New boundary point detection algorithm NPRIM in hospital information security intrusion detection system

Xingshan LI [a], Xiangyang SHAO

Luohe medical college, Luohe, 462000, China

[a]email:604141388@qq.com

Keywords: NPRIM algorithm, triangulation graph, threshold of boundary value, boundary point detection algorithm

Abstract. In this paper, the advantages and disadvantages of various kinds of clustering analysis and boundary point detection algorithms are analyzed, Combined with intrusion detection to deal with the characteristics of the data object containing noise, varied types,uneven density, and large amount of data. The Non parametric boundary point detection algorithm NPRIM algorithm is used by the project group to add the clustering function. It is proposed that the algorithm can not only be able to cluster and extract the boundary points without input parameters, but also can be applied to intrusion detection system.

Related concepts of NPRIM algorithm

In any data set, the distance between the interior point and the nearest neighbor is small, but the distance between the two clustering boundaries or between the cluster boundary and the noise are relatively large. So in the triangle subdivision graph, the edge which regard clustering boundary point as the endpoint, short and long (Short side are connected to the cluster internal point, long edge connected with noise or other clusters), As shown in Figure 1, the point A is the clustering boundary point, point B, E is the noise point, point C is the inner point of the cluster, and the point D is the boundary point of the noise class, Can be seen from the diagram,the difference in the long side and short which regard A,D as the endpoint, however, the difference is small with point B, C, E as the endpoint[1].The algorithm using boundary point in the triangle subdivision graph characteristics of the distribution to detect the boundary points.

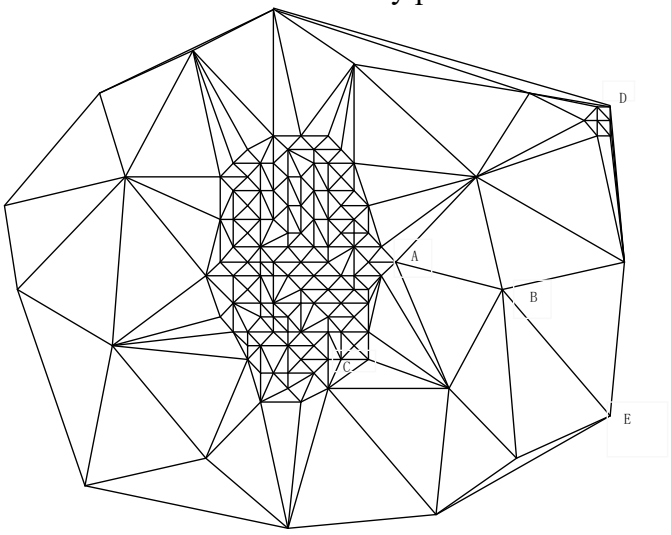

Fig.1 Triangulation graph

Definition 1: In the triangle subdivision graph, p and q are directly connected if between p and q has only one side. The neighbor of the point P is a set of points that are directly connected to the P.

In this paper, NE (P) is represented by a set of edges with P as the endpoint, NR (P) represents the neighbor set of point P (including P).

In order to determine whether the data object is a boundary point, we use the concept of the boundary degree to reflect the degree of a point. The greater the value of the boundary value, the

greater the possibility of the point is the boundary point.

Definition 2: Triangle subdivision graph, the boundary of the point p degrees can be defined as a point p and its neighbor's DL value of the ratio of standard deviation and average：

$$Bf\ (p) = \frac{\sqrt{\sum_{q \in NR\ (p)} (DL\ (q) - MeanDL\ (p))^2 / |\ NR\ (p)\ |}}{MeanDL\ (p)} \tag{1}$$

$DL(p)$ is the difference between the maximum and minimum values of the Euclidean distance $|L_i|$ at the point P as the end point: $DL(p) = \max\{|L_i|\,|\,L_i \in NE(p)\} - \min\{|L_i|\,|\,L_i \in NE(p)\}$,

$$meanDL(p) = \sum_{q \in NR(p)} DL(q) / |\ NR(p)\ |$$

As shown in Figure 1, the DL value of the points near the clustering boundary is not uniform, then the boundary of these points is larger than that of the clustering interior point and the noise points, So it is possible to deduce that the larger the boundary degree is, the more likely it is to be a boundary point. But if the data set containing no noise/isolated point, degree of clustering boundary of the border points may similar to internal points , if only Bf > T point as a boundary point was extracted and the boundary point can' be detected completely and accurately.

In this algorithm, the boundary value is defined as the candidate boundary point, which is larger than the threshold value:

Definition 3 If Bf (P) > T , then point P as the candidate boundary point. T is the threshold value of the boundary.

In the data set with noise, there are two kinds of noises in the candidate boundary points:

（1）More concentrated distribution of noise points, In Figure 1, the noise points in the D points are small clusters, and the number of small cluster points is less than $\sqrt{n}$, which is called the noise class[2].Boundary degree of the border of points near noise class will be greater than T.

（2）By clustering boundary points for the noise of the neighbors, As shown in figure 1 B attached directly to the boundary point near the border, because the DL value of the neighbor A is larger, the DL value of other noise neighbors (such as point E) is small, and the boundary of B is more than T.

From figure 1, we can see that the interior point and most of the noise points are smaller than the candidate boundary points. The noise points are sparse, and the length of the shortest profile is relatively large（As shown in Figure 1, point B and point E）.A large number of experiments show that if p subdivision edge is greater than the length of the shortest all candidate boundary point average side length of the shortest subdivision of $1+T/2$ times, point p as noise points. So the internal points and boundary point is defined as:

Definition 4 If Bf(p)<T and $\min\{|\ L_i\ ||\ L_i \in NE(p)\} < Avm*(1+T/2)$, p is the interior point. Avm is the mean value of the shortest split edge for all candidate boundary point .

Definition 5 Assuming point P is not the point of the noise class, if $\min\{|L_i|\,|\,L_i \in NE(p)\} < Avm*(1+T/2)$ and Bf(p)>T, the point P is the boundary point.

Method for calculating the threshold of the boundary degree

The existing boundary point detection algorithms need to input parameters, but in practical applications, users are often difficult to determine the correct parameters of the algorithm. In order to solve this problem, K-means was used to calculate the threshold value automatically.

K-means is a clustering technique based on the prototype, the prototype data points defined it mean, usually used for object clustering n-dimensional continuous space. Can be seen from the definition of boundary clustering that boundary degrees of near boundary points is greater than the average of the boundary of the internal point and noise point degrees, boundary degrees will be

considered a one dimensional data objects by k – means was divided into two classes (that is, the value of 2 k), thus the points with the internal boundary points/apart from the noise[3]. The boundary points correspond to the Bf value of the larger class, the internal points and noise points corresponding to the smaller Bf value of the class. Bf value larger category marked as c1, the smaller values of the Bf that marked c2，the boundary threshold value T is $T = \min\{Bf \mid Bf \in c1\}$.

The specific steps of the algorithm of boundary points detection

The main idea of this algorithm is：Through the establishment of a triangular diagram shows 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. It consists of the following six steps: the subdivision graph establishment, the boundary degree calculation, the boundary degree threshold calculation, marking the candidate boundary points, remove the noise points and extract the boundary.

Step 1 to establish a section map: Using the method of document [4] to establish the partition map. Its main idea is the discrete points according to a first coordinate values (e.g., y value), and then use a plane perpendicular to the axis of the linear scanning point set, and stopped in the event (event points) and dealing with every event, add triangle, change the event queue, until the end. The worst-case time complexity of the algorithm is up to O (N log N), and the memory requirement is not high.

Step 2 boundary degree calculation：The boundary of each point is calculated according to the definition of 2.

Step 3 boundary threshold calculation：The boundary of all points is divided into two classes by using K-means algorithm, and then the lower bound of the class of the larger Bf value is used as the boundary threshold value T.

Step 4 mark candidate boundary points：According to the definition of 3, the boundary is larger than the T point mark as the candidate boundary points.

Step 5 remove the noise points：There may be some noise points in the candidate boundary points. Only the noise points can be removed to accurately extract the boundary points. Before removing the noise points, the interior points are identified based on the definition 4, and the following two methods are used to identify and remove the noise points:

（1）From each of the unclassified internal point p begin depth-first traversal, the adjacent interior points and internal candidates for the adjacent boundary point to, clustering is less than $\sqrt{n}$ the number of points will be marked as noise, when not extract the boundary points are extracted noise points of a class, so to extract the boundary points of definition does not include the 3 mentioned in the first kind of noise.

（2）By scanning edge to extract the boundary point. will meet the partition of A will meet the partition of $\min\{|L_i| \mid L_i \in NE(P)\} > Avm*(1+T/2)$ point boundary marked Deleted in the whole data set while scanning, scanning is not marked as Deleted, this will not be scanned into the definition of second kinds of noise points mentioned in 3, in addition to second kinds of noise.

Step 6 extract the boundary point：When the data set does not contain noise and isolated points, boundary and internal point of the boundary points of the cluster boundary have little difference, if only Bf more than T points as the boundary points extracted ,the boundary points are extracted to may be incomplete. In addition, there are some cluster boundary points directly connected to the internal point, they may also be the boundary degree is greater than a threshold value T, if the simple boundary is greater than the threshold of T at all as the boundary points, it may put these points within the mistaken boundary point. In order to get complete and accurate boundary, this algorithm first scan the original subdivision graph of each edge e (p, q), if it only belongs to a triangle or two endpoints belong to different categories, the endpoint p and q markers for the edge

of the boundary points, finally from each a marked boundary point depth first traversal search to meet the definition of boundary point 5, until all 5 points are to meet the definition of ergodic, this can reduce the thickness of the border and the border to more complete extraction.

Results of boundary point detection

Fig2 Typical integrated data set

In order to verify the effectiveness of the algorithm, the algorithm of multiple comprehensive data sets and multiple real data sets were tested, we selected six typical comprehensive data set from (Fig 2) and the real data of two typical set (Fig4).

Figure 2 data set DS2[4]，DS3、DS4[5]. The DS1 has 22180 data points, the data set contains a star shape with uneven density clustering, without noise, DS2 contains 9993 data points, the data set contains four different shapes of the density of the cluster and some noise points, DS3 contains 7832 points, including a small amount of noise points, it is composed of two symmetrical diamond cluster, the two data points within the cluster from middle to gradually reduce, DS4 has 5034 data points, from 5 different shapes, different sizes and different density of clustering and a large number of noise, while the cluster has a "short bridge" , DS5 containing 11680 data points from six different shape and size and density of uneven clustering and with noise interference lines between clustering, DS6 contains 11399 data points, the data set contains 12 different shapes, the density is not uniform clustering, with a lot of noise points.

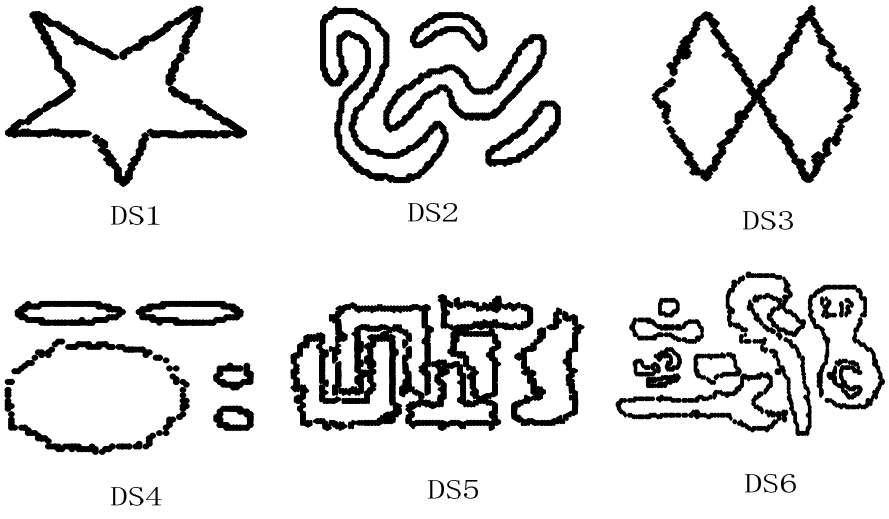

Fig 3 Boundary point detection results

Test results can be seen from the figure 3 boundary point NPRIM algorithm in the case of don't need to input parameters, not only in excluding noise, density of homogeneous data sets, can effectively extract the boundary point, in the data set which contains a large amount of noise, the density is not uniform, and the data set of arbitrary shape clusters can be effectively extracted to the boundary points, and the accuracy is higher.

In order to further verify the effectiveness of the NPRIM algorithm, the algorithm is still on the real data set to do some experiments, as shown in Fig 4. Fig 4 (b) is NPRIM test results on the character image, figure 4 (d) for NPRIM on image containing hand test results.

（a）原始图像 1　（b）原始图像 1的检测结果　（c ）原始图像 2　（d ）原始图像 2 的检测结果

Fig4. NPRIM real data set and the test results of the algorithm

As can be seen from the fig4 (b) NPRIM accurately detects the eyes, nose, mouth, and face the main outline, as can be seen from the fig4 (d) NPRIM can accurately it detects and outline of the cup. From the results it can be seen that the NPRIM algorithm can effectively detect the image contours. The image object contour in image retrieval, plays an important role in the fields of virtual reality, we have the NPRIM algorithm is applied to the field of virtual reality, and achieved good results.

Acknowledgement

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

Reference

[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] Lee W, Stolfo S J. A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security (TiSSEC), 3(4),2000: 227-261.