MEASURING THE QUALTITY OF SAMPLES IN THE SUPERVISED CLASSIFCATION OF REOTELY SENSED IMAGERY

Yong Ge [a, *], Hexiang Bai [a, b], Sanping Li [a, b], Ruifang Duan [a, b], Deyu Li [b]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences &
Natural Resources Research, Chinese Academy of Sciences Beijing 100101, China –
(gey, baihx, lisp, duanrf) @lreis.ac.cn
[b] School of Computer and Information Technology, Shanxi University
Taiyuan 030006, China - lidy@sxu.edu.cn

KEY WORDS: Supervised classification, measuring the sample quality, rough set

ABSTRACT:

In the supervised classification process of remotely sensed imagery, the quantity of samples is one of the important factors affecting the accuracy of the image classification as well as the keys used to evaluate the image classification. In general, the samples are acquired on basis of prior knowledge, experience and higher resolution images. With the same size of samples and the same sampling model, several sets of training sample data can be obtained. In such sets, which set reflects perfect spectral characteristics and ensure the accuracy of the classification can be known only after the accuracy of the classification has been assessed. So, before classification, it would be a meaningful research to measure and assess the quality of samples for guiding and optimizing the consequent classification process. Then, based on the rough set, a new measuring index for the sample quality is proposed. The experiment data is the Landsat TM imagery of the Chinese Yellow River Delta on August 8th, 1999. The experiment compares the Bhattacharrya distance matrices and purity index Δ and Δ_X based on rough set theory of 5 sample data and also analyzes its effect on sample quality.

1. INTRODUCTION

Sample points play an important role in supervised classification of remotely sensed imagery. It provides training data for classifier and test data for classification result. Therefore, it is important to study how to choose sample points, determine sample volume and guarantee quality of sample points. Generally, sample points are acquired through prior knowledge or experience. Under same sample pattern and equivalent sample volume, the "real effect" of the sample data sets, which is used as training area for the classifier, can only be validated and appraised after the image is classified. So, it is a meaningful work that how we can measure sample data as well as guide and optimize the classification process before classification.

Some statistical methods have been used to evaluate sample data quality in classification for remotely sensed imagery, such as Mahalanobis distance, Bhattacharrya distance and transformed divergence（Richards,1986; PCI Geomatica, 2003; Sun, 2003; ENVI, 2007）. These methods effectively describe the sample quality from the perspective of statistics and they also have wide application. Recent years, the rough set theory is developed greatly. It is based on data-driven, and attracts widespread concern by its merits, such as "it needs no priori assumptions on data" and "it can provide non-complete, non-coordination uncertainty knowledge acquisition method" (Liang and Li, 2005). Therefore this paper measures the sample data quality based on rough set theory.

In the supervised classification for remotely sensed imagery, training data can be acquired from higher resolution imagery or be assigned by user. So the class information of each pixel is known beforehand. This class information can be taken as decision attribute in the decision information system, and the gray value can be seen as conditional attribute in the decision information system. Sample dataset, conditional attribute and decision attribute constitute the decision table of a rough set. In the decision table, no matter if it is complete or non-complete, decision attribute can determine a partition of the set. Decision attribute only has positive area and boundary area. The positive area of the decision attribute is the union of some basic knowledge grain. Every element in the positive area can derive a harmonized rule and every element in the boundary area can derive a non-harmonized rule. In the practical application, the decision table constructed mostly is non-harmonized one. So it is necessary to use certain index to measure these rules. Certainty factor is an index for measuring rule, and it reflects the ratio to obtain a decision under same condition attribute.

Based on these, this paper aims at measuring sample quality from rough set perspective. New indexes named purity indexes Δ and Δ_X are proposed and compared with Bhattacharrya distance in an empirical study on five sets of sample data. The experiment data is the Landsat TM imagery of the Chinese Yellow River Delta on August 8th, 1999.

[*] Corresponding author. The State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences. Beijing 100101, China. Tel: +86 10 64888967; Fax: +86 10 64889630. Email address: gey@lreis.ac.cn

2. SAMPLE MEASURING BASED ON ROUGH SET

2.1 Certainty Factor

For a decision table $S = (U, A, V, F)$, U called discourse universe is the non-empty finite set of objects, A is the non-empty finite set of attribute, $A = C \bigcup D$ and $C \bigcap D = \varnothing$, C is called conditional attribute set, D is called decision attribute set, $V = \bigcup_{a \in A} V_a$, V_a presents the range of attribute a, f : $U \times A \to V$ is a information function, i.e. $f(x, a) \in V_a$, $\forall x \in U, a \in A$. In the decision table, $(a_1, v_1) \wedge (a_2, v_2) \wedge \cdots \wedge (a_n, v_n)$ is called the basic formula P , where $v_i \in V_{a_i}$, $\{a_1, a_2, \cdots, a_n\} \in P$, $P \subseteq C$. If A is P basic formula and $B = (d, d_i), d \in D, d_i \in V_D$, then $A \to B$ is called a decision rule. The certainty factor of $A \to B$ is denoted as $CF(A \to B)$,

$$CF(A \to B) = \frac{|X \bigcap Y|}{|X|} \qquad (1)$$

where $X = \{x \mid x \in U \wedge A_x\}$, $Y = \{x \mid x \in U \wedge B_x\}$. A_x means that conditional attribute value of element x satisfies formula A and B_x means that decision attribute value of element x satisfies formula B. Then the set X is the set of elements which conditional attribute value satisfies A, then the set Y is the set of elements which conditional attribute value fits B (Wang, 2001; Liang and Li, 2005).

2.2 Measuring method

During the classification process for remotely sensed imagery, the rough set theory can be used to measure the uncertainty of sample data. First, one can obtain several sample sets from a remotely sensed imagery. And then the sample data need to be converted into decision table in order to be handled by rough set easily. After that, the index which can reflect the sample quality can be calculated using the rough set theory. Finally, the index calculated is used to evaluate the sample quality. The whole process is shown in Figure 1, and the detailed description of the process can be divided into following 4 steps:

(1) Sampling

When sampling, there may be only one sample set or several sample sets. The Sample should reflect the spectral feature of every class. When there is only one sample set, the sample appraisal result can be used as a reference for the classification result. And when there are several sample sets, the sample measuring result can be used as a reference index to select the better sample set.

(2)Preparation process for sample data

When a sample set is obtained, the spectral attribute is set to conditional attribute of the decision table and the class information is set to decision attribute of the decision table. Furthermore, it can be changed into the format which can be handled by rough set theory easily.

(3) Calculating measurement

First, many decision rule can be extracted from the decision table which is prepared. For every decision rule, its certainty factor can be calculated by using formula (1). Then for every class in the decision attribute, which is denoted as X , the rules which certainty factor are larger than 0.9 (denoted as $r_{0.9}$)is selected among all the rules separately. And then for every X , the total number of instance for $r_{0.9}$ is denoted as $sum_X(r_{0.9})$, and this number divide X 's $|X|$, i.e., $\Delta_X = sum_X(r_{0.9}) / |X|$, which is called the purity of X . Finally, all the $sum_X(r_{0.9})$ is summed up and the result divides $|U|$, i.e. $\Delta = \sum_{i=1}^{n} sum_{Xi}(r_{0.9}) / |U|$, which is called the purity of sample. Now for every class X and the population, the measurement is calculated.

(4) Evaluating sample quality

In this step, the sample quality is evaluated according to the two indices Δ and Δ_X . Generally speaking, the sample quality is getting better, when the Δ and Δ_X get larger.

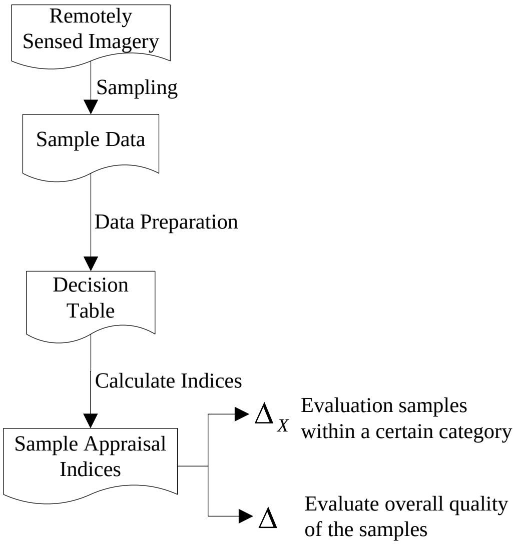


Figure 1. Sample measuring process

3. AN EMPIRICAL STUDY

3.1 Experiment area and experiment data

The study area was selected from a Landsat TM image which was taken over the Chinese Yellow River Delta on August 8th, 1999. The image area is located at the intersection of the terrain between Dongying and Binzhou, Shandong Province. The image size is 515 × 515 and its resolution is 30 m. Its left-upper latitude and longitude coordinates are 118^{o}0'34.07"E and 37^{o}22'24.00"N, respectively, and its right-lower latitude and longitude coordinates are 118^{o}10'52.83"E and 37^{o}13'58.13"N, respectively.

3.2 Experiment course

Step.1 Sampling

Five sets of sample data numbered as s-1, s-2, s-3, s-4 and s-5 are collected from this image. The sample strategy and sample volume are the same when sampling. The sample results are shown in figure 3. All the samples has 6 classes, they are water, agriculture I, agriculture II, buildings, bottomland and bare ground. Figure 3(a) to 3(e) represent samples numbered s-1 to s-5 respectively. For every sample data set of the image, the corresponding decision table can be gained by setting the spectral information as conditional attribute and the class as decision attribute.

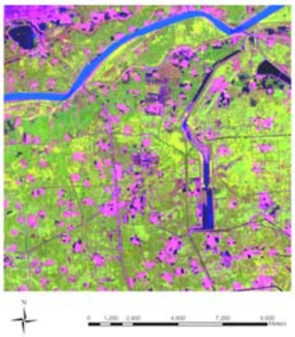


Figure 2. 5, 4, 3-band pseudo-color composition image of the experimental area

Step.2 Data preparation

There are many methods to implement data preparation, such as various data discretization methods. This paper performs k-means cluster for the remotely sensed imagery by using PCI Geomatica 9.0. The whole image is clustered into 50 classes. The cluster result is show in figure 4 and each color represents 1 class.

Step.3 Calculating Δ_X and Δ

According to the method introduced in section 2.3(3), the Δ_X for every category of surface object and the Δ for the population can be calculated, as is show in table 1. The first column is the number of the sample data sets, column 2 to column 7 is the Δ_X corresponding to each category of surface object respectively, and the last column is the Δ of the population.

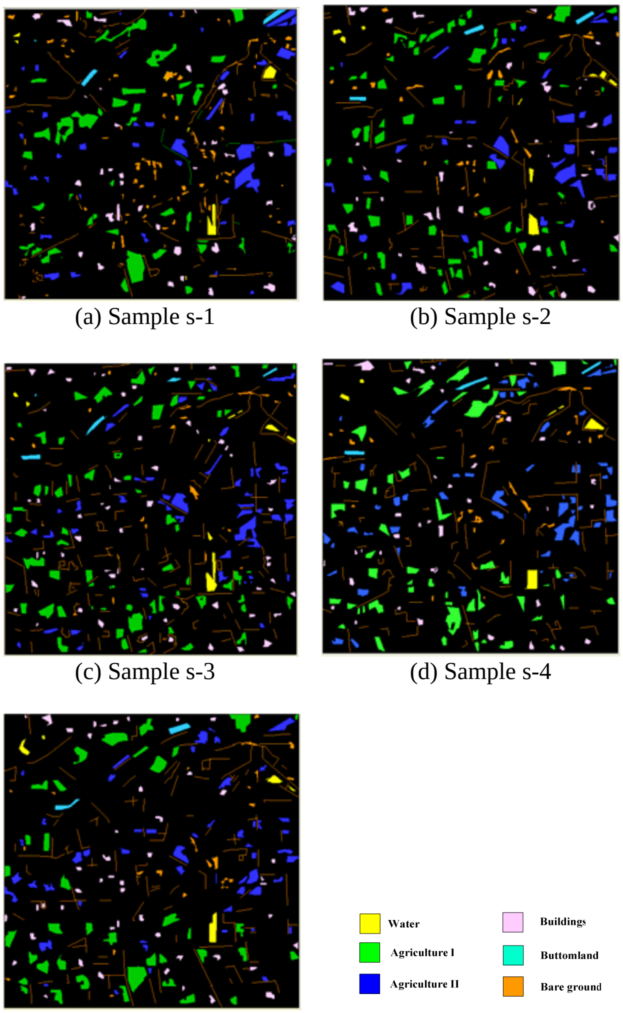
Step.4 Calculating Bhattacharrya distance

Bhattacharrya distance matrix can be calculated by using PCI Geomatica 9.0. Table 2 to 6 is the Bhattacharrya distance matrix of sample s-1 to s-5. The format of the table can be found in PCI online help (2003). And the bottom line under each table shows the average value, maximum and minimum of the whole sample. In the next section Δ and Δ_X are compared carefully with the Bhattacharrya distance, and they have statistical significance linear relationship.

3.3 Analysis

By comparison we found that sample's Δ value has some relation with Bhattacharrya distance. The Δ value and average Bhattacharrya distance formed a scatter plot, as is shown in figure 5. It can be seen that Δ value is increasing when the average Bhattacharrya distance is increasing except sample s-2 which is signed by triangle, though there are three sets of data values changed little.

The relationship between purity and Bhattacharrya distance can also be viewed within one sample. Sample s-5 as an example, the average Bhattacharrya distance of each category of surface object and Δ_X of each category of surface object constitutes a scatter plot, which is shown in figure 6. It can be seen there is a line relationship between Δ_X and average Bhattacharrya distance. When all the samples are made the same analysis, similar result can be obtained. And all the scatter plot putting together forms figure 7. By observation and comparison, we can speculate that when the number of the sample tends to infinity, a linear function between Δ_X and average Bhattacharrya distance can be gotten using linear regression method.



(a) Sample s-1

(b) Sample s-2

(c) Sample s-3

(d) Sample s-4

(e) Sample s-5

Figure 3 Illustration of 5 samples

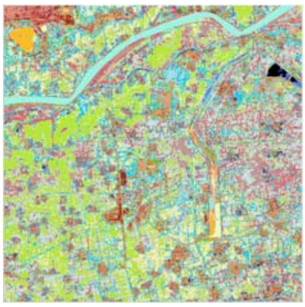


Figure 4. k-means cluster result for the image

Table 1. Δ and Δ_X for 5 sample data sets

Number	Δ_{water}	$\Delta_{AgricultureI}$	$\Delta_{AgricultureII}$	$\Delta_{Building}$	$\Delta_{bottomland}$	$\Delta_{Bareground}$	Δ
s-1	0.967992	0.860312	0.880444	0.957096	0.998366	0.847952	0.880556
s-2	0.856618	0.759749	0.454731	0.345859	0.96087	0.19716	0.553759
s-3	0.955455	0.501804	0.538087	0.685045	0.949081	0.388766	0.542334
s-4	0.934199	0.466245	0.571196	0.74633	1	0.462244	0.552677
s-5	0.973077	0.435526	0.610999	0.774953	0.942997	0.488859	0.555837

Table 2. Bhattacharrya distance matrix of sample s-1

	Water	Agriculture I	Agriculture II	Building	Bottomland
Agriculture I	1.999913				
Agriculture II	1.999994	1.76806			
Building	2	1.999998	2		
Bottomland	2	2	2	2	
Bare ground	1.998557	1.897925	1.961505	1.95302	2

Where Bhattacharrya distance's average=1.9719, maximum=2, minimum=1.76806.

Table 3. Bhattacharrya distance matrix of sample s-2

	Water	Agriculture I	Agriculture II	Building	Bottomland
Agriculture I	1.999997				
Agriculture II	1.999991	1.525386			
Building	2	1.999958	1.999963		
Bottomland	2	2	2	2	
Bare ground	1.995881	1.666782	1.753598	1.616956	2

Where Bhattacharrya distance's average=1.9039, maximum=2, minimum=1.525386.

Table 4. Bhattacharrya distance matrix of sample s-3

	Water	Agriculture I	Agriculture II	Building	Bottomland
Agriculture I	1.999999				
Agriculture II	2	1.685667			
Building	1.999999	1.999989	1.999999		
Bottomland	2	2	2	2	
Bare ground	1.998278	1.591221	1.776549	1.738729	2

Where Bhattacharrya distance's average=1.9194, maximum=2, minimum=1.591221.

Table 5. Bhattacharrya distance matrix of sample s-4

	Water	Agriculture I	Agriculture II	Building	Bottomland
Agriculture I	2				
Agriculture II	2	1.692465			
Building	2	1.999918	1.999996		
Bottomland	2	2	2	2	
Bare ground	1.99545	1.582299	1.740689	1.867771	2

Where Bhattacharrya distance's average=1.9252, maximum=2, minimum=1.582299.

Table 6. Bhattacharrya distance matrix of sample s-5

	Water	Agriculture I	Agriculture II	Building	Bottomland
Agriculture I	2				
Agriculture II	2	1.705036			
Building	2	1.999998	2		
Bottomland	2	2	2	2	
Bare ground	1.999677	1.578688	1.786735	1.904678	2

Where Bhattacharrya distance's average=1.9317, maximum=2, minimum=1.578688.

Thus, the sample can be appraised from another point of view, and this new appraisal method is all data-driven. The sample quality can be calculated only from the sample data itself. And the new indices can be used to supervise sample selection; moreover it can provide objectively quantitive evaluation to the final classification.

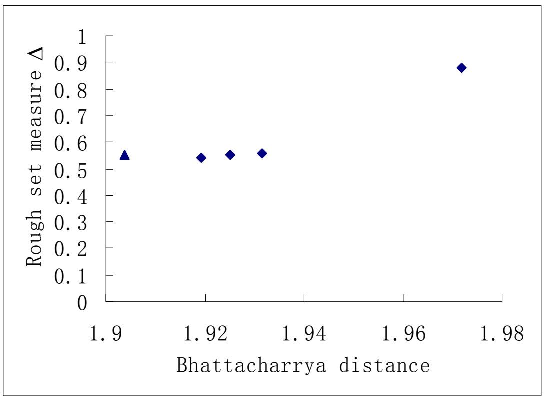


Figure 5. Scatter plot of Δ and average Bhattacharrya distance

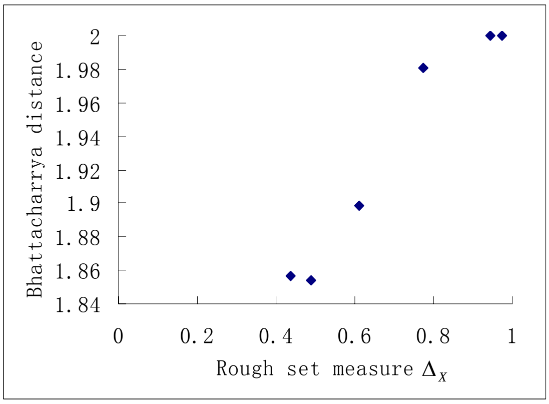


Figure 6. Scatter plot of Δ_X and average Bhattacharrya distance within sample s-5

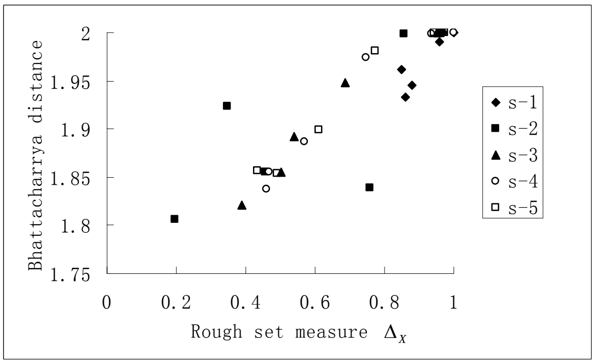


Figure 7. Scatter plot of Δ_X and average Bhattacharrya distance of all samples

4. DISCUSSION AND CONCLUSION REMAREKS

This paper mainly discussed the sample appraisal issue in the classification of remotely sensed imagery. The method introduced in this paper studies the issue from a view different from the traditional statistical method. This method calculates the purity Δ_X of every class in a sample and overall purity Δ of a sample to evaluate the sample quality. Moreover, the new indices are compared with the traditional Bhattacharrya distance index in order to validate the effectiveness of the purity. Furthermore, in future works, the sample quality can be measured by the overlap degree between classes by using rough set theory, and corresponding index can be got. For example the upper and lower approximation of the two classes can be calculated according to the conditional attribute, and then the overlap between the approximations can be used to evaluate sample quality. And other rough set measure, rough entropy as an example, can be used to evaluate sample quality, i.e. the non-harmonized degree. And also we can visualize the sample quality by using these indices, so the sample quality can be observed more intuitively.

ACKNOWLEDGEMENTS

This work was supported in part in part by the National 973 Program (Grant No.2006CB701305), the National Natural Science Foundation of China (40671136)， Open Research Fund from State Key Laboratory of Remote Sensing Science (LRSS0610) and Innovation Grant of Chinese Academy of Sciences (V36400).

References

ITT Inc., 2007. ENVI Online Help, Version 4.3. USA:ITT Corporation, USA.

Richards, J. A., and Jia, X. P., 1999. Remote sensing digital image analysis: an introduction, Springer-Verlag, 3rd ed, Berlin, New York.

PCI Inc. 2003. PCI Online Help, Version 9.0. Canada: PCI Enterprises, Canada.

Liang, J.Y and Li D.Y., 2005. Uncertainty and knowledge acquisition in the information system. Beijing: Science Press (in Chinese). 118 p.

Sun, J.X., 2003. Modem Pattern Recognition. Changsha: The Defence University of Science and Technology Publishing House (in Chinese). 460 p.

Wang, G.Y., 2001. Rough Set Theory and its Knowledge Acquisition. Xi'an: Xi'an Jiaotong University Press (in Chinese). 226 p.