A new algorithm for clustering analysis and boundary detection in hospital information security

Yifan Zhang, Xingshan Li [a], Xiangyang Shao

Luo He Medical College, Luohe, 462000, China
[a]Email:604141388@qq.com

Keywords: Clustering, analysis, boundary detection, NPRIM algorithm

Abstract: This paper expounds the intrusion detection based on clustering and the boundary point detection of the specific models, and the clustering analysis technology has carried on the detailed description, and then 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. Introduction

Intrusion detection model based on clustering and boundary detection is designed in this project the NPRI clustering and boundary detection algorithm applied to intrusion detection system, the whole process is shown in figure 1.

Can be seen from the figure 1 the program design of intrusion detection system consists of data processing, clustering analysis, intrusion, intrusion response, warehousing five most.

2. Clustering analysis technology

With the development of the cluster analysis technology, researchers have put forward a lot of clustering algorithm, different clustering algorithm is applicable to different data types and application scenarios. The existing clustering algorithms can be broadly divided into five categories: partition based method, hierarchical method, density based method, grid based method, model-based method.

2.1. Partition based clustering method

The clustering method based on partitioning is actually the clustering problem into an optimization problem, it is by optimizing a criterion function, including the N object data set from an initial partition gradually transformed into optimal clustering requirements, its output is K (k<=n) disjoint sub division. A sub partition corresponds to a cluster, which satisfies the similarity between the inner points of the same cluster, height difference between different clusters. At present, the widely used partition based clustering algorithms are k-means algorithm [1] and K- mediod algorithm[2]. The method based on classification is the earliest and the most widely used type, through the improvement, it can carry on the clustering of large database. But it also has many disadvantages: before clustering, it requires the user to determine the number of clusters K, however, K is usually difficult to determine in large scale data sets. In contains clusters of arbitrary shape, size and density data sets, the algorithms can not effectively clustering. The algorithms of outliers and the choice of the starting point is more sensitive.

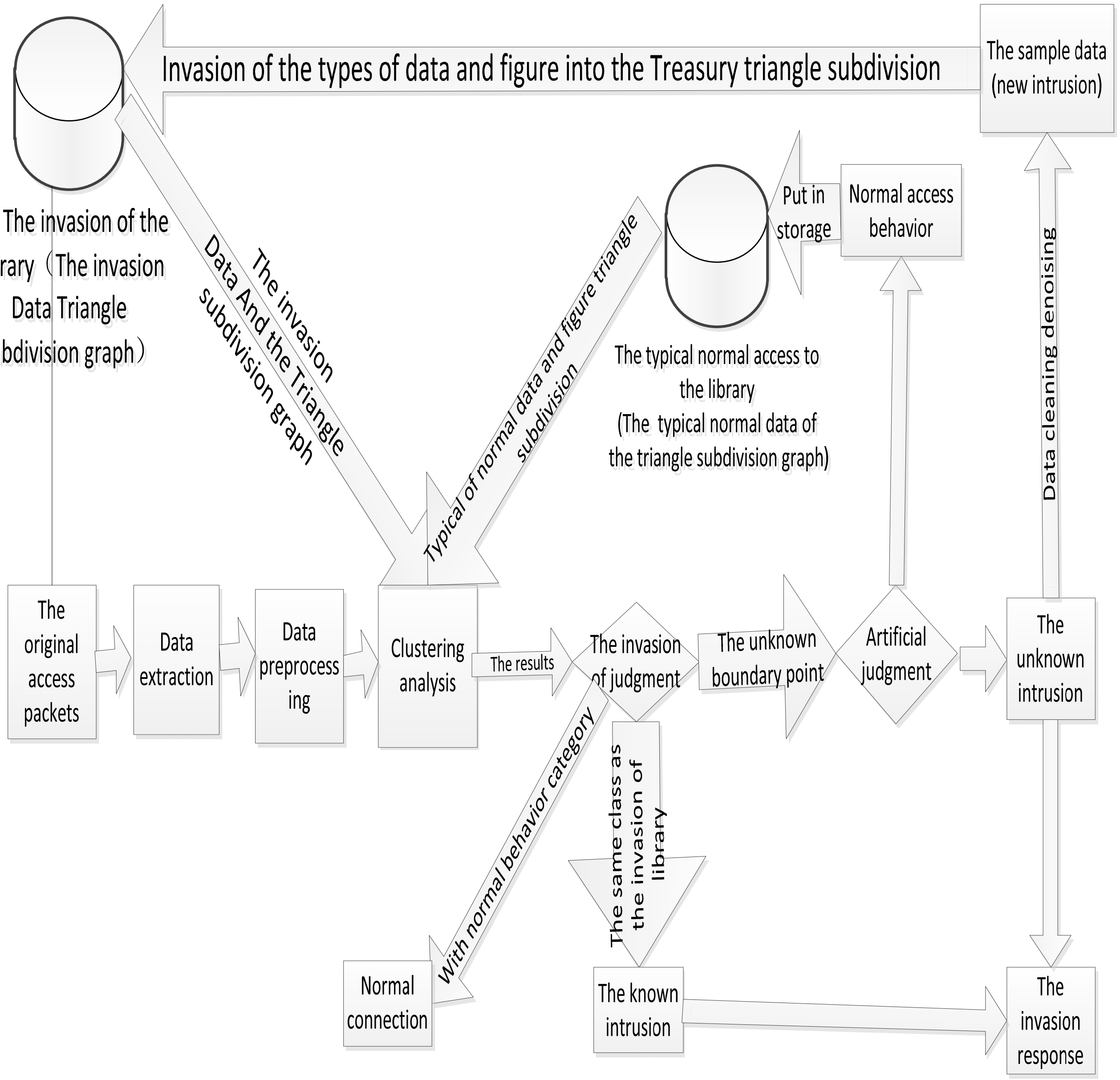
Invasion of the types of data and figure into the Treasury triangle subdivision
The sample data (new intrusion)
The invasion of the library（The invasion Data Triangle subdivision graph）
The invasion Data And the Triangle subdivision graph
Typical of normal data and figure triangle subdivision
Put in storage
Normal access behavior
The typical normal access to the library (The typical normal data of the triangle subdivision graph)
Data cleaning denoising
The original access packets
Data extraction
Data preprocess ing
Clustering analysis
The results
The invasion of judgment
The unknown boundary point
Artificial judgment
The unknown intrusion
With normal behavior category
The same class as the invasion of library
Normal connection
The known intrusion
The invasion response

Figure 1. The whole flow chart of intrusion detection model

2.2. Hierarchical clustering method

Same as partition method, hierarchical clustering is a relatively old and widely used clustering method. According to the similarity (dissimilarity) between the objects, the method decomposes the data set into a hierarchical cluster tree. According to the different method of hierarchical decomposition, hierarchical clustering methods can be divided into condensed split hierarchical clustering and hierarchical clustering. Based on hierarchy clustering method can identify the complicated shape, different sizes of clusters, and it avoids the problem of difficult to solve combinatorial optimization, in addition, this method need not choose the initial points, also need not solve the problem of local minimum, the method is relatively simple. However, once the two clusters are merged or split, it can not be modified, which makes the computation time and storage space higher. Now, it is necessary to combine the hierarchical clustering method and other clustering techniques in order to improve the clustering quality, such as BIRCH [3], CURE[4], CHAMELEON[5] and so on.

2.3. The clustering method based on density

In order to overcome the dividing method and hierarchical method cannot accurately identify the shortcoming of the spherical clusters, can discover clusters of arbitrary shape, the researchers put forward the clustering method based on density. The clustering method based on density from the density of distribution of data objects, for any area, as long as it contains some of the density is greater than a certain threshold, will add it to close to the cluster. Common density based clustering algorithms are DBSCAN[6], OPTICS[7] and DENCLUE, etc. This kind of clustering algorithm can not only obtain the random shape clusters according to the density of the data objects, but also can effectively remove the noise. However, this method is sensitive to the input parameters, and the sparse clusters are easily treated as noise in the multi density data sets.

2.4. Grid based clustering method

Clustering method based on grid data space is first divided into a finite number of grid cell, form a grid structure, and then map the data to the grid, and finally to a single grid cell clustering for the unit. The advantages of clustering algorithm is fast processing speed, the processing time has nothing to do with the number of data objects, only depends on the data on each dimension in the space of the number of grid unit. The disadvantage is that the clustering quality and precision are general, we can only find the boundary is horizontal or vertical clustering, but can not detect the oblique boundary, which requires special treatment of the boundary. The result of clustering is highly dependent on the granularity of K, And K value is not easy to determine.

2.5. Clustering method based on the model

The clustering method based on the model of the first hypothesis for each cluster a data model, and then with the model data to the corresponding cluster, the algorithm tries to optimize the fitting the given data and a data model, the best fit between the two. The clustering algorithm can reflect the distribution of the data object is built with density function to locate the clustering, so the poor scalability. The clustering method generally contains two methods: statistical method and neural network method.

3. 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 the algorithm, the algorithm is divided into four categories: density based boundary detection algorithm, grid based boundary detection algorithm, and angle based boundary detection algorithm.

3.1. The boundary point detection algorithm based on density

The boundary point detection algorithm based on density near the border is the use of clustering characteristics of the uneven distribution of density data objects to extract 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.

3.2. 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 proposed algorithm can detect boundary points quickly and efficiently on noisy data sets with arbitrary shape and density. The typical representation is GRIDEN, EDGE.

The GRIDEN algorithm can eliminate the interference of noise points in the noisy data set with arbitrary shape and different size. Quickly and efficiently detect the boundary point. At the same time, the operating speed is much higher than BORDER, BRIM and other algorithms. 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, which will bring great difficulty in getting the user to select parameters. On the other hand, the GRIDEN algorithm is not ideal, and still contains a small amount of noise.

3.3. Edge detection algorithm based on angle

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 uses some concepts such as GRIDEN, such as data space, grid unit, grid density, neighbor of grid, dense mesh and sparse grid, etc.

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.

3.4. Boundary detection algorithm based on figure

The boundary point detection algorithm based on graph by using many important features to figure reflects the similarity relationship between data objects, and then according to the special distribution characteristics of boundary point to extract boundary point. 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 belong to this kind of algorithm, in which the TRICLUST focus on the clustering function does not give detailed boundary detection methods and results.

In order to eliminate the dependence of the boundary detection algorithm on the input parameters, and to improve the detection accuracy and the efficiency of the algorithm, NPRIM algorithm is proposed. The algorithm through the establishment of the triangle profile reflects the similarity between data objects, calculating the boundary of each object, according to the boundary of extract clustering boundary point, in the process, using the K - means automatic computing boundary threshold, eliminate the influence of parameters on 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.

4. Conclusion

The clustering analysis based on intrusion detection technology and the boundary point detection techniques of concrete model were expound in this paper, and the clustering analysis technology are described in detail; Then analyzed the result of the experiment environment and experiment, application in hospital information security system, further validation of this project is based on the improved NPRIM algorithm applied to intrusion detection is effective.

Acknowledgements

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

References

[1] J. McQueen. Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, 281-297.

[2] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[3] T. Zhang, R. Ramakrishnan, M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. SIGMOD ’96, ACM Press, 1996, 103-114.

[4] S. Guha, R. Rastogi, K.Shim. CURE: an efficient clustering algorithm for large databases. Information Systems Vol.26 No.1, 2001, pp35-38.

[5] George Karypis,Eui-Hong(Sam)Han, Vipin Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling[C]. IEEE Computer 32(8):68-75, August 1999

[6] M. Ester, H. P. Kriegel, J. Sander, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of KDD ’96, AAAI Press, 1996, pp.226-231.

[7] M. Ankerst , M. Breunig , et al, “OPTICS: Ordering points to identify the clustering structure”, In Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD ’99), PP.49-60, Philadelphia, PA, June 1999.