Journal of Southeast University (English Edition) Vol. 22, No. 2, pp. 222 – 227 June 2006 ISSN 1003—7985

Knowledge discovery method for feature-decision level fusion of multiple classifiers

Sun Liang[1,2] Han Chongzhao[1]

([1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

([2] Department of Electrical Information Engineering, Institute of Information Science and Technology, Zhengzhou, 450001, China)

Abstract: To improve the performance of the multiple classifier system, a new method of feature-decision level fusion is proposed based on knowledge discovery. In the new method, the base classifiers operate on different feature spaces and their types depend on different measures of between-class separability. The uncertainty measures corresponding to each output of each base classifier are induced from the established decision tables (DTs) in the form of mass function in the Dempster-Shafer theory (DST). Furthermore, an effective fusion framework is built at the feature-decision level on the basis of a generalized rough set model and the DST. The experiment for the classification of hyperspectral remote sensing images shows that the performance of the classification can be improved by the proposed method compared with that of plurality voting (PV).

Key words: multiple classifier fusion; knowledge discovery; Dempster-Shafer theory; generalized rough set; hyperspectral

The multiple classifier fusion is an important research topic with various names in different research fields. For example, it can be called the classifier combination in pattern analysis or the classifier ensemble in neural networks. Together there are two fusion scenarios[1]. In the first scenario, all base classifiers operate in different parts of one feature space. And in the second scenario, each base classifier may be allowed to operate in different measurement/feature spaces. Obviously, different types of features can be used effectively in the latter case. It is important to improve the performance of the classification in the case of lacking training samples and only a few classifiers being available.

In general, if only class labels are available, a voting fusion[2] is often used. However, it is difficult to make an effective fusion when only a few classifiers are available. In this case, it is appropriate to view an output of base classifiers as the condition that is associated with an uncertainty measure, such as the mass function in the Dempster-Shafer theory (DST).

This paper focuses on a feature-decision level fusion method in which the mass functions mentioned above are mined by using the proposed knowledge discovery approach from different feature spaces. Although one can obtain the mass function by using a classical rough set method [3], we must adopt a generalized rough set model because between-class separability being used in the classification relates to a nonequivalence relation.

Received 2005-11-10.

Foundation item: The National Basic Research Program of China (973 Program) (No. 2001CB309403).

Biographies: Sun Liang (1961—), male, graduate, associate professor, sun_liang@people.com.cn; Han Chongzhao (1943—), male, professor, czhan@mail.xjtu.edu.cn.

1 Generalized Rough Set and DST

Rough set[4] can give the set approximation based on equivalence relation on U. To nonequivalence relation on U, researchers have proposed several generalized rough set models[5–7]. The random rough set model is one of those taking into consideration that information may be obtained by some random experiments.

Let U and W be two finite nonempty sets, and $I: U \to \mathscr{P}(W) \setminus \{\varnothing\}$ be a set-valued mapping. For any $X \in \mathscr{P}(W)$, $x \in U$, the upper and lower approximations of X with respect to I can be defined, respectively, by

$$\underline{I}(X) = \{x: I(x) \subseteq X\}, \bar{I}(X) = \{x: I(x) \cap X \neq \varnothing\}$$

Furthermore, define a mapping $j: \mathscr{P}(W) \to \mathscr{P}(U)$ as

$$j(Y) = \{x \in U: I(x) = Y\} \qquad Y \in \mathscr{P}(W)$$

called the relational partition function of I, where Y is called a focus element iff $j(Y) \neq \varnothing$. Denote a family of subsets of U as

$$\mathscr{J} = \{j(Y) \in \mathscr{P}(U): j(Y) \neq \varnothing, Y \subseteq W\}$$

It is easy to verify that $\mathscr{J}$ forms a partition of U. Since U is finite, the family of all measurable sets formed from $\mathscr{J}$ is σ-algebra, denoted by $\sigma(\mathscr{J})$.

Theorem 1[5,7] Let $I: U \to \mathscr{P}(W) \setminus \{\varnothing\}$ be a set-valued mapping, and j be the relational partition function of I, then the following properties hold for any X

$\subseteq W$.

$$\underline{I}(X) = \cup\{j(Y): Y\subseteq X\}$$
$$\overline{I}(X) = \cup\{j(Y): Y\cap X\neq\varnothing\}$$
$$j(X) = \underline{I}(X)\setminus\cup\{\underline{I}(Y): Y\subset X\}$$

Suppose that Σ and Σ' ($\Sigma\subseteq\mathscr{P}(U)$, $\Sigma'\subseteq\mathscr{P}(W)$) are set algebra, then (U,Σ) and (W,Σ') are measurable spaces. The set-valued mapping $I: U\to\mathscr{P}(W)\setminus\{\varnothing\}$ is called a random set if $\{x\in U: I(x)\cap Y\neq\varnothing\}\in\Sigma$ holds for any $Y\in\Sigma'$.

Research[3,7] has shown that there exists a natural connection between the rough set theory and the DST. Based on the random rough set model, the connection between the two theories can be described as the following theorem.

Theorem 2[7] Let $I: U\to\mathscr{P}(W)\setminus\{\varnothing\}$ be a set-valued mapping and j the relational partition function of I. For $\mathscr{J}=\{j(Y)\in\mathscr{P}(U): j(Y)\neq\varnothing, Y\subseteq W\}$, let $\Sigma=\sigma(\mathscr{J})$ and P be probability measure on $(U,\sigma(\mathscr{J}))$. Then I is Σ-$\mathscr{P}(W)$ random set. If

$$m(Y)\triangleq P(j(Y))$$
$$\mathrm{Bel}(X)\triangleq P(I(X)) = \sum_{Y\subseteq X} m(Y)$$
$$\mathrm{Pl}(X)\triangleq P(\overline{I}(X)) = \sum_{Y\cap X\neq\varnothing} m(Y)$$

then m is the mass function over W, Bel and Pl are probability and plausibility functions on W.

2 Condition Mass Function (CMF) Generated by Classification Approximating Expert Decisions

Let $U=\{x_1, x_2, \ldots, x_n\}$ be a finite universe of discourse or object set, $A=\{a_1, a_2, \ldots, a_m\}$ be a finite condition attribute set, and $F=\{f_t: t\leqslant m\}$ be the set of relations between U and A, where f_t is the mapping $f_t: U\to V_t$ (V_t is the value set of attributes a_t, $t\leqslant m$). Then (U,A,F) is called information system, denoted by (U,A) for short. Decision table (DT) is a system having the form $(U, A\cup\{d\}, F\cup\{g\})$, where d is called decision attribute and $g: U\to V_d$ (V_d is the value set of d). In general, the DT is often denoted by $(U, A\cup\{d\})$ for short.

2.1 Approximate classification with respect to the tolerance relations on U

Based on binary equivalence relations on U, Skowron et al.[3] investigated to induce mass function from the classification approximating expert decisions. However, a generalized rough set model is needed for the computation of the mass function if the binary relation on U is nonequivalent.

Let $(U, A\cup\{d\})$ be a DT, τ_A be the tolerance relation on U with respect to A, denoted by

$$\tau_A=\{(x,y)\in U\times U: d_A(x,y)\leqslant h\} \quad (1)$$

where $d_A(x,y)$ is called between-object distance related to A and h is a fixed threshold. Let $I_A: U\to\mathscr{P}(W)$ be a set-valued mapping and $I_A(x)\triangleq\{y\in U: (y,x)\in\tau_A\}$. If we let $W=U$ and denote by j_A, the relational partition function of I_A, then for any $X\subseteq U$ the upper and lower approximations of X with respect to I_A can be defined by theorem 1 as

$$\underline{I_A}(X) = \cup\{j_A(Y): Y\subseteq X\} \quad (2)$$

$$\overline{I_A}(X) = \cup\{j_A(Y): Y\cap X\neq\varnothing\} \quad (3)$$

Furthermore, the boundary region of X with respect to I_A can be defined as

$$\mathrm{BN}_A(X) = \overline{I_A}(X)\setminus\underline{I_A}(X) \quad (4)$$

Otherwise, $I_A(x)\neq\varnothing$ for any $x\in\underline{U}$ because of the reflexivity of τ_A.

Let $V_d=\{1,2,\ldots,r(d)\}$, $H=\{h_1, h_2, \ldots, h_{r(d)}\}$ be the frame of discernment in the DST, where $r(d)=|V_d|$ ($|\cdot|$ denotes set cardinality). Let $\phi: \mathscr{P}(H)\to\mathscr{P}(V_d)$ be the standard bijection, i.e., $\phi(h_i)=i$ for $i=1,2,\ldots,r(d)$ and $\phi(\Delta)=\{i: h_i\in\Delta\}$ ($\Delta\subseteq H$). The equivalence relation on U with respect to d is defined as $R_d=\{(x,y)\in U\times U: g(x)=g(y)\}$. Furthermore, we denote

$$X_d = U/R_d = \{X_1, X_2, \ldots, X_{r(d)}\}$$
$$\mathrm{BH}_A(\Delta) = \left(\bigcap_{k\in\phi(\Delta)}\mathrm{BN}_A(X_k)\right)\cap\left(\bigcap_{k\notin\phi(\Delta)}\sim BN_A(X_k)\right)$$
$$K(Y)=\{k\in V_d: Y\cap X_k\neq\varnothing\} \qquad Y\in\mathscr{P}(U)$$
$$G_A(\Delta)=\cup\{j_A(Y): K(Y)=\phi(\Delta)\}$$

Theorem 3 Let $(U, A\cup\{d\})$ be a DT, $I_A: U\to\mathscr{P}(U)\setminus\{\varnothing\}$ be a set-valued mapping and $j_A: \mathscr{P}(U)\to\mathscr{P}(U)$ be the relational partition function of I_A. For any $\Delta\in\mathscr{P}(H)$,

$$G_A(\Delta)\triangleq\begin{cases}\underline{I_A}(X_k) & \Delta=\{h_k\}, X_k\in\mathscr{X}_d\\ \overline{\mathrm{BH}_A}(\Delta) & |\Delta|>1\\ \varnothing & \Delta=\varnothing\end{cases}$$

Then $\{G_A(\Delta)\neq\varnothing: \Delta\subseteq H\}$ is a partition of U.

Proof $|\Delta|=1$, then $\phi(\Delta)=K(Y)=\{k\}$ ($k\leqslant r(d)$) iff $Y\subseteq X_k$, $X_k\in\mathscr{X}_d$, so

$$G_A(\Delta)=\cup\{j_A(Y): K(Y)=\{k\}\}=\cup\{j_A(Y): Y\subseteq X_k\}=\underline{I_A}(X_k)$$

$|\Delta|>1$, then $\phi(\Delta)=K(Y)\subseteq\overline{V_d}$ iff $\forall k\in K(Y)$, $Y\not\subset X_k$ and $Y\cap X_k\neq\varnothing$, so $j_A(Y)\subseteq\bigcap_{k\in\phi(\Delta)}\mathrm{BN}_A(X_k)$. Moreover, $Y\cap X_k=\varnothing$ when $k\notin\phi(\Delta)$, i.e., $j_A(Y)\subseteq\bigcap_{k\notin\phi(\Delta)}\sim\mathrm{BN}_A(X_k)$, then

$$j_A(Y)\subseteq\mathrm{BH}_A(\Delta)\Rightarrow G_A(\Delta)\subseteq\mathrm{BH}_A(\Delta)$$

On the contrary, from Eqs. (3) and (4) we have

$$\mathrm{BH}_A(\Delta) \subseteq \bigcap_{k\in\phi(\Delta)} \mathrm{BN}(X_k) \subseteq \bigcap_{k\in\phi(\Delta)} \overline{I_A}(X_k) = \bigcap_{k\in\phi(\Delta)} \cup\{j_A(Y): Y\cap X_k \neq \varnothing\} = \cup\{\bigcap_{k\in\phi(\Delta)} \{j_A(Y): Y\cap X_k \neq \varnothing\}\} = \cup\{j_A(Y): K(Y) = \phi(\Delta)\} = G_A(\Delta)$$

Therefore, $G_A(\Delta) = \mathrm{BH}_A(\Delta)$. From the definition of $\mathrm{BH}_A(\Delta)$, if $\Delta_1 \neq \Delta_2 (\Delta_1, \Delta_2 \subseteq H)$, then $\exists i \in \phi(\Delta_2), i \notin \phi(\Delta_1)$ such that

$$\mathrm{BH}_A(\Delta_2) \subset \mathrm{BN}_A(X_i),\ \mathrm{BH}_A(\Delta_1) \subset \sim \mathrm{BN}_A(X_i)$$

i. e., $\mathrm{BH}_A(\Delta_1) \cap \mathrm{BH}_A(\Delta_2) = \varnothing$. Otherwise, since for $\forall x \in U$, $\exists i \in \phi(H)$ such that $x \in X_i$, then $x \in \underline{I_A}(X_i)$ or $x \in \bigcup_{\Delta\supset h_i} \mathrm{BH}_A(\Delta)$, i. e., $x \in \bigcup_{\Delta\subseteq H} G_A(\Delta)$. Therefore, $\{G_A(\Delta) \neq \varnothing: \Delta \subseteq H\}$ is a partition of U.

By the concept of supervised classification, we can call $\mathscr{X}_d = \{X_1, X_2, \ldots, X_{r(d)}\}$ the expert classification, and

$$\{G_A(\Delta) \neq \varnothing: \Delta \subseteq H\} = \{\underline{I_A}(X_1), \underline{I_A}(X_2), \ldots, \underline{I_A}(X_{r(d)})\} \cup \{\mathrm{BH}_A(\Delta): \Delta \subseteq H, |\Delta| > 1\}$$

the classification approximating the expert decisions.

2.2 CMF based on classification approximating expert decisions

Definition 1 Let m be the mass function over the frame of discernment H. There exists $E \subseteq H$, $E \neq \varnothing$ such that $\mathrm{Bel}(E) > 0$, if

$$m(A \mid E) \triangleq \begin{cases} \dfrac{m(A)}{\mathrm{Bel}(E)} & \text{if } A \subseteq E \\ 0 & \text{otherwise} \end{cases}$$

then the set function $m(\cdot \mid E): \mathscr{P}(H) \to [0, 1]$ is called the CMF over H with respect to the evidence E.

For any $i \in V_d$, let us denote

$$\mathscr{Y}_i = \{Y \in \mathscr{P}(U): Y = I_A(x), x \in U, Y \cap X_i \neq \varnothing\}$$

$$\mathscr{X}_i = \{X_k \in \mathscr{X}_d: \exists Y \in \mathscr{Y}_i, X_k \cap Y \neq \varnothing, k \leqslant r(d)\}$$

$$U_i = \bigcup_{X_k\in\mathscr{X}_i} X_k$$

$$H_i = \{h_k: X_k \in \mathscr{X}_i\}$$

$$K_i(Y) = \{k \in V_d: Y \cap X_k \neq \varnothing\} \qquad Y \in \mathscr{Y}_i$$

Theorem 4 Let $(U, A \cup \{d\})$ be a DT, $I_A: U \to \mathscr{P}(U) \setminus \{\varnothing\}$ be a set-valued mapping, j_A be the relational partition function of I_A and $\mathscr{J}_A = \{j_A(Y) \in \mathscr{P}(U): j_A(Y) \neq \varnothing, Y \subseteq U\}$. Let $\Sigma_A = \sigma(\mathscr{J}_A)$ and P be the probability measure on (U, Σ_A), then I_A is Σ_A-$\mathscr{P}(U)$ random set. For any $H_i \subseteq H, \Delta \subseteq H$, if

$$m_A(\Delta \mid H_i) = \begin{cases} \dfrac{P(G_A(\Delta))}{P(\underline{I_A}(U_i))} & \Delta \subseteq H_i, P(\underline{I_A}(U_i)) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{Bel}_A(\Delta \mid H_i) = \sum_{\Delta'\subseteq\Delta} m_A(\Delta' \mid H_i)$$

$$\mathrm{Pl}_A(\Delta \mid H_i) = \sum_{\Delta'\cap\Delta\neq\varnothing} m_A(\Delta' \mid H_i)$$

Then $m_A(\cdot \mid H_i)$ is the mass function over H, called the CMF with respect to the evidence H_i. $\mathrm{Bel}_A(\cdot \mid H_i)$, $\mathrm{Pl}_A(\cdot \mid H_i)$ are the belief and plausibility functions induced from $m_A(\cdot \mid H_i)$.

Proof Let $P(G_A(\Delta)) = m_A(\Delta)$. From theorem 3, we have

$$\sum_{\Delta\subseteq H} m_A(\Delta) = \sum_{\Delta\subseteq H} P(G_A(\Delta)) = P\left(\bigcup_{\Delta\subseteq H} G_A(\Delta)\right) = P(U) = 1$$

then $m_A(\cdot)$ is the mass function over H. Let $\mathrm{Bel}_A(H_i)$ denote the belief function induced from $m_A(\cdot)$, then we should prove $P(\underline{I_A}(U_i)) = \mathrm{Bel}_A(H_i)$ further. Since for any $Y \in \mathscr{Y}_i$, $j_A(Y) \subseteq \underline{I_A}(\bigcup_{X_k\in\mathscr{X}_i} X_k) = \underline{I_A}(U_i)$ and if

① $|K_i(Y)| = 1$, i. e., $K_i(Y) = \{k\}$, then $\exists X_k \in \mathscr{X}_i$ such that $Y \subseteq X_k$, so $j_A(Y) \subseteq \underline{I_A}(X_k) = G_A(h_k)$.

② $|K_i(Y)| > 1$, then $K_i(Y) = \phi(\Delta) \subseteq \phi(H_i)$, then $Y \not\subset X_k$ and $Y \cap X_k \neq \varnothing$ for any $k \in \phi(\Delta), X_k \in \mathscr{X}_i$, so

$$j_A(Y) \subseteq \mathrm{BH}_A(\Delta)$$

Therefore,

$$\underline{I_A}(U_i) = \underline{I_A}(\bigcup_{X_k\in\mathscr{X}_i} X_k) \subseteq \bigcup_{\Delta\subseteq H_i} G_A(\Delta)$$

Considering

$$G_A(\Delta) \subseteq \underline{I_A}(\bigcup_{X_k\in\mathscr{X}_i} X_k) \qquad \Delta \subseteq H_i$$

we have

$$\underline{I_A}(\bigcup_{X_k\in\mathscr{X}_i} X_k) \supseteq \bigcup_{\Delta\subseteq H_i} G_A(\Delta)$$

$$\underline{I_A}(U_i) = \underline{I_A}(\bigcup_{X_k\in\mathscr{X}_i} X_k) = \bigcup_{\Delta\subseteq H_i} G_A(\Delta)$$

$$P(\underline{I_A}(U_i)) = P\left(\bigcup_{\Delta\subseteq H_i} G_A(\Delta)\right) = \sum_{\Delta\subseteq H_i} P(G_A(\Delta)) = \sum_{\Delta\subseteq H_i} m_A(\Delta) = \mathrm{Bel}_A(H_i)$$

Therefore, $m_A(\cdot \mid H_i)$ is the CMF over H, $\mathrm{Bel}_A(\cdot \mid H_i)$ and $\mathrm{Pl}_A(\cdot \mid H_i)$ are the belief and plausibility functions induced from $m_A(\cdot \mid H_i)$.

3 Knowledge Discovery-Based Multiple Classifier Fusion (KD-MCF)

To c-class problems, let $U = \{x_1, x_2, \ldots, x_n\}$ be the set of symbols corresponding to the c-class training sample sets, $V_d = \{1, 2, \ldots, c\}$ $(c = r(d) \leqslant n)$ be the label set of the c classes, $H = \{h_1, h_2, \ldots, h_c\}$ be the frame of discernment and $A_j (j = 1, 2, \ldots, N)$ be the feature set for the j-th classifier. Then the N DTs corresponding to N classifiers can be established, denoted by $\mathrm{DT}_1, \mathrm{DT}_2, \ldots, \mathrm{DT}_N$. From each DT, one can acquire the c CMFs over H.

For an input pattern, let $L = \{l_1, l_2, \ldots, l_N\}$ denote the output of the classifier set, $K_L = \{k: k = l_j, j = 1, 2, \ldots, N\}$ be the class label set corresponding to L, where

$l_j \in V_d$ is the output of the j-th classifier. To the classifier set, the decision assigning an input pattern to a class s ($s \in K_L$) depends on the combination of all the CMFs relating to L. Therefore, the fusion method using the CMFs may be viewed as a feature-decision level fusion and the corresponding algorithm can be designed as follows:

① Establish N DTs with N feature sets. Suppose that pattern feature vectors include m independent standard normal variables. To avoid loss of classification information, the values of attributes ought to be taken statistical parameters such as $(\mu_p(t), \sigma_p(t))$ or $(\mu_p(t), \sigma_p(t))$ instead of discrete values in DT. Here, $\mu_p(t)$, $\sigma_p(t)$ represent, respectively, the mean and standard deviation of the t-th feature corresponding to the p-th sample set.

② To the DT_j, define the tolerance relation τ_{A_j} and I_{A_j} with an appropriate measure of between-class distance.

③ Select the appropriate classifiers for each kind of between-class distance. For example, the maximum-likelihood classifier (MLC) may be used corresponding to Bhattacharyya distance[8].

④ Compute $I_{A_j}(j = 1, 2, \ldots, N)$ and the approximate classifications $\{G_{A_j}(\Delta) \neq \varnothing : \Delta \subseteq H\}$ from the N DTs.

⑤ Given the probability distribution on U, then $m_{A_j}(\Delta \mid H_{l_j})$ $(j = 1, 2, \ldots, N;\ l_j = 1, 2, \ldots, c)$ can be calculated for each classifier by using the approximate classifications and theorem 4. For example, let P be the uniform probability distribution on U, then for any $\Delta \in \mathscr{P}(H)$ we have

$$P(G_{A_j}(\Delta)) = \frac{|G_{A_j}(\Delta)|}{|U|}$$

$$P(\underline{I}_{A_j}(U_i)) = \frac{\sum_{\Delta \subseteq H_i} |G_{A_j}(\Delta)|}{|U|}$$

so

$$m_{A_j}(\Delta \mid H_{l_j}) = \begin{cases} \dfrac{|G_{A_j}(\Delta)|}{\sum_{\Delta \subseteq H_i} |G_{A_j}(\Delta)|} & \Delta \subseteq H_{l_j} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

⑥ Fuse the CMFs corresponding to L using Dempster-Shafer's rules. Let m_L denote the orthogonal sum of these mass functions and e_L the evidence corresponding to L, then

$$m_L(\Delta \mid e_L) = m_{A_1}(\Delta \mid H_{l_1}) \oplus \ldots \oplus m_{A_N}(\Delta \mid H_{l_N}) \quad (6)$$

⑦ Assign an input pattern to a class label s ($s \in K_L$) if

$$F_L(h_s) = \max_{k \in K_L}(F_L(h_k)) \quad (7)$$

where $F_L(h_k)$ is called the heuristic fusion function[9] corresponding to e_L, which is defined as

$$F_L(h_k) = \alpha \mathrm{Bel}(h_k \mid e_L) + (1 - \alpha)\mathrm{Pl}(h_k \mid e_L) \quad (8)$$

where $\alpha \in (0.5, 1]$, $\mathrm{Bel}(h_k \mid e_L)$ and $\mathrm{Pl}(h_k \mid e_L)$ are the belief and plausibility functions induced from $m_L(\Delta \mid e_L)$.

⑧ Assign an input pattern as unclassified when the corresponding m_L cannot be obtained.

Fig. 1 illustrates the multiple classifier fusion framework, where $m_{jl_j}(l_j \in \{1, 2, \ldots, c\})$ denote the CMF corresponding to what the j-th classifier outputs class label $l_j \in K_L$, i.e., $m_{A_j}(\Delta \mid H_{l_j})$.

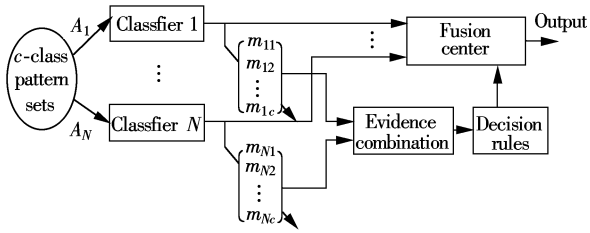

Fig. 1 Framework of the KD-MCF method based on DST and CMFs

4 Experiment

The experiments are conducted on the hyperspectral image cube (provided by China Coal-ARSC, Xi'an, China) that includes 80 bands with wavelength from 455.7 to 1 642.4 nm. According to the ground truth, three vegetation types and nine regions of interest (ROI) were selected from the image cube and three of them remained as test regions. The label set of the three vegetations and the number of corresponding samples are shown in Tab. 1.

Tab. 1 Lable set of three vegetation types and the number of corresponding samples

Land cover	Class label	Signs of training pixel sets	Number of training pixels	Number of testing pixels
Vegetation 1	1	x_1, x_2	$91(x_1) + 108(x_2)$	386
Vegetation 2	2	x_3, x_3	$114(x_3) + 103(x_4)$	333
Vegetation 3	3	x_5, x_6	$76(x_5) + 85(x_6)$	341

Applying PCA to the original image cube and its second derivative images[10], two feature sets A_1, A_2 were generated. Here, the second derivative images relate to 10 bands with wavelengths from 663.9 to 767.1 nm. A_1 consists of five dominant principle components (PCs) extracted from the second derivative images and A_2 includes three dominant PCs extracted from the original image cube. Another feature set A_3 consists of the original 80 bands and a statistical parameter. Thus, three DTs corresponding to the feature sets can be established as Tabs. 2, 3 and 4. The set

values corresponding to random sets I_{A_1}, I_{A_2} and I_{A_3} are calculated, respectively, with the tolerance relation.

Tab. 2 DT_1 corresponding to MLC_1 and PCs (A_1)

U	$a_1/10^{-3}$	$a_2/10^{-3}$	$a_3/10^{-3}$	$a_4/10^{-3}$	$a_5/10^{-3}$	d
x_1	(-27.5, 4.1)	(-5.5, 2.4)	(1.6, 1.3)	(0.2, 1.1)	(0.6, 1.0)	1
x_2	(-18.1, 2.9)	(-6.7, 1.2)	(-0.6, 0.7)	(0.0, 0.7)	(0.1, 0.6)	1
x_3	(-16.9, 2.8)	(2.4, 1.3)	(3.1, 1.0)	(1.7, 0.8)	(0.5, 0.5)	2
x_4	(-15.3, 3.8)	(3.5, 0.9)	(2.6, 1.1)	(2.0, 1.0)	(1.1, 0.5)	2
x_5	(-26.8, 5.7)	(2.5, 2.0)	(4.8, 1.8)	(1.7, 1.6)	(0.5, 1.3)	3
x_6	(-26.5, 4.4)	(2.4, 1.1)	(3.5, 1.3)	(2.0, 1.4)	(0.7, 1.0)	3

Tab. 3 DT_2 corresponding to MLC_2 and PCs (A_2)

U	$a_1/10^{-3}$	$a_2/10^{-3}$	$a_3/10^{-3}$	d
x_1	(670.9, 92.2)	(25.1, 20.6)	(-0.6, 7.6)	1
x_2	(438.7, 61.5)	(-5.0, 14.6)	(-2.1, 3.8)	1
x_3	(464.4, 60.9)	(-5.6, 11.1)	(26.0, 4.0)	2
x_4	(360.1, 91.9)	(-0.8, 16.4)	(6.2, 4.6)	2
x_5	(597.4, 127.6)	(132.4, 34.7)	(-35.0, 9.0)	3
x_6	(609.7, 86.0)	(100.0, 15.1)	(-26.9, 8.5)	3

Tab. 4 DT_3 corresponding to SAM and A_3

U	a_1	…	a_{80}	a_{81}	d
x_1	0.21	…	0.38	0.022	1
x_2	0.23	…	0.39	0.017	1
x_3	0.21	…	0.38	0.018	2
x_4	0.27	…	0.39	0.020	2
x_5	0.22	…	0.37	0.035	3
x_6	0.21	…	0.37	0.021	3

Note: $a_l(x_p) = \mu_p(l)$ and $a_{81}(x_p) = \rho_p(\text{rad})$ ($l = 1, 2, \ldots, 80$; $p = 1, 2, \ldots, 6$).

$$\tau_{A_1} = \{(x_p, x_q) \in U \times U: d_{A_1}(x_p, x_q) \leqslant 1.5\}$$
$$\tau_{A_2} = \{(x_p, x_q) \in U \times U: d_{A_2}(x_p, x_q) \leqslant 1.5\}$$
$$\tau_{A_3} = \{(x_p, x_q) \in U \times U: d_\beta(x_p, x_q) \leqslant 1\}$$

The results are shown in Tab. 5. Here, d_{A_1}, d_{A_2} are the Bhattacharyya distance with respect to feature sets A_1, A_2, respectively, which have the form in Ref. [8]. d_β is the similarity measure defined by

$$d_\beta(x_p, x_q) = \frac{\beta(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q)}{\rho_p + \rho_q},\ \beta(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q) = \cos^{-1}\left(\frac{\langle \boldsymbol{\mu}_p, \boldsymbol{\mu}_q \rangle}{\|\boldsymbol{\mu}_p\| \cdot \|\boldsymbol{\mu}_q\|}\right)$$

where $\boldsymbol{\mu}_p = \{\mu_p(1), \mu_p(2), \ldots, \mu_p(m)\}$ is the mean vector of training set x_p. ρ_p is the spectral angle[8] bound between the training pattern vector and $\boldsymbol{\mu}_p$, in which 70% training samples corresponding to x_p is included. The values of ρ_p ($p = 1, 2, \ldots, n$) can be obtained by statistical experiment. Suppose P is the uniform probability distribution on U, then the CMFs over $H = \{h_1, h_2, h_3\}$ are induced from the three DTs and shown in Tab. 6, respectively.

Tab. 5 Set values of I_{A_1}, I_{A_2} and I_{A_3}

U	I_{A_1}	I_{A_2}	I_{A_3}
x_1	$\{x_1\}$	$\{x_1\}$	$\{x_1, x_5, x_6\}$
x_2	$\{x_2\}$	$\{x_2, x_4\}$	$\{x_2, x_3\}$
x_3	$\{x_3, x_4, x_5, x_6\}$	$\{x_3\}$	$\{x_2, x_3, x_5\}$
x_4	$\{x_3, x_4, x_6\}$	$\{x_2, x_4\}$	$\{x_4\}$
x_5	$\{x_3, x_5, x_6\}$	$\{x_5, x_6\}$	$\{x_1, x_3, x_5, x_6\}$
x_6	$\{x_3, x_4, x_5, x_6\}$	$\{x_5, x_6\}$	$\{x_1, x_5, x_6\}$

Tab. 6 CMFs induced from DT_1, DT_2 and DT_3

Δ	m_{11}	m_{12}	m_{13}	m_{21}	m_{22}	m_{23}	m_{31}	m_{32}	m_{33}
$\{h_1\}$	1	0	0	1/4	1/4	0	0	0	0
$\{h_2\}$	0	0	0	1/4	1/4	0	1/6	1/6	1/6
$\{h_3\}$	0	0	0	0	0	1	0	0	0
$\{h_1, h_2\}$	0	0	0	1/2	1/2	0	1/6	1/6	1/6
$\{h_1, h_3\}$	0	0	0	0	0	0	1/3	1/3	1/3
$\{h_2, h_3\}$	0	1	1	0	0	0	1/6	1/6	1/6
$\{h_1, h_2, h_3\}$	0	0	0	0	0	0	1/6	1/6	1/6

The classifier set consists of two MLCs and a spectral angle mapping (SAM)[8] because the Bhattacharyya distance and spectral angle are used. To simplify, all land covers are classified into three categories by the MLCs with no threshold and by SAM with a large angle threshold (0.1 rad) such that all training samples are classified.

The classification accuracy comparisons are shown in Tabs. 7 and 8, where $PC_{1\text{-}20}$ represent the twenty dominant PCs extracted from the original image cube. Coefficient α in Eq. (8) is taken 0.6. Compared with MLC and plurality voting (PV)[2], the proposed KD-MCF is the best one in classification performance.

Tab. 7 Classification accuracy on training areas %

Classifier / Feature set	MLC_1 (A_1)	MLC_2 (A_2)	SAM (A_3)	MLC $PC_{1\text{-}20}$	PV	KD-MCF
Unclassified	0.00	0.00	0.00	0.00	1.21	0.00
Vegetation 1	90.95	100.00	94.47	100.00	96.98	100.00
Vegetation 2	93.09	97.70	89.86	100.00	98.16	100.00
Vegetation 3	100.00	90.68	91.93	100.00	96.27	100.00
Overall accuracy	94.28	96.53	92.03	100.00	97.23	100.00

Tab. 8 Classification accuracy on test areas %

Classifier / Feature set	MLC_1 (A_1)	MLC_2 (A_2)	SAM (A_3)	MLC $PC_{1\text{-}20}$	PV	KD-MCF
Unclassified	0.00	0.00	0.67	0.00	10.74	0.86
Vegetation 1	54.15	97.71	87.82	98.96	94.56	91.71
Vegetation 2	94.89	67.27	41.44	28.53	77.18	98.20
Vegetation 3	99.38	54.63	59.88	100.00	71.60	94.44
Overall accuracy	81.21	72.39	64.33	76.80	81.88	94.63

5 Conclusion

The experiment for the classification of hyperspectral remote sensing images shows that the proposed KD-MCF method surpasses PV in classification performance. The KD-MCF method has two attractive characteristics. First, the discovered knowledge is not ambiguous but transparent. And the other one is that

the CMFs are induced automatically from DTs. However, the research of this paper is merely preliminary. Some problems, such as finding diverse feature sets that complement each other and estimating the probability distribution on U need to be researched further.

References

[1] Kittler J, Hatef M, Duin RPM, et al. On combining classifiers [J]. *IEEE Trans Pattern Analysis and Machine Intelligence*, 1998, **20**(3): 226 - 239.

[2] Parhami B. Voting: a paradigm for adjudication and data fusion in dependable systems [A]. In: Diab H B, Zomaya A Y, eds. *Dependable Computing System* [C]. Wiley & Sons, Inc, 2005.

[3] Skowron A, Graymala-Busse J. From rough set to evidence theory [A]. In: Yager R R, Kacprzyk J, Fedrizzi M, eds. *Advances in the Dempster-Shafer Theory of Evidence* [C]. New York: Wiley, 1994. 193 - 236.

[4] Pawlak Z. *Rough set*: *theoretical aspects of reasoning about data* [M]. Boston: Kluwer Academic Publishers, 1991.

[5] Yao Y Y. Generalized rough set models [A]. In: Polkowski L, Skowron A, eds. *Rough Sets in Knowledge Discovery* [C]. Heidelberg: Physica-Verlag, 1998. 286 - 318.

[6] Slowinski R, Vanderpooten D. A generalized definition of rough approximations based on similarity [J]. *IEEE Trans Knowledge and Data Engineering*, 2000, **12**(2): 331 - 336.

[7] Wu W Z, Leung Y, Zhang W X. Connections between rough set theory and Dempster-Shafer theory of evidence [J]. *International Journal of General System*, 2002, **31**(4): 405 - 430.

[8] Richards J A, Jia X P. *Remote sensing digital image analysis*: *an introduction* [M]. 3rd ed. Berlin, Springer-Verlag, 1999. 322; 331.

[9] Ma Y, Chandler J S, Wilkins D C. On the decision making problem in Dempster-Shafer theory[A]. In: *Proc of the International Symposium on Artificial Intelligence* [C]. Cancun, Mexico, 1992. 7 - 11.

[10] William D P. The derivative ratio algorithm: avoiding atmospheric effects in remote sensing [J]. *IEEE Trans Geoscience and Remote Sensing*, 1991, **29**(3): 350 - 357.

特征-决策层多分类器融合的知识发现方法

孙 亮[1,2] 韩崇昭[1]

([1] 西安交通大学电子与信息工程学院,西安 710049)

([2]解放军信息工程大学理学院,郑州 450001)

摘要:为进一步提高多分类器系统的分类性能,提出了一种基于知识发现的特征-决策层多分类器融合新方法.各分类器工作于具有互补分类信息的不同特征空间且其类型由不同的类间可分性度量决定.各分类器输出的不确定性度量从建立的多个决策表中导出,并具有条件 mass 函数的形式.进而基于广义粗集模型和 Dempster-Shafer 理论(DST)构造了一种新颖的特征-决策层融合框架.高光谱遥感图像的分类实验表明,与多数表决融合(PV)相比,所提出的方法可有效提高多分类器系统的分类性能.

关键词:多分类器融合;知识发现;Dempster-Shafer 理论;广义粗集;高光谱

中图分类号:TP391