Exploring the sample quality using rough sets theory for the supervised classification of remotely sensed imagery

Yong Ge, Hexiang Bai, Sanping Li & Deyu Li

To cite this article: Yong Ge, Hexiang Bai, Sanping Li & Deyu Li (2008) Exploring the sample quality using rough sets theory for the supervised classification of remotely sensed imagery, Geo-spatial Information Science, 11:2, 95-102, DOI: 10.1007/s11806-008-0020-0

To link to this article: http://dx.doi.org/10.1007/s11806-008-0020-0

Copyright Taylor and Francis Group, LLC

Published online: 14 Aug 2012.

Submit your article to this journal

Article views: 229

View related articles

Citing articles: 2 View citing articles
Exploring the Sample Quality Using Rough Sets Theory for the Supervised Classification of Remotely Sensed Imagery

GE Yong  BAI Hexiang  LI Sanping  LI Deyu

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 the 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 $\Delta$ and $\Delta_X$ based on rough set theory of 5 sample data and also analyzes its effect on sample quality.

Keywords  supervised classification; measuring the sample quality; rough set

CLC number  P237.3

Introduction

Sample points play an important role in supervised classification of remotely sensed imagery. It provides both 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 the 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 estimated after the classification by using error matrix or Kappa coefficient. Thus, it is a meaningful work to determine how we can measure sample data as well as guide and optimize the classification process before classification[1].

Some existing 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[2-5]. They effectively describe the sample qual-
ity from the perspective of statistics. These methods, however, only describe and evaluate the separation degree between different class types in sample data, and have not discussed much on the quality of sample data in a certain class type. In addition, sample data need to meet the assumption of a certain probability distribution.

In recent years, the rough sets theory has been applied to many fields of data analysis. It is data-driven, and the advantage of this approach is that it needs no a priori assumptions on data. The rough sets theory, as a tool for data mining, can discover knowledge concealed in incomplete and inconsistent data, and make the knowledge easy to understand. In order to give the more comprehensive description on the separation degree of entire sample data and between different class types, as well as the quality of sample data in a certain class type, a new measurement index is proposed in this paper to explore and estimate the sample data quality based on the rough sets theory. To analyze and test the validity of the method, a comparison between the Bhattacharyya distance, as a representative of existing methods to evaluate sample data quality and the new measurements has been made on a Landsat 5 TM image that covers China’s Yellow River Delta on August 8th, 1999. The experiment compared the Bhattacharyya distance matrices and purity index $\Delta$ and $\Delta_x$ based on rough set theory of 5 sample data and also analyzes its effect on sample evaluation. It has been demonstrated from experiment results that there is a positive correlation between the purity index and the Bhattacharyya distance.

In this paper, the Bhattacharyya distance is first discussed briefly in section 2. The derivation of the measurement of purity degree is presented in section 3. In section 4, the validity of the new method is analyzed by a comparison between the purity degree and the Bhattacharyya distance. We conclude our paper with a summary and outlook for further research in section 5.

1 Measurement based on Bhattacharyya distance

Currently, the existing statistical methods of assessing the sample data quality in remote sensing include Mahalanobis distance, Bhattacharyya distance and transformed divergence\cite{2,5}. These methods effectively describe the sample quality from the perspective of statistics and also have a broad application. For example, the Bhattacharyya distance and transformed divergence evaluate the quality of sample data based on the probability distribution. For convenience without loss of generality, we select the Bhattacharyya distance to compare with the new measurement based on the rough sets theory.

The Bhattacharyya distance, assessing the divergence degree by sample probability distribution, indicates the case of overlap between various classes. Given two classes $\omega_1$ and $\omega_2$ with the same a priori probability distribution. For all the points, if $p(x | \omega_1) = 0$ and $p(x | \omega_2) \neq 0$ stand, two classes are considered to be completely separable; on the contrary, if $p(x | \omega_1) = p(x | \omega_2)$ stands for every $x$, two classes are completely inseparable. The overlap degree can be measured by $J_p$, which denotes the distance between the distribution density functions $p(x | \omega_1)$ and $p(x | \omega_2)$. The equation $J_p(\cdot) = \int g[p(x | \omega_1), p(x | \omega_2), P_1, P_2]dx$, where $P_1$ is the a priori probability of the first class, $P_2$ is the a priori probability of the first class, and $g$ is a function which uses $p(x | \omega_1), p(x | \omega_2), P_1, P_2$ as its variable, stands, if

1. $J_p \geq 0$;
2. The value of $J_p$ can be maximized when two classes are not overlapping, that is, $p(x | \omega_1) = 0$ if $p(x | \omega_2) \neq 0$ for all $x$;
3. $J_p = 0$ if $p(x | \omega_1) = p(x | \omega_2)$;

Consequently, the criterion of Bhattacharyya distance can be defined as:

$$J_b = -\ln \left[ (p(x | \omega_1)p(x | \omega_2))^\frac{1}{2} \right] dx$$

According to the criteria of minimum misjudgment probability\cite{3}, minimum misjudgment probability can be denoted as:

$$P_e(\epsilon) \leq \frac{1}{2} \left[ (p(\omega_1)p(\omega_2))^\frac{1}{2} \exp[-J_b] \right]$$

In order to be more sensitive for small values, some improved modification has been made on the criterion when calculating the Bhattacharyya distance between the sample classes $i$ and $j$\cite{3}:

$$BD(i, j) = 2[1 - \exp[-a(i, j)]]$$

where

$$a(i, j) = 0.125 \times [\mu(i) - \mu(j)]^T \times [A(i, j)]^\frac{1}{2} \times [\mu(i) - \mu(j)]$$
and $A(i, j) = 0.5 \times [\text{cov}(i) + \text{cov}(j)]$; $\mu(i)$ is the mean vector of class $i$ with $n$ entries ($n$ is the number of bands); $\text{cov}(i)$ is the covariance matrix of class $i$.

The value of $BD(i, j)$ is within $[0, 2]$. The greater the value of $BD(i, j)$, the higher the divergence degree between different samples, and then the better the sample quality. If $BD(i, j) = 2$, the two different classes can be completely separated, that is, the sample quality is the best; if $BD(i, j) = 0$, the two different classes are completely overlapped, that is, the sample quality is the worst.

2 Measurement based on rough sets theory

Presently, there are many approaches to assess sample data. For example, exploratory data analysis method is commonly used to extract eigenvector by probabilistic means or depict the data pattern by graphs or charts; the Bhattacharyya distance and divergence are used to measure the divergence degree between different classes on the basis of probabilistic methods. These approaches effectively describe the sample quality from the perspective of statistics and also have a broad application. These methods, however, just describe and evaluate the separate degree between various class types in sample data, and do not estimate the quality of sample data in a particular class type. In addition, sample data have to meet the assumption of a certain probability distribution.

Because of the data-driven feature, there is no a priori knowledge or assumption needed before data analysis; the rough sets theory is introduced to assess the quality of sample data in remote sensing through purity degree for every class and the whole sample data respectively. In the rough set theory, a decision table is complete or non-complete, the decision attribute can determine a partition of the sample dataset. The decision attribute is only composed by positive area and boundary area. The so-called positive area is the union of basic knowledge granules. Every element in the positive area can produce a harmonized rule and every element in the boundary area can derive a non-harmonized rule. While most constructed decision tables are non-harmonized in practical application, it is necessary to use some index to measure these rules. Reliability, an index for measuring rule, reflects the ratio of obtaining a decision under same conditional attributes.

2.1 Reliability

For a decision table $S = (U, A)$, $U$, called the discourse universe, is the non-empty finite set of objects, $A$ is the non-empty finite set of attribute, $A = C \cup D$ and $C \cap D = \emptyset$, $C$ is called conditional attribute set, $D$ is called decision attribute set.

Suppose $U / IND(C) = \{X_1, X_2, \cdots, X_n\}$ and $U / IND(D) = \{Y_1, Y_2, \cdots, Y_m\}$ represent the division of $U$ which is derived from the indiscernibility of attribute set $C$ ($IND(C)$) and attribute set $D(IND(D))$ respectively, where $X_i$ and $Y_i$ is the equivalent class determined by $IND(C)$ and $IND(D)$ respectively. If $X_i \cap Y_i \neq \emptyset$, $r_{ij} : Des_c(X_i) \rightarrow Des_d(Y_i)$ represents a $C \rightarrow D$ decision rule in the decision table $S$, where $Des_c(X_i)$ and $Des_d(X_i)$ are the sole description of $X_i$ and $Y_i$ ($i = 1, 2, \cdots, n; j = 1, 2, \cdots, m$) respectively. For every decision class $Y_j$ ($j = 1, 2, \cdots, m$), the set of decision rule is defined as $\{r_{ij}\} = \{Des_c(X_i) \rightarrow Des_d(Y_j)\}$ $\forall i = 1, 2, \cdots, n$.

A decision rule is certain if and only if $X_i \cap Y_j = X_i$, otherwise $r_{ij}$ is uncertain. The reliability of the decision rule is defined as:

$$\alpha_{x_i}(Y_j) = \frac{|X_i \cap Y_j|}{|X_i|}$$

(1)

The reliability of a decision rule measures the rule’s sufficiency$^{6,7}$.

2.2 Procedure of calculation

During the classification process for remotely sensed imagery training data can be acquired from high-resolution imagery or be assigned by user which means the class information of each training pixel is known, this class information can be taken as a decision attribute, and the gray value can be considered as a conditional attribute. In the decision table, no matter if it is complete or non-complete, the decision attribute can determine a partition of the sample dataset. The decision attribute is only composed by positive area and boundary area. The so-called positive area is the union of basic knowledge granules. Every element in the positive area can produce a harmonized rule and every element in the boundary area can derive a non-harmonized rule. While most constructed decision tables are non-harmonized in practical application, it is necessary to use some index to measure these rules. Reliability, an index for measuring rule, reflects the ratio of obtaining a decision under same conditional attributes.
sensed imagery, the rough set theory can be used to measure the uncertainty of the sample data. First, the sample data is converted into a decision table. After data preparation, 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 detailed description of the process can be divided into the following 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. When there are several sample sets, the sample measuring result can be used as a reference index to select the better sample set.

2. Transforming sample data into a decision table

For every sample datum in the sample data set, there are two kinds of attributes, the spectrum attributes and class attribute. The spectrum attributes include gray level of every band and can be regarded as the conditional attributes in the decision table; the landtype information can be viewed as the decision attribute in the decision table. For example, in a Landsat-5 TM image, each pixel has seven spectral values representing seven spectral channels denoted as band1, band2, band3, band4, band5, band6, band7. These seven spectral values are taken as conditional attributes, while decision attributes are their corresponding classes. Consequently, a decision table can be constructed.

3. Preparation process for sample data

As the conditional values representing the spectral bands are within 0~255, a discretization processing may need to be introduced to improve the learning efficiency, accuracy and reduce the complexity of solving problem[6]. In this paper, we proposed to discrete the sample data by a clustering method. For convenience without loss of generality, K-means is selected to cluster the sample data into the number of classes designated by the user. After this, seven spectral values in conditional attributes are replaced by the clustered value.

4. Calculating measurement

First, many decision rules can be extracted from the decision table that is prepared. For every decision rule, its reliability can be calculated by using Eq.(1).

Then for every class in the decision attribute, which is denoted as $X$, the rules whose certainty factor are larger than one specified threshold $\lambda$ are selected among all the rules separately, denoted as $R_{X}$. Then for each $X$, the total number of instance for $R_{X}$ is denoted as $\text{sum}(R_{X}(X))$, and this number divide the cardinality of $X$, i.e., $\Delta_{X}=\text{sum}(R_{X}(X))/|X|$, which is called the purity degree of set $X$. The physical meaning of $R_{X}(X)$ is the sample data whose value of decision attribute is $X$ and their reliability of decision rules is equal or greater than $\lambda$; in which $0 \leq \Delta_{X} \leq 1$, $\Delta_{X}=1$ indicates that the reliability of decision rules are not smaller than $\lambda$ for all sample data whose decision attribute is $X$; $\Delta_{X}=0$ indicates that the reliability of decision rules are smaller than $\lambda$ for all sample data whose decision attribute is $X$. Therefore, the greater the value of $\Delta_{X}$, the greater the purity degree of sample data, and the better the sample quality.

Finally, summarize all $\text{sum}(R_{X}(X)), i=1...n$ and then divides |$U$|. The result is called the purity degree of the whole sample data, i.e., $\Delta=\sum_{i=1}^{n}\text{sum}(R_{X}(X))/|U|$.

3 An empirical study

In the case study, we compare the matrix of Bhattacharyya distance and the measurements of $\Delta$ and $\Delta_{X}$ by calculating the five different sample data, and analyze the assessing effect for sample data with $\Delta$ and $\Delta_{X}$.

3.1 Study area and data

The study area was selected from a Landsat TM image which was taken over China’s 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 $515 \times 515$ and its resolution is 30 m. Its left-upper latitude and longitude coordinates are 118°03’40.07”E and 37°22’24.00”N, respectively, and its right-lower latitude and longitude coordinates are 118°10’52.83”E and 37°13’58.13”N, respectively. Fig.1 is the 5, 4, 3-band pseudo-color composition image of the ex-
perimental area.

![Experimental area](image1)

Fig.1 5, 4, 3-band pseudo-color composition image of the experimental area

3.2 Experiment procedure

In order to compare and analyze from a statistical point of view, five sets of sample data are collected from this study area. They are numbered as $s_1, s_2, s_3, s_4, s_5$. The sampling strategy and size are the same in the five sets of sample data. The sample result is shown in Fig.2(a) to 2(e). In each sample set, samples are classified into six land types which are water, agriculture I, agriculture II, buildings, bottom-land and bare ground. For every sample data set, the corresponding decision table can be obtained by setting the spectral information as conditional attribute and the class as decision attribute. For example, the decision table of the sample data set $s_1$ is shown in Table 1. Next, these decision tables with respect to sample data sets will be prepared by applying $k$-means cluster to the remote sensing image. 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 Fig.3 and each color represents 1 class. Thus, the number of fields in the decision table has changed to 3, as shown in Table 2.

According to the method introduced in section 3.2, the $\Delta_\lambda$ for every category of surface object and the $\Delta$ for the population can be calculated, as is show in

![Illustration of 5 samples](image2)

(a) Sample $s_1$  (b) Sample $s_2$  (c) Sample $s_3$  (d) Sample $s_4$  (e) Sample $s_5$

Fig.2 Illustration of 5 samples

| Sample point | Band1 | Band2 | Band3 | Band4 | Band5 | Band6 | Band7 | Class |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1            | 107   | 43    | 46    | 84    | 92    | 133   | 43    | 6     |
| 2            | 104   | 38    | 37    | 96    | 73    | 130   | 23    | 2     |
| 3            | 104   | 38    | 38    | 94    | 70    | 130   | 23    | 2     |
| ...          | ...   | ...   | ...   | ...   | ...   | ...   | ...   | ...   |

Table 1 The decision table transferred from the sample data set $s_1$

![Decision table transferred from sample data set](image3)

![Clusters result for the image](image4)

![Decision table after pretreatment of k-means method](image5)

![Clusters result for the image](image6)

Table 2 The decision table after pretreatment of $k$-means method

| Sample point | Clustered class | Class |
|--------------|-----------------|-------|
| 1            | 26              | 6     |
| 2            | 19              | 2     |
| 3            | 19              | 2     |
| ...          | ...             | ...   |
Table 3. The first column in Table 3 is the number of the sample data sets, the 2nd column to 7th column is the $\Delta_i$ corresponding to each category of surface object respectively, and the last column is the $\Delta_t$ of the population. We can note that the purity degree of sample data set $s_i$ is the greatest according to the value of $\Delta$; in every sample, the values of $\Delta_i$ of the bottom land and water are greater than the other classes, that is, the purity degree is better. To get a clear description and test the validity of the method, the Bhattacharrya distance will be calculated in the next discussion.

| Number | $\Delta_{\text{water}}$ | $\Delta_{\text{Agriculture I}}$ | $\Delta_{\text{Agriculture II}}$ | $\Delta_{\text{Building}}$ | $\Delta_{\text{Bottomland}}$ | $\Delta_{\text{Bareground}}$ | $\Delta$ |
|--------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|--------|
| $s_1$  | 0.967 992              | 0.860 312              | 0.880 444              | 0.957 096              | 0.998 366              | 0.847 952              | 0.880556 |
| $s_2$  | 0.856 618              | 0.759 749              | 0.454 731              | 0.345 859              | 0.960 870              | 0.197 160              | 0.553759 |
| $s_3$  | 0.955 455              | 0.501 804              | 0.538 087              | 0.685 045              | 0.949 081              | 0.388 766              | 0.542334 |
| $s_4$  | 0.934 199              | 0.466 245              | 0.571 196              | 0.746 330              | 1                      | 0.462 244              | 0.552677 |
| $s_5$  | 0.973 077              | 0.435 526              | 0.610 999              | 0.774 953              | 0.942 997              | 0.488 859              | 0.555837 |

The Bhattacharrya distance matrix can be calculated by using PCI Geomatica 9.0. The following Table 4 is the Bhattacharrya distance matrix of sample data set $s_1$, in which, the first column is the Bhattacharrya distance between the water and other classes, the second column is the Bhattacharrya distance between agriculture I and other classes, the third column is the Bhattacharrya distance between agriculture II and other classes, the fourth column is the Bhattacharrya distance between the urban and other classes, the fifth column is the Bhattacharrya distance between the bottomland and other classes. Because of the symmetry of the Bhattacharrya distance matrix, only half of the calculated results are listed. The Bhattacharrya distance between the bareground and other classes is not listed in the matrix because it can be obtained from the sixth row of the matrix. The last row of the matrix is the mean, minimum and maximum of Bhattacharrya distance, respectively. Table 5 to 8 are the Bhattacharrya distance matrices of

**Table 4 Bhattacharrya distance matrix of sample $s_1$**

|                  | Water     | Agriculture I | Agriculture II | Building | Bottomland |
|------------------|-----------|---------------|----------------|----------|------------|
| Agriculture I    | 1.999 913 | 1.768 065     |                |          |            |
| Agriculture II   | 1.999 994 | 1.999 998     | 2              |          |            |
| Building         | 2         | 1.999 998     |                | 2        | 2          |
| Bottomland       | 2         | 2             | 2              | 2        | 2          |
| Bare ground      | 1.998 557 | 1.897 925     | 1.961 505      | 1.953 02 | 2          |

Where Bhattacharrya distance’s average=1.971 9, maximum=2, minimum=1.768 06.

**Table 5 Bhattacharrya distance matrix of sample $s_2$**

|                  | Water     | Agriculture I | Agriculture II | Building | Bottomland |
|------------------|-----------|---------------|----------------|----------|------------|
| Agriculture I    | 1.999 997 | 1.525 386     |                |          |            |
| Agriculture II   | 1.999 991 | 1.999 958     | 1.999 963      |          |            |
| Building         | 2         | 1.999 958     | 1.999 963      |          | 2          |
| Bottomland       | 2         | 2             | 2              | 2        | 2          |
| Bare ground      | 1.995 881 | 1.666 782     | 1.753 598      | 1.616 956| 2          |

Where Bhattacharrya distance’s average=1.903 9, maximum=2, minimum=1.525 386.

**Table 6 Bhattacharrya distance matrix of sample $s_3$**

|                  | Water     | Agriculture I | Agriculture II | Building | Bottomland |
|------------------|-----------|---------------|----------------|----------|------------|
| Agriculture I    | 1.999 999 | 1.685 667     |                |          |            |
| Agriculture II   | 1.999 999 | 1.999 989     | 1.999 999      |          |            |
| Building         | 2         | 2             | 2              | 2        | 2          |
| Bottomland       | 2         | 2             | 2              | 2        | 2          |
| Bare ground      | 1.998 278 | 1.591 221     | 1.776 549      | 1.738 729| 2          |

Where Bhattacharrya distance’s average=1.919 4, maximum=2, minimum=1.591 221.
Table 7  Bhattacharrya distance matrix of sample $s_4$

|             | Water | Agriculture I | Agriculture II | Building | Bottomland |
|-------------|-------|---------------|----------------|----------|------------|
| Agriculture I | 2     |               |                |          |            |
| Agriculture II | 2     | 1.692 465    |                |          |            |
| Building     | 2     | 1.999 918    | 1.999 996      |          |            |
| Bottomland   | 2     | 2            | 2              |          | 2          |
| Bare ground  | 1.99545 | 1.582 299    | 1.740 689      | 1.867 771 | 2          |

Where Bhattacharrya distance’s average=1.925 2, maximum=2, minimum=1.582 299.

Table 8  Bhattacharrya distance matrix of sample $s_5$

|             | Water | Agriculture I | Agriculture II | Building | Bottomland |
|-------------|-------|---------------|----------------|----------|------------|
| Agriculture I | 2     |               |                |          |            |
| Agriculture II | 2     | 1.705 036    |                |          |            |
| Building     | 2     | 1.999 998    | 2              |          |            |
| Bottomland   | 2     | 2            | 2              |          | 2          |
| Bare ground  | 1.999677 | 1.578 688    | 1.786 735      | 1.904 678 | 2          |

Where Bhattacharrya distance’s average=1.931 7, maximum=2, minimum=1.578 688.

In the next section, $\Delta$ and $\Delta_r$ are compared carefully with the Bhattacharrya distance, and they have a statistical significance linear relationship.

### 3.3 Analysis of the experimental results

By comparison we found that the sample’s $\Delta$ value has some relation with Bhattacharrya distance. The $\Delta$ value and average Bhattacharrya distance formed a scatter plot, as shown in Fig.4. It can be seen that $\Delta$ value is increasing when the average Bhattacharrya distance is increasing except sample $s_2$ which is indicated by a triangle, though there are three sets of data values that did not change much. The $\Delta_r$ of water and bottomland are obviously greater than the $\Delta_r$ of other classes in the later four groups of sample data, so the Bhattacharrya distances between the water or bottomland and other classes are greater in the later four groups of sample data. The re

![Fig.4 Scatter plot of $\Delta$ and average Bhattacharrya distance](image)

Table 9  the mean of Bhattacharrya distance of each class in sample data set $s_5$

|             | Water | Agriculture I | Agriculture II | Urban | Bottomland | Bareground |
|-------------|-------|---------------|----------------|-------|------------|------------|
| Mean        | 1.999 935 | 1.856 744    | 1.898 354      | 1.980 935 | 2           | 1.853 956  |

It can be seen that there exists a linear relationship between purity and Bhattacharrya distance can also be viewed within one sample. In sample $s_5$, the average Bhattacharrya distance of each category of surface object and $\Delta_r$ of each category of surface object (shown in Table 9) constitutes a scatter plot, which is shown in Fig.5.

It can be seen that there exists a linear relationship between $\Delta_r$ and average Bhattacharrya distance. When all the samples are given the same analysis, similar result can be obtained. All the scatter plots

![Fig.5 Scatter plot of $\Delta_r$ and average Bhattacharrya distance within sample $s_1$](image)
putting together forms Fig. 6. By observation and comparison, we can speculate that when the number of the sample tends to infinity, a linear function or quadratic function between $\Delta_x$ and average Bhattacharyya distance can be gotten using linear regression method.

The Bhattacharyya distance describes the divergence degree between classes from the point view of statistics, and the $\Delta_y$ and $\Delta$ assess the purity degree of each class from the point view of rough sets theory, including the whole sample data set, the divergence degree between classes and the quality inside the class. Unlike the traditional method such as probability, in which the experience assumptions usually need to be given, for example normal distribution, the method based on the rough sets theory can calculate the sample quality from the data itself. Thus it can guide the selection of sample point and provide a more objective quantitative evaluation for classification. From the experimental results, it can be shown that the better the purity degree inside a class, the better the divergence degree between classes.

4 Discussion and conclusions

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 another view different from the traditional statistical method. This method calculates the purity $\Delta_x$ of every class in a sample and overall purity $\Delta$ of a sample to evaluate the sample quality. Moreover, the new indices are compared with the traditional Bhattacharyya 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 a corresponding index can be obtained. 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. Other rough set measures, rough entropy as an example, can be used to evaluate sample quality, i.e., the non-harmonized degree. Also, we can visualize the sample quality by using these indices, so the sample quality can be observed more intuitively.

References

[1] Ge Y, Bai H X, Li S P, et al (2007) Measuring the quality of samples in the supervised classification of remotely sensed imagery[ C]. The 5th International Symposium for Spatial Data Quality, ITC, Enschede
[2] Richards J (1986) Remote sensing digital image analysis[M] New York: Springer-Verlag
[3] PCI Inc (2003) PCI Online Help, Version 9.0 [OL], http://www.pcigeomatics.com/services/support_center/pcihlp_gateway.html
[4] Sun J X (2003) Modern pattern recognition[M]. Changsha: The Defence University of Science and Technology Publishing House (in Chinese)
[5] ITT Inc. (2007) ENVI Online Help, Version 4.3[OL]. http://www.gisits.com/otros/ onlguide.pdf
[6] Liang J Y, Li D Y (2005) Uncertainty and knowledge acquisition in the information system[M]. Beijing: Science Press (in Chinese)
[7] Wang G Y (2001) Rough set theory and its knowledge acquisition[M]. Xi’an: Xi’an Jiaotong University Press (in Chinese)