The Spatial Outlier Mining Algorithm based on the KNN Graph

Lijun Cao
Hebei Normal University of Science & Technology
Email: misscao6666@163.com

Xiyin Liu, ZhiPing Wang, Zhongping Zhang
Hebei Normal University of Science & Technology, China
Foreign Language College of Dalian Jiaotong University
College of Information Science and Engineering Yanshan University, China

***Abstract*—In order to solve the defect in the spatial outlier mining algorithm that the spatial objects may be affected by their surrounding abnormal neighbors, a Based K-Nearest Neighbor (BKNN) algorithm was proposed based on the working principle of KNN Graph, which could effectively identify the spatial outliers by using cutting edge strategies. The core idea of BKNN is to calculate the dissimilarity of the non-space attribute values the between adjacent objects, and to find the find the largest local outlier or outlier regions by cropping off the edges with the largest dissimilarity. The experiments for the spatial outlier mining algorithm BKNN based on the KNN Graph were carried out in the real datasets FMR and WNV. The example of the algorithm and the time complexity were analyzed and the results were compared to those of the existing classical algorithms, which verified that this algorithm could improve the accuracy of spatial outlier mining and simultaneously mine spatial region outliers.**

***Index Terms*—Spatial outlier; Spatial region outliers; KNN Graph; BKNN algorithm**

I. Introduction

The definition of the spatial outlier is different with that of the traditional outliers, which could attribute its complexity [1]. For the study of the spatial data, it is necessary to consider the spatial attributes and their mutual relations, as well as the non-spatial attributes and their inter-relationships. Considering the unique nature of spatial data, a definition of the spatial outlier was proposed by Shekhar et al. [2]: The spatial outliers are the spatial objects whose non-spatial attribute values are significantly different with those of other spatial objects in the neighborhood. The degree of difference between two spatial objects is usually represented as dissimilarity. With the in-depth research work, a lot of new relevant definitions and concepts for spatial outliers have been proposed [3, 4]. However, the existing definitions of the spatial outliers are all based on the basic idea of Shekhar. Spatial outliers are the objects with unstable local non-spatial attribute values or with extreme performance comparing to the neighbor reference objects, even if they were not much difference with the entire population. For example, in a community of the residents with bungalows, a high-rise building is one of space outliers based on non-spatial attribute - number of housing floors. Spatial outlier mining is very useful in the geographic information systems and spatial database, including the fields of transportation, ecology, public safety, public health, climate, and location-based services [5].

Spatial outlier mining is an expansion of the outlier mining in spatial datasets. The early spatial outlier mining algorithms are all performed for the global datasets, which has no distinction between spatial attributes and non-spatial attributes, leading to the mining results of global outliers. Subsequently, a series of improved algorithm were proposed, which could distinct the spatial and non-spatial attributes, and thus mine the local spatial outlier. However, there are still some problems. For example, for the calculation of the difference between the spatial object and its surrounding neighbors during the spatial outlier mining process, if one of the neighborhood object has too large or too small non-spatial attributes, the whole neighborhood non-spatial attribute values of the object, as well as the outlier degree calculation for the spatial objects would be affected, resulting in false detection and undetection. Thus the true outliers would be ignored, while some non-outlier would be mistaken for spatial outlier.

The development and innovation works in this paper are in in the following areas:

A spatial outlier mining algorithm - Based K-Nearest Neighbor (BKNN) algorithm was proposed in this paper based on the KNN Graph. Based on the traditional spatial outlier mining algorithm and combining the

Manuscript received January 1, 2013; revised May 1, 2013
Accepted June 1, 2013

doi:10.4304/jsw.8.12.3158-3165

aufbau principle of the KNN Graph, this algorithm is performed as follows: Firstly, the spatial neighborhood of the spatial objects was determined by their spatial attributes, and the difference of the non-space attribute values between adjacent objects were set as weight value of edge formatted by the two adjacent points. Then the edges of maximum weight were cropped through cutting edge strategies. After several iterations, the points and areas zero out-degree were set as the spatial outliers and spatial outlier regions for mining. The algorithm could not only avoid the influence of the abnormal neighbors on the spatial object, but also identify the spatial outlier regions during the spatial outlier mining procedure.

In order to further verify the effectiveness of the algorithm, the experiments for the spatial outlier mining algorithm based on the KNN Graph were performed in the real datasets FMR and WNV. All the results of the simulation experiments were compared with those of the existing classical algorithm. From the time complexity point of view, the BKNN spatial outlier mining algorithm based on KNN Graph did not show sufficient superiority. However, from the perspective of the mining results, it can be seen that the algorithm could improve the precision of spatial outlier mining and simultaneously mine spatial region outliers.

II Spatial Outlier Mining

A. The Characteristics of the Spatial Data

Compared with traditional relational data, spatial data has two important features: Spatial Autocorrelation and Spatial Heterogeneity. The so-called Spatial Autocorrelation refers to the mutual relations of the adjacent objects, namely that an attribute value of a spatial object depends on those of the adjacent ones. As described in the famous "The First Law of Geography" by Tobler, all spatial objects are associated. The closer the distance between the two objects in the space, the stronger they associate with each other, and vice versa. This spatial dependence or spatial autocorrelation are local properties for space objects, which is contrary to geographic independence of the traditional data mining theory. The so-called spatial heterogeneity refers to that the attribute values of the objects in different regions have different variation trends, showing the diversity and locality of spatial objects.

The attributes of a spatial object can be divided into two categories according to their nature: spatial attributes and non-spatial attributes. Spatial attributes include shape, position, orientation, spatial adjacency relationships and other geometric or topological properties; while the non-spatial attributes include name, age, length, ownership, etc. The non-spatial attributes are inherent attributes of the spatial object, which essentially described the spatial data objects; while the spatial attributes not inherent for the spatial object space, but provide the position index of the spatial object, which is used to determine the neighborhood relationship of each the spatial object. The spatial neighborhood determining is the prerequisite for successful implementation of the spatial outlier mining algorithm. The spatial outlier detection is calculated based on the non-spatial attribute values between the space objects, and the difference of the non-spatial attribute values will cause the localized instability.

Because of the above mentioned characteristics of the space data, the characteristics of spatial data should be well considered for the spatial outlier mining algorithm, in order to mine the ture spatial outliers.

B. The Defects of Traditional Spatial Outlier Mining Algorithms

Despite the existing algorithms could achieve relatively good results in mining accuracy and computational complexity, there are still some defects. In order to better illustrate the defects of the existing algorithms, a real spatial dataset was applied for the description in detail.

Figure 1 shows a spatial dataset, where the X-Y axes represent two-dimensional space coordinates, while the Z-axis represents the non-spatial attributes of the spacal object, assuming that the non-spatial attributes is one-dimensional variation in this case for convenience. Known that the Figure 1 contains s18patial object, wherein O1(180) a spatial outlier, O2(20) and O3(20) forming the space outlier region, and E1(20) a normal space object. Both the spatial objects O1 and E1 would be detected as spatial outliers by mistake using the traditional spatial outlier mining algorithm. The reason is that non-spatial attribute value of the spatial objects O1 is too large comparing with those of other neighborhood space objects, which virtually increases the non-spatial attribute values of space region surrounding E1, lead to the large calculated differences between E1 and its surroundings. Hence, spatial objects E1 would be mistaken for spatial outlier. Moreover, the existing spatial outlier mining algorithm is only able to identify one of the spatial object O2 or O3, while the spatial outlier regions composed by O2 and O3 could not be of identified. There are two reasons: On the one hand, as O2 and O3 are spatial neighbors and have small non-spatial attribute values difference, any space object would influence the average neighborhood non-spatial attribute values of another one, thereby affecting the final mining results. On the other hand, most of the existing spatial outlier algorithms can not identify the spatial outlier regions.

It can be seen that the average non-spatial attribute values of the whole neighborhood under study would increase or decrease when one of the spatial objects has too large or too small non-spatial attribute values, leading to the false detection and undetection during the spatial outlier mining procedure, as well as the inaccurate results of the spatial outlier mining algorithm

In order to solve the defect in the existing spatial outlier mining algorithm, a Based K-Nearest Neighbor (BKNN) algorithm was proposed in this paper based on KNN Graph.

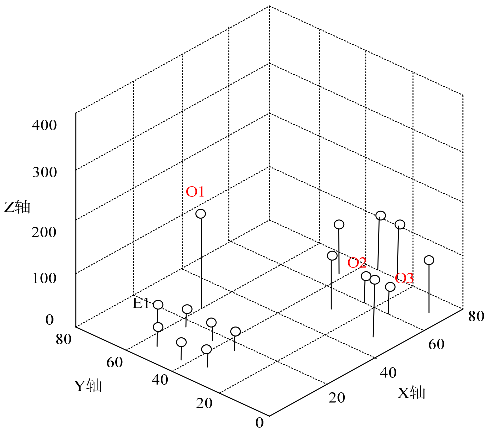

Fig. 1 A set of spatial data

III THE DESIGN SPATIAL OUTLIER MINING ALGORITHM BASED KNN GRAPH

A. *The Description of the Problem*

Given a spatial dataset X, where each spatial object could be represented by a single or multi-dimensional non-spatial attributes, denoted by $x_i \in R^d$. It is assumed that each spatial object x_i is constituted by d non-spatial attribute values, denoted by $(y_1, y_2, ... y_d)(d \ge 1)$, where T the transpose operation symbols.

Given an integer k, which represents the neighbor objects number of each object in the its spatial neighborhood, namely that NNk (xi) is the k nearest neighbor set of xi, where i=1,2,…,n.

Given an integer m, which represents the detectable number of outliers, generally m << n.

The KNN graph based spatial outlier mining algorithm has two main objectives: One is to construct KNN Graph of the spatial data sets. Firstly, the spatial neighborhood of each spatial object should be determined through space attributes, and then KNN Graph could be constructed according to the k neighbor relationships of the spatial object. The other is to find a data set Z containing m spatial outliers by using cutting edge strategies, satisfying the condition of $Z \subset X$ and $\forall x_i \in Z$, where xi has no connection with other objects.

B. *The Data Structure of KNN Graph*

The K-Nearest Neighbor Graph (KNN Graph) is a weighted digraph. Assuming that there are n vertices in the graph, for each vertex vi ($1 \le i \le n$), its k nearest neighbors v1, v2, ..., vk, could be found through spatial neighborhood calculation. Then the vertex vi was connected to the k neighbor objects, respectively, forming the k directed edges, where each edge was pointed from vi to its neighbor object. Set the difference value between the vertex vi and its k neighbor objects as the weight values of the edges, a KNN Graph could be constructed.

Set Figure 2 for example, points vi and vj represent two spatial objects of the spatial dataset, respectively, and vj is the nearest neighbor objects of vi, constructing a simple KNN Graph, where eij represents the directed edge between two space objects, whose pointer indicates that vj is an object of vi in its neighborhood space. The wij in the graph represents the weight of the edge eij, indicating the difference value between the two spatial objects vi and vj.

The data structure of KNN Graph is shown in Table 1. Each point vi is represented by an array consisting of 3k+3 elements, Vi=<Id,value,n1,diff1,flag1,…,nk,diffk,flagk,Cid>, where Id the unique identifier for each point, Value the value of non-spatial attributes of vi, nj (1≤j≤k) the j-th neighbor of vi, diffj the dissimilarity between vi and vj, namely the weight of edge eij, flagj the state of edge eij (flagj = 1 represents the edges have been cut, flagj = 0 represents the edge is still connected), Cid the number of the subgraph where vi belongs to. the neighbor list is in descending order according to the weights of the edges, namely diff1≥diff2≥…≥diffk.

C. *The Ideas of BKNN Algorithm*

The core idea of the KNN graph based spatial outlier mining algorithm BKNN is to calculate the dissimilarity of the non-space attribute values the between adjacent objects, and to find the find the largest local outlier or outlier regions by cropping off the edges with the largest dissimilarity. For the spatial dataset X, the spatial neighbor relationship of each object in the dataset was firstly calculated. After determining the k nearest neighbor of each object, the dissimilarities between each object and its k neighbors were calculated and then the dissimilarity between the two objects was set as the weight of the edge, forming the KNN Graph G. As the weight value depends on the dissimilarity of non-spatial attribute values between two adjacent objects, the greater weight value, the larger difference between two adjacent objects. Given a threshold T to perform cutting edge functions, all the edges with the weight values greater than the threshold T were cut off. Then returned to the starting point of the cutting edge and put it in to the suspected outliers-set for further verification.

To determine whether a point is the outlier, it is mainly to investigate whether the point is isolated, that is, whether the point has output edges. If the point has no output edge, this point could be considered outlier. If the point still has output edges, this point is not the real spatial outlier.

Real space outlier regions is a region connecting by less than k spatial objects, whose inner attributes are very similar while extremely different with those of the outside neighborhood. To determine whether an area is a real spatial outlier region, The findcandidateregions (G) function should be firstly called to find the suspected outlier regions. Then, the further verification includes three steps:

(1) For a selected outlier regions, determine its spatial neighborhood, namely the set that disjoint k neighbors of each object in the spatial neighborhood. If the number of

spatial neighborhoods of this region is smaller than that of the objects, this suspected outlier region is not the real one, which could be excluded. Otherwise, enter the next step to continue verification.

(2) Check the similarity of the objects in the suspected outlier region. If the the similarity is less than a given threshold Tsim, this suspected outlier region is also not the real one, which could be excluded. Otherwise, enter the next step to continue verification.

(3) Calculate the dissimilarity between the suspected outlier region and its spatial neighborhoods. If the dissimilarity is smaller than a given threshold value Tdiff, the candidate outlier region is not the real one, which could be excluded. Only the above three conditions are all satisfied after the verification, the outlier region could be considered as a real one.

D. Description of BKNN Algorithm

In this section, the BKNN algorithm will be introduced in detail according to the idea in the last section which applies the spatial outlier mining based on KNN diagram.

Algorithm 1: the spatial outlier mining algorithm based on KNN diagram.

Input: Spatial data set X, non spatial attributes set Y, the number of neighbors k, the edge threshold T, regional similarity threshold Tsim, the dissimilarity threshold between a region and its neighborhood Tdiff;

Output： Outlier set Os, Outlier region Or

```
BKNN(X,Y,k,T,Tsim,Tdiff)
Begin
(1) Scan the entire spatial data set X, and calculate the spatial neighbor relation of each spatial object;
(2) Construct KNN diagram according to the neighbor relationship obtained in (1)
(3) Scan the entire KNN diagram;
(4) Sort the edge with their weights in descending order
(5) While weight of an edge: edge_ij>threshold: T do
(6) delete edge_ij, free its starting point xi to O
(7) EndWhile
(8)   While O is not empty do
(9)   if there is no edge connecting to xi then
(10)          xi is marked as an outlier, and sent to Os
(11)  Else
(12)     { Calculate area of the region by function findcandidateregions(G), and find our the candidate outlier regions of k
(13)     For each candidate outlier region do
(14)        {  if the similarity of the region > given threshold Tsim Then
(15)        { calculate its neighbor region
(16)     if the dissimilarity between the current region and its neighbor region > given threshold Tdiff Then
(17)        Mark the current region as an outlier region, and sent it to Or
(18)        }
(19)     }
(20)     EndFor
(21)  }
(22)  EndIf
(23) EndWhile
(24) Output{Os, Or}
End
```

E Analysis on the Examples

In order to better describe the spatial outlier mining algorithm BKNN based on KNN diagram, the spatial data set in Figure 1 is taken as an example for a detailed analysis of the algorithm.

From the main idea of spatial outlier mining algorithm BKNN based on KNN diagram, we first use analysis on spatial attributes to determine the neighbor relationship of objects in the spatial data set. The mining calculated on spatial outliers is conducted by comparing the non-spatial attributes between adjacent space objects. If the neighbor number k is equal to 3, according to Figure 1, the spatial neighbor relationship can be shown as Figure 2 as by KNN diagram. In Figure 3, each circle represents a space object in this data set, and each value is non-spatial attribute of a spatial object. The bidirectional arrow arcs indicate that two spatial objects are spatial neighbors for each other, while single arrow arcs represent that two spatial objects are arc-tail connected spatial neighbors. Arc weights mean the difference between two adjacent space objects. For example, there is a single arrow arc between E1 and O1, and tE1 is tail, O1 is the head. This means that E1 is one of k nearest neighbors of O1, but O1 is not one of k nearest neighbors of E1. If there are two other arcs which are bidirectional, E1 and the other two connected objects are spatial k nearest neighbors.

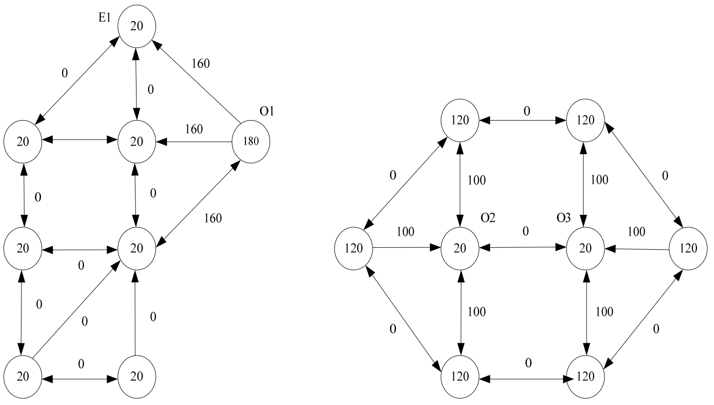

Fig. 2 The KNN graph representation of the spatial data set

Sort the weights for each arc in descending order, execute the function to cut the maximum edge, and send the object in the arc head to outlier set O for detection of spatial outliers and outlier. Figure 3 shows the KNN diagram obtained by executing spatial outlier mining algorithm BKNN. In Figure 3, because the non spatial attribute O1 and its spatial neighbors differ greatly, edge cutting function will cut all edges connecting to its spatial neighbors, and return O1. At the same time, because the degree of space object O1 is zero, which means it is an isolated point, the space object O1 is sent to the spatial outlier set Os. Since O2 and O3 are mutual k nearest neighbors, and they have obvious difference with their spatial neighbors, also higher internal similarity, they form a spatial outlier region, and can be sent into the spatial outlier region set Or.

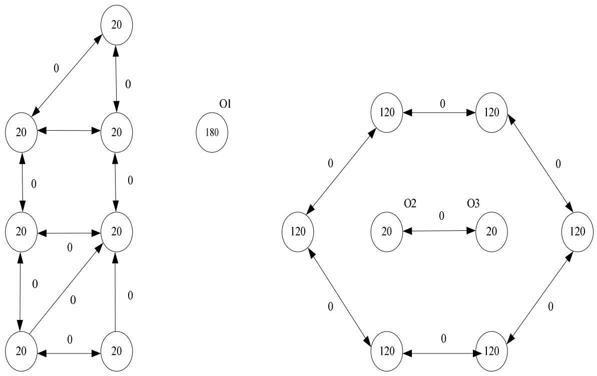

Fig. 3 The result of KNN based graph spatial outlier detection

F. Analysis of BKNN Algorithm

The principle of BKNN algorithm based on KNN diagram is to utilize the method of calculating the degree of difference between adjacent objects and cutting down the largest edge for a comprehensive spatial outlier mining. The edge cutting strategy is mainly to solve the shortcoming of current algorithms which are always influenced by abnormal neighbors and thus lead to mistakenly outlier identification. The algorithm proposed conquers this shortcoming and can avoid the spatial outlier identification error in theory.

Computation complexity of BKNN algorithm contains six parts as below.

(1) In calculation of each space object's k nearest neighbors, the computation complexity depends on the k nearest neighbor query. If the R*- tree index query is used, the complexity is O (nlogn);

(2) KNN graph structure requires complexity of O (kn);

(3) the edge weights sorting is with complexity of O ((kn) log (kn));

(4) when the total number of cutting edges is n, complexity in edge cutting is O (n);

(5) complexity in checking whether each zero degree point is the outlier or not is O (KN);

(6) complexity in identifying the candidate outlier region is O (n);

(7) complexity in verification of outlier region is O (MK), when assuming the number of object in outlier region is M (m<<n),

In total, the computation complexity if O(nlogn)+O(kn)+O(n)+ O((kn)log(kn)) + O(kn) + O(n)+O(mk)。Since k<<n and m<<n， the total complexity is about O(nlogn)。

From the computation complexity, although BKNN algorithm does not show superiority compared with the traditional spatial outlier mining algorithms, this algorithm solves the problem in existing algorithms and avoids the affects by abnormal spatial neighbors. It can find out the spatial outliers mining area when searching for outliers, and therefore improve the accuracy of mining.

IV. The Experiment and Analysis of the Spatial Outlier Mining Algorithm based on KNN Diagram

A. The Experiment Environment and Experiment Data Set

The spatial outlier mining algorithm BKNN based on KNN diagram is conducted with experimental hardware environment of the Intel Core2 Duo CPU E4500 2.20GHz and 2G of memory, and software environment for the operating system Microsoft Windows XP Professional SP2, while Microsoft Visual C++ 6 is used as the coding environment..

In order to verify the effectiveness and feasibility of the proposed algorithm, the experiment compares it with current algorithms in terms of running time and accuracy of outlier region judgment. Data set is 2008 FMR (Fair Market Rent) data set provided by the United States housing rental housing and City Development Department of the United States of America, which records housing rental in 3095 counties. Each record is composed of different living room rental prices in the various counties of US, such as the one-bedroom, two-bedroom, three-bedroom, and four-bedroom apartment. In addition, each record also includes some other important information, such as the name of the area counties, district name and counties codes.

B. Results and Analysis

In BKNN, the spatial position of each county is provided by the geographical position of the United States Census Bureau Web site files. The center positions of spatial objects are calculated through the longitude and latitude. According to the geographical location relations, the value of k is set to 8, which means there are 8 geographical spatial neighbors for each county, namely east, south, west, north, northeast, southeast, southwest and northwest. The distance between a county and its surrounding neighbors is obtained by calculating the Euclidean distance between their center positions. For the sake of simplicity, only one-bedroom apartment rent is set as the object for computing the difference.

In order to effectively verify the effectiveness and correctness, the experiment results are divided into outliers and outlier region, which are then compared with results of other algorithms.

In the identification of individual spatial outlier results, Z-value algorithm, Moran algorithm, POD algorithm, are compared with our proposed algorithm in Table 1.

TABLE 1

THE COMPARISON OF EXPERIMENTAL RESULTS(O_S)

Z-value	Moran	POD	BKNN
Pitkin Co.,CO(1178)	Pitkin Co.,CO(1178)	Pitkin Co.,CO(1178)	Pitkin Co.,CO(1178)
McDonald Co.,MO(424)	Wilbarger Co.,TX(443)	York Co.,SC(712)	York Co.,SC(712)
Ravalli Co.,MT(530)	Blbine Co.,NE(474)	Tammany Co.,LA(898)	Baker Co.,FL(498)
Dakota Co.,NE(516)	Gallatin Co.,MT(596)	East Carroll Co.,LA(416)	Glascock Co.,GA(393)
York Co.,SC(712)	Mercer Co.,MO(445)	Glascock Co.,GA(393)	Johnson Co.,KY(400)
Franklin Co.,MO(617)	East Carroll Co.,LA(416	Bristol Co.,MA(1130)	East Carroll Co.,LA(416)
Bristol Co.,MA(1130)	Tammany Co.,LA(898)	Humphreys Co.,MS(386)	Tammany Co.,LA(898)
San Francisco Co.,CA(1338)	Monroe Co.,FL(1096)	Johnson Co.,KY(400)	Bristol Co.,MA(1130)
Moore Co.,TX(480)	San Francisco Co.,CA(1338)	Baker Co.,FL(498)	Humphreys Co.,MS(386)
Tammany Co.,(898)	York Co.,SC(712)	Franklin Co.,MO(617)	Franklin Co.,MO(617)

From Table 2, it can be seen that in the detection of single spatial outliers, the proposed algorithm can reach similar result with the other three algorithms, which proves the its correctness and effectiveness.

The main idea of the Z-value algorithm and the Moran scatter diagram is to compare the non spatial attributes of the object with its surrounding neighbors' average value to determine whether the object is an outlier. Therefore, the above two algorithms are dependent on non-spatial attributes values of all neighborhood objects in the neighborhood regions. If the non-spatial attributes of a spatial object is too large or too small, the average non-spatial attributes will change, which brings the influence of spatial outlier identification, even leads to false outlier identification. The proposed algorithm utilizes the maximum edge cutting strategy, which is suitable to distinguish the effects of each neighbor objects. If the non-spatial attributes of a neighbor object value is too large or too small, the edges between them will be cut down. In this manner, normal objects are not influenced by outliers, which can avoid the mistake of outlier identification. For example, in the result of Z-value algorithm, the eighth spatial outlier with its 8 neighborhood objects is shown in Figure 4.

Figure 4 describes one-bedroom apartment rental data in San Francisco county with its 8 neighbors in California. The threshold T in BKNN algorithm is assumed as is set to 150$, which means that it's normal when the difference between an object and its adjacent counties is less than 150$. From the observation and comparison, we can see that the rent difference between San Francisco and most of its neighbors are less than 150$. Therefore, San Francisco County is a normal record, and it is not a spatial outlier. However, because the Z-value algorithm is the first to calculate the average of San Francisco County's 8 neighbors, and then compares it with the rental value of San Francisco, it is recognized as an spatial outlier while the difference between San Francisco's neighbor Mendocino (783) and Lake (661) is vary large. In the calculation of average value, the average neighborhood result is too small, and thus has a negative impact on spatial outlier detection, which is the reason of the mistaken identification of San Francisco County.

Mendocino(783)	Lake(661)	Solano(1184)
Marin(1338)	San Francisco(1338)	Alameda(1204)
San Mateo(1338)	Santa Cruz(1229)	Santa Clara(1171)

Fig. 4 The one-bedroom rent data of San Francisco county and its neighbors

In results of identifying outlier region, the Z-value algorithm, Moran scatter diagram algorithm, ROD algorithm, and our proposed algorithm are selected in comparison, among which, the ROD algorithm is selected to evaluate the outlier region algorithm, whose validity and correctness have been validated. During the evaluation of candidates, the proposed algorithm first evaluates whether the number of objects in a candidate region is less than k. If yes, go to the next evaluation. Otherwise, delete this candidate. Secondly, the formula $Rsim = \frac{Max - Min}{Mean}$ is used to calculate the suspected region similarity of Rsim, which is then compared with the similarity threshold Tsim. If Rsim<Tsim, go to the next detection step. Otherwise, delete the candidate outlier regions. For example, Tsim in the experiments is assumed as 0.15, in other words, only when the difference between the candidate region's largest and smallest non-spatial attributes is less than 15% of the average value of all objects in this region, the next detection step will be proceeded. The final detection step is to calculate difference between outlier region and its surrounding neighborhood: Rdiff by the formula $Rdiff = \frac{|m_r - m_{nbr}|}{(m_r + m_{nbr})/2}$. If Rdiff>Tdiff, it is a real spatial outliers region. Otherwise, the candidate outlier regions are excluded. For example, in the experiment, Tdiff is assumed as 0.4, in other words, only when the difference between the candidate region and its surrounding neighbor regions is greater than the average value of 40%, the candidate region is the true spatial outliers region. Otherwise, the candidate region is excluded. Table 2 shows the comparison of outlier region by the four algorithms.

TABLE 2

THE COMPARISON OF EXPERIMENTAL RESULTS(OR)

	Algorithm			
	Z-value	Moran	ROD	BKNN
1	---	---	Forsyth Co.,GA(777)	Forsyth Co.,GA(777) Fulton Co.,GA(777)
			Fulton Co.,GA(777)	Dougherty Co.,GA(777) Fayette Co.,GA(777)
			Dougherty Co.,GA(777) Fayette Co.,GA(777)	Gwinnett Co.,GA(777)
			Gwinnett Co.,GA(777)	
2	---	---	Gallatin Co.,IL(439) Hamilton Co.,IL(440)	Gallatin Co.,IL(439) Hamilton Co.,IL(440)
			Edgar Co.,IL(440)	Edgar Co.,IL(440) Hardin Co.,IL(443)
			Hardin Co.,IL(443)	

Table 2 lists the results on first two groups of spatial outliers detection. The first outlier region includes 5 counties, while the second region includes 4 counties. Among them, five neighbor counties in the first outlier region have one-bedroom apartment rent value of 777$. This region has 12 neighbor regions with one-bedroom rent range from 331$ to 562$. After calculation, the region similarity Rsim is 0%<15% (Tsim), and the difference degree with its neighborhood Rdiff is 52%>40% (Tdiff). Therefore, the region is a spatial outlier region. From Table 3, Z-value algorithm and Moran scatter diagram algorithm cannot detect spatial outliers region, only the ROD algorithm and the proposed BKNN algorithm have the capability to well detect spatial outliers region.

The above comparison is based on the accuracy analysis on BKNN algorithm. Next the algorithms will be compared respect to the computation complexity.

It is not difficult to find that the four algorithms in experiment use spatial attributes and spatial relationship to determine the spatial neighborhood, which are with complexity of O(nlogn) in Figure 6. From Figure 5, compared with Z-value algorithm and Moran algorithm, the proposed algorithm uses less time when N is increasing due to the application of edge cutting strategy. However, it spends more time than POD because it needs to judge outlier region.

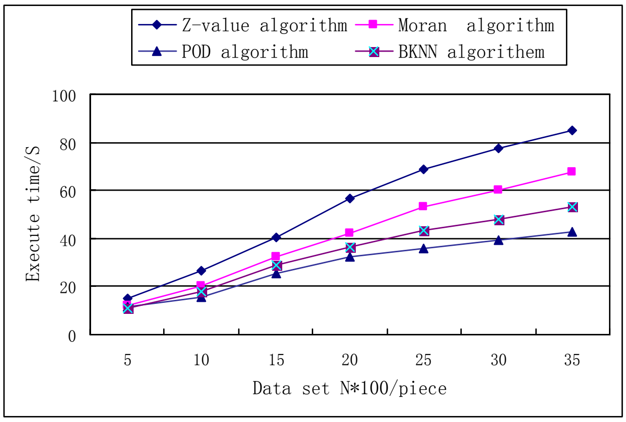

Fig.5 The runtime of BKNN algorithm

From the analysis of the computation complexity, the proposed algorithm does not show enough superiority. However, from the perspective of mining results, it can be seen that the proposed algorithm improves the spatial outlier mining accuracy, and can find out the spatial outlier region in the same time.

V CONCLUSIONS

This paper analyzes the shortcoming of the traditional spatial outlier mining algorithms, and proposes a new mining algorithm: BKNN based on the KNN diagram structure principle and the edge cutting strategy. The algorithm makes the spatial objects avoid the effects of its abnormal neighbors, and effectively identifies the spatial outlier region during the outlier identification.

ACKNOWLEDGEMENT

This work was financially supported by the Project of Major Reform to the Network Enginnering Programme in Hebei Normal University of Science and Technology, which is assigned by Hebei Education Department in 2012.

Human resources needs of the IT industry in Hebei Province and the adaptability of college computer education model, Hebei Human Resources and Social Security Department project, project number: JRS-2012-1103

REFERENCES

[1] Ma Ronghua, He Zengyou “from the GIS database mining spatial outlier An efficient algorithm “,Wuhan University, 31 (8) , pp.679-682,2006

[2] Li Deren, Shuliang, Li Deyi. “Spatial data mining theory and application”, Beijing: Science Press, pp.19-37, 2006

[3] Sun Pei, Chawla S. On Local Spatial Outliers. In Proc. 4th IEEE International Conference on Data Mining, Brighton, pp.209-216,2004

[4] Hu Tianming, Sung S Y. “A Trimmed Mean Approach to Finding Spatial Outliers”. Intelligent Data Analysis, pp.79-95, 2004 (8)

[5] Shashi Shekhar, Sanjay Chawla. “Spatial Databases: A Tour.Upper Saddle River”, N.J.:Prentice Hall, pp.435-445,2003

[6] Edward H. Herskovits, Joan P. Gerring. “Application of A Data-Mining Method Based on Bayesian Networks to Lesion-Deficit Analysis”, NeuroImage, 19(4) , pp.1664-1673,2003

[8] M.Kanevski, R. Parkin, A. Pozdnukhov, et al.,”Environmental Data Mining and Modeling Based on Machine Learning Algorithms and Geostatistics”, Environmental Modelling&Software, 19(9) , pp. 845-855,2004

[9] Anthony K.H. Tung, Jean Hou, Jiawei Han. “Spatial Clustering in The Presence of Obstacles”, In Proc. Of the 17th International Conference on Data Engineering, IEEE Computer Society, Washington, DC, U.S.A., 11, pp.359-369, 2001

[10] James Malone, Ken McGarry, Chris Bowerman,”Automated Trend Analysis of Proteomics Data Using An Intelligent Data Mining Architecture”,Expert Systems with Applications, 30(1) , pp.24-33,2006

[11] Li Ai Bing,”Evaluation on Slope Stability Using Visual Quick Nerval Network”, Kuangye Yanjiu Yu Kaifa, 21(5) , pp.8-11,2001

[12] Charu C. Aggarwal, Philip S. Yu.,”Finding Generalized Projected Clusters in High Dimensional Spaces”, In Proc. Of the ACM SIGMOD International Conference on Management of Data, Dallas, Texas: ACM Press, pp.70-81,2000

[13] Hu Tianming, Sung S Y,”A Trimmed Mean Approach to Finding Spatial Outliers”,Intelligent Data Analysis, pp.79-95 ,2004(8)

[14] Dechang Chen, Chang-Tien Lu, Yufeng Kou, Feng Chen,”On Detecting Spatial Outliers”, Geoinformatica, 12(4) , pp.455-475,2008

[15] Anrong Xue, Shiguang Ju, “Algorithm for Spatial Outlier Detection Based on Outlying Degree”, In Proc. Of the WCICA, Dalian, 12(7) , pp.6005-6009, 2006

[16] Chang-Tien Lu, Dechang Chen, Yufeng Kou, ” Algorithms for Spatial Outlier Detection”,In Proc. Of the 3rd International Conference on Data Mining(ISDM 2003), Melbourne, Florida, U.S.A., IEEE Computer Society, pp.597-600, 2003

[17] Jun Wang, WeiRu Chen,” A Reliability-oriented Evolution Method of Software Architecture Based on Contribution Degree of Component ”, Journal of Software, Vol 7, No 8 (2012), pp.1744-1750, Aug 2012

[7] Vladimir Estivill-Castro, Ickjai Lee. “Clustering With Obstacles for Geographical Data Mining”, ISPRS Journal of Photogrammetry and Remote Sensing, 59(1-2) , pp.21-34,2004

[18] Qinghua Lu, Phillip John McKerrow, Zhiquan Zhou,” Design and Performance of a Minimal Real-Time Operating System in a Safe Language: Experience with Java on the Sun SPOT ”, Journal of Software, Vol 7, No 8 (2012), pp.1835-1844, Aug 2012

[19] Mingcheng Qu, Xiang-hu Wu, Xiao-zong Yang,” A Comprehensive Optimization Model Based on Time and Cost Constraints for Resource Selection in Data Grid”, Journal of Software, Vol 6, No 12 (2011), pp.2472-2478, Dec 2011

[20] Jackson Phiri, Tie-Jun Zhao, Jameson Mbale,” Identity Attributes Mining, Metrics Composition and Information Fusion Implementation Using Fuzzy Inference System”, Journal of Software, Vol 6, No 6 (2011), pp.1025-1033, Jun 2011

[21] Richard Lai, Sajjad Mahmood, Shaoying Liu,” RAAP: A Requirements Analysis and Assessment Process Framework for Component-Based System (Invited Paper)”, Journal of Software, Vol 6, No 6 (2011), pp.1050-1066, Jun 2011

[22] Jianguo Chen, Hao Chen,”,A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning”, Journal of Software, Vol 8, No 1 (2013), pp.55-62, Jan 2013

Lijun Cao, Born in 1971 , master, associate professor, college of mathematics and information science of Hebei Normal University of Science and Technology, main research directions are: data mining, grid technology.

Xiyin Liu, Born in 1970 , master, associate professor, College of mechanical and electrical engineering, of Hebei Normal University of Science and Technology, main research directions are: data mining, grid technology.