Clustering Boundary Detecting Algorithm for Each Cluster

Kun Wang[1,a], Baozhi Qiu[1,b], Xiangdong Shen[2,c]

[1]School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China.

[2]Henan Provincial Institute of Scientific& Technical Information , Zhengzhou 450001, China.

[a]kunw2012@163.com, [b]qbzzzu@163.com, [c]sxd371@126.com.

Keywords: Boundary detecting; K Nearest Neighbors (KNN); Reverse K Nearest neighbors(RKNN)

Abstract. Detecting the boundary of each cluster in a data set is a tough problem for many existed boundary detecting algorithms. In order to solve that problem, a clustering boundary detecting algorithm based on KNN and RKNN named CBDEC(Clustering Boundary Detecting Algorithm for Each Cluster: CBDEC) is proposed. Firstly, the KNN and the RKNN for each object in the data set will be calculated. And the boundary degree of each object will be calculated according to its RKNN value. Then, a concept named Reached Neighbors(RN) is proposed according to the neighbors' relationship between the objects. And an edge will be put between the objects which are satisfied the concept of RN. Many connected undirected graphs will be constituted in this way, and each one of them represents a cluster. Finally, The boundary of the whole data set or each cluster can be detected by boundary degree combined with the boundary percent and the cluster division. The experimental results on many data sets with noises show that CBDEC algorithm can obtain the boundary of the whole data set or each cluster with different size or shapes effectively.

1 Introduction

Advances in information technologies have led to the continual collection and rapid accumulation of data in repositories. Patterns, like cluster[1], classification[2] and outlier analysis[3], are used to find the interesting models from the repositories to help us extract the useful information. Besides, boundary detection[4] is also an emerging pattern which has quickly developed during these years. Boundary detection aims at finding the boundary objects which are located in the edge of the clusters[5]. Compared with the other objects in a cluster, boundary objects have their unique features. For example, the boundary of the patients with benign tumor may means the patients who are easily developed into malignant tumor in the field of medicine. Searching and finding those patients will contribute to the early prevention and diagnosis of malignant tumors.

Now days, methods, like BORDER[6], BOUND[7], have been proposed in the field of boundary detection. BORDER algorithm can receive a preferable result on the data set without noise compared with the data set with noises. On the noisy data set, BORDER will see all the noisy objects as the boundary objects.That is to say: BORDER can not avoid the interfere of the noisy objects. BOUND algorithm can apply to the data sets with noises or without noise, and it will get good results. But BOUND can not apply to the data set with high dimension data.

Although BORDER and BOUND algorithms can obtain the whole boundary of the data set effectively, they cannot get the boundary of each cluster in the data set. In order to extract the boundary of each cluster in a data set, a clustering boundary detecting algorithm based on K Nearest Neighbors(KNN) and Reverse K Nearest Neighbors(RKNN) named CBDEC is proposed.

The paper is organized as follows: Section 2 introduces CBAEC algorithm in detail. Section 3 compares CBDEC algorithm with other boundary detection algorithms to validate the validity of CBDEC algorithm. Section 4 time complexity analysis of CBDEC algorithm. Section 5 parameter discussion of CBDEC algorithm. Section 6 concludes the paper.

2 CBDEC Algorithm

First of all, we introduce the definition of CBDEC algorithm.

Definition 1 ***K Nearest Neighbors(KNN)***[8]: Given a data set *D*, distance metric *M*, $\forall p \in D$, *p*'s k nearest neighbors, denotes as *KNN(p)*, is a set which is formed by the closest k points to point *p*. Here the distance metric *M* is Euclidean Distance.

Definition 2 ***Reverse K Nearest Neighbors(RKNN)***[9]: Given a data set *D*, distance metric *M*, $\forall p \in D$, *p*'s reverse k nearest neighbors, denoted as *RKNN(p)*, is a set of points which point p is a KNN point to each one of them. The number of *p*'s reverse k nearest neighbors is denoted as *|RKNN(p)|*.

Definition 3 ***Reached Neighbor(RN)***: $\forall p,q \in D$, if $q \in KNN(p)$, and $q \in RKNN(p)$, then point *p* and point *q* are mutual reached neighbors.

Definition 4 ***Noisy Percent***: noisy percent, denoted as α, means the percent of the noisy objects in a data set.

Noisy objects are located far away from the clusters or the dense part of the data set. And the numbers of their reverse k nearest neighbors are less than the objects which are located inside the clusters[10]. According to the value of α and the reverse k nearest neighbors numbers of each object, We can find the noisy objects from a data set.

Definition 5 ***Border Degree***: $\forall p \in D$, *p*'s border degree, denoted as *BD(p)*, is the value of *RNKKD* divide $RKNN_P$.

$$BD(p)=RKNND\ /RKNN_P \quad (1)$$

RKNND represents the largest reverse k nearest neighbors number subtract the smallest reverse k nearest neighbors number in a data set.

$$RKNND=MAX\_RKNN-MIN\_RKNN \quad (2)$$

MAX_RKNN represents the largest reverse k nearest neighbors number.

$$MAX\_RKNN=max\{|RKNN(q)| \mid \forall q \in D\} \quad (3)$$

MIN_RKNN represents the smallest reverse k nearest neighbors number.

$$MIN\_RKNN=min\{|RKNN(r)| \mid \forall r \in D\} \quad (4)$$

RKNNp represents *p*'s reverse k nearest neighbors number subtract *MIN_RKNN* and plus 1(The reason of plus 1 is to avoid the value of $RKNN_P$ equals 0).

$$RKNN_P=|RKNN(p)|-MIN\_RKNN+1 \quad (5)$$

Definition 6 ***Border***: the boundary of the whole data set.

$$Border=bor(C_1) \cup bor(C_2) \cup ..... \cup bor(C_K) \quad (6)$$

k represents the number of the clusters in a data set. C_i is the ith cluster. $bor(C_i)$ represents the boundary set of the ith cluster. Each object from the cluster C_i has its border degree value. $bor(C_i)$ has *RANK* objects, and it is comprised by the objects which have the largest *RANK* border degree values.

$$RANK=bp*|Ci| \quad (7)$$

bp represents boundary percent, it is used to control the thickness of the boundary, and $bp \in (0,1)$. The bigger the *bp* is, the thicker the boundary of the *Ci*. *|Ci|* represents the number of objects in the cluster *Ci* .

Then the specific steps of CBDEC algorithm are as follows:

CBDEC Algorithm

Input:data set *D*, neighbor's number *k*, noisy percent α, boundary percent *bp*.

Output:boundary of each cluster in the data set, boundary of the whole data set.

Step 1: According to definition 1 and 2, calculating the KNN and RKNN set of each object in the data set. And calculating the RKNN number of each object.
Step 2: According to noisy percent α, calculate the noisy number S. Marking the minimum RKNN number objects of S as noise. Then emptying the KNN set and the RKNN set of the noisy objects and removing the noisy objects from the sets of non noisy objects. Then recalculating the RKNN number of the non noisy objects, and calculating the border degree value of every non noisy objects according to the definition 5.
Step 3: According to definition 3, an edge will be put between the points which are satisfied the concept of RN. Many connected undirected graphs will be constituted in this way. And the cluster numbers and the clustering division of the data set can be known by those connected undirected graphs for each one of them represents a cluster and has its unique class label.
Step 4: According to border degree, cluster division, boundary percent and definition 6, calculating the boundary set of each cluster. The boundary of the whole data set is the union of every cluster's boundary set.

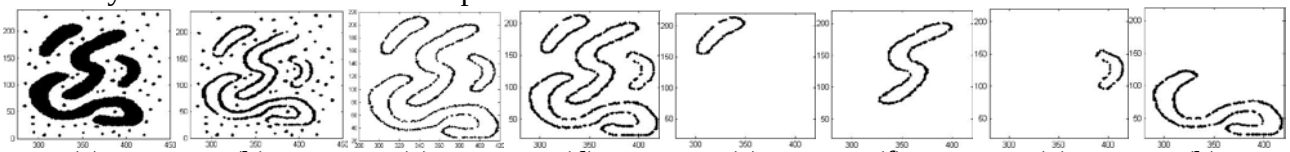
3 Experimental Results and Analysis

In order to validate the effectiveness of CBDEC algorithm, we performed experiments on multiple data sets. The data sets include comprehensive data set and real data set. First, we compare CBDEC with BORDER and BOUND boundary detecting algorithms to validate the effectiveness of CBDEC; The data of the real data set are high dimension data, BOUND fails to check the boundary of high dimension data set. So, on the real data set, we only compare CBDEC with BORDER to validate the effectiveness of CBDEC.

Experimental Environment: CPU: Intel(R) Core(TM) i3-2130 3.40GHz; Memory:4GB; Operating System: Microsoft Windows 7; Algorithm Writing Environment: MATLAB2012.

3.1 Experimental on Comprehensive Data Set

There are 4997 objects (including noisy objects) in the comprehensive data set which is showed in Fig1(a). As is shown in Fig1(a), there are 4 different clusters with different sizes and shapes in the data set. The result of BORDER(k=65, n=1400) is shown in Fig1(b); The result of BOUND (Eps=6, $minpts$=30, δ=3) is shown in Fig1(c); The result of CBDEC(k=40, α=0.04, bp=0.15) is shown in Fig1(d)-Fig1(h), Fig1(d) is the whole boundary of the comprehensive data set, Fig1(e)-Fig1(h) is the boundary of each cluster in the comprehensive data set.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 1: CBDEC compares with BORDER and BOUND

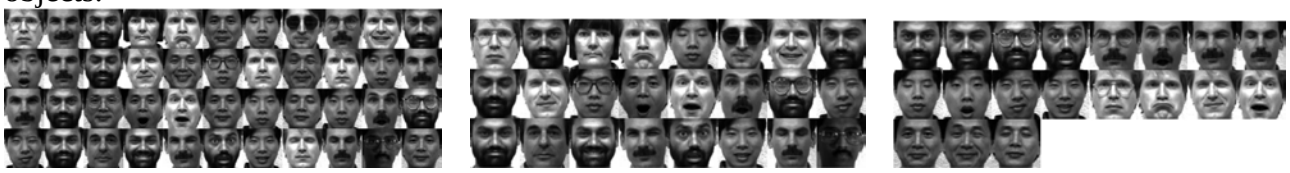
The experiment shows: BORDER, BOUND and CBDEC algorithm can find the boundary of the whole data set. BORDER can not avoid the interfere of the noises, and all the noises in the data set will be the boundary under the algorithm of BORDER. BOUND and CBDEC can avoid the interfere of noises, and they can get the boundary of the whole data set clearly. What's more, unlike the BOUND, CBDEC also can get the boundary of each cluster in the data set.

3.2 Experimental on Real Data Set

The real data set in Fig2(a) comes from yale face database. Fig2(a) contains 44 face pictures, each one of them is consisted by 100*100 pixels. The value of each pixel is between 0 and 255. Here, we treat every pixel as an attribute. So, there are 44 objects in the real data set, and each object has 10000 attributes. In this way, we transfer the picture objects into numerical objects and get the numerical real data set of yale face.

In the real data set, there are 44 different objects(including 4 noisy objects) and 5 different

classes. Each class contains 8 objects and 4 boundary objects. And the boundary of each class is: wear glasses, close eyes, wink and open mouth. So there are 20 boundary objects in this real data set. The result of BORDER(k=6, n=24) is shown in Fig2(b); BORDER algorithm obtain 24 boundary objects(including 4 noisy objects and 15 correct boundary objects).The result of CBDEC(k=5, α=0.11, bp=0.5) is shown in Fig2(c). CBDEC algorithm obtain 5 classes, and get 16 boundary objects for the first four classes(every class gets 4 boundary objects), and get 3 boundary objects for the fifth class. So, CBDEC algorithm get 19 boundary objects totally, and 15 of them are the correct boundary objects.



(a) (b) (c)

Figure 2: CBDEC compares with BORDER

In order to quantitatively evaluate BORDER and CBDEC algorithm, we use Accuracy[11], Recall[11] and F-measure[11] to evaluate the effectiveness of BORDER and CBDEC. And the result of them is shown in Table1. The experiment shows: CBDEC algorithm has higher Accuracy and F-measure. And unlike the BORDER algorithm, CBDEC algorithm can avoid the interfere of noises and find the boundary of the real data set.

Table 1: the quantitative result of BORDER and CBDEC

Algorithm	Class Number	Boundary Number	Accuracy	Recall	F-measure
BORDER	null	24(15)	62.5%	75%	68.75%
CBDEC	5	19(15)	78.95%	75%	76.97%

Notes: the number in the brackets are the boundary objects which is correctly detected.

4 Time Complexity Analysis

CBDEC algorithm has 3 phases: First, it calculates the KNN and RKNN of each object, and the time complexity of this phase is $O(kN^2)$; k is the neighbors of the object. N is the number of the objects in the data set. Second, it builds undirected graphs to determine the cluster division of the data set, and the time complexity of this phase is $O(Nk^2)$; Third, according to border degree, cluster division and boundary percent to obtain the boundary of each cluster and the whole data set, and the time complexity of this phase is $O(cN)$; In summary, the time complexity of CBDEC is $O(kN^2)$. The time complexity of BORDER is $O(kN^2)$. The time complexity of BOUND is $O(kN^2)$. So the time complexity of these 3 algorithms is equal.

5 Parameter Discussion

CBDEC algorithm has 3 parameters: K neighbor's number k, noisy percent α, boundary percent bp. The value of k determines the KNN number and the RKNN number of each object. The bigger of the k, the more KNN or RKNN numbers of each object. α means the percent of the noisy objects in a data set. If there is no noise in a data set, α equals 0. bp determines the thickness of the boundary. The bigger of the bp, the thicker of the boundary.

6 Summary

This paper proposes CBDEC algorithm on the basis of KNN, RKNN, undirected graphs and border degree. CBDEC can extract the boundary of each cluster and the whole data set with the function of avoiding the interference of noises in the data set. This is the feature of this algorithm. But CBDEC also has its limitation. It is only applying to the numerical data sets, and it can not applying to the categorical data sets or the mixed data sets. So how to solve the boundary detection problems on the categorical data sets and the mixed data sets are the following work of our research.

References

[1] Alex R, Alessandro L. Clustering by fast search and find of density peaks [J]. Science, Vol. 344 (2014) No. 6191, p. 1492-1496.

[2] Soumadip G, Sushanta B, Debasree S,Partha P S, A novel Neuro-fuzzy classification technique for data mining[J]. Egyptian Informatics Journal, Vol. 15 (2014) No. 3, p. 129-147.

[3] Mohamed B. A practical outlier detection approach for mixed-attribute data[J]. Expert Systems With Applications, Vol. 42 (2014) No. 22, p. 8637-8649.

[4] Qiu B Z, Wang B. Cluster boundary detection technology for categorical data[J]. Journal of Computer Applications, Vol. 32 (2012) No. 6, p. 1654-1656.

[5] Qiu B Z, Yang Y, Du X W. BRINK: An Algorithm of Boundary Points of Clusters Detection Based On Local Qualitative Factors[J]. Journal of Zhengzhou University, Vol. 33 (2012) No. 3, p. 117-121.

[6] Xia C, Hsu W, Lee M L, et al. BORDER: An efficient computation of boundary points[J]. Knowledge and Data Engineering, IEEE Transactions on, Vol. 18 (2006) No. 3, p. 289-303.

[7] Yue F, Qiu B Z. Boundary Points Detecting Algorithm for Clusters in Noisy Dataset[J].Computer Engineering, Vol. 33 (2007) No. 19, p. 82-84.

[8] JEV Ferreira, CHSD Costa, RMD Miranda, AFD Figueiredo. The use of the k nearest neighbor method to classify the representative elements[J]. Educación Química, Vol. 20 (2015) No. 3, p. 195-201.

[9] Xie F F, Xun L C, Niu B R. An improved detection algorithm based on reverse k-nearest neighbor[J]. Computer Applications and Software Vol. 31 (2014) No. 6, p. 267-270.

[10] Yue F, Qiu B Z. Outlier detection algorithm based on Reverse K Nearest Neighbors[J]. Computer Engineering and Application, Vol. 43 (2007) No. 7, p. 182-184.

[11] Li X L, Geng P, Qiu B Z. Clustering Boundary Detection Technology for Mixed Attributes Data Set [J]. Control and Decision, Vol. 30 (2015) No. 1, p. 171-175.