International Journal of Applied Earth Observation and Geoinformation 26 (2014) 432–440

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Spatial data discretization methods for geocomputation

Feng Cao, Yong Ge*, Jinfeng Wang

State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, A11 Datun Road, Beijing 100101, China

ARTICLE INFO

Article history:
Received 27 September 2012
Accepted 11 September 2013

Keywords:
Geocomputation
Spatial data
Discretization
Spatial autocorrelation
Spatial heterogeneity

ABSTRACT

Geocomputation provides solutions to complex geographic problems. Continuous and discrete spatial data are involved in the geocomputational process; however, geocomputational methods for discrete spatial data cannot be directly applied to continuous or mixed spatial data. Therefore, discretization methods for continuous or mixed spatial data are involved in the process. Since spatial data has spatial features, such as association, heterogeneity and spatial structure, these features cannot be handled by traditional discretization methods. Therefore, this work develops feature-based spatial data discretization methods that achieve optimal discretization results for spatial data using spatial information implicit in those features. Two discretization methods considering the features of spatial data are presented. One is an unsupervised method considering autocorrelation of spatial data and the other is a supervised method considering spatial heterogeneity. Discretization processes of the two methods are exemplified using neural tube defects (NTD) for Heshun County in Shanxi Province, China. Effectiveness is also assessed.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Geocomputation uses computational methods and tools to explain geographic phenomena and solve geographic problems (Couclelis, 1998; Clarke, 2003; Fotheringham et al., 1997). Classification, prediction, modeling, knowledge discovery, and visualization are the main topics (Fischer and Leung, 2001; Gahegan, 1999; Openshaw, 1998; Openshaw and Abrahart, 2000), which have wide application (Murgante et al., 2009, 2011). Different types of spatial data are involved in the geocomputational process, such as nominal, ordinal, interval and ratio data. Nominal and ordinal data are discrete, and interval and ratio data are continuous (Haining, 2003). Geocomputational methods for continuous spatial data are only suitable for analysis of continuous data and cannot be applied to mixed spatial data, including continuous and discrete data together. To the contrary, the methods for discrete data can be applied to discrete, continuous and mixed spatial data, after continuous data are converted to discrete data with appropriate discretization methods. Since different discretization methods may produce different results, use of an optimal method is a critical issue. In practical applications, researchers always use a user-defined discretization method to discretize continuous data based on their experience (Liao et al., 2010; Wang et al., 2010; Hu et al., 2011). However, user-defined methods are subject to weaknesses such as randomness and subjectivity. To overcome such drawbacks, discretization methods based on statistical characteristics of the spatial data are used (Du et al., 2010; Fischer and Wang, 2011; Ge et al., 2011). Among those methods, unsupervised and supervised are two commonly used types, based on the taxonomy of having class information or not.

A common disadvantage of current discretization methods for spatial data discretization is that data features are commonly ignored in the discretization process. In practice, user-defined discretization is used to discretize continuous spatial data and select the cut point set according to experience (Leung et al., 2007). However, there are no fixed rules to determine whether that set is optimal. Unsupervised and supervised methods have been proposed and widely used for the discretization of spatial data by considering their statistical characteristics. For example, unsupervised methods such as equal interval (EI), quantile (QU), natural breaks (NB) and standard deviation (SD) are common in spatial data mapping and geovisualization (Fischer and Wang, 2011; Stewart and Kennelly, 2010). Unsupervised methods assisted research into classed choropleth maps when Jenks and Caspall (1971) introduced methods to find optimal values of class breaks after the first unclassed choropleth map appeared in 1826 (Robinson, 1982). Monmonier (1972) created a taxonomic clustering-based unsupervised discretization method to select proper class intervals for creating clear and simple choropleth maps. Armstrong et al. (2003) used a genetic algorithm to obtain optimal discretization intervals based on a multi-criteria framework.

Compared with unsupervised methods, supervised methods use class information to assist selection of optimal cut points during discretization. Supervised methods are common in the research

* Corresponding author. Tel.: +86 10 64888053; fax: +86 10 64889630.
E-mail addresses: gey@lreis.ac.cn, gey@igsnrr.ac.cn (Y. Ge).

1569-8432/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jag.2013.09.005

fields of classification, prediction and data mining. For example, Berger (2004) used a supervised method called the Minimum Description Length Principle (MDLP) to discretize continuous environmental attributes and assess crop suitability for agricultural soils with rough set rule induction. Bai et al. (2010) also used MDLP to discretize continuous risk factors of neural tube defects (NTD), and mined underlying rules between NTD and its risk factors. Lustgarten et al. (2011) proposed an efficient supervised Bayesian discretization method to give better classification results from a high-dimensional biomedical dataset. Ge et al. (2011) compared the impacts of three supervised discretization methods on remote sensing classification. Those works directly used supervised methods for spatial data discretization. Compared to non-spatial data, spatial data has special features characterizing spatial information that can help achieve a reasonable data discretization result.

Therefore, we investigate two spatial data discretization methods, taking into account spatial features of the data. One is an unsupervised local indicator of spatial association-based discretization method (LISABD) and the other is a supervised spatial stratification-based discretization method (SSBD). LISABD is suitable for discretization of a single continuous attribute with no additional decision class. SSBD can simultaneously discretize multiple continuous attributes having spatial heterogeneity. The discretization processes of LISABD and SSBD are constructed and exemplified with NTD data from Heshun County in Shanxi Province, China. Analysis results based on two geocomputational methods, geographical detectors and rough set, demonstrate the effectiveness of the two discretization methods.

2. Discretization methods

The aim of discretization is to convert continuous into discrete data. Compared with continuous data, discrete data are easier to understand, use and explain, and are closer to a knowledge-level representation (Dougherty et al., 1995; Liu et al., 2002). Data discretization is a process whereby continuous data are divided into intervals with selected cut points, and each interval is mapped to a qualitative symbol.

2.1. Traditional discretization methods: unsupervised and supervised

2.1.1. Unsupervised discretization methods

Unsupervised methods do not consider class information during the discretization process. If no such information is available, unsupervised discretization is the only choice. Unsupervised methods are very simple, and directly partition continuous data using user-defined parameters. The methods can be executed easily and rapidly, even for large amounts of spatial data. However, one needs to predefine the number of intervals before discretization. These are determined according to the user's domain knowledge, which is somewhat subjective. The discretization result is difficult to understand.

2.1.2. Supervised discretization methods

Supervised methods relate class information to the selection of cut points. Appropriate cut points are chosen so that data instances in the same interval have the same class label; the labels vary across consecutive intervals. Most current discretization methods are supervised. Supervised methods mainly include those based on dependency, entropy, rough set, and chi-square measure.

Dependency-based discretization mainly uses statistical characteristics to measure the strength of association between the class and a continuous attribute. This type evaluates the importance of the cut point according to dependency of the class and discretization scheme of the continuous attribute. Zeta, Class-Attribute Interdependence Maximization (CAIM), Class-Attribute Contingency Coefficient (CACC) and Class-Attribute Dependent Discretizer (CADD) are the main dependency-based discretization methods (Ching et al., 1995; Ho and Scott, 1998; Kurgan and Cios, 2001; Tsai et al., 2008). Entropy-based discretization is widely used in continuous data analysis, which evaluates the distribution of class in the discretization scheme of continuous data with types of entropies. This type uses heuristics to select the optimal cut point set, which has good time complexity and is easily parallelizable. This type mainly includes Interactive Dichotomizer 3 (ID3), Discretiser 2 (D2), Minimal Description Length Principle (MDLP) and Mantaras Distance (MD) (Cerquides and Mantaras, 1997; Dougherty et al., 1995; Fayyad and Irani, 1992; Quinlan, 1986). Rough set-based discretization considers the indiscernibility relation based on rough set theory. This type largely maintains the indiscernibility relation unchanged during the discretization process, and includes Boolean, Naïve, Semi-naïve and Discretization Based on Genetic Algorithm (DBGA) (Ge et al., 2011; Nguyen and Skowron, 1995; Son et al., 1996). Chi-square measure-based discretization evaluates the importance of cut points based on chi-square measure and uses heuristics to select the optimal cut point set, which includes a series of supervised discretization methods such as ChiMerge, Chi2, ConMerge, Modified Chi2 and Extended Chi2 (Kerber, 1992; Liu and Setiono, 1995; Su and Hsu, 2005; Tay and Shen, 2002; Wang and Liu, 1998).

2.2. Spatial data discretization methods: unsupervised and supervised

Although traditional discretization methods are used to discretize spatial data, we do not consider them spatial data discretization methods. The latter methods should take into account the spatial features of spatial data. Geographic objects have spatial position, direction and distance, causing spatial data to have a specific spatial distribution and relationship features; these are currently ignored in traditional discretization schemes. Therefore, some useful spatial information is ignored by traditional discretization methods in spatial data discretization. Here, we investigate two potential study areas of spatial data discretization, considering spatial autocorrelation and spatial heterogeneity. This constitutes exploratory research on spatial data discretization.

2.2.1. Unsupervised local indicator of spatial association-based discretization

The traditional unsupervised discretization converts a continuous single attribute into a discrete attribute without using any additional information. The discretization process is a mapping f from a continuous attribute A to discrete attribute A_d:

$$f : A \rightarrow A_d. \quad (1)$$

Each value of A_d corresponds to an interval of A after the continuous attribute is discretized. The traditional unsupervised methods have two main drawbacks when used to discretize spatial data. The first is that they require the user to predefine the number of intervals. The other is that they ignore spatial association that is prevalent in spatial data (Murgante and Danese, 2011; Lanorte et al., 2013). Spatial association provides additional spatial information that can be used to assist the data discretization.

The local indicator of spatial association (LISA) proposed by Anselin (1995) is commonly used to assess the autocorrelation of spatial data. Here, we investigate an unsupervised LISA-based discretization method (LISABD). LISABD takes into account the spatial autocorrelation with LISA during data discretization. Compared to traditional unsupervised discretization, LISABD uses additional spatial autocorrelation clustering to achieve a better discretization result. The LISABD discretization process is a mapping f from a

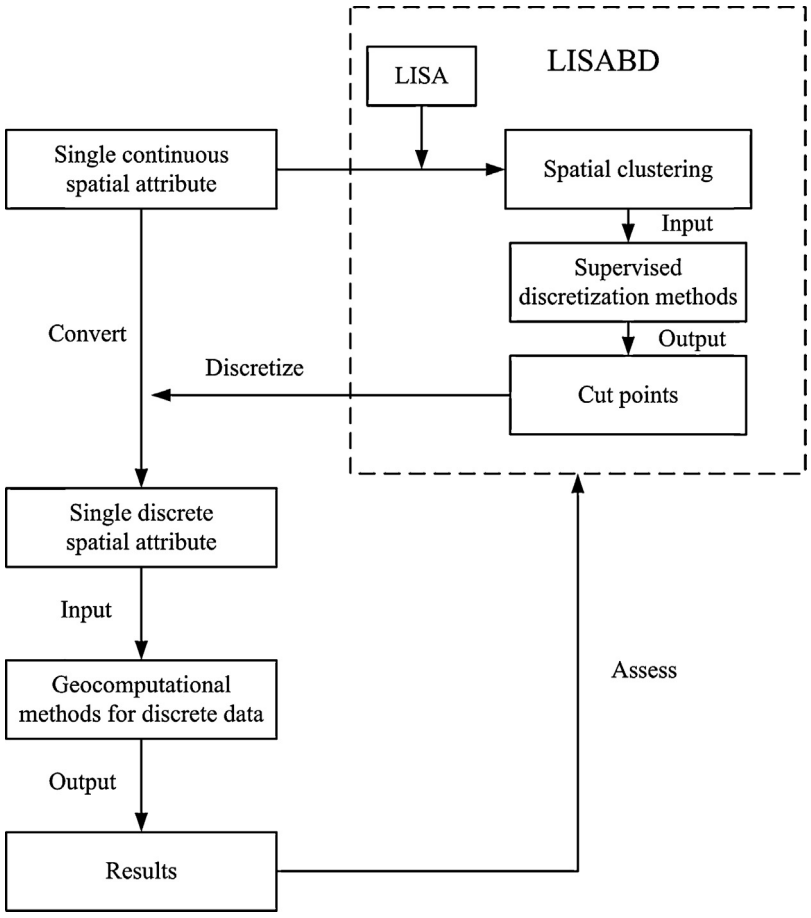

Fig. 1. Flowchart of discretization and assessment process for LISABD.

continuous attribute A to a discrete attribute A_d, and the spatial autocorrelation clustering is used to supervise the process:

$$f : A|S_A \rightarrow A_d, \tag{2}$$

where S_A is the spatial autocorrelation clustering of continuous attribute A.

To further explain LISABD and assess its effectiveness, its flowchart of discretization and assessment is shown in Fig. 1.

The spatial autocorrelation clustering of a single continuous spatial attribute is obtained first. The type of spatial clustering is classified into high–high, high–low, low–low, low–high, and not significant. The type is designated as the class, which is the input of supervised discretization methods selected from those commonly used. Then, the cut points of the continuous spatial attribute are obtained. Afterward, those points are used to discretize the single continuous spatial attribute and convert it into a discrete attribute. To assess the effectiveness of LISABD, analysis results of certain geocomputational methods for discrete data can be compared, after the continuous attribute is discretized with LISABD and other discretization methods.

2.2.2. Supervised spatial stratification-based discretization

The traditional supervised discretization methods of multiple attributes convert multiple continuous attributes $A = \{A_1, A_2, \ldots, A_n\}$ into discrete attributes $A_d = \{A_{d1}, A_{d2}, \ldots, A_{dn}\}$, using class information. The discretization process is a mapping f from multiple continuous attributes A to multiple discrete attributes A_d, and class information is used to supervise the discretization process:

$$f : A|C \rightarrow A_d, \tag{3}$$

where C is the class information.

Spatial heterogeneity is common in geographic phenomena, and it reflects the distribution variability of geographic objects. Discretization considering such spatial heterogeneity can generate good performance of geocomputational methods for discrete data when they are applied to continuous spatial data with strong spatial heterogeneity. Here, we investigate the SSBD method for discretization of multiple continuous spatial attributes with spatial heterogeneity. Compared to traditional supervised discretization methods of multiple attributes, SSBD discretizes multiple spatial attributes with spatial stratification, except for class information. The SSBD discretization process is a mapping f from multiple continuous spatial attributes $A = \{A_1, A_2, \ldots, A_n\}$ to multiple discrete attributes $A_d = \{A_{d1}, A_{d2}, \ldots, A_{dn}\}$, and class information and spatial stratification are both used to supervise the discretization process:

$$f : A|C, S \rightarrow A_d, \tag{4}$$

where C is the class information and S is the spatial stratification.

For explanation and effectiveness assessment of SSBD, the flowchart of discretization and assessment is shown in Fig. 2.

First, the class information of spatial stratification is combined to be the new class information. Then the latter information is used to supervise the discretization of multiple spatial attributes with the genetic algorithm (GA) to find the optimal cut points. For multiple spatial attributes, the optimal discretization is to find the minimum cut points among all the attributes, by maintaining strong consistency for class. However, this has been shown to be an NP difficult problem (Wang, 2001). The notation NP stands for "nondeterministic polynomial time". Therefore, a GA optimization method was used to find a sub-minimal cut point set (Nguyen et al., 1996). The

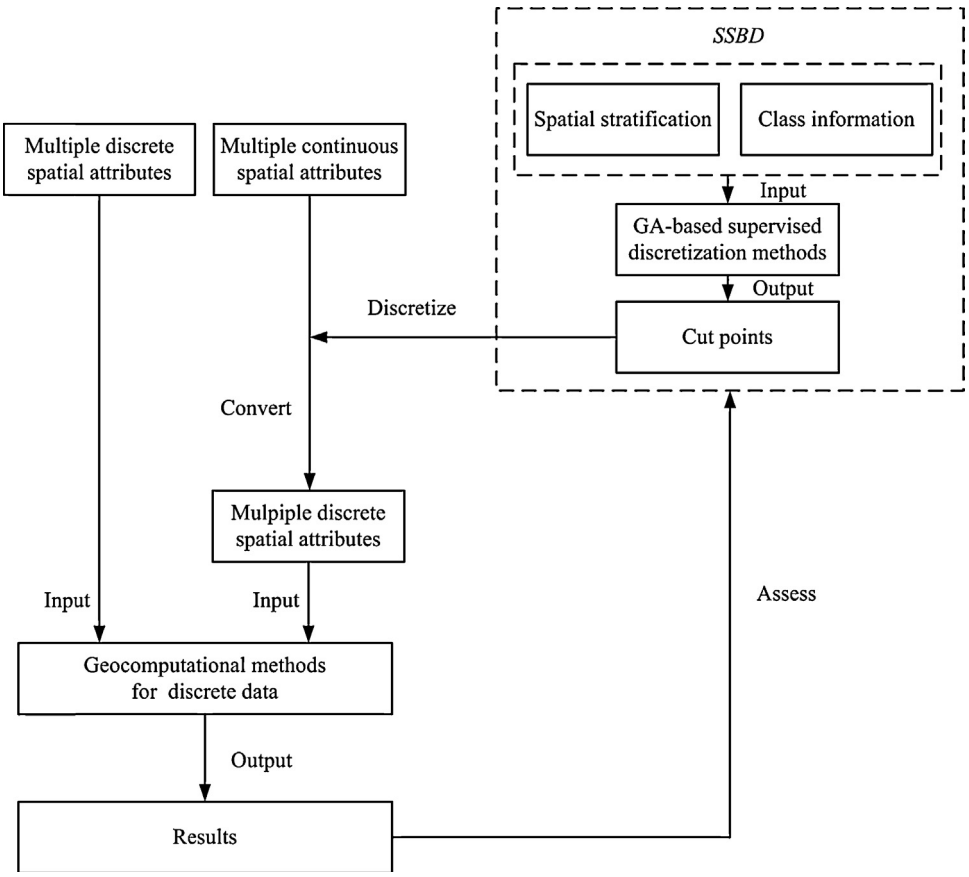

Fig. 2. Flowchart of discretization and assessment process for SSBD.

length of individual (chromosomes) equals the number of all candidate cut points of all continuous attributes for the GA method. The individuals are encoded with 1 and 0, to represent whether the candidate cut point belongs to the optimal cut points set or not. The fitness function is defined as

$$f(x) = N_{cuts} / \exp(-N_{inconsistent}), \tag{5}$$

where N_{cuts} represents the number of cut points encoded by 1 in the individual, and $N_{inconsistent}$ the number of inconsistent objects. For multiple attributes A, class information C and spatial stratification S,

$$N_{inconsistent} = \text{card}\{U' | a(x) = a(y) \wedge (C(x) \neq C(y) \text{ or } S(x) \neq S(y)), \forall x, y \in U' \subseteq U, \ a \in A\}, \tag{6}$$

where U is the set of geographic objects. The individuals with smaller fitness values remain in the new generation. The method searches the minimal cut points set, maintaining the least number of inconsistent objects. The cut points obtained with GA are used

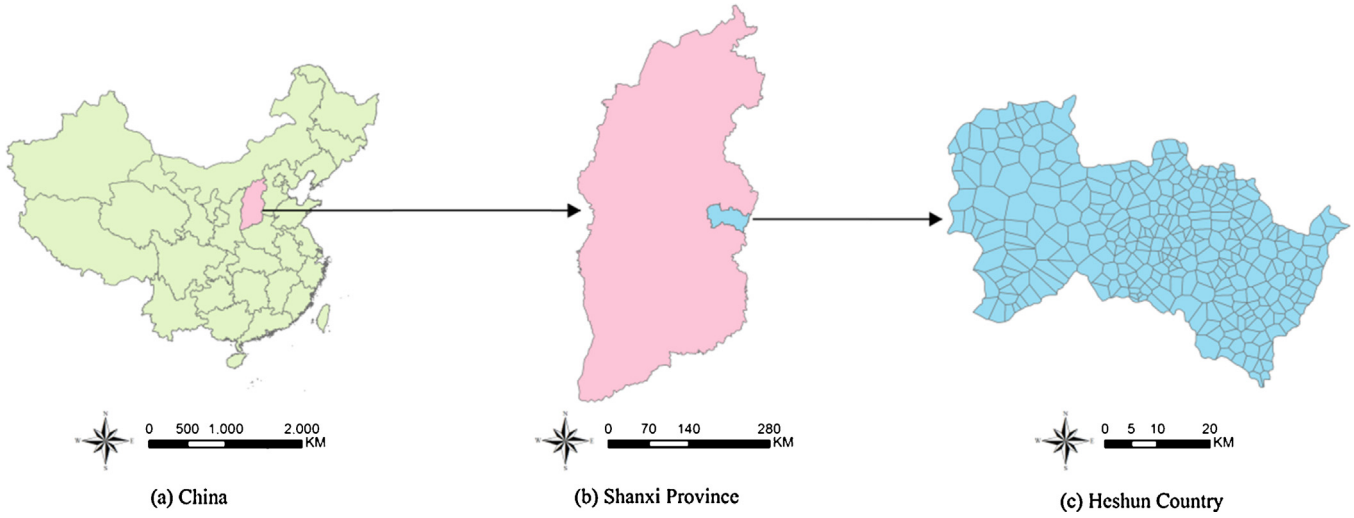

Fig. 3. Study area (a) China, (b) Shanxi Province, and (c) Heshun Country.

Table 1
Summary of NTD risk factors.

Social risk factors			Physical risk factors		
Risk factors	Meaning	Data type	Risk factors	Meaning	Data type
GDP	Average GDP during eight years	Continuous	Gradient	Main gradient type	Discrete
Doctor	Number of doctors	Continuous	Watershed	Main watershed type	Discrete
Fruit	Total fruit production during eight years	Continuous	Elevation	Village altitude	Continuous
Fertilizer	Total fertilizer used during eight years	Continuous	Soil	Main soil type	Discrete
Vegetable	Total vegetable production during eight years	Continuous	Lithology	Main lithology type	Discrete
Net income	Total net income during eight years	Continuous	Land cover	Main land cover type	Discrete
Pesticide	Total pesticide used during eight years	Continuous			

to discretize the multiple continuous spatial attributes. To assess SSBD effectiveness, analysis results of geocomputational methods for multiple discrete attributes can be compared for SSBD and other discretization methods.

3. Experiments

3.1. Data description

Neural tube defects (NTD) are one of the most common forms of birth defects, and are a major cause of stillbirth and infant mortality (Carmona, 2005; Liao et al., 2010). However, the etiology of NTD remains unknown. In previous studies, many social and physical risk factors were believed to have a close relationship to NTD prevalence (Wang et al., 2010). This provides important clues regarding the etiology, for determining the relationship between risk factors and NTD prevalence and helping prevent NTD occurrence. Shanxi Province has the highest rate of NTD in the world (Liao et al., 2010). Heshun County in northern Shanxi has one of the highest incidence of NTD. Therefore, this county is selected as the study area. Since there were no administrative boundaries, we drew them for each village using Thiessen polygons generated from village location points (Fig. 3). This gave 326 administrative villages. There were 187 NTD cases reported during 1998–2005.

Thirteen NTD risk factors were obtained, including seven social and six physical factors (Table 1). Among the 13, five are discrete-valued and the others are continuous-valued.

3.2. Unsupervised discretization of GDP and elevation with LISABD

GDP and elevation are two important risk factors among all physical and social environment risk factors of NTD (Wang et al., 2010). Thus, we selected these two variables to exemplify the discretization process of LISABD. First, their spatial autocorrelation clustering was obtained with LISA, in which the spatial relationship was measured with the rook contiguity weight matrix. The distribution of spatial clustering is shown in Fig. 4. Second, GDP and elevation were discretized with the selected supervised discretization method by designating the spatial clustering as class information. Here, five supervised discretization methods were selected for comparison, which are Chi2, CAIM, ChiMerge, CACC and MDLP. The numbers of cut points for GDP and elevation discretized with LISABD are shown in Table 2. To assess LISABD effectiveness, its impacts on geographical detectors-based relationships between GDP, elevation and NTD incidence are analyzed.

The geographical detectors model is a new geocomputational method for health and environmental risk assessment (Wang et al., 2010). This model has been applied to environmental risk assessment of NTD occurrence in Heshun County, and under-five mortality in the 2008 Wenchuan earthquake in China (Wang et al., 2010; Wang and Hu, 2012). In the model, continuous risk factors must be converted into discrete intervals before relationships between continuous risk factors and health prevalence are analyzed. The power of determinant (PD) is used to assess the impact

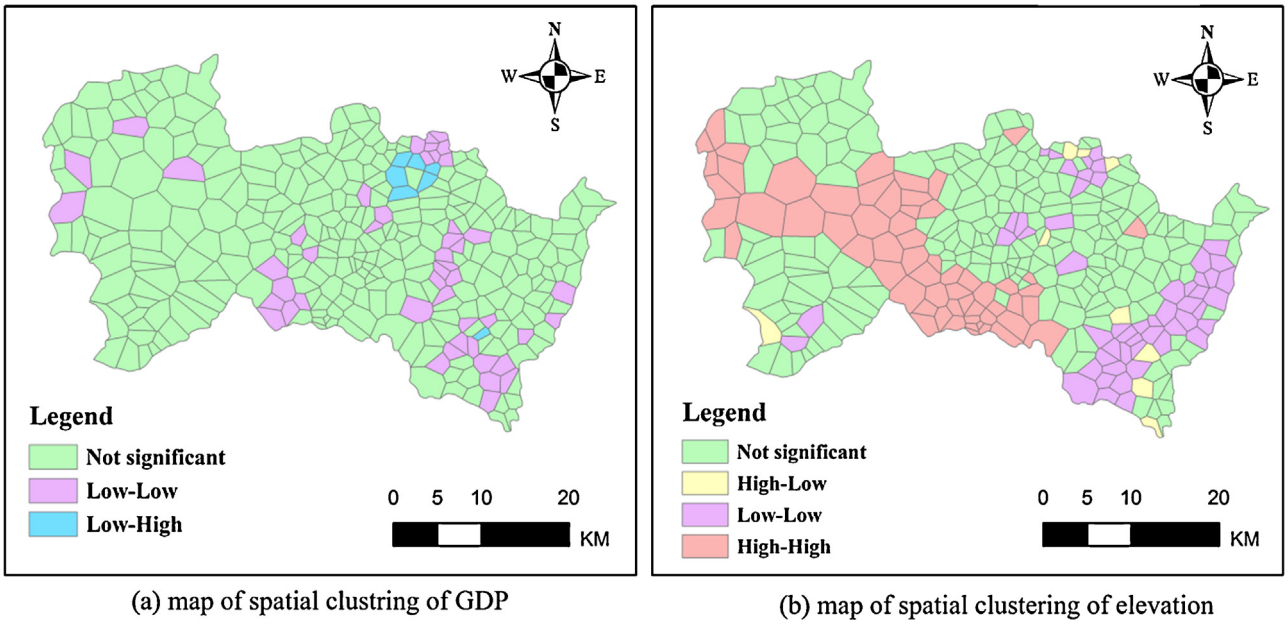

Fig. 4. Map of spatial autocorrelation clustering (a) GDP and (b) elevation.

Table 2
Numbers of cut points of GDP and elevation discretized with LISABD.

Risk factors	Number of cuts				
	LISABD (Chi2)	LISABD (CAIM)	LISABD (ChiMerge)	LISABD (CACC)	LISABD (MDLP)
GDP	53	2	28	2	0
Elevation	35	3	8	0	0

Table 3
PD values of GDP and elevation discretized with various discretization methods.

Discretization methods	PD values					
	Number of intervals for GDP			Number of intervals for elevation		
	2	28	53	3	8	35
EI	0.000	0.000	0.32	0.137	0.191	0.23
QU	0.028	0.282	0.326	0.194	0.155	–
NB	0.047	0.338	0.364	0.194	0.225	0.243
GI	0.046	0.261	0.32	0.190	0.176	0.241
SD	–	–	–	–	–	–
LISABD (Chi2)	–	–	0.357	–	–	0.268
LISABD (CAIM)	0.056	–	–	0.163	–	–
LISABD (ChiMerge)	–	0.281	–	–	0.168	–
LISABD (CACC)	0.051	–	–	–	–	–
LISABD (MDLP)	–	–	–	–	–	–

of risk factors on disease spatial pattern in the model. A higher PD value means that the risk factor has a stronger contribution to disease occurrence. PD values of GDP and elevation discretized with LISABD were calculated by GeogDetector software (Wang and Hu, 2012) and are shown in Table 3. For comparison, we also calculated those values discretized with five other commonly used unsupervised methods with the same intervals as LISABD (Table 3).

3.3. Supervised discretization of multiple continuous risk factors with SSBD

For the 13 risk factors of NTD shown in Table 1, it is meaningful to determine the relationship rules between multiple risk factors and NTD prevalence. These rules can be obtained with geocomputational methods such as decision-tree and rough set, after the continuous risk factors are discretized. Those factors of NTD are spatially heterogeneous. For example, the distribution intervals of elevation can vary with land-use type. Therefore, SSBD was used to discretize the continuous risk factors, for reasonable mining of relationship rules between multiple risk factors and NTD prevalence. Here, we are concerned whether a village has NTD cases; hence, the class has 0 or 1 binary value. The value is 1 when the village has an NTD case, otherwise it is 0. To assess SSBD effectiveness, we analyzed its impacts on rough set-based rule induction between multiple risk factors and binary class.

The rough set proposed by Pawlak (1982) is an extension of classical set theory and can derive classification or decision rules without any prior information. This set has been applied to research fields of spatial analysis (Murgante et al., 2008), spatial classification and uncertainty analysis (Ahlqvist et al., 2000, 2003; Leung et al., 2007; Ge et al., 2011), and geographic knowledge discovery (Beaubouef et al., 2007; Bai et al., 2010). The rough set was chosen because it is a powerful rule-based geocomputational classification method for discrete and continuous data, in which continuous data are discretized prior to rule induction. The rough set-based assessment includes five main parts: (1) Decision table construction. This table is two-dimensional, whose rows correspond to the Heshun County villages. The columns are divided into two parts; the first is called the condition attribute, which contains all the risk factors. The last two columns are called the decision attribute, which corresponds to the class information and spatial stratification. (2) Discretization, which is a data preprocessing procedure that converts continuous into discrete risk factors. For the SSBD method, discretization is performed with the MATLAB GA toolbox. The combination of class and spatial stratification is used to supervise the selection of cut point set for continuous risk factors. For the GA algorithm, the origin generation is 20, and crossover and mutation rates are 0.8 and 0.01, respectively. The fitness function is defined by Eq. (5). (3) Reduction. Redundant values in the decision table can be reduced with the reduction method of Øhrn (1999) after

Table 4
Numbers of cut points of each continuous risk factor, discretized with different discretization methods.

Continuous risk factors	Discretization methods									
	SSBD (GR[a])	SSBD (LC[b])	SSBD (LI[c])	SSBD (SO[d])	SSBD (WA[e])	DBGA	MDLP	Boolean	Naïve	Semi-naïve
Fertilizer	18	15	18	18	10	16	9	6	62	19
Pesticide	54	50	52	54	49	38	13	1	116	76
Fruit	57	50	49	57	51	58	43	3	122	93
Vegetable	61	60	66	61	64	61	75	2	133	129
Net Income	47	53	61	47	62	60	109	3	115	115
GDP	54	61	51	54	61	56	64	3	128	91
Elevation	16	26	26	16	17	22	34	2	55	45
Number of cut points	307	315	323	307	314	311	347	20	731	568

[a] GR, gradient.
[b] LC, land cover.
[c] LI, lithology.
[d] SO, soil.
[e] WA, watershed.

Table 5
Ten-fold cross validation rough set-based classification accuracies for different discretization methods (%).

	Discretization methods									
ID	SSBD (GR[a])	SSBD (LC[b])	SSBD (LI[c])	SSBD (SO[d])	SSBD (WA[e])	DBGA	MDLP	Boolean	Naïve	Semi-naïve
1	58.06	61.29	64.52	64.52	61.29	64.52	61.29	61.29	64.52	61.29
2	67.74	70.97	61.29	58.04	64.52	61.29	64.52	64.52	67.74	70.97
3	67.74	70.97	67.74	67.74	74.19	70.97	70.97	67.74	77.40	54.84
4	70.97	70.97	70.97	67.74	74.19	70.97	70.97	67.74	74.19	74.19
5	87.10	83.87	83.87	83.87	83.87	83.87	74.19	84.65	77.40	83.87
6	84.38	71.88	78.13	65.63	100.00	65.63	53.12	59.38	53.13	53.13
7	68.75	65.63	78.13	71.88	71.88	75.00	78.13	59.38	65.63	78.13
8	59.38	56.25	65.63	50.00	59.38	56.25	56.25	40.63	46.88	62.50
9	68.75	75.00	78.13	68.75	78.13	75.00	75.00	75.00	71.88	78.13
10	84.38	84.38	87.5	84.38	87.50	87.50	87.50	87.50	90.63	68.75
Average accuracy	71.73	71.12	73.59	68.26	75.50	71.10	69.19	66.78	68.94	68.58

[a] GR, gradient.
[b] LC, land cover.
[c] LI, lithology.
[d] SO, soil.
[e] WA, watershed.

discretization. (4) Rule induction. The decision rules are easily constructed by reading values from the training decision table, with only class as the decision attribute after reduction. (5) Classification. The decision rules obtained can then be used to classify the testing decision table, with only class as the decision attribute.

We compared the SSBD discretization results and other commonly used discretization methods in rough set-based applications. Five discrete physical risk factors were used for spatial stratification. The number of cut points for each continuous risk factor is shown in Table 4. Ten-fold cross validation was done for accuracy assessment of rough set-based classification, and classification accuracies are shown in Table 5.

4. Discussion

LISABD discretizes a continuous single attribute by supervising the selection of cut points with the type of spatial autocorrelation clustering acquired with LISA. A supervised discretization method selected from current such methods was used to discretize the spatial data, by designating the type of spatial clustering as class information. LISABD ensures that data instances in the same interval have nearly the same type of spatial clustering, and these types vary across consecutive intervals. For LISABD, the type of spatial clustering is used to reflect similarities of data instances in the same interval and differences across consecutive intervals, which are subjectively measured in traditional unsupervised discretization methods. Therefore, compared with the latter methods, the discretization result of LISABD should have improved spatial data discretization performance.

In our experimental study, LISABD was used to discretize GDP and elevation, and discretization results were compared. Five distinct supervised discretization methods were used in LISABD. The numbers of cut points of GDP and elevation discretized with LISABD are shown in Table 2. Compared to the traditional unsupervised methods, the cut points were obtained without the user predefining the number. To assess LISABD performance, the geographical detectors-based PD values of GDP and elevation were calculated and compared, as shown in Table 3. This table shows that PD values are higher for LISABD than for five unsupervised discretization methods. Especially when the Chi2 method was used in LISABD, the PD values of GDP and elevation were nearly maximized. However, the PD values varied when the five different supervised discretization methods were used in LISABD. For example, the Chi2 method was clearly superior to the other supervised discretization methods. These results show that the selection of supervised discretization method affects PD values for LISABD. Therefore, this selection is important in real applications of LISABD. An appropriate method can be selected from popular supervised methods by comparing the results of geocomputational methods.

SSBD discretizes multiple attributes by supervising the selection of cut points with the combination of spatial stratification and class information. Since traditional multiple-attribute supervised methods only use class information to supervise the selection of cut points and convert continuous data to discrete intervals representing different knowledge about varying types of class, data instances in the same interval may be spatially heterogeneous. The knowledge about varying types of class should be more meaningful if the data instances are not spatially heterogeneous. For SSBD, spatial stratification is also used to supervise the selection of cut points except for class information, causing data instances in the same interval to have nearly the same class labels and no spatial heterogeneity, and the class labels vary across consecutive intervals. SSBD should also have better performance for further geocomputational classification, prediction and modeling when spatial data have heterogeneity.

SSBD was used to discretize multiple continuous risk factors of NTD, and the five discrete physical risk factors were used for spatial stratification. To assess SSBD effectiveness, the number of cut points and rough set-based classification accuracies were compared between SSBD and other commonly used discretization methods. The numbers of cut points of these discretization methods are shown in Table 4. This table shows that the numbers of cut points of continuous risk factors discretized with SSBD are almost the same, although the spatial stratification is distinct. However, compared to SSBD, the numbers of cut points are significantly different for the Boolean, Naïve and Semi-naïve methods. There were only 20 cut points for Boolean method, much less than SSBD methods. In contrast, the numbers for Naïve and Semi-naïve were 731 and 568 respectively, much more than SSBD. Discretization generates useful geographic knowledge from spatial data by reducing data size. However, neither too many nor too few cuts points assist geographic knowledge discovery, since such knowledge is not representative in those cases. Knowledge is too fine with too many cut points, and too rough with too few. Table 4 indicates that SSBD has an appropriate number of cut points. The rough set was used to assess SSBD performance. The ten-fold cross validation was done, and classification accuracies are shown in Table 5. This table shows that average classification accuracies based on SSBD exceeded 70%, except for SSBD (SO). Compared to SSBD, average classification accuracies based on MDLP, Boolean, Naïve and Semi-naïve were <70%. The highest average classification accuracy was 75.5%, based

on SSBD (WA), and the lowest was 66.78%, based on Boolean. The Boolean method had only 20 cut points, producing much loss of information in the discretization process, which results in lesser accuracy. The highest accuracy, based on SSBD (WA), indicates that the watershed has greater influence on NTD prevalence than other physical risk factors. This agrees with another study on NTD (Wang et al., 2010). The relationship rules between multiple risk factors and NTD mined with the rough set after continuous risk factor discretization with SSBD (WA) can improve the relevance of those rules.

5. Conclusions and future work

We investigated two types of spatial data discretization methods, called LISABD and SSBD. These consider spatial autocorrelation and spatial heterogeneity in their discretization processes, respectively. Those processes were exemplified using NTD data. The geographic detector-based and rough set-based analysis results of NTD data demonstrated the effectiveness of the two discretization methods, respectively. LISABD enables the use of supervised methods for continuous spatial data discretization, by considering spatial autocorrelation. Comparing with traditional unsupervised methods, a predefined number of intervals is not necessary for LISABD. However, the number of intervals is relevant to the supervised discretization methods examined. Thus, the selection of appropriate supervised method requires additional accuracy comparison in real applications. For SSBD, spatial stratification is used to help achieve the optimal discretization result for multiple continuous spatial attributes with spatial heterogeneity. However, discretization results are influenced by the selection of spatial stratification. Therefore, the selection of suitable spatial stratification is important in real applications. The GA optimization method used in SSBD has a relatively high temporal computation requirement, especially for large amounts of spatial data. Particle swarm optimization (PSO) is also a global optimization method, which is easily implemented and computationally inexpensive (Parsopoulos and Vrahatis, 2002). To overcome the computation requirement drawback, we will replace the GA method with PSO, because of its fast computation and easy convergence.

Aside from spatial autocorrelation and spatial heterogeneity, spatial data has other important features, including spatial structure. This structure implies important spatial information and is useful in the analysis of geographic issues (Brunsell and Anderson, 2011; Ge and Bai, 2011). In future work, we will incorporate spatial structure information in the discretization process of spatial data, toward acquisition of discretization results.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant Nos. 40971222, 41271404, 41023010), the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KZCX2-EW-QN303) and the National Basic Research Program (Grant No. 2012CB955503) of China's Ministry of Science and Technology (973).

References

Ahlqvist, O., Keukelaar, J., Oukbir, K., 2000. Rough classification and accuracy assessment. International Journal of Geographical Information Science 14, 475–496.
Ahlqvist, O., Keukelaar, J., Oukbir, K., 2003. Rough and fuzzy geographical data integration. International Journal of Geographical Information Science 17, 223–234.
Anselin, L., 1995. Local indicators of spatial association—LISA. Geographical Analysis 27, 93–115.
Armstrong, M.P., Xiao, N.C., David, A., Bennett, D.A., 2003. Using genetic algorithms to create multicriteria class intervals for choropleth maps. Annals of the Association of American Geographers 93, 595–623.
Bai, H.X., Ge, Y., Wang, J.F., Liao, Y.L., 2010. Using rough set theory to identify villages affected by birth defects: the example of Heshun, Shanxi, China. International Journal of Geographical Information Science 24, 459–476.
Beaubouef, T., Petry, F.E., Ladner, R., 2007. Spatial data methods and vague regions: a rough set approach. Applied Soft Computing 7, 425–440.
Berger, P.A., 2004. Rough set rule induction for suitability assessment. Environmental Management 34, 546–558.
Brunsell, N.A., Anderson, M.C., 2011. Characterizing the multi-scale spatial structure of remotely sensed evapotranspiration with information theory. Biogeosciences 8, 2269–2280.
Carmona, R.H., 2005. The global challenges of birth defects and disabilities. Lancet 366, 1142–1144.
Cerquides, J., Mantaras, R.L., 1997. Proposal and empirical comparison of a parallelizable distance-based discretization method. In: Proceeding of the Third International Conference on Knowledge Discovery and Data Mining, Newport beach, California, pp. 139–142.
Ching, J.Y., Wong, A.K.C., Chan, K.C.C., 1995. Class-dependent discretization for inductive learning from continuous and mixed-mode data. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 641–651.
Clarke, K.C., 2003. Geocomputation's future at the extremes: high performance computing and nanoclients. Parallel Computing 29, 1281–1295.
Couclelis, H., 1998. Geocomputation in context. In: Longley, P.A., et al. (Eds.), Geocomputation: A Primer. John Wiley, Chichester, pp. 17–29.
Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Proceedings of the Sixth International Conference on Hybrid Intelligent systems, Adelaide, pp. 194–202.
Du, Y., Wen, W., Cao, F., Ji, M., 2010. A case-based reasoning approach for land use change prediction. Expert Systems with Applications 37, 5745–5750.
Fayyad, U.M., Irani, K.B., 1992. On the handling of continuous-valued attributes in decision tree generation. Machine Learning 8, 87–102.
Fischer, M.M., Leung, Y., 2001. Geocomputational Modelling: Techniques and Applications. Springer Verlag, Berlin.
Fischer, M.M., Wang, J.F., 2011. Spatial Data Analysis Models, Methods and Techniques. Springer, New York.
Fotheringham, S., Clarke, G., Abrahart, B., 1997. Geocomputation and GIS. Transactions in GIS 2, 199–200.
Gahegan, M., 1999. What is geocomputation? Transactions in GIS 3, 203–206.
Ge, Y., Bai, H.X., 2011. Multiple-point simulation-based method for extraction of objects with spatial structure from remotely sensed imagery. International Journal of Remote Sensing 32, 2311–2335.
Ge, Y., Cao, F., Duan, R.F., 2011. Impact of discretization methods on the rough set-based classification of remotely sensed images. International Journal of Digital Earth 4, 330–346.
Haining, R.P., 2003. Spatial Data Analysis: Theory and Practice. Cambridge University Press, London.
Ho, K.M., Scott, P.D., 1998. An efficient global discretization method. Research and Development in Knowledge Discovery and Data Mining 1394, 383–384.
Hu, Y., Wang, J.F., Li, X.H., Ren, D., Zhu, J., 2011. Geographical detector-based risk assessment of the under-five mortality in the 2008 Wenchuan earthquake, China. PLoS ONE 6, e21427.
Jenks, G.F., Caspall, F.C., 1971. Error on choroplethic maps: definition, measurement, reduction. Annals of the Association of American Geographers 61, 217–244.
Kerber, R., 1992. Chimerge: discretization of numeric attributes. In: Proceeding of the Tenth National Conference on Artificial Intelligence, San Joe, California, pp. 123–128.
Kurgan, L., Cios, K.J., 2001. Discretization algorithm that uses class-attribute interdependence maximization. In: Proceedings of the 2001 International Conference on Artificial Interlligence, Las Vegas, Vevada, pp. 980–987.
Leung, Y., Fung, T., Mi, J.S., Wu, W.Z., 2007. A rough set approach to the discovery of classification rules in spatial data. International Journal of Geographical Information Science 21, 1033–1058.
Lanorte, A., Danese, M., Lasaponara, R., Murgante, B., 2013. Multiscale mapping of burn area and severity using multisensor satellite data and spatial autocorrelation analysis. International Journal of Applied Earth Observation and Geoinformation 20, 42–51.
Liao, Y.L., Wang, J.F., Guo, Y.Q., Zheng, X.Y., 2010. Risk assessment of human neural tube defects using a Bayesian belief network. Stochastic Environmental Research And Risk Assessment 24, 93–100.
Liu, H., Hussain, F., Tan, C.L., Dash, M., 2002. Discretization: an enabling technique. Data Mining and Knowledge Discovery 6, 393–423.
Liu, H., Setiono, R., 1995. Chi2: feature selection and discretization of numeric attributes. In: Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, Herndon, Virginia, pp. 388–391.
Lustgarten, J.L., Visweswaran, S., Gopalakrishnan, V., Cooper, F.G., 2011. Application of an efficient Bayesian discretization method to biomedical data. BMC Bioinformatics 12, 309.
Monmonier, M.S., 1972. Contiguity-biased class-interval selection: a method for simplifying patterns on statistical maps. Geographical Review 62, 203–228.
Murgante, B., Las Casas, G., Sansone, A., 2008. A spatial rough set for extracting periurban fringe. In: Revue des Nouvelles Technologies de l'Information, vol. 857. Éditions Cépaduès, Toulouse, France, pp. 101–125.
Murgante, B., Borruso, G., Lapucci, A., 2009. Geocomputation and Urban Planning. Springer-Verlag, Berlin.

Murgante, B., Borruso, G., Lapucci, A., 2011. Geocomputation, Sustainability and Environmental Planning. Springer-Verlag, Berlin.
Murgante, B., Danese, M., 2011. Urban versus Rural: the decrease of agricultural areas and the development of urban zones analyzed with spatial statistics. International Journal of Agricultural and Environmental Information Systems 2, 16–28.
Nguyen, S.H., Nguyen, S.H., Skowron, A., 1996. Searching for features defined by hyperplanes. Foundations of Intelligent Systems 1079, 366–375.
Nguyen, S.H., Skowron, A., 1995. Quantization of real value attributes: rough set and boolean reasoning approach. In: Proceedings of the International Workshop on Rough Sets Soft Computing at Second Annual Joint Conference on Information Sciences, Wrightsville Beach, North Carolina, USA, pp. 34–37.
Openshaw, S., 1998. Building automated geographical analysis and explanation machines. In: Longley, P.A., Brooks, S.M., McDonnell, R., Macmillan, B. (Eds.), Geocomputation: A Primer. John Wiley and Sons, Chichester, pp. 95–115.
Openshaw, S., Abrahart, R.J., 2000. GeoComputation. Taylor & Francis, London.
Parsopoulos, K.E., Vrahatis, M.N., 2002. Recent approaches to global optimization problems through Particle Swarm Optimization. Natural Computing 1, 235–306.
Pawlak, Z., 1982. Rough sets. International Journal of Computer and Information Sciences 11, 341–356.
Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81–106.
Øhrn, A., 1999. Discernibility and Rough Sets in Medicine: Tools and Applications, Computer and Information Science. Norwegian University of Science and Technology, Trondheim.
Robinson, A.H., 1982. Early Thematic Mapping in the History of Cartography. University of Chicago, Chicago.
Son, N., Hoa, N., Skowron, A., 1996. Searching for features defined by hyperplanes. Foundations of Intelligent Systems, 366–375.
Stewart, J., Kennelly, P.J., 2010. Illuminated choropleth maps. Annals of the Association of American Geographers 100, 513–534.
Su, C.T., Hsu, J.H., 2005. An extended Chi2 algorithm for discretization of real value attributes. IEEE Transactions on Knowledge and Data Engineering 17, 437–441.
Tay, F.E.H., Shen, L., 2002. A modified chi2 algorithm for discretization. IEEE Transactions on Knowledge and Data Engineering 14, 666–670.
Tsai, C.J., Lee, C.I., Yang, W.P., 2008. A discretization algorithm based on Class-Attribute Contingency Coefficient. Information Sciences 178, 714–731.
Wang, G.Y., 2001. Rough Set Theory and Knowledge Acquisition. Xi'an Jiaotong University Press, Xi'an.
Wang, J.F., Hu, Y., 2012. Environmental health risk detection with GeogDetector. Environmental Modelling & Software 33, 114–115.
Wang, J.F., Li, X.H., Christakos, G., Liao, Y.L., Zhang, T., Gu, X., Zheng, X.Y., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. International Journal of Geographical Information Science 24, 107–127.
Wang, K., Liu, B., 1998. Concurrent discretization of multiple attributes. In: The Pacific Rim International Conference on Artificial Intelligence, Singapore, pp. 250–259.