of mesh with each dimension k, grid density threshold Minpts, boundary point threshold E-Minpts)and boundary detection results of the algorithm are very much dependent on the parameters, this will bring great difficulty in getting the user to select parameters[3]. In addition, on the multidimensional data set containing noise, GRIDEN test results of the algorithm is not very ideal, still contain a small amount of noise.

1.3. Edge detection algorithm based on angle

The boundary point detection algorithm based on angle is the angle of the use of the space vector relation to detect the boundary point. Boundary points and other points constitute the Angle between the vector distribution as shown in figure 1: Point p as the boundary point, point q for internal points, can be seen from the diagram, point p in the neighborhood and other points within the angle between the vector consisting of less than180°, and point q is equal to 180° and the angle degree evenly distributed. This kind of boundary detection algorithm detects the boundary point is more accurate, and its main parameters for the angle, relatively easy to determine the scope (0°, 180°).

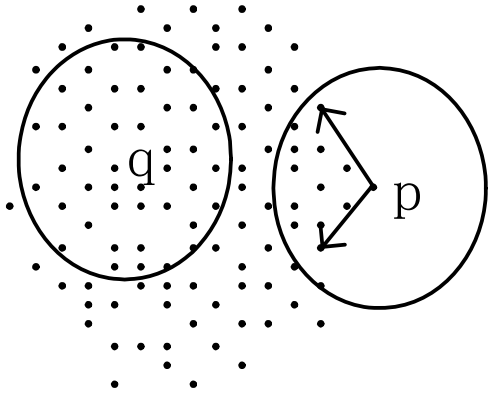

Figure 1 Characteristics of boundary points

FRINGE algorithm uses the grid technology and the characteristics of the angle, is based on the angle of the boundary point detection algorithm in a better detection algorithm. The algorithm like GRIDEN algorithm with the help of some of the concept of grid, such as data space, grid cell, the density of grid, grid, neighbors, dense grids and sparse grid, etc[4].

The FRINGE algorithm can accurately detect the boundary points on the noisy data sets with arbitrary shape clusters, and also reduce the difficulty of the user's choice of parameters. However, we still need to input 3 parameters, without eliminating the dependence of the boundary detection results on the parameters. When the distance between the cluster and the cluster is detected in the data set, the FRINGE algorithm will lose the connection part, so that the detection result is not as good as BRIM and GRIDEN.

1.4. Graph based edge detection algorithm

The graph based boundary detection algorithm firstly uses the many important features of the graph to reflect the similarity between the data objects, and then extracts the boundary points according to the special distribution characteristics of the boundary points. This algorithm is different from the first three kinds of algorithms, it can not only extract the clustering of the boundary points, but also to cluster, effectively combine the two. TRICLUST, DTBOUND all belong to this kind of algorithm, in which TRICLUST focuses on the clustering function without boundary detection method and the results are given in detail[5].

DTBOUND algorithm using the triangle subdivision technology and variation coefficient testing boundary point. First using triangular subdivision to reflect the similarity relationship between data objects, and then the variation coefficient is used to measure the variation of distribution of data objects, according to the degree of variation to extract the boundary point.

DTBOUND algorithm has only one parameter and the value is relatively easy, can contain arbitrary shapes, different density and size of the noise data sets on clustering and boundary detection. Compared with BORDER, BRIM boundary detection algorithm, it blends clustering and boundary detection and boundary detection precision is high. But it still requires the user to input parameters, it will not only affect the efficiency of algorithm implementation, and in setting parameters is introduced as prejudice which affect the accuracy of the boundary detection results.

NPRIM algorithm is to eliminate the boundary point detection algorithm run results dependence on input parameters, and improve the efficiency of the boundary detection precision and algorithm, and puts forward a parameter boundary detection algorithm. The algorithm through the establishment of triangulation diagram reflect the similarity between data objects, calculating the boundary of each object, according to the degree of boundary extraction of boundary points of clusters, in this process, the automatic calculation of threshold by K-means, to eliminate the

influence of parameters on the boundary detection results. The algorithm does not need to enter any parameters and low time complexity, and can contain arbitrary shape, different density and different size clustering boundary detection and containing noise data sets, compared with the existing boundary detection algorithm does not need to enter any parameters, and boundary detection of the advantages of high accuracy, high speed, but it can't clustering algorithm

2. Application of clustering analysis in intrusion detection

At present, there are many clustering algorithms used in intrusion detection system, such as: K-means algorithm based on partition clustering algorithm, ART neural network algorithm based on Fuzzy model, density based clustering algorithm CADGB, etc. However, there are few researches on the application of the boundary point detection technology in Intrusion Detection System.

2.1. Data processing of intrusion detection system based on Clustering

In general, the process of data processing based on clustering intrusion detection system is: data acquisition, data standardization, clustering, marking, intrusion detection. The first is the system by Tcpdupm, sniffer, Wireshark and other tools to obtain the original data capture. Because the original data capture tool to obtain the data types, measurement standards are not identical and contain noise, which requires the adoption of standardized pretreatment on the data to eliminate the interference factors. The second is on the processed data according to certain clustering algorithm to cluster analysis; Through cluster analysis, the data object is divided into multiple classes, these classes as part of the normal or belong to the exception class, need the clustering analysis results are normal or abnormal tag operation. Finally is using tagged has good characteristics of a data set to complete the intrusion detection.

2.2. The requirement of clustering algorithm for intrusion detection

Clustering is an unsupervised learning method, and it should be based on the following assumptions:

(1) Abnormal data distribution deviates from the normal data distribution, and accounts for a small part of the whole data set(<=2%);

(2) Test data can represent the entire data event area;

(3) Under the standard of reasonable calculation, the data events with the same class are very similar, but different kinds of data events are very different.

Intrusion detection system on the clustering algorithm requirements are the following:

① Able to deal with massive data sets

At present, many clustering algorithms have high accuracy in small data sets, but the results are not satisfactory on the large scale data sets. However, the actual network data is just a huge amount of data, which requires that the clustering algorithm applied to intrusion detection must have a high degree of scalability, and can effectively deal with massive data sets

② Eliminate the influence of parameters

In clustering analysis, a lot of clustering algorithms require the user to manually input some parameters, the clustering result is sensitive to the input parameters. Usually these parameters are more difficult to determine, and artificial input parameters not only increase the burden on the user, and it is difficult to adapt to the needs of new applications. Design does not need to manually input parameters and algorithm parameters can be determined adaptive clustering algorithm and the current academic research hot spot and the difficulty.

③ The order of input data is not sensitive

Some existing clustering algorithms are very sensitive to the input order of data. For the same data set, the clustering results are very different when they are submitted to the same clustering algorithm in different order。However, the practical application scenarios of intrusion detection require that the clustering algorithm must be insensitive to the order of input data.

④ Able to handle different types of data

The existing clustering algorithms are usually only suitable for numerical data, and the data of intrusion detection are both numerical and non numerical, only the clustering algorithm which can handle multiple data types can be applied to intrusion detection.

⑤ Ability to handle high dimensional data

Most clustering algorithms have high processing accuracy for low dimensional data (usually involving only two to three dimensions), but they can not deal with high dimensional data accurately. Intrusion detection data is high dimensional data with multiple attributes. Therefore, the clustering algorithm applied to intrusion detection must have the ability of high dimensional data processing.

⑥ Ability to manage noise data

In the application of intrusion detection, the vast majority of the data contains outliers (such as abnormal data, incomplete data, or even wrong data), so those sensitive to outlier data clustering algorithm cannot be applied in Intrusion Detection.

⑦ Interpretability and availability

In practical applications, users usually want the clustering results to be interpretable, understandable and usable. In other words, clustering needs to be associated with specific semantic interpretations and Applications

⑧ Ability to discover arbitrary shape clusters

Many clustering algorithms tend to find globular clusters with similar density or size. In view of the diversity of the actual data, the clustering algorithm applied to intrusion detection is the best way to find any shape clustering.

At present, researchers have achieved good results in the application of clustering analysis in intrusion detection, but so far, there is not a clustering algorithm which can meet the requirement of intrusion detection. This project through the comprehensive analysis of various kinds of clustering analysis and the advantages and disadvantages of boundary point detection algorithm, the improved parameterless boundary point detection algorithm NPRIM algorithm applied to intrusion detection system.

3. Summary

This article from the cluster analysis technology, boundary point detection technology, the clustering analysis technology application in intrusion detection system in three aspects clustering technology is introduced and the boundary point detection technology and its application in intrusion detection system. The cluster analysis technology in the five types of clustering analysis algorithm: based on the division method, based on the hierarchy method and the method based on density and grid based method, based on the method of the mode. Boundary point detection technology are introduced in detail in the technology of four types of boundary point detection algorithm: boundary detection algorithm based on density, boundary detection algorithm based on grid, the boundary point detection algorithm based on angle. Finally, the paper introduces the data processing process and the requirement of clustering algorithm in Intrusion Detection System.

Acknowledgements

In this paper, the research was sponsored by Medical Science Research project of Henan Province (Project No. 201404065).

References

[1] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 2014: 1690-1700.

[2] Khor K C, Ting C Y, Phon-Amnuaisuk S. A cascaded classifier approach for improving detection rates on rare attack categories in network intrusion detection. Applied Intelligence, 36(2), 2012: 320-329.

[3] Davis J J , Clark A J. Data preprocessing for anomaly based network intrusion detection: A review. Computers & Security, 30(6), 2011: 353-375.

[4] Lin S W, Ying K C, Lee C Y, et al. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. Applied Soft Computing, 12(10), 2012: 3285-3290.

[5] Louvieris P, Clewley N, Liu X, Effects-based feature identification for network intrusion detection. Neurocomputing, 121 ,2013 ; 265-273 .