THE STUDY OF KNOWLEDGE EXPRESSION ON IMAGE CLASSIFICATION BY USING ROUGH SETS THEORY AND PCA

*T. Y. Chou **T. C. Lei ***I. L. Shih
*T. Y. Chou
*Director, GIS Research Center,
Feng Chia University.
Email: jimmy@gis.tw

**T. C. Lei
**Assistant Professor, Graduate Institute of ESIST,
Feng Chia University.

***I. L. Shih
***Master of Institute of ESIST,
Feng Chia University.
100 Wenhwa Rd., Taichung, Taiwan, R.O.C.
Tel:(886)-4-2451666#127 Email: excellentsunny@gmail.com

KEY WORDS: Remote sensing, Rough Sets Theory, Knowledge Engine

ABSTRACT: Using texture information in image classification is usually an extensive solution to obtain enhanced accuracy. However, the whole mass of texture information is not all that useful in improving the accuracy of the classified image. On the other hand, it brings a lot of noise from the texture information. Traditionally, Principle Components Analysis (PCA) has been employed to extract vital knowledge. But, PCA is still controversial, as both influenced and non-influenced attributes were taken into account. Hence, this research employs a new method, Rough Sets Theory (RST), for image classification. RST efficiently deals with the uncertain problems by using Core and Reduce techniques on the surrounding world.

Setting up the structure of image classification using this new method involves several steps. Firstly, the texture information should be calculated from the image and included. Then, comparing with PCA, statistic normal distribution analysis is added. With Rough Set, using segment points by logical judgment in image classification improves accuracy. This aims at address two issues:

1. Selecting the core textures in image classification, and
2. Finding the segment point of core textures in image classification.

1. INTRODUCTION

In Taiwan, a great deal of efforts are made by the government to construct the rice area data and
estimate rice production correctly. The employment of both man-made processes and digitized data from air photo has not yielded an effective solution in the past. Therefore, the related research efforts to find a methodology for auto-investigation of the area of rice paddy using remote-sense images. The use of satellite imagery for observation provides an immediate and effective method for continually gathering surface information.

With time, the technological developments in improved texture information and spectral resolution were expected to lead to better classification of images. Unfortunately, the mass of improper additional texture information may also increase the noise in the classification process. Furthermore, the high dimensionality of multispectral data necessitates new analytical methods to avoid mistakes while sifting through voluminous information (Landgrebe, 2001).

In this paper, a new data mining technology namely Rough Sets Theory is introduced. Its goal is to reduce the dimensionality substantially without sacrificing significant information. While finding significant information, the steps involve putting cuts’ information of image classified into ERDAS Image – Knowledge Engineer.

2. THE STRUCTURE OF THE RESEARCH

This paper combines the original multispectral image and panchromatic image for the advantage of image information using PCA. Then, two strategies are established simultaneously. First, calculated texture information is integrated into the image spectral data. Secondly, the extraction of core attributes on the image relies on PCA and Rough Sets Theory. The next step utilizes the software of ERDAS Image – Knowledge Engineer to operate the classification on the pattern of paddy rice. As mentioned earlier, an increased efficiency results due to the combinations of ‘PCA+ Knowledge Engineer’ and ‘Rough Set + Knowledge Engineer’. The discrepancies in the classification accuracy via PCA and Rough Set techniques are compared.

2.1. Study Area

The study area is located at Tanzi County, Taichung, Taiwan. The complexity features of paddy rice, grass, bare land, building, asphalt road, water body, shadow area and others are shown in Figure 1. In the next section, the material is introduced.

2.2. Material

The material used is Quickbird fusion image, 2096×2096 Pixel by the area of 215 ha taken on 25/10/2003. The calculation of the texture information consists of three parts: (1) NDVI image (2) Fractal Dimension and (3) Semi-Variogram. The NDVI and Fractal Dimension algorithm can be obtained from each other and the Semi-Variogram was attained from Semi-Madogram Variogram and Pseudo-cross Semi-Variogram. During the course of texture information algorithms, the Semi-Madogram Variogram algorithm creates 4 images: MV(B), MV(G), MV(R) and MV(IR) and Pseudo-cross Semi-Variogram algorithm can create 7 images: PCV(B-G),
PCV(G-B), PCV(G-R), PCV(G-IR), PCV(R-G), PCV(IR-G) and PCV(IR-R).
It includes 17-bands to express an analysis material for knowledge rules of paddy rice.

3. METHODOLOGY

3.1. Quantization Problems of Rough Sets Theory (Nguyen and Skowron, 1995)
The construction of quantization problems of Rough Sets Theory will be as follows:
Let $A = (U, A \cup \{d\})$ be a decision table with a large number of values of objects from $U$ for some $a \in A$. We assume $V_a = [l_a, r_a] \subset \mathbb{R}$ for any $a \in A$ where $\mathbb{R}$ is the set of real numbers. We assume that $A$ is a consistent decision table. Let $P_a$ be a partition on $V_a$ (for $a \in A$) into subintervals i.e.

$$P_a = \{[C_0^a, C_1^a), [C_1^a, C_2^a), \ldots, [C_{k-1}^a, C_k^a]\}$$

for some integer $k$, where $l_a = C_0^a < C_1^a < C_2^a < \ldots < C_k^a = r_a$ and,

$$V_a = [C_0^a, C_1^a) \cup [C_1^a, C_2^a) \cup \ldots \cup [C_{k-1}^a, C_k^a) .$$

Any $P_a$ is uniquely defined by the set $C_a = \{C_0^a, C_1^a, C_2^a, \ldots, C_{k-1}^a, C_k^a\}$ called the set of cuts on $V_a$. In the sequel we often identify $P_a$ with the set of cuts on $V_a$ defined by $C_a$. Any family $\{P_a : a \in A\}$ where $P_a$ is a partition on $V_a$ is called a partition on $A$. Any pair $(a, c) \in P$ will be called a cut on $V_a$.

Any family $P = \{P_a : a \in A\}$ of partitions on $A$ defines from $A = (U, A \cup \{d\})$ a new decision table $A^p = (U, A^p \cup \{d\})$, where $A^p = (a^p : a \in A)$ and $a^p(x) = i \iff a(x) \in [C_i^a, C_{i+1}^a)$ for any $x \in U$ and $i \in \{0, \ldots, k\}$. The table $A^p$ is called P-quantization of $A$. The quantization problems of real value attributes of $A$ can be described as decision problem.

3.2. Complexity of quantization problems (Nguyen and Skowron, 1995)
Let $A = (U, A \cup \{d\})$ be a decision table where $U = \{x_1, x_2, \ldots, x_n\}$. An arbitrary attribute $a \in A$ defines a sequence $V_1^a < V_2^a < \ldots < V_{n_a}^a$, where $\{V_1^a, V_2^a, \ldots, V_{n_a}^a\} = \{a(x) : x \in U\}$ and $n_a \leq n$. Let $P_k^a$ be a propositional variable corresponding to the interval $[v_k, v_{k+1})$ for any $k \in \{1, \ldots, n_a - 1\}$ and $a \in A$. By $BV(A)$, we denote the set of all propositional variables of the above form.

Any partition $P \subseteq \bigcup_{a \in A} \{a\} \times V_a$ defines a valuation val-p of propositional variables $P_k^a$ by
valp \( P^a_k \) \( = \text{true} \) iff there exists a cut \((a, ca)\in P\) satisfying \( v^a_k \leq c_a < v^a_{k+1} \). Instead of \( \text{valp} ( P^a_k ) \) \( = \text{true} \) we will also write \( P \models v^a_k \).

By \( \varphi \{a, i, j\} \) we denote a disjunction of all Boolean variables from the set

\[ \{ P^a_k : [v^a_k, v^a_{k+1}] \subseteq [\min(a(x_i), a(x_j)); \max(a(x_i), a(x_j))] \} \]

Hence \( \text{val-p} \{ \varphi \ (a,i,j) \} = \text{true} \) iff there is a cut in \( P \) on \( Va \) between \( a(x_i) \) and \( a(x_j) \).

By \( \psi \ (i, j \) we denote a disjunction of all \( \varphi \{a, i, j\} \), where \( a \in A \) and \( a(x_i) \neq a(x_j) \).

Formula \( \psi \ (i, j \) is called the discernibility formula for objects \( x_i , x_j \) (we assume the disjunction of the empty set of variables to be equivalent to \( \text{true} \)). The discernibility Boolean prepositional formula of \( A \) is defined by:

\[ \Phi^d = \land \{ \psi(i, j) : d(x_i) \neq d(x_j) \} \]

Any non-empty set \( S = \{ P^a_{k_1}, P^a_{k_2}, ..., P^a_{k_r} \} \) of Boolean propositional variables from \( BV(A) \) defines a family of partition \( P(S) \) as follows:

\[ P(S) = \left\{ \left( a_1, \frac{v^a_{k_1} + v^a_{k+1}}{2} \right), \left( a_2, \frac{v^a_{k_2} + v^a_{k+2}}{2} \right), ..., \left( a_r, \frac{v^a_{k_r} + v^a_{k+r+1}}{2} \right) \right\} \]

3.3. The Rough Set Exploration System Software

In this paper, we utilized the “Rough Set Exploration System (RSES 2.2)” software to create the knowledge of image classification. This software was developed by Andrzej Skowron and his R&D team in Warsaw University (RSER 2.2 User’s Guide, 2005).

3.4. Knowledge expression

The ERDAS Expert Classifier interface is designed to handle the process that an expert in a particular field of expertise would use to analyze spatial data and infer information within a given location. In this study, the computer classifies the data in the knowledge database via Rough Set Method and PCA. A decision-making rule approach is used to present classified knowledge (ERDAS, 1997).

4. PROCEDURE

4.1. Rough Sets Theory + Rice Spatial Knowledge Classifier (RST + RSKC)

The knowledge rules of paddy rice can be determined using Rough Sets. The process involves (1) Training Sample from spectral image, (2) Computing the separated point, (3) Creating the
new information table, (4) Developing the engine of paddy rice knowledge rules classifier. As the process proceeds, 7 bands will be extracted, and the range of knowledge rule pattern presented on the NDVI, MV(R), MV(IR), PCV(G-B), PCV(IR-G), PCV(IR-R) and PCV(R-G). Then, the final result will be plugged into the ERDAS image classifier. The result is presented in Figure 1(a).

4.2. PCA + Rice Spatial Knowledge Classifier (PCA + RSKC)
This study extracts 3 PCA components which has an accumulated information content of over 98.51% by using the Threshold of Cumulative Variance Proportion (TCVP) method. In addition, minimum and maximum range on the paddy rice knowledge rules are displayed using the mean values of training sample attributes ( +/- two standard deviation values). Consequently, the knowledge database engine is constructed and the result is shown on Figure 1(b).

4.3. Result
Figure 1 (a) and Figure 1 (b) present the image classification results of RST + RSKC and PCA + RSKC. The yellow parts represent paddy rice and the black parts represent non-paddy rice. Table 1 displays the outcomes from two different analyses. There are 300 sample point check selected randomly. It includes 150 sample points of paddy rice and 150 sample points of non-paddy rice. The overall accuracy of Rough Set + RSKC (96.67%; Table 1(a); Figure 1(a)) is better than the conventional PCA + RSKC (86.00%; Table 1(b); Figure 1(b)).

5. DISCUSSIONS AND CONCLUSIONS
The results in this study can be expressed as three parts: (1) PCA + Rice Spatial Knowledge Classifier (2) Rough Set + Rice Spatial Knowledge Classifier and (3) The comparison on (1) and (2). The methodology of PCA is to consider the linear combinatorial relationship of all variables in the Information System.

The drawback of PCA is that all the input variables are taken into account. That is, the chaotic information of the target of interest may be incorporated into the classification computation. The advantage of Rough Set is to sieve out the core factors that are influenced by the classification outputs. This way, it can successfully eliminate the unnecessary information of the input variables and prevent the uncertainty component in image classification. The overall accuracy of Rough Set + RSKC (96.67%) is better than the conventional PCA + RSKC (86.00%). In addition, two main contributions are also drawn using Rough Set method. That is, the separate points and core factors are also found for the target categories. It is believed that the concept of separate points and core factors are very crucial for the related research.

6. Reference:
Table 1 Accuracy Report

(a) RS

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Producers Accuracy</th>
<th>Users Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>rice</td>
<td>95.45%</td>
<td>96.33%</td>
<td>0.9421</td>
</tr>
<tr>
<td>non-rice</td>
<td>97.88%</td>
<td>96.86%</td>
<td>0.9151</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall Classification Accuracy = 96.67%

(b) PCA

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Producers Accuracy</th>
<th>Users Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>rice</td>
<td>99.09%</td>
<td>72.67%</td>
<td>0.5684</td>
</tr>
<tr>
<td>non-rice</td>
<td>78.42%</td>
<td>99.33%</td>
<td>0.9818</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall Classification Accuracy = 86.00%