International Journal of Digital Earth
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t777764757

Impact of discretization methods on the rough set-based classification of remotely sensed images

Y. Ge[a]; F. Cao[a]; R. F. Duan[a]
[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing, P.R. China

First published on: 29 June 2010

To cite this Article Ge, Y. , Cao, F. and Duan, R. F.(2011) 'Impact of discretization methods on the rough set-based classification of remotely sensed images', International Journal of Digital Earth, 4: 4, 330 — 346, First published on: 29 June 2010 (iFirst)
To link to this Article: DOI: 10.1080/17538947.2010.494738
URL: http://dx.doi.org/10.1080/17538947.2010.494738

International Journal of Digital Earth,
Vol. 4, No. 4, July 2011, 330–346

Taylor & Francis
Taylor & Francis Group

Impact of discretization methods on the rough set-based classification of remotely sensed images

Y. Ge*, F. Cao and R.F. Duan

State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, P.R. China

(*Received 24 December 2009; final version received 14 May 2010*)

In recent years, the rough set (RS) method has been in common use for remote-sensing classification, which provides one of the techniques of information extraction for Digital Earth. The discretization of remotely sensed data is an important data preprocessing approach in classical RS-based remote-sensing classification. Appropriate discretization methods can improve the adaptability of the classification rules and increase the accuracy of the remote-sensing classification. To assess the performance of discretization methods this article adopts three indicators, which are the compression capability indicator (CCI), consistency indicator (CI), and number of the cut points (NCP). An appropriate discretization method for the RS-based classification of a given remotely sensed image can be found by comparing the values of the three indicators and the classification accuracies of the discretized remotely sensed images obtained with the different discretization methods. To investigate the effectiveness of our method, this article applies three discretization methods of the Entropy/MDL, Naive, and SemiNaive to a TM image and three indicators for these discretization methods are then calculated. After comparing the three indicators and the classification accuracies of the discretized remotely sensed images, it has been found that the SemiNaive method significantly reduces large quantities of data and also keeps satisfactory classification accuracy.

Keywords: remote sensing; classification; rough set; discretization; image processing; data mining

1. Introduction

Digital Earth originally proposed by Gore (1998) is an information expression of the real Earth and is a new way of understanding the Earth in the twenty-first century (Guo *et al.* 2009). It is mainly composed of the following five phases: data extraction, information extraction, knowledge extraction, modeling, and decision making (Chen and van Genderen 2008). Remote-sensing technology provides a strong technical support for the phase of data extraction, while information extraction techniques, such as image classification, geo-statistical analysis, and data mining can extract relevant information from these huge data archives and data bases. In recent years, rough set (RS) theory, proposed by Pawlak (1982, 1991), has already been used in geographical fields, such as spatial analysis, information extraction, and uncertainty

*Corresponding author. Email: gey@lreis.ac.cn

ISSN 1753-8947 print/ISSN 1753-8955 online

DOI: 10.1080/17538947.2010.494738
http://www.informaworld.com

analysis and geo-knowledge discovery (Ahlqvist *et al.* 2000, Bittner 2001, Bittner and Stell 2001, Ahlqvist *et al.* 2003, Berger 2004, Wang *et al.* 2004, Beaubouef *et al.* 2007, Ge *et al.* 2009, Bai *et al.* 2010). Especially, some applications focus on remotely sensed data preprocessing and on the RS-based classification (Pal and Mitra 2002, Wu 2004, Ouyang and Ma 2006, Li *et al.* 2007, Leung *et al.* 2007, Li *et al.* 2008, Xiao and Zhang 2008). For example, Wu (2004), in research on remote-sensing classification using RS method, investigated data preprocessing, classifier designing, and classification evaluation. Leung *et al.* (2007) used RS to extract classification rules of remotely sensed data. The experimental results demonstrate that it can effectively discover in remotely sensed data the optimal spectral bands and optimal rule set for a classification task. Lei *et al.* (2008) used discrete RS to extract the texture information rules of remotely sensed image and added in its classification. The overall accuracy of classification with texture information extracted by discrete RS is higher than the overall accuracy of classification with texture information extracted by principal components analysis (PCA).

In addition to apply RS to the above studies, RS can also be integrated with other methods, such as support vector machines, neural network, and fractal to improve the accuracy of classification of remote sensing (Wu 2001, Liu *et al.* 2004, Ma and Hasi 2005, Zhang *et al.* 2005, Das *et al.* 2006, Zhan *et al.* 2007). This is particularly attractive because it combines the advantages of RS and other methods in data mining to improve the accuracy of classification of remote sensing. Commonly, the gray values of spectral bands for an 8-bit gray image are within 0–255, therefore, the gray values are considered as continuous (numerical). The term 'continuous' is used to indicate both real- and integer-valued attributes. However, the RS theory assumes that all attributes are nominal, so continuous-valued attributes must be discretized (Fayyad and Irani 1992). Discretization of attributes is an important data preprocessing approach in machine learning, particularly for the classification problem. Empirical results have shown that the quality of classification methods depends on the discretization method used in the preprocessing step (Nguye 1998). In past decades, discretization of attributes has received significant attention and many discretization methods have been developed, such as 1RD (Holte 1993), C4.5 (Quinlan 1993), Entropy/MDL (Fayyad and Irani 1992, 1993, Dougherty 1995), Naive (Øhrn 1999), and SemiNaive (Øhrn 1999). In the rule extraction of remote-sensing information based on classical RS, the discretization of remote-sensing data plays an important role. Reasonable discretization method can reduce the data size of remote sensing and improve its quality. The rules extracted with RS after reasonable discretization are then more understandable and concise. However, in the current applications of the discretization methods, few discussions on the selection and comparison of different discretization methods are given in the information extraction from remotely sensed images (Duan *et al.* 2007, Zhang *et al.* 2008).

This article investigates the differences between different discretization methods and uses three indicators, which are data compression capability indicator (CCI), consistency indicator (CI), and number of the cut points (NCP) to assess the efficiencies of these methods for the given data (Yue 2006). The impact of different discretization methods on classification accuracies is then implemented. The result shows that these three indicators integrating with the analysis of their influences on the classification accuracy can help the user determine the choice of discretization method for remote-sensing classification.

2. Discretization of attributes

The value of attributes mainly consists of nominal (categorical) or continuous (numerical). The nominal value mainly contains string and enum, while continuous value mainly contains integer numbers and float numbers (Wang 2001). For example, the color attribute value of remotely sensed images is nominal, which can be expressed by enum, such as red, green, and blue. The gray value of pixels is continuous, which is integer. To adapt to intelligence method used in the image procession of remote sensing, the continuous values of the remote sensing should be discretized into nominal values. Sometimes, the nominal values will be discretized further to acquire more abstract discretization values. In general, discretization is a process of searching for partition of attribute domains into intervals and unifying the values over each interval (Nguyen 1998). Hence discretization can be defined as a problem of searching for a suitable set of cuts (i.e. boundary points of intervals) on attribute domains (Nguyen 1998). According to different criteria, the discretization methods can be roughly classified into global/local, supervised/unsupervised, and static/dynamic. The commonly used discretization methods include supervised and unsupervised. The supervised discretization methods use the decision class information in setting cut points, so the discretization result can provide effective help for further classification. And the commonly used supervised methods include 1RD, Entropy/MDL, Naive, and SemiNaive (Dougherty 1995, Øhrn 1999). The unsupervised discretization methods do not consider the decision class information in setting cut points and they mainly contain Equal Interval and Width. In this article, Entropy/MDL, Naive, and SemiNaive are taken and compared by using these three indicators to exemplify the effect of discretization on the remote-sensing classification result. The three discretization methods used will be introduced as follows.

2.1. Discretization methods of continuous attributes

2.1.1. Entropy/MDL method

Entropy/MDL method (Fayyad and Irani 1992, 1993, Dougherty *et al.* 1995) uses the class information entropy of candidate partitions to select bin boundaries for discretization. For a set of instances S, Let there be k classes $C_1,\ldots, C_k$. Let $P(C_i, S)$ be the proportion of instances in S that have class C_i. The class entropy of S is defined as:

$$\text{Ent } (S) = -\sum_{i=1}^{k} P(C_i,\ S)\ \log(P(C_i,\ S)). \tag{1}$$

Given a set of instances S, a feature A, and a partition boundary T, the class information entropy of the partition induced by T, denoted $E(A, T; S)$ is given by:

$$E(A,\ T;\ S) = \frac{|S_1|}{|S|}\text{Ent}(S_1) + \frac{|S_2|}{|S|}\ \text{Ent } (\mathrm{S}_2), \tag{2}$$

where $|S|$ is the number of instances in the set S, $\mathrm{S}_1 \subset S$ and $S_2 = S–S_1$.

For a given feature A, the boundary $T_{\min}$ which minimizes the entropy function over all possible partition boundaries is selected as a binary discretization boundary. This method can then be applied recursively to both of the partitions induced by

$T_{\min}$ until the stopping condition minimal description length principle defined by Fayyad and Irani is achieved, thus creating multiple intervals on the feature A.

Recursive partitioning within a set of values S stops iff

$$\text{Gain } (A, T; S) < \frac{\log_2 (N-1)}{N} + \frac{\Delta (A, T; S)}{N}, \tag{3}$$

where N is the number of instances in the set S,

$$\text{Gain } (A, T; S) = \text{Ent } (S) - E (A, T; S), \tag{4}$$

$$\Delta (A, T; S) = \log_2 (3^k - 2) - [k \times \text{Ent}(S) - k_1 \times \text{Ent}(S_1) - k_2 \times \text{Ent}(S_2)], \tag{5}$$

and k_i is the number of class labels represented in the set S_i. Since the partitions along each branch of the recursive discretization are evaluated independently using this criteria, some areas in the continuous spaces will be partitioned very finely whereas others (which have relatively low entropy) will be partitioned coarsely.

2.1.2. Naive method

The value for each condition attribute 'a' is sorted in Naive method (Øhrn 1999). And then, the instances in the universe are scanned. For two adjacent instances x_i and x_j in the universe, the average value of the two instances is set the value of the cut point, when $a(x_i) \neq a(x_i)$ and $d(x_i) \neq d(x_i)$ (which means that the values and class types are different for the two instances). Naive method doesn't need any extra parameters and sets cut point between two instances which have a different attribute value and decision value, regarding the cut point is very important. But it does not consider the indiscernibility among instances. Consequently, many important cut points will be ignored, which should be chosen for keeping the indiscernibility unchanged, having great contribution for classification. At the same time, the cut point will have a large distinction when the attribute values of the instances are sorted according to different order. Naive increases the cut points step-by-step and usually gets a large set of cut points.

2.1.3. SemiNaive method

SemiNaive method is similar to naive method, but has more logic to handle the case where value-neighboring instances belong to different decision classes (Øhrn 1999). The set of cut points found by SemiNaive method is a subset of the cut points found by naive method. SemiNaive method scans the cut points found by naive method and decides which cut points are needed further. It is supposed that c is a cut point of attribute a. Also, x_i and x_j are two neighbor values of cut c, and D_i and D_j are the dominant decision value set. D_i corresponds to the equivalence class containing D_i while D_j corresponds to the equivalence class containing x_j. The cut point c is deleted from the set of the cut points found by naive method when $D_i \subseteq D_j$ or $D_j \subseteq D_i$, otherwise c is considered as an important cut point in the set of the cut points. SemiNaive method is considered as the optimization of the naive method due to reducing some redundant cut points. However, compared with Naive method, SemiNaive method might cause more inconsistent data.

2.2. Indicators for assessing the discretization method

The purpose of discretization is a process of grouping the values of the attributes in intervals in such a way that the knowledge content or the discernibility is not lost (Roy and Pal 2003). Discretization of the attributes can reduce the redundant data of the data base and then achieve the purpose of data compression. In order to find an appropriate discretization method for classical RS-based remote-sensing classification, in this article, three indicators of CCI, CI, and NCP are adopted to evaluate discretization methods for classical rough set-based remote-sensing classification. Yue (2006) used these indicators to compare the different discretization methods and the experimental results have shown that these indicators are effective in analyzing the difference between discretization methods. Therefore, these three indicators are used in this article to evaluate the discretization method for remote-sensing classification. By comparing the values of these three indicators and the classification accuracies of the discretized remotely sensed images with the different discretization methods, the appropriate discretization method will be acquired. The spectral bands of Landsat TM image are used to exemplify the method. First, the spectral bands of the Region of Interest (ROI) are discretized with different discretization methods and the cut points of all spectral bands are then acquired. Three indicators, which are defined as follows, are calculated.

Let $A_{\text{Band\_1}}$, $A_{\text{Band\_2}}$, $A_{\text{Band\_3}}$, $A_{\text{Band\_4}}$, $A_{\text{Band\_5}}$, $A_{\text{Band\_6}}$, $A_{\text{Band\_7}}$ denote the seven bands and D_{class} denote the class for ROI image. In RS theory, $A_{\text{Band\_1}}$, $A_{\text{Band\_2}}$, $A_{\text{Band\_3}}$, $A_{\text{Band\_4}}$, $A_{\text{Band\_5}}$, $A_{\text{Band\_6}}$, $A_{\text{Band\_7}}$ are called condition attributes and D_{class} is called decision attribute.

Let $A = \wedge A_{\text{Band\_j}}$

$$A(i) = \wedge A_{\text{Band\_j}}(i)$$

$$= A_{\text{Band\_1}}(i) \wedge A_{\text{Band\_2}}(i) \wedge A_{\text{Band\_3}}(i) \wedge A_{\text{Band\_4}}(i)$$

$$\wedge A_{\text{Band\_5}}(i) \wedge A_{\text{Band\_6}}(i) \wedge A_{\text{Band\_7}}(i)\ i = 1, \ldots, n, j = 1, \ldots, 7$$

where n is the number of pixels within ROI, $A_{\text{Band\_j}}(i)$ denotes the spectral bands values of pixel i in band_j and $D_{\text{class}}(i)$ denotes the class value of pixel i. For a set S, $|S|$ is the number of instances in the set.

2.2.1. Compression capability indicator (CCI)

$$\text{CCI} = 1 - \frac{|\{A \wedge D_{class}(i) | i = 1, 2, \ldots, n\}^d|}{|\{A \wedge D_{class}(i) | i = 1, 2, \ldots, n\}|} \tag{6}$$

In formula (6), d denotes that the set which is derived from the discretized ROI Image. Discretization will bring about a reduction of the data size and loss of information, but it can generate useful knowledge or rules from the large quantity of data. The CCI reflects the data processing ability of different discretization methods.

2.2.2. Consistent indicator (CI)

$$\text{CI} = 1 - \frac{|\{A \wedge D_{class}(i) | i = 1, \ldots, n\}^{dI}|}{|\{A \wedge D_{class}(i) | i = 1, \ldots, n\}^{d}|} \tag{7}$$

In formula (7), *dI* denotes that the set which is derived from the discretized ROI image and objects of the set are inconsistent. If two objects in the set have the same $A(i)$ values, but the $D_{\text{class}}(i)$ values are different, the two objects are called inconsistent objects. The consistent indicator can reflect the degree of the loss of category information owing to the discretization method.

2.2.3. Number of the cut points (NCP)

The NCP of all spectral bands is calculated when different discretization methods are used to discrete each spectral band. Discretization of the spectral bands is a process of searching for partition of spectral bands domains into intervals. For example, for a band of TM image, the gray values are within 16–142. Suppose 25.5, 70.5, and 125.5 are all the cut points of the band, the band is divided into four intervals [16, 25], [26, 70], [71, 125], and [126, 142]. The gray values in the same intervals are regarded as indiscernibility and usually designated the same value. The NCP is an important feature of discretization method.

3. Experimental study

These three indicators defined in Section 2.2 will be exemplified to evaluate the discretization method used in remote-sensing classification. The impact of different discretization methods on the classification of remotely sensed images is then analyzed. After the analysis of these three indicators and the impact of different discretization methods on the classification, the appropriate discretization method is acquired. The flow chart of this experiment is shown in Figure 1. The Entropy/MDL, Naive, and SemiNaive methods are used in this example.

3.1. Data description

A Landsat TM image of the Yellow River Delta in China on 28 August 1999 is used to substantiate the conceptual discussion and demonstrate the application of the above-discussed method. A verification data obtained by fusing PANchromatic (PAN) band of Systeme Probatoire d'Observation de la Tarre (SPOT) two images acquired on 16th October 2002 and Enhanced Thematic Mapper (ETM) on 9 August 2001 is applied to test the analytical result. The spatial resolution of this verification image is 10 m. The TM image size is 515×515 pixels and the resolution is 30 m except that the spatial resolution of band six is 120 m. The size of verification data is 1545×1161 pixels. Figure 2 is the 4, 3, 2-band pseudo-color composition image.

There are 26,639 pixels selected as the ROI from the study area by using a random-sampling scheme according to prior knowledge and each pixel has seven different spectral values and a class value. The ROI image is shown in Figure 3.

3.2. Discretization of the Region of Interest (ROI) and study area

The spectral bands of the ROI are discretized with Entropy/MDL, Naive, and SemiNaive methods and the cut points of all spectral bands are then acquired. The cut points acquired are sorted and the values of each pixel of the ROI are divided into several intervals for each spectral band. With these sorted cut points, the whole

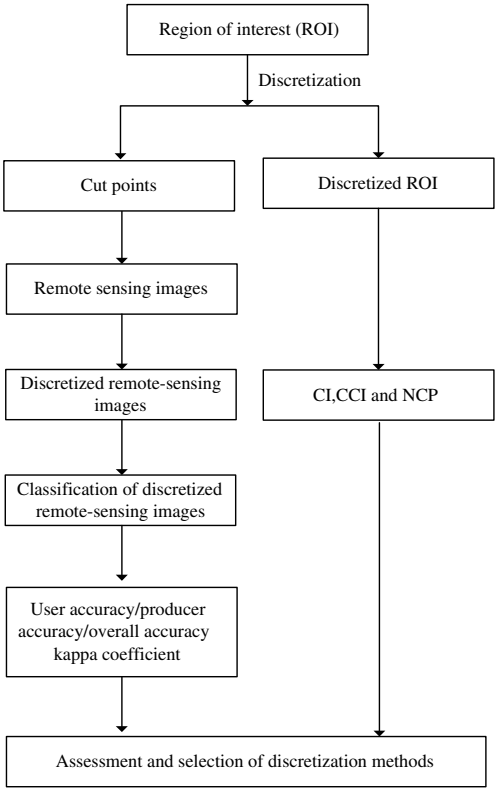

Figure 1. The flow chart of the experiment.

image is discretized. The values of the pixels in each interval are then set to the same value. The pixels with values in the same interval are indiscernibility and their values are designated as the average value in that interval in this article. The pseudo-color composition images of the study area discretized with different discretization methods are shown in Figure 4.

CCI, CI, and NCP of the ROI with different discretization methods are calculated and shown in Tables 1 and 2. In Table 1, the values of fields of original data, discretization data, and inconsistent data are obtained by removing the repeating items. For example, if the values of condition attributes and decision attributes of pixel i and pixel j are completely identical, these two items are counted into one item.

3.3. Results

After comparing the CCI, CI, and NCP of different discretization methods, the impact of the three discretization methods on the classification result of remotely sensed images will be analyzed. Here, the original and discretized remotely sensed images with Entropy/MDL, Naive, and SemiNaive discretization methods are classified with RS. Also, the results with classical RS classifier are then compared with the result with maximum likelihood classifier (MLC) method. The chosen

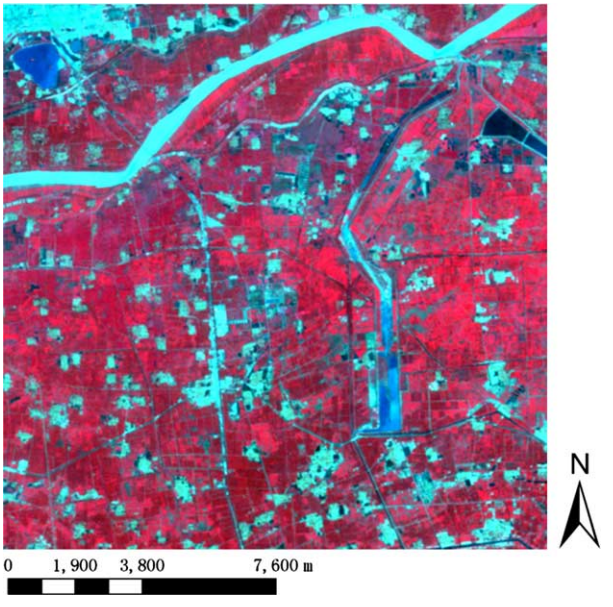

Figure 2. Landsat TM pseudo-color composition image (RGB 4, 3, 2) of the study area acquired on 28 August 1999.

supervised classifier MLC is one of the most popular tools for classification in remotely sensed images processing and discussed much in the literature. The MLC and RS classification results are shown in Figure 5.

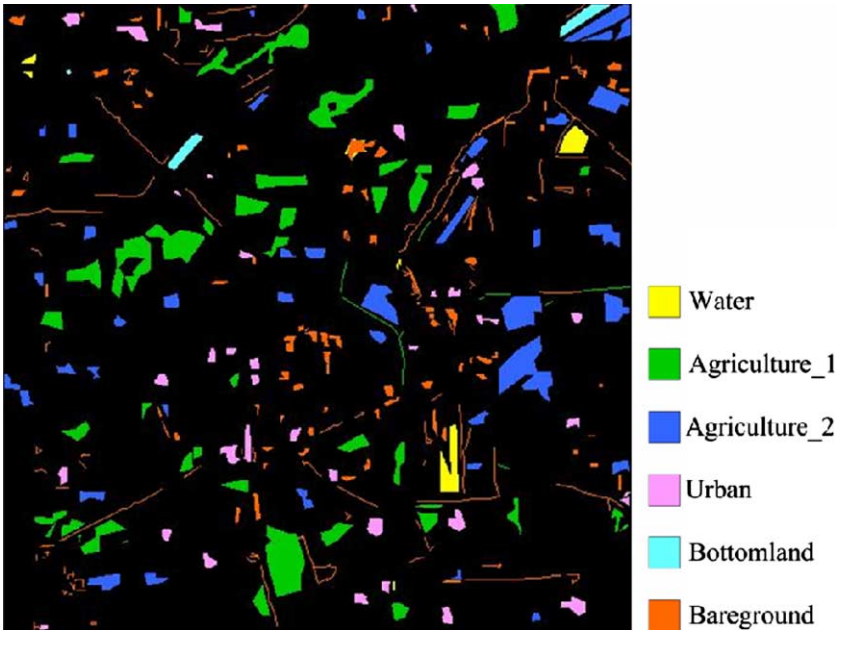

Figure 3. Sample data collected by using stratified random sampling scheme.

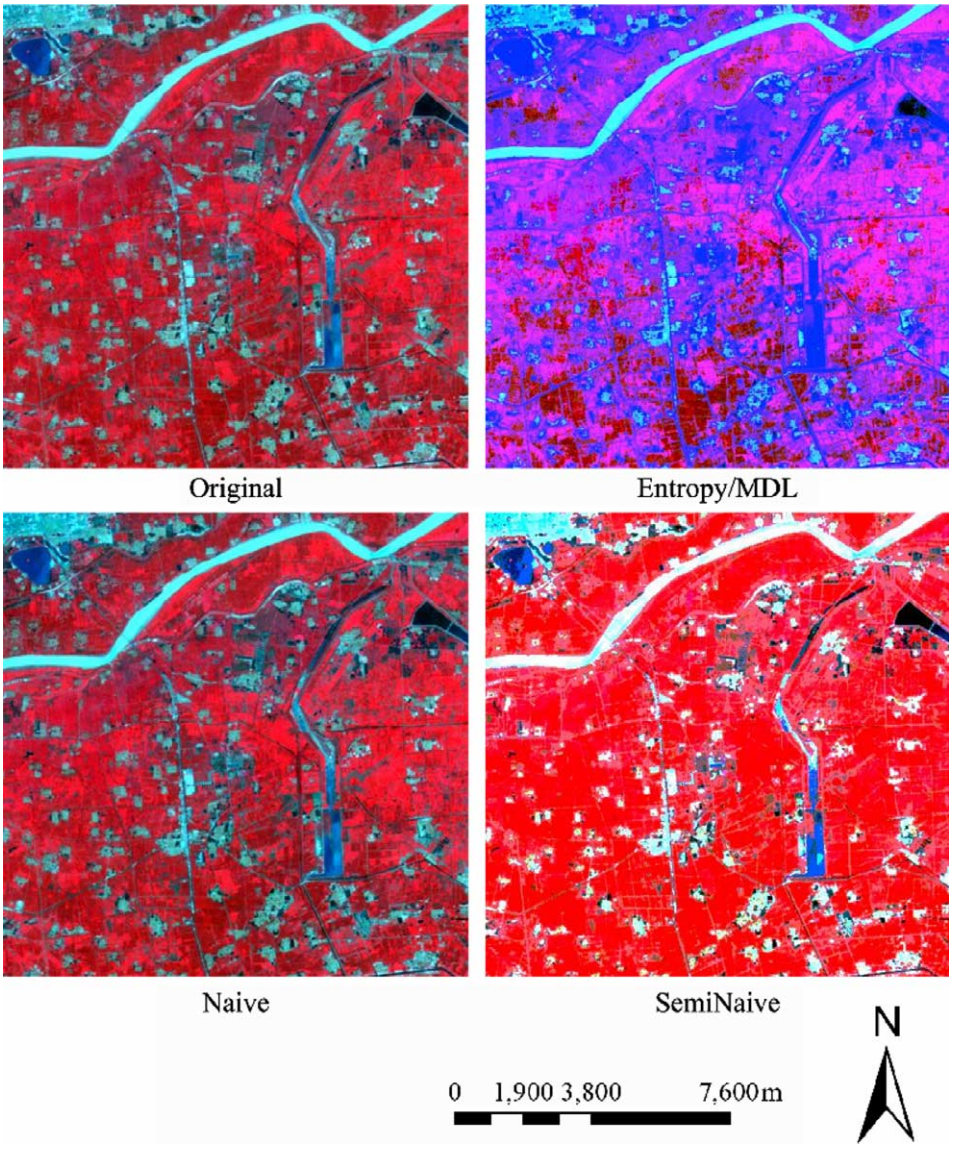

Figure 4. Pseudo-color composition image of the study area discretized by Entropy/MDL, Naive, and SemiNaive methods.

To validate the accuracy of RS classification results of original and discretized remotely sensed images with different discretization methods, this article presents a group of error matrices. The reference data are selected according to prior knowledge of the same area from the SPOT image with resolution 2.5 m random sampling scheme is used. The sample unit is a single pixel in the SPOT image, representing a 10×10 m^2 area of the ground data. According to Edwards *et al.* (1998), at least $n = u^2_{1-\alpha/2}/d^2 p \times (1-p)$ samples are requested to be chosen, where n is the minimal number of samples required, α is a parameter determining the confidence level, u is a

Table 1. Data compression situation and consistent situation of the ROI with Entropy/MDL, Naive, and SemiNaive methods.

Discretization method	Original data	Discretization data	Inconsistent data	CCI (%)	CI (%)
Entropy/MDL	26065	12243	505	53.03	95.88
Naive	26065	25700	4	1.40	99.98
SemiNaive	26065	1278	497	95.10	61.12

value corresponding to α in the Gaussian distribution, d is the desired precision, and p is the estimated accuracy of the classification result. When $\alpha = 0.05$, $u = 1.96$, and $p = 0.5$, there should be at least 384 samples. Here 1000 samples comprising 22 samples for water, 290 samples for agriculture_1, 202 samples for agriculture_2, 121 samples for urban, 18 samples for bottomland, and 347 samples for bareground are collected. The confusion matrixes, classification accuracy and kappa coefficients are acquired. The confusion matrixes of MLC and RS classification results are shown in Table 3, respectively. The producer' accuracy, user' accuracy, overall accuracy, and kappa coefficients of MLC and RS classification results are shown in Table 4.

To analyze the effect of different discretization methods on the RS classification results more clearly, the comparison of the producer' accuracy, user' accuracy, overall accuracy, and kappa coefficients are depicted in Figure 6.

4. Discussion

It can be seen from Figure 4 that the discretized remotely sensed image with Naive method is most similar to the original image, having least loss of the value information of the spectral bands. The discretized remotely sensed images obtained by the SemiNaive and Entropy/MDL methods look different to the original image. The differences between discretized and original images intuitively show the spectral information with different discretization methods. The comparison of the CCI, CI, and NCP for different discretization methods are shown in Tables 1 and 2. From Tables 1 and 2, it can be seen that the CI of the SemiNaive method is lower than the other two discretization methods, while the CI of Naive method is the highest. As to the CCI of the SemiNaive is the highest while that of Naive is the lowest. As to NCP, it has the same change means as that of CCI. Furthermore, although the CI of Naive method is the highest, its CCI is relatively low. It means the Naive method can not significantly reduce the amount of data on the database and improve the efficacy of RS classification. Relatively, the SemiNaive method can compress the data size better, but the CI is relatively lower.

The confusion matrixes of the MLC and RS classification results of the original and discretized images with the Entropy/MDL, Naive, and SemiNaive methods are

Table 2. Number of cut points of the ROI.

Discretization method	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7	NCP
Entropy/MDL	13	5	24	53	62	14	47	218
Naive	34	23	51	71	92	18	70	359
SemiNaive	5	5	5	8	13	4	8	48

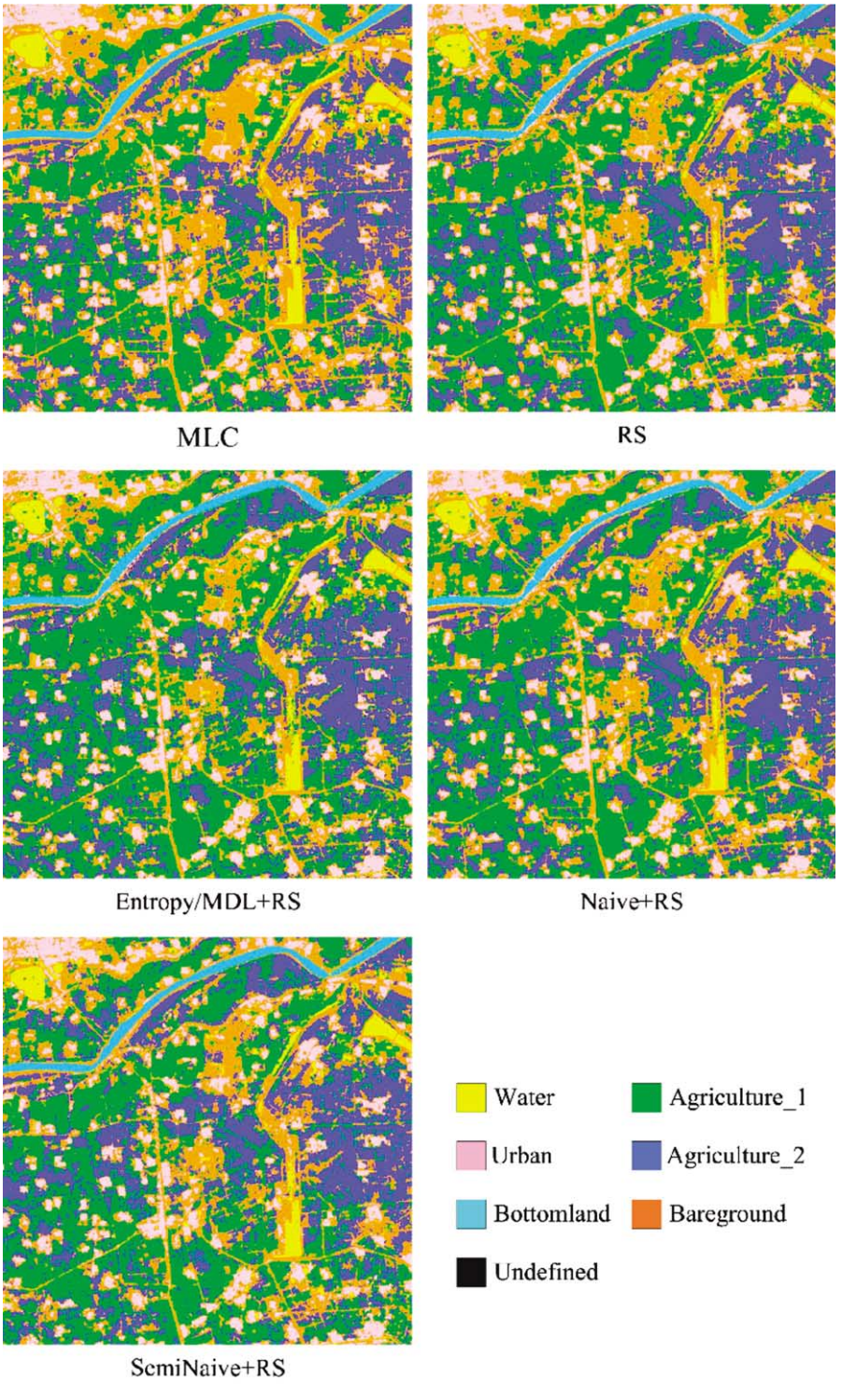

Figure 5. MLC classification results and RS classification results from original and discretized remotely sensed images with Entropy/MDL, Naive, and SemiNaive methods.

shown in Table 3. The producer's accuracy, user's accuracy, overall accuracy, and kappa coefficients acquired from the confusion matrixes are shown in Table 4. The relationship between producer's accuracy, user's accuracy, overall accuracy, kappa coefficients, and different discretization methods, is clearly shown in Figure 6a–d. From Figure 6a, it shows that the producer's accuracy of the water, agriculture_1 and bottomland is the highest for RS classification with the Entropy/MDL method. Only

Table 3. Confusion matrixes for MLC classification result and RS classification result of original and discretized classified remotely sensed images with Entropy/MDL, Naive, and SemiNaive methods.

		Reference data						
	Classified date	WT	AG1	AG2	UB	BT	BR	Column Sum
MLC	WT	18	0	0	0	1	2	21
	AG1	0	215	16	0	0	45	276
	AG2	0	38	174	0	0	22	234
	UB	0	0	1	88	0	6	95
	BT	0	0	0	0	16	0	16
	BR	4	37	11	33	1	272	358
	Row sum	22	290	202	121	18	347	1000
RS	WT	16	0	0	0	2	1	19
	AG1	2	230	21	0	0	90	343
	AG2	0	37	174	0	0	24	235
	UB	0	0	1	81	2	9	93
	BT	0	0	0	0	14	0	14
	BR	4	23	6	40	0	223	296
	Row sum	22	290	202	121	18	347	1000
Entropy/MDL + RS	WT	20	0	0	0	0	2	22
	AG1	1	232	24	0	1	96	372
	AG2	0	39	173	0	0	26	238
	UB	0	0	0	0	17	0	17
	BT	1	18	3	38	0	189	249
	BR	0	1	1	0	0	6	8
	Row sum	22	290	202	121	18	347	1000
Naive + RS	WT	17	0	0	0	3	1	21
	AG1	2	230	21	0	0	90	343
	AG2	0	37	174	0	0	24	235
	UB	0	0	1	81	2	10	94
	BT	0	0	0	0	14	0	14
	BR	3	23	6	40	0	222	294
	Row sum	22	290	202	121	18	347	1000
SemiNaive + RS	WT	18	0	0	0	0	3	21
	AG1	2	222	32	0	0	78	334
	AG2	0	45	157	0	0	46	248
	UB	0	0	1	87	0	14	102
	BT	0	0	0	0	16	0	16
	BR	2	23	12	34	2	206	279
	Row sum	22	290	202	121	18	347	1000

Note: WT, water; AG1, agriculture_1; AG2, agriculture_2; UB, urban; BT, bottomland; BR, baregroud.

Table 4. Producer' accuracy, user' accuracy, overall accuracy and kappa coefficients of MLC classification result and RS classification results of original and discretized remotely sensed images with Entropy/MDL, Naive, and SemiNaive methods.

	Indicators	WT	AG2	AG3	UB	BT	BR
MLC	Producer' accuracy (%)	81.81	74.14	86.13	72.3	88.89	78.39
	User' accuracy (%)	85.71	77.90	74.36	92.63	100.00	75.98
	Kappa	0.85	0.69	0.68	0.92	1.00	0.63
	Overall accuracy	= 78.30%			Kappa = 0.705		
RS	Producer' accuracy (%)	72.73	79.31	86.14	66.94	77.78	64.27
	User' accuracy (%)	84.21	67.06	74.04	87.10	100.00	75.34
	Kappa	0.84	0.54	0.67	0.85	1.00	0.62
	Overall accuracy	= 73.08%			Kappa = 0.65		
Entropy/MDL + RS	Producer' accuracy (%)	90.91	80.00	85.64	68.60	94.44	54.47
	User' accuracy (%)	90.91	62.37	72.69	88.30	100.00	75.90
	Kappa	0.91	0.47	0.66	0.87	1.00	0.63
	Overall accuracy	= 71.40%			Kappa = 0.62		
Naive + RS	Producer' accuracy (%)	77.27	79.31	86.14	66.94	72.22	63.98
	User' accuracy (%)	80.95	67.06	74.04	86.17	100	75.51
	Kappa	0.81	0.54	0.67	0.84	1.00	0.63
	Overall accuracy	= 73.70%			Kappa = 0.64		
SemiNaive + RS	Producer' accuracy (%)	81.82	76.55	77.72	71.09	88.89	59.37
	User' accuracy (%)	85.71	66.47	63.31	85.29	100	73.84
	Kappa	0.85	0.53	0.54	0.83	1.00	0.60
	Overall accuracy	= 70.60%			Kappa = 0.60		

Note: WT, water; AG1, agriculture_1; AG2, agriculture_2; UB, urban; BT, bottomland; BR, baregroud.

the producer's accuracy in the terrains of bareground is the highest for MLC classification. The producer's accuracy in agriculture_1 is the lowest for MLC classification. Figure 6b and c were also analyzed and showed that the user's accuracy and kappa of water, agriculture_2, bottomland and bareground are very close for RS classification compared with MLC classification, although they are lower than MLC classification in agriculture_1 and urban lands.

From Figure 6d, it can be seen that the overall accuracies of RS classification with different discretization methods are lower than MLC classification. The overall accuracies of RS classification with different discretization methods are very close. Although the overall accuracies are very close, the CCI of the SemiNaive method is much higher than the CCI of the Entropy/MDL and Naive methods. It shows that the data capability of the SemiNaive is much greater than the Entropy/MDL and Naive methods. The discretized remotely sensed images with the SemiNaive method still have relatively higher overall accuracy, while reducing the large quantity of remotely sensed data. Although the accuracy of RS classification with SemiNaive method is lower than the other two discretization methods, SemiNaive method reduces large quantity of image data, and improves the efficacy of the classification.

5. Conclusion and future works

In this article, the CCI, CI, and NCP are used to evaluate the discretization method which is the indispensable procedure of classical RS-based classification for a given

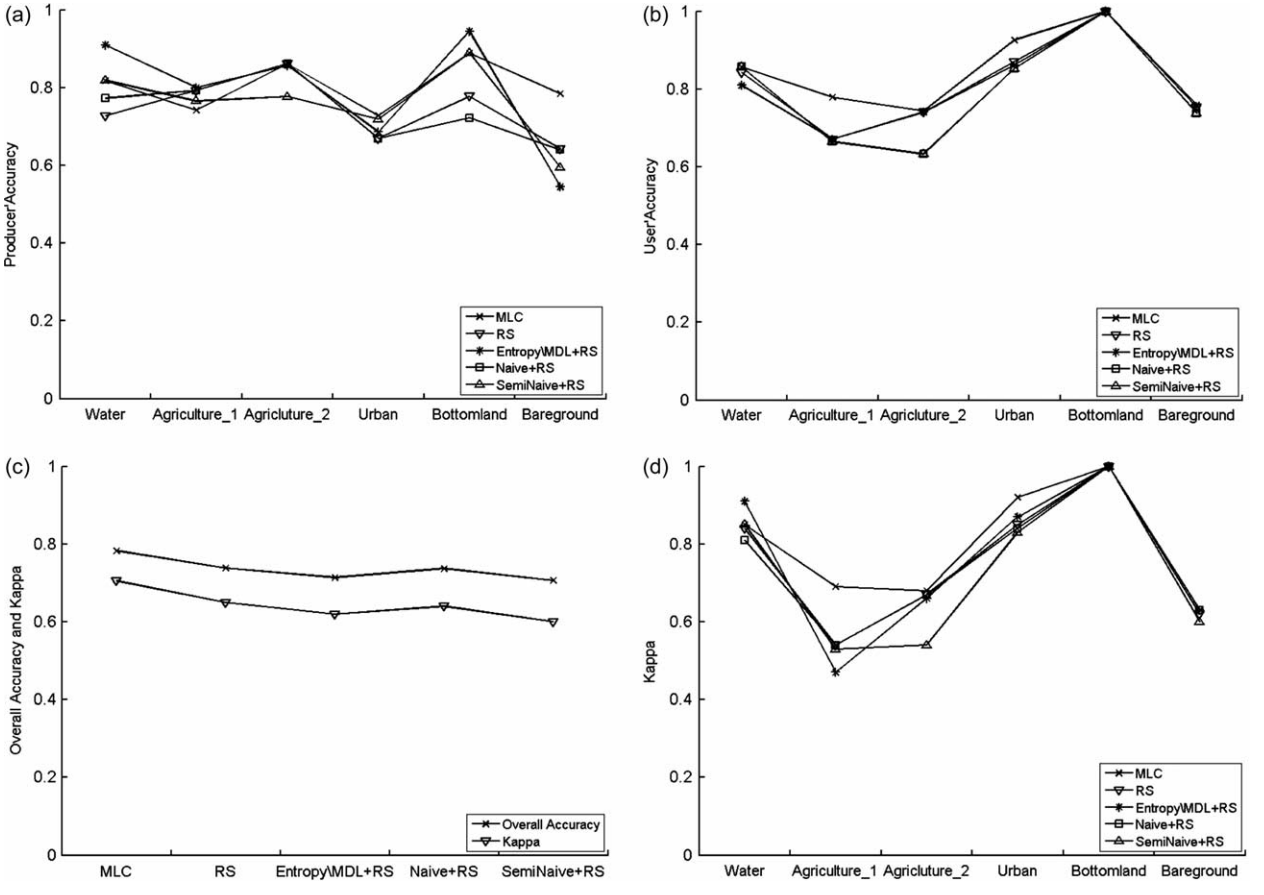

Figure 6. Assessment accuracies of RS classification results for original and discretized remotely sensed images with different discretization methods; (a) the producer' accuracy of each category; (b) the user' accuracy of each category; (c) the overall accuracy; (d) the kappa coefficients of each category.

remotely sensed image. The impact of different discretization methods on remote-sensing classification results was then analyzed. From the experimental results, it can be seen that the CCI of the SemiNaive method is much higher than the CCI of the Entropy/MDL and Naive methods. Although the CI of the SemiNaive is lower than the CI of Entropy/MDL and Naive methods, the overall accuracy of RS classification with SemiNaive method is very close to the overall accuracy of RS classification with Entropy/MDL and Naive methods, and the accuracy of MLC classification. At the same time, the NCP of the SemiNaive is lower than that of the other two methods. Also, the SemiNaive method reduces the spectral band values at magnitude level and greatly improved the efficacy of RS classification. There are still some problems about the discretization methods used in the remote-sensing classification: (1) most discretization methods discretize each attribute independently and the design of a discretization method that discretizes all attributes simultaneously needs further study, (2) remotely sensed data usually has its own features. For example, the values of the spectral bands obey to normal distributions or have spatial correlations. Therefore, the design of a new discretization method that can reflect those features is an important future research topic. The accuracies of RS classification with Entropy/MDL, Naive, and SemiNaive methods are lower than the accuracy of MLC classification in this article. It is because the used discretization methods are traditional ones which do not consider the features of remotely sensed images. In future work, we will study the discretization methods considering the features of remotely sensed images and improve the RS classification accuracy.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 40971222) and the National High Technology Research and Development Program of China (Grant No. 2006AA120106).

Notes on contributors

Yong Ge received BS and MS degrees on surveying and mapping from Wuhan University, Wuhan, PRC, in 1995 and 1998, respectively, and the Ph.D. degree in cartography and geographical information system from Chinese Academy of Sciences in 2001. Since July 2001, she has been with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences where she is currently an associate professor.

Feng Cao is a Ph.D. candidate in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. He received a master's degree at the School of Computer and Information Technology, Shanxi University in 2009. His current research is the application of RS theory in GIS and RS.

Ruifang Duan studied GIS at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences from 2006 to 2008. She received a master's degree at the School of Computer and Information Technology, Shanxi University in 2008.

References

Ahlqvist, O., *et al.*, 2000. Rough classification and accuracy assessment. *International Journal Geographical Information Science*, 14 (5), 475–496.

Ahlqvist, O., *et al*., 2003. Rough and fuzzy geographical data integration. *International Journal Geographical Information Science*, 17 (3), 223–234.
Bittner, T., 2001. Rough sets in spatio-temporal data mining. *In*: J.F. Roddick and K. Homsby, eds. *Temporal, spatial, and spatio-temporal data mining*. Berlin, Heidelberg: Springer, 89–104.
Bittner, T. and Stell, G.J., 2001. Rough sets in approximate spatial reasoning. *In*: W. Ziarko and Y. Yao, eds. *Rough sets and current trends in computing*. Berlin, Heidelberg: Springer, 445–453.
Berger, A.P., 2004. Rough set rule induction for suitability assessment. *Environmental Management*, 34 (4), 546–558.
Beaubouef, T., *et al*., 2007. Spatial data methods and vague regions: a rough set approach. *Applied Soft Computing*, 7 (1), 425–440.
Bai, H.X., *et al*., 2010. Using rough set theory to identify villages affected by birth defects: the example of Heshun, Shanxi, China. *International Journal Geographical Information Science*, 24 (4), 559–576.
Chen, S.P. and van Genderen, J., 2008. Digital Earth in support of global change research. *International Journal of Digital Earth*, 1 (1), 43–65.
Dougherty, J., *et al*., 1995. Supervised and unsupervised discretization of continuous features. *In*: *Proceedings of the twelfth international conference on machine learning*, 9–12 July, Tahoe City, California. Morgan Kaufmann, 194–202.
Das, S., *et al*., 2006. A hybrid rough set particle swarm algorithm for image pixel classification. *In*: *Proceedings of the sixth international conference on hybrid intelligent systems*, 13–15 December, Auckland, New Zealand, 26–30.
Duan, R.F., *et al*., 2007. Experimental study on the discretization on remote sensing data. *In*: *7th international workshop on geographical information system*, 14–15 September, Beijing, China, 383–388.
Edwards, T.C., *et al*., 1998. Assessing map accuracy in a remotely sensed, ecoregion-scale cover map – a user's perspective. *Remote Sensing of Environment*, 63, 73–83.
Fayyad, U.M. and Irani, K.B., 1992. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8, 87–102.
Fayyad, U.M. and Irani, K.B., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. *In*: *Proceedings of the 13th international joint conference on artificial intelligence*, 28 August–3 September, Chambery, France. Morgan Kaufmann, 1022–1027.
Ge, Y., *et al*., 2009. Rough set-derived measures in image classification accuracy assessment. *International Journal of Remote Sensing*, 30 (20), 5323–5344.
Gore, A., 1998. The digital earth: understanding our planet in the 21st century. *In*: *Presented at the Californian Science Center*. 31 January Los Angeles, CA, USA.
Guo, H., *et al*., 2009. A digital earth prototype system: DEPS/CAS. *International Journal of Digital Earth*, 2 (1), 3–15.
Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11, 63–90.
Liu, H.J., *et al*., 2004. Rough neural network of variable precision. *Neural Processing Letters*, 19, 73–87.
Leung, Y., *et al*., 2007. A rough set approach to the discovery of classification rules in spatial data. *International Journal of Geographical Information Science*, 21 (9), 1033–1038.
Li, L.W. *et al*., 2007. Tolerant rough set on satellite remote sensing data classification. *Computer Egineering and Applications*, 43 (20), 11–13. (In Chinese)
Li, L.W., *et al*., 2008. Tolerant rough set processing on uncertainty of satellite remote sensing data classification. *Computer Engineering*, 34 (6), 2–6. (In Chinese)
Lei, T.C., *et al*., 2008. The comparison of PCA and discrete rough set for feature extraction of remotely sensed imagery classification – a case study on rice classification, Taiwan. *Computers & Geosciences*, 12, 1–14.
Ma, J.W. and Hasi, B., 2005. Remote sensing data classification using tolerant rough set and neural networks. *Science in China Ser. D Earth Science*, 48 (12), 2251–2259.
Nguyen, H.S., 1998. Discretizaiton problem for rough sets methods. *In*: L. Polkowski and A. Skowron, eds. *Rough sets and current trends in computing*. London: Springer-Verlag, 545–552.

Ouyang, Y. and Ma, J.W., 2006. Land cover classification based on tolerant rough set. *International Journal of Remote Sensing*, 27 (14), 3041–3047.

Pawlak, Z., 1982. Rough sets. *International Journal of Computer and Information Sciences*, 11 (5), 341–356.

Pawlak, Z., 1991. *Rough sets: theoretical aspects of reasoning about data*. Boston: Kluwer Academic.

Pal, S.K. and Mitra, P., 2002. Multispectral image segmentation using the rough-set-initialized EM algorithm. *IEEE Transactions on geoscience and remote sensing*, 40 (11), 2495–2501.

Quinlan, J.R., 1993. *C4.5: programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

*Ø*hrn, A., 1999. *Discernibility and rough sets in medicine: tools and applications, computer and information science*. Thesis (PhD). Norwegian University of Science and Technology.

Roy, A. and Pal, R.K., 2003. Fuzzy discretizaiton of feature space for a rough set classifier. *Pattern Recognition Letters*, 24 (6), 859–902.

Wang, G.Y., 2001. *Rough set theory and knowledge acquisition*. Xi an: Xi'an Jiaotong University Press. (In Chinese)

Wu, Z.C., 2001. Research on remotely sensed imagery classification using neural network based on rough sets. *In*: *International conferences on info-tech and info-net*, 29 October–1 November, Beijing, China, 279–284.

Wu, Z.C., 2004. *Rough sets approach to remotely sensed imagery processing and classification*.Thesis (PhD). Wuhan University. (In Chinese)

Wang, S.L., *et al*., 2004. Rough spatial interpretation. *In*: S. Tsumoto *et al*., eds. *Rough sets and current trends in computing*. Berlin, Heidelberg: Springer, 435–444.

Xiao, H. and Zhang, X.B., 2008. Comparison studies on classification for remotely sensed imagery based on data mining method. *WSEAS Transactions on computers*, 5 (7), 552–558.

Yue, X.D., 2006. *Research on discretization of continuous features based on rough set theory*. Thesis (MA). Shanxi University. (In Chinese)

Zhang, G.X., *et al*., 2005. A hybrid classifier based on rough set theory and support vector machines. *In*: L. Wang and Y. Jin, eds. *Fuzzy systems and knowledge discovery*. Berlin, Heidelberg: Springer, 1287–1296.

Zhan, Y.J., *et al*., 2007. Hyperspectral RS image classification based on fractal and rough set. In *Second International Conference on Space Information Technology*, Wuhan, 6795 (3), 67954F.1–67954F.6.

Zhang, G.F., *et al*., 2008. A remote sensing feature discretization method accommodating uncertainty in classification systems. *In*: *Proceedings of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences*, 25–27 June, Shanghai, China, 195–202.