ISSN : 0974 - 7435 Volume 10 Issue 24

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(24), 2014 [15129-15135]

Spatial neighborhood classifiers

Liao Wei-Hua*

Guangxi Colleges and Universities Key Laboratory of Mathematics and Its Applications, Guangxi University, Nanning, 530004, (CHINA)

E-mail: gisliaowh@163.com

ABSTRACT

Spatial data classification is a high-frequency spatial decision evaluation method. It can only choose according to experience frequently. When there were many spatial decision evaluation conditional attributes, such as hyperspectral image, it has obviously been lack of strong mathematics foundation. As an uncertainty mathematical method, Pawlak rough set can only dispose discrete data formerly, so we must discrete spatial continuous data when using this method, it would bring the profits and losses of the information in the transformation process. We used neighborhood rough set concept, and put forward a spatial continuous data classification method based on neighborhood rough set, when conditional attribute is continuous data and decision attribute is discrete data.

KEYWORDS

Classification; Neighborhood rough set; Spatial continuous data; Granulation.

INTRODUCTION

Spatial data are rich in content, and contains a lot of information. Spatial data classification is a kind of important data analysis form, and can be used to extract description of important data classification model [1]. The purpose of spatial data classification is to learn to a classification function or classification model (Also called classifier). This model can divide the space data in a database mapping to a given category of a certain category [2]. The structure of the classifier method has statistical method, machine learning method, neural network method, etc. Specific classification method has Bayesian method and the nonparametric method, decision tree method and rule induction, BP algorithm, etc [3]. Since the 1982, Pawlak put forward the rough set theory, it received more and more attention in the research of data integrity and accurate knowledge representation, learning and induction, etc[4]. Rough set theory can also be used in classification, found imprecise data or noise data for inner structure contact. Before, for rough set research mainly focus on the classical rough set, and data attribute value must be discrete. Therefore, whether the general relational database or spatial database, continuous value must be discreted before classification [5]. And spatial analysis data, such as remote sensing data are mostly some with continuous value of the attribute.

Some classification algorithms based on neighborhood were proposed, where a new sample is associated with a neighborhood, rather than some nearest neighbors. Owen developed a classifier which uses information from all data points in a neighborhood to classify the point at the center of the neighborhood [6]. The neighborhood-based classifier is shown to outperform linear discriminant analysis on some LANDSAT data. Salzberg proposed a family of learning algorithms based on nested generalized exemplars (NGE), where an exemplar is a single training example, and generalized exemplars is an axis-parallel hyperrectangle that may cover several training examples. Once the generalized exemplars are learned, a test example can be classified by computing the Euclidean distance between the example and each of the generalized exemplars [7].

In fact, neighborhoods and neighborhood relations are a class of important concepts in topology. Lin pointed out that neighborhood spaces are more general topological spaces than equivalence spaces and introduced neighborhood relation into rough set methodology, which has shown to be a powerful tool to attribute reduction, feature selection, rule extraction and reasoning with uncertainty [8]. Many researchers discussed the properties of neighborhood approximation spaces. However, few applications of the model were reported in these years [9,10,11]. Since Lin pointed out neighborhood model in 1988 [8], the model through the granulating theory field of the space point neighborhood to the formation of neighborhood structure. And it approach and describe some of the concepts of data space through neighborhood being taken as a basic information particle. The proposed technique combines the advantages of feature subset selection and neighborhood-based classification. It is conceptually simple and is straightforward to implement. Some experimental analysis is conducted on a spatial regional data set. Three kinds of norms, 1-norm, 2-norm and infinite-norm, are tried. The results show that the proposed classification system is outperformed.

Neighborhood rough set theory, as a kind of numerical information granularity computation model, neighborhood system in relational database and spatial database research and have been ignored. This paper will use the neighborhood topological space of spherical neighborhood structure to build the spatial data neighborhood system classification method. This method is intuitive easy to understand to be able to deal with continuous value of the spatial data, without the space data discretization of data pretreatment. Therefore and based on the classical rough set of spatial data classification method, omitting the discretization process, and reducing the loss of spatial data information [12]. Compared with K-step neighbourhood model, it can not establish sample space neighborhood diagram [9].

CLASSIFICATION AND NUMERICAL SPACE NEIGHBORHOOD GRANULATING

Given a database $D=\{t_1,t_2,...t_n\}$ and a set of classes $C=\{C_1,C_2,...C_m\}$, classification problem is to determine a mapping $f:D\to C$, and each tuple t_i is assigned to a class. A class C_i contains mapping to the all the tuple, that is[13] $C_j=\{t_i \mid f(t_i)=C_j,1\le i\le n,且 t_i\in D\}$. The general data classification is divided into two steps, modeling and use. Modeling is the use of training data are guided learning training (also called supervised classification). General learning model is provided by classification rule, decision tree or equality rules type and other forms. These rules can be used for the following data sample classification, and the contents of database can provide better understanding. If you think that model accuracy can be accepted. It can be used to unknown data classification.

Generally for spatial data classification analysis, we should be the first to measure the test sample's distance between the different attribute. Given dij is distance for two attributes Xi and Xj. The frequently used distance formulas are Minkowski distance, Mahalanobis distance, Canberra distance [14]. We used Mahalanobis distance to define distance of two examples:

$$d_{ij}(q)=\left(\sum_{a=1}^{p}\left|x_{ia}-x_{ja}\right|^{q}\right)^{1/q} \quad (1)$$

When q=1 that is Absolute distance:

$$d_{ij}(1)=\sum_{a=1}^{p}\left|x_{ia}-x_{ja}\right| \quad (2)$$

When q=2 that is Euclidean distance:

$$d_{ij}(2)=\left(\sum_{a=1}^{p}(x_{ia}-x_{ja})^2\right)^{1/2} \quad (3)$$

When $q=\infty$ that is Chebyshev distance:

$$d_{ij}(\infty)=\max_{1\le a\le p}\left|x_{ia}-x_{ja}\right| \quad (4)$$

Then, we can obviously see diamond is absolute distance, roundness is Euclidean distance, square is Chebyshev distance.

Example 1: Now we consider it has a GIS map level that composed of nine basic units in fig.1, B and C stand for different attribute. Then it should use absolute distance for measure distance x1 and x2 in attribute B, that we can compute d(x1,x2)=0.2. It should use Euclidean distance for measure distance x1 and x2 in attribute B,C, that we can compute d(x1,x2)=0.45. We should dispose source data pretreatment in practice, owing to space constraints, the details will not be dealt with here.

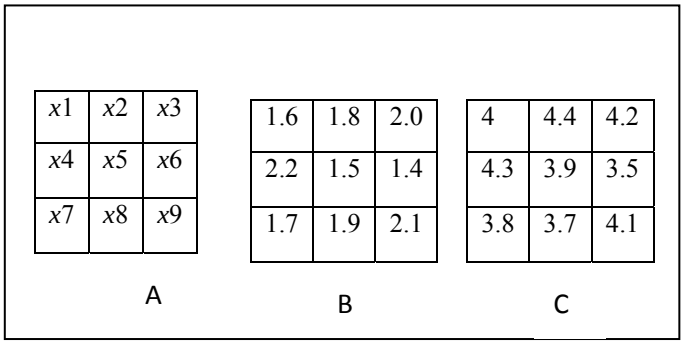

FIGURE 1. Grid data attribute value map

Approximation and granulating is basic problems for granular computing and the rough set theory. The classical rough set model is a space discrete value based on clear equivalence relation model. Universe space granulating is the equivalence relation partition for universe. But for spatial data application in real space, attribute value is continuous, such as DEM elevation value etc. Therefore, aiming at this kind of spatial data real space, the equivalence relation will be a fitting individual numerical attributes. Sequence structure and neighborhood structure is real number space important structure, this paper will be built on the basis of neighborhood structures.

Definition 1 Given a N dimension real number space Ω, we call d is a measurement of RN, it usually satisfy follows properties:

1) non-negativity: d(x1,x2)≥0, d(x1,x2)=0, if and only if x1=x2, $\forall x1, x2 \in R^N$;
2) symmetry: d(x1,x2)= d(x2,x1), $\forall x1, x2 \in R^N$;
3) triangle inequality: d(x1,x3)≤d(x1,x2)+ d(x2,x3), $\forall x1, x2, x3 \in R^N$.

Then we called (Ω, d) is real number space. And Euclidean distance is a common measurement tool for real number space.

Definition 2 Given a non-null limited set U{x1, x2, x3, …… xn} in real number space, for every object xi in U, then the δ-neighborhood definition is as follows [11]:

$$\delta(x_i)=\{x|x\in U, d(x,x_i)\le\delta\} \quad (5)$$

Where δ>0, $\delta(x_i)$ is δ neighborhood information granulation from xi, it for short called as xi neighborhood granulation.

From the measurement properties, we can get three properties about neighborhood information granulation:

1) $\delta(x_i)\neq\infty$, because of $x_i\in\delta(x_i)$;
2) $x_j\in\delta(x_i)\Rightarrow x_i\in\delta(x_j)$;
3) $\cup\delta(x_i)=U$.

So Given a measurement space (Ω, d) and a non-null limited set U{x1, x2, x3, ……xn}, if δ1≤δ2, then we can get these properties:

1) $\forall x_i\in U:\delta 1(x_i)\subseteq\delta 2(x_i)$;
2) $N_1\subseteq N_2$.

All universe object neighborhood constitute object universe granulation. And the basic concept system of universe space is constituted of neighborhood particle family. We can use these basic concepts to approximate any concept for spatial universe data.

Example 2: Nine polygons are seen in fig.1, U={x1, x2, x3,……x9 }, and B and C are respectively stand for two attribute level value (such as slope, aspect etc), when we choose value in one dimension attribute, we can use absolute distance. We use f(x, b) to express the value in attribute B for example x , then we can get f(x1,b)=1.6，f(x2,b)=1.8…f(x9,b)=2.1. if we assigned the neighborhood threshold is 0.2, because of

|f(x1,b)−f(x2,b)|=0.2≤0.2, then $x_2 \in \delta(x_1), x_1 \in \delta(x_2)$. In this case, we can get $\delta(x_1)=\{x_1,x_2,x_5,x_6,x_7\}$, $\delta(x_2)=\{x_1,x_2,x_7,x_8\}$, …. $\delta(x_9)=\{x_3,x_4,x_8,x_9\}$. When we get value in two dimension attribute, we should use Euclidean distance, we used f(x, b) to express the value for attribute B, C for example x, if the neighborhood threshold is 0.3. Then we can compute each polygon's neighborhood in two dimension space, $\delta(x_1)=\{x1,x5,x7\}$? $\delta(x2)=\{x2,x3\}$? $\delta(x3)=\{x2,x3,x4,x9\}$? $\delta(x4)=\{x3,x4,x9\}$? $\delta(x5)=\{x1,x5,\}$? $\delta(x6)=\{x6\}$? $\delta(x7)=\{x1,x5,x7,x8\}$? $\delta(x9)=\{x3,x4,x9\}$. If it has many attributes, we can compute the distance for examples, and computed the neighborhood for examples.

Definition 3 Given a set of objects U{x1, x2, x3,……xn} and a neighborhood relation R, called D={U,R} is a neighborhood approximation space [15].

Definition 4 Given D= {U, R} and X⊆U. For any X⊆U, two subsets of objects, it is called lower and upper approximations of X in D= {U, R}, that are defined as follows:

$$\begin{cases} \underline{apr}X = \{x_i \in U | \delta(x_i) \subseteq X, x_i \in U\} \\ \overline{apr}X = \{x_i \in U | \delta(x_i) \cap X, x_i \in U\} \end{cases} \quad (6)$$

Obviously, $\underline{apr}X \subseteq X \subseteq \overline{apr}X$. The positive region of X (pos (X)), negative region of X (neg (X)) and boundary region of X in the approximation space are defined as follows:

$$\begin{cases} pos(X) = \underline{apr}X \\ neg(X) = \sim \overline{apr}X \\ bn(X) = \overline{apr}X - \underline{apr}X \end{cases} \quad (7)$$

A sample in the decision system belongs to either the positive region or the boundary region of decision. Therefore, the neighborhood model divides the samples into two subsets: positive region and boundary region. Positive region is the set of samples which can be classified into one of the decision classes without uncertainty, while boundary region is the set of samples which can not be determinately classified. Intuitively, the samples in boundary region are easy to be misclassified. In data acquirement and preprocessing, one usually tries to find a feature space in which the classification task has the least boundary region. It is as summarized in Zhang.

Example 3: We given two sets X= {x1, x2, x3, x5, x7} and Y={x2, x4, x6} in fig.1, one sets stand for a group continuous value. Then we can get pos(X) ={x1, x2, x5}, pos(Y) ={x6}, accordingly, we can get the negative region and boundary region for two sets.

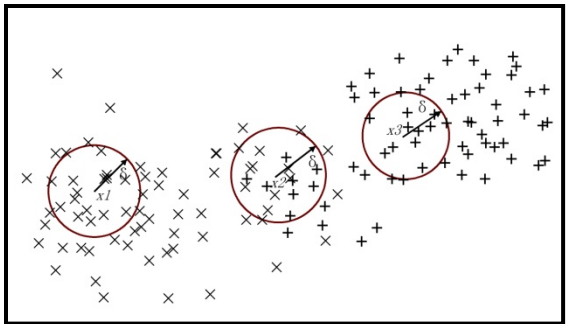

FIGURE 2. Neighborhood rough approximation in continuous numerical value spaces

Then we can get a map that shown binary classification in a 2-D numerical space in fig.2. The first example is labeled with "X" identification, the second example is labeled with "+" identification. So we can see x1 is belongs to the lower approximations of the first example, x3 is belongs to the lower approximations of the second example because of its neighborhood are from the second number, x2 is boundary example because of its neighborhood is belongs to the first example and the second example too. The definition is according to our intuitive recognition for classification problem in real world.

NEIGHBORHOOD ATTRIBUTE REDUCTION ALGORITHM

Definition 5 Given a neighborhood decision system $NDT=<U,A,D>$, discrete decision class D divided universe U into N equivalence class: $X_1, X_2, \ldots X_N, \forall B \subseteq A$, then the lower and upper approximations of the decision D with respect to attribute B are defined as:

$$\underline{N_B}D = \bigcup_{i=1}^{N} \underline{N_B}X_i, \overline{N_B}D = \bigcup_{i=1}^{N} \overline{N_B}X_i \quad (8)$$

Where

$$\begin{aligned} &\underline{N_B}X = \{x_i | \delta_B(x_i) \subseteq X, X_i \in U\}, \\ &\overline{N_B}X = \{x_i | \delta_B(x_i) \cap X \neq \Phi, X_i \in U\} \end{aligned} \quad (9)$$

$\delta_B(x_i)$ is the neighborhood information granules that generated by measurement d and attribute B, $B \subseteq A$. The lower approximation of the decision, also called positive region of decision, denoted by $POS_B(D)$, is the subset of objects whose

neighborhoods consistently belongs to one of the decision classes. The bigger the $POS_B(D)$, the less cross area. We can more detailed describe classification problem through this attribute property. Then the dependency degree of D to B is defined as the ratio of consistent objects:

$$\gamma_B(D) = Card(\underline{N_B}D)/Card(U) \quad (10)$$

Where $\gamma_B(D)$ reflects the ability of B to approximate D. obviously, $0 \le \gamma_B(D) \le 1$. We say that D completely depends on B if $\gamma_B(D) = 1$. If the value of $\gamma_B(D)$ is bigger, we can say the dependency of D to condition B is strong. Such as in remote image classification, land classification type depends on several bands, not on all bands. Given a neighborhood decision system $NDT = < U, A, D >$, if $\gamma_{B-a}(D) < \gamma_B(D)$, then we can say a relative to B, it is indispensable, otherwise $\gamma_{B-a}(D) < \gamma_B(D)$ we can say a relative to B, it is redundant. In spatial data classification, it has many conditional attributes, such as remote bands. But not all the bands are useful to remote image classification, we can get rid of useless band. In a given spatial neighborhood decision system, the union set of all reduction is the core of this spatial neighborhood decision system, such as remote image classification, all bands combination that can affect classification. We can get different core for different threshold, so we need some algorithm to check these cores for the classification accuracy in different threshold.

It has hundreds of condition attributes in spatial data classification some times. So if we want to all reductions in spatial decision system, it requires 2N-1 tests. When there are 100 conditional attributes, we should require 1.26+E30 tests, we can imagine the calculation amount. The researchers judge classification band according to the experience in some remote image classification. So how to find attribute core for spatial classification, based on dependency function, construct a spatial data pretreatment forward the greedy algorithm, it would greatly simplify the works. Given a neighborhood decision system $NDT = < U, A, D >, B \subseteq A, a \in A - B$, then the importance degree of a relative to B is:

$$SIG(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \quad (11)$$

So any spatial data attribute is function that it is constructed of the attribute itself, attribute in the attribute subset and spatial decision variable. We can build a spatial data forward greedy attribute reduction algorithm based on spatial data attribute importance degree formula. The conditional attribute is continuous value, and the decision attribute is discrete value in our study. This algorithm is based on empty sets starting point, calculate all other attribute importance degree, choose the biggest attribute importance degree value into the attribute reduction set until all other residual attribute importance degree is 0. That is to say, adding a new attribute, the dependency degree would not change in spatial decision system. The algorithm description is seen in follow:

Input: spatial neighborhood decision system $NDT = < U, A, D >, B \subseteq A, a \in A - B$ and δ // δ is the threshold to control the size of the spatial neighborhood

Specify the norm to be used

Output: spatial data attribute reduction red

Step 1: calculate the neighborhood relation of $\forall a \in A$

Step 2: $\Phi \to red$ // red is the pool to contain the selected spatial attributes

Step 3: For each $a_i \in A - red$

Compute $SIG(a_i, red, D) = \gamma_{red \cup a}(d) - \gamma_{red}(D)$ // Here we define $\gamma_\Phi(D)$

end

Step 4: Select the attribute ak which satisfies $SIG(a_k, red, D) = \max_i(SIG(a_i, red, B))$

$SIG(a_k, red, D) > 0,$

Step 5: If $red \cup a_k \to red$

go to step 3

else

return red

Step 6: end

EXPERIMENTAL ANALYSIS

In order to test the proposed spatial classification model, we choose an administrative region to experiment. It has 178 administrative units in our experimental region. In economic decision-making evaluation, there are many economic evaluation indexes, such as GDP, population density etc. there are 13 conditional indexes in this experiment, all these conditional indexes are continuous value. There is one decision index in this experiment, and the decision index value is discrete, the value domain is {1, 2, 3}.

We had a normalization processing for all conditional attribute before calculation each object's neighborhood, the normalization formula is:

y=(x-MinValue)/(MaxValue-MinValue) (12)

X, Y is the value of pre/post-transition, MaxValue、MinValue is the maximum and minimum value, this work can reduce the influence of attribute dimension inconsistency to evaluation result. We respectively set different threshold value δ from 0.05 to 0.3 step 0.05, extended an object's neighborhood from small to large. We find when the threshold value δ is bigger, the spatial data attribute core tend to all conditional attributes through our experiment. The overlarge threshold value δ has not help to spatial classification attribute reduction, and it can not get rid of unrelated attributes, so when there are

hundreds of condition attribute in a spatial decision evaluation, how to find an appropriate threshold value δ and spatial data attribute core is very important.

TABLE 1 Different reduction core for different threshold

threshold value δ	Spatial data attribute core	threshold value δ	Spatial data attribute core
0.05	10,13,5,1	0.2	13,10,7,1,5,2
0.1	13,10,7,4,2	0.25	13,10,7,1,5,2,6
0.15	13,10,7,5,2,1	0.3	13,10,11,1,12,5,2,3,7

Table2 experimental classification data confusion matrix

classification	actual classification			accuracy
	1	2	3	
1	59	0	0	100%
2	2	66	3	93%
3	1	0	47	98%

We had tested spatial data classification for all spatial attribute core results from above table 1. We should transform the test region data into Grid data, because only Grid data can carry through statistic analysis in GIS data evaluation. We have 43466 grids cell. We find when threshold value δ=0.1, namely attribute core is 13, 10, 7, 4, 2, the classification result is the most accurate in our experiment (spatial value distributed map for attribute seen in fig.3 and fig.4). Relatively speaking, the classification is better when the threshold value δ in [0.1, 0.2]. Of course the specific analysis and evaluation result is according to the different region and evaluation object data. So for different evaluation object and region, we should repeatedly test different threshold value δ through test data and train data, then it can get a reasonable spatial data classifier.

We choose attribute 2,4,7,10,13 to test, through the training data training and testing the region test data. We get the comparative result for classification result and actual classification, specific seen Fig 3. , Fig 4. We can get that the result of classification is very ideal. The statistic analysis data for GIS are all GRID data, so one administrative region is always transformed into multi-grid, namely the same data in GIS grid operation would be converted into a lot of the same value. So we finally had converted the grid data into polygon for calculation confusion matrix. Finally, we tested this method classification accuracy from confusion matrix, specific seen in table 2 we can see the classification result is very ideal.

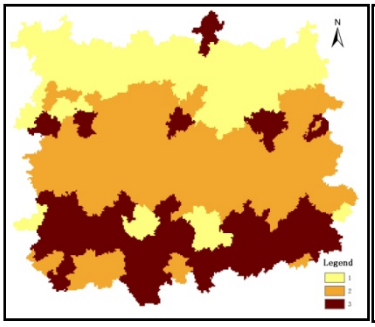

FIGURE 3. neighborhood rough set classification result map

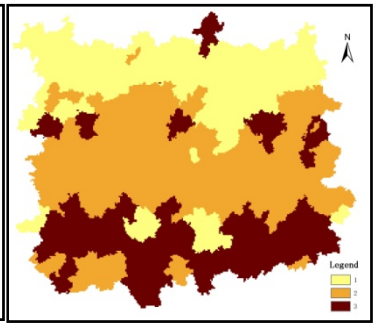

FIGURE 4.practical value distributed map

CONCLUSION

Spatial data classification is a regular method for spatial decision evaluation. It has produced more and more information data in spatial data classification with the rapid development of the information age. How to find the appropriate band for hyperspectral remote sensing imaging in a spatial decision evaluation, extract useful condition attribute data for

decision-making evaluation object, is an issue that is worth studying. This method used decision attribute participation, combined with neighborhood rough set, can improve the continuous value of the spatial data classification accuracy and provide strong mathematical foundation for the choice of condition attributes.

REFERENCES

[1] Bellotti F, Capra L, Sarocchi D 2010 Journal of Volcanology and Geothermal Research 191(1-2) 117-128.
[2] Browna G, Brabyn L, 2012 Landscape and Urban Planning 107(3) 317-331.
[3] Bagan H, MA J W, Li Q Q 2004 SCIENCE IN CHINA SERIES D 34(10) 967-974.
[4] Kumar P, Radha K P, Raju S 2007 Data & Knowledge Engineering 63(1) 183–199.
[5] Peters J F, Suraj Z, Shan S, 2003 Pattern Recognition Letters 24(6) 911–920.
[6] Pan X, Zhang S Q, Zhang H Q 2010 Computers & Geosciences 36(12)1466-1473.
[7] Hu Q H, An S, Yu X, 2011 Fuzzy Sets and Systems 183(1) 26–43.
[8] Zhang W X, Qiu G F, Wu W Z 2007 Sci China Ser F-Inf Sci 50(2)188—197.
[9] Hu Q H, Yu D R, Xie Z X, 2006 Pattern Recognition Letters 27(5) 414–423.
[10] Hu Q H, Yu D R,.Xie Z X, Liu J F, 2006 IEEE Transactions on Fuzzy Systems 14(2) 191–201.
[11] Jensen R, Shen Q, 2004 IEEE Transactions of Knowledge and Data Engineering 16(12) 1457–1471.
[12] Liao W. H., 2012 Journal of Geographic Information System, 4(1) 37-45.
[13] Chen Y S, Cheng C H, 2012 Applied Soft Computing, 12(1) 462–475.
[14] Zhang Y H, Zhang F, 2014 COMPUTER MODELLING & NEW TECHNOLOGIES 18(1) 44-51
[15] Zadeh L A, 1996 IEEE Trans on Fuzzy Systems, 4(2) 103−111.