QUATORZIEME COLLOQUE GRETSI - JUAN-LES-PINS - DU 13 AU 16 SEPTEMBRE 1993

AMELIORATION DE LA CLASSIFICATION SUPERVISEE DES DONNEES SATELLITAIRES PAR UN CHOIX OPTIMAL DE CARACTERISTIQUES.

S.CHITROUB, A.BELHADJ-AISSA

Laboratoire Traitement d'Images, Institut d'Electronique.
Université des Sciences et de la Technologie Houari Boumedienne.
BP 32, EL Alia, Bab Ezzouar, Alger, Algérie.

RÉSUMÉ

La mise au point de méthodes susceptibles de fournir des caractéristiques spectrales précises est primordiale pour la classification supervisée des données satellitaires. Lorsque l'on n'a que la forme analytique des fonctions de densités de probabilités des classes connues à priori, le problème de la classification se pose en termes d'estimation exacte des paramètres des densités et de discrimination spectrale des classes. Pour une meilleure classification, on doit établir un compromis entre la réduction des données et la sélection des caractéristiques discriminantes.

ABSTRACT

The development of statistical methods which provide accurate spectral features is necessary for the supervised classification of Remote Sensing data. If the analytical form of the probability density function of the classes is a priori known then the problem of classification should be reduced to an accurate estimation of the P.D.F parameters and spectral discrimination of classes. A best classification is then obtained by establishing compromise betwen the reduction of data and the selection of discriminant features.

INTRODUCTION

Lorsqu'on ne dispose pas de suffisament de données d'apprentissage à partir desquelles le modèle décisionnel est élaboré, la réduction de la dimension des données devient nécessaire. Cependant, une réduction trés poussée peut provoquer une perte d'information importante pour la discrimination des classes. Le but de cet article est de montrer comment on effectue un choix optimal des caractéristiques pour améliorer le résultat d'une classification supervisée.

REDUCTION DE CARACTERISTIQUES PAR TRANSFORMATION DES DONNEES

Avec cette approche nous voulons obtenir une représentation des proximités des points images dans un espace de variables décorrélées de faible dimension. Un critère à retenir consiste à assurer une redistribution essentielle des variabilités internes des classes le long des axes qui forment le nouvel espace.

I>ANALYSE EN COMPOSANTES PRINCIPALES (ACP):

Dans l'ACP, le choix de l'espace de projection s'effectue selon deux critères: déformer le moins possible les données en projection **[1] [4]** et choisir les axes à retenir qui assurent une reproduction totale de la variance globale de l'image.

1- CRITERE DE CHOIX DU REFERENCIEL:

Pour un sous-espace vectoriel affine E_c qui possède comme origine le point c, la déformation globale des données en projection est mesurée par la quantité ξ_c appelée inertie autour de E_c **[4]** et définie par:

$$\xi_c = \sum_{i=1}^{n} P(x_i).d^2(x_i , \underline{x}_i)$$

n: le nombre de points en projection
$P(x_i)$: probabilité d'apparition de x_i
$d^2(x_i,\underline{x}_i)$: carré de la distance entre x_i et sa projection orthogonale $\underline{x}_i$ sur E_c.
Si g et E_g sont respectivement le centre de gravité des points images dans l'espace de départ et le sous-espace parallèle associé, $\underline{y}_i$ et g les projections orthogonales de x_i et g sur E_g et E_c, alors on a:

$$\xi_c = \sum_{i=1}^{n} P(x_i).\|x_i-\underline{x}_i\|^2$$

$$= \xi_g + \sum_{i=1}^{n} P(x_i).\left(\|\underline{y}_i-\underline{x}_i\|^2 + 2.\langle x_i-\underline{y}_i,\underline{y}_i-\underline{x}_i\rangle\right)$$

$$\text{d'où:}\quad \xi_c = \xi_g + d^2(g , E_c)$$

Pour rendre la quantité ξ_c minimale nous devons choisir comme origine du sous-espace d'arrivée le centre de gravité des données images dans l'espace de départ.

2- NOUVELLE BASE DE PROJECTION:

La richesse de l'information contenue dans une image multispectrale est caractérisée par la variance totale de cette image. Cette variance se répartit entre toutes les variables caractéristiques originales de l'image, qui sont plus ou moins corrélées, ce qui se traduit par une redondance d'information. Par contre si l'on souhaite concentrer le maximum d'information dans un nombre réduit de nouvelles variables non corrélées, on a intérêt à sélectionner le sous-ensemble des caractéristiques qui maximisent la variance expliquée par l'image originale. Soit G la transformation linéaire qui permet d'obtenir ces nouvelles variables:

$\mathbf{y = G.x}$ x: données de départ centrées

Par définition, la matrice de covariance des données y est:

$$\Sigma_y = E\left(y.y^t\right) = E\left(G.x.x^t.G^t\right) = G.\Sigma_x.G^t$$

Σ_x= matrice de covariance des données x.
Nous voulons que Σ_y soit diagonale (variables décorrélées) avec les éléments diagonaux classés dans l'ordre décroissant (maximisation de la variance totale dans les premières composantes).
La décomposition spectrale de la matrice Σ_x (symétrique) **[1]** est donnée par:

$$\Sigma_x = \sum_{i=1}^{k} \lambda_i.d_i.d_i^t$$

k: rang de la matrice Σ_x
λ_i: valeurs propres de Σ_x
d_i: vecteurs propres normalisés associés avec: $d_i^t.d_j = 1$ pour i=j et 0 pour $i \neq j$.
Soit D la matrice orthogonale formée par les vecteurs propres: D = [d1 d2.....dk], on a alors:

$$\Sigma_x = D.\Sigma.D^t$$

où Σ est la matrice diagonale des valeurs propres de Σ_x. Comme $D^t.D = D.D^t = I$ matrice idendité, en posant $G = D^t$ on obtient:

$$\Sigma_y = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & .\lambda_k \end{bmatrix} \quad \textbf{avec: } \lambda_1>\lambda_2>...>\lambda_k>0$$

Ainsi la nouvelle base orthonormée est la matrice transposée des vecteurs propres normalisés de la matrice de covariance globale des données images en projection. Les coefficients de combinaison permettent de donner l'importance de chaque variable originale dans la construction de chaque composante principale. La composante à coefficients élevés présente une partie importante de la variance totale de l'image **[3]**. Si les variables originales sont affectées du même bruit de variance σ_b^2, de moyenne nulle et décorrélé des variables, on a alors:

$\Sigma_x = \Sigma_{x0} + \sigma_b^2.I$ et $\Sigma_{x0}.d_i = (\lambda_i - \sigma_b^2).d_i = \lambda_{i0}.d_i$

Σ_{x0}: matrice de covariance des données images non bruitées. Cela signifie que chaque composante principale est affectée du même bruit de variance σ_b^2. Le rapport signal à bruit est: $S/B = \lambda_i/\sigma_b^2$, par conséquent les dernières composantes de variances petites sont trés bruitées.

II>ANALYSE CANONIQUE:

L'idée consiste à rechercher de nouvelles variables (discriminantes) correspondant à des directions séparant le mieux possible en projection les M classes de l'échantillon d'apprentissage.

1-FORME LINEAIRE DISCRIMINANTE DE FISHER:

La variance totale de l'image se décompose en variances à l'intérieur des classes et variances entre classes. En analyse canonique, on utilise les matrices de covariance intra-classes et inter-classes pour formuler le critère de séparation de FISHER. Si Σ_W et Σ_B sont respectivement matrices de covariance intra-classes et inter-classes, on a:

$$\Sigma_W = \sum_{i=1}^{M} P(C_i).E\left(x.x^t|C_i\right) = \sum_{i=1}^{M} P(C_i).\Sigma_i$$

$$\Sigma_B = E\left((m_i-m_0).(m_i-m_0)^t\right), \quad m_0 = \sum_{i=1}^{M} P(C_i).m_i$$

C_i: la i^{em} classe
$P(C_i)$: sa probabilité à priori
Σ_i: sa matrice de covariance
m_i: la moyenne de la i^{em} classe
m_0: la moyenne globale des données x

$$\lambda_i = \sigma_B^2 / \sigma_W^2 = (d^t.\Sigma_B.d) / (d^t.\Sigma_W.d)$$

est le rapport variance inter-classes sur la variance intra-classes. Si d est le vecteur qui maximise le rapport λ, la forme $y = d^t.x$ est la fonction linéaire discriminante de FISHER.

2- EQUATION AUX VALEURS PROPRES GENERALISEE:

Soit $\mathbf{y = D^t.x}$ la transformation qui nous permet d'obtenir des variables donnant une séparation optimale entre les classes. Les matrices de covariance Σ_W, Σ_B dans le nouvel espace ont donc la forme:

$$\Sigma_{Wy} = D^t.\Sigma_W.D \text{ et } \Sigma_{By} = D^t.\Sigma_B.D$$

La fonction $\lambda = f(d)$ à maximiser étant homogène de degré 0 en d (invariante en d), le maximum de cette fonction est atteint pour un vecteur d tel que: $\partial\lambda / \partial d = 0$

$$\frac{\partial \lambda}{\partial d} = \frac{\partial}{\partial d}\left[(d^t.\Sigma_B.d).(d^t.\Sigma_W.d)^{-1} \right]$$

$$= 2.\Sigma_B.d\ (d^t.\Sigma_W.d)^{-1}$$
$$- 2.\Sigma_W.d\ (d^t.\Sigma_B.d).(d^t.\Sigma_W.d)^{-2} = 0$$
$$\Sigma_B.d - \Sigma_W.d\ (d^t.\Sigma_B.d).(d^t.\Sigma_W.d)^{-1} = 0$$
$$(\ \Sigma_B - \lambda.\Sigma_W\ ).d = 0$$

Cette équation aux valeurs propres généralisée devient:

$$(\ \Sigma_B - A.\Sigma_W\ ).D = 0$$

Avec A la matrice des valeurs propres. Les vecteurs propres trouvés sont donc les vecteurs unitaires des axes discriminants cherchés. Le pouvoir discriminant de ces axes est fourni par les valeurs propres λ_i. Cependant, une contrainte additionnelle, pour avoir une homogénéité des classes, est que:

$$\Sigma_{Wy} = D^t.\Sigma_W.D = I$$

REDUCTION DE CARACTERISTIQUES PAR MESURES DE SEPARABILITE.

La sensibilité des caractéristiques originales aux réponses spectrales des objets terrestres à classifier est variable. Notre approche, est de trouver un sous-ensemble des variables originales qui donnent un pouvoir discriminant maximal des classes. Les variables n'aidant pas à la discrimination seront éliminées.

I>DIVERGENCE ET DIVERGENCE TRANSFORMEE:

Le rapport des densités des probabilités de deux classes:

$$R_{ij}(x) = P(x|C_i)\ /\ P(x|C_j)$$

est une mesure instantanée de chevauchement entre les deux classes au point x.

Soit: $\tilde{R}_{ij}(x) = \log(P(x|C_i)) - \log(P(x|C_j))$

$$E\left(\tilde{R}_{ij}(x)|C_i\right) = \int_x \tilde{R}_{ij}(x).P(x|C_i).dx$$

est la valeur moyenne du produit du rapport des densités par la distribution de la classe. Donc, la divergence d'une paire de distributions est mesurée par:

$$D_{ij} = E\left(\tilde{R}_{ij}(x)|C_i\right) + E\left(\tilde{R}_{ji}(x)|C_j\right)$$
$$= \int_x (P(x|C_i)-P(x|C_j)).\log(P(x|C_i)/P(x|C_j)).dx$$

Si les classes sont normales on a alors:

$$D_{ij} = 0.5.Tr\left((\Sigma_i - \Sigma_j).(\Sigma_j^{-1} - \Sigma_i^{-1})\right) +$$
$$0.5.Tr\left((\ \Sigma_i^{-1} + \Sigma_j^{-1}\ ).(m_i - m_j).^t(m_i - m_j)\right)$$

le second terme étant en quelque sorte la distance normalisée (par covariance) entre les moyennes des classes. On formule la valeur moyenne de la séparabilité par:

$$D_{moy} = \sum_{i=1}^{M} \sum_{i=1}^{M} P(C_i).P(C_j).D_{ij}(x)$$

Pour une large séparation entre deux classes (une grande valeur de D_{ij}), le comportement de D_{moy} ne traduit pas forcément une classification correcte. Par contre, avec la divergence transformée:

$$DT_{ij} = 2.(1 - \exp(-D_{ij}/8))$$

le comportement de saturation coincide avec une classification correcte.

II>DISTANCE DE BHATTACHARYYA ET DISTANCE DE JEFFRIES-MATUSITA:

Avec la distance de Bhattacharyya on mesure la séparabilité entre deux densités de probabilités des deux classes:

$$B_{ij} = -\log\left(\int_x (P(x|C_i).P(x|C_j))^{1/2}.dx\ \right)$$

Pour des densités gaussiennes B_{ij} admet comme expression:

$$B_{ij} = (m_i - m_j).^t\left((\Sigma_i + \Sigma_j)/2\right)^{-1}.(m_i - m_j)/8$$
$$+ 0.5.\log\left(|\Sigma_i + \Sigma_j|^{1/2} / (\ |\Sigma_i|^{1/2}.|\Sigma_j|^{1/2}\ )\right)$$

Le premier terme ressemble donc au carré de la distance normalisée entre les moyennes des classes. La distance de Jeffries - Matusita est définie comme:

$$JM_{ij} = 2.(1 - \exp(-B_{ij}))$$

Le comportement de saturation avec cette mesure n'est pas différent de celui espéré pour une classification correcte. L'indication moyenne de cette mesure est:

$$JM_{moy} = \sum_{i=0}^{M} \sum_{i=0}^{M} P(C_i).P(C_j).JM_{ij}(x)$$

LE CHOIX OPTIMAL DES CARACTERISTIQUES

Le choix optimal consiste à retenir un nombre minimum de variables, tout en assurant la discrimination des classes. Avec l'approche de réduction par transformation des données, l'ACP permet d'expliquer et d'interpréter la structure variance- covariance de l'ensemble global des données images en termes de premières composantes principales. Seulement les dernières composantes peuvent présenter des variances utiles à la discrimination des classes. Dans ce cas, L'ACP n'est donc pas sensible à la structure interne des classes [2]. L'analyse canonique optimise l'aspect discriminant de la transformation en tenant compte de la variabilité interne des classes, mais elle reste dépendante de la qualité des données d'apprentissage sur lesquelles elle est basée [2] [5]. Une approche, moins coûteuse, consiste à mesurer le degré de séparabilité entre les classes fournies par les variables originales. La divergence et la distance **B** restent les distances statistiques les plus convenables pour sélectionner les variables qui donnent une séparation maximale entre deux classes seulement [3]. Cependant, pour plus de deux classes la **DT** et La distance de **JM** possèdent un comportement de saturation qui coincide avec une classification correcte. La **DT** présente en

plus l'avantage d'être moins coûteuse du point de vu calcul [3] [2].

RESULTATS

Les images ci-contre représentent les résultats des classifications, six classes, d'une scène (300x255) de la région de sud d'Algérie (Laghouat). Les images TM sont prises le 1/ 1/ 89. L'algorithme utilisé est la classification de Mahalanobis. Nous remarquons que la meilleure classification est obtenue à partir des trois axes principaux de l'ACP. Les zones humides, la végétation, le sol nu et les chainons de montagnes sont bien discriminés. La classification à partir des deux premiers axes canoniques, correspondant aux deux plus grandes valeurs propres, présente des chevauchements entre les classes. Ces chevauchements sont dûs à la sélection des échantillons. Les résultats obtenus à partir des combinaisons sélectionnées par les mesures des distances sont moins fiables à cause de la représentativité des échantillons d'apprentissage (le tableau).

combinaison des canaux	Dmoy	DTmoy	Bmoy	JMmoy
1 3 5	1.35	228.50	0.096	0.390
1 3 7	*4.80	*1907.0	0.520	0.650
1 5 7	15.8	1578.0	0.360	*0.930
3 5 7	15.2	951.20	*0.600	0.630

CONCLUSION

Dans cet article nous avons abordé le problème d'influence de la dimension et de la nature des données sur la taille des échantillons d'apprentissage. L'approche de l'établissement d'un compromis entre réduction et discrimination des données peut optimiser l'erreur et le temps de calcul de la classification. Cependant nous n'avons pas pris en considération la représentativité des échantillons.

BIBLIOGRAPHIE

[1] Richard A.Johnson & Dean W.Wichern.
Applied Multivariate Statistical Analysis. Prentice hall internatonal NEW-JERSEY 1990.

[2] John.A.Richards.
Remote Sensing Digital Image Analysis. Edition Spring-Verlag 1986.

[3] P.W.Mausel, W.J.Krambre,& K Lee.
Optimum Band Selection for Supervised Classification of Multispectral Data. P.E.R.S, January 1990 pp 55-60.

[4] Gillex Celeux et al.
Classification Automatique Des Données. Dunod Informatique, 1989.

[5] Byungyong Kim and David A.Landgrebe.
Hieraachical Classifier Design in High-Dimensional, Numerous Class Cases. IEEE Transaction On Geoscience And Remote Sensing, Vol, 29, NO, July 1991.

image 1 - résultat à partir des 3 premiers axes de l'ACP.

image 2 - résultat à partir de 2 premiers axes canoniques.

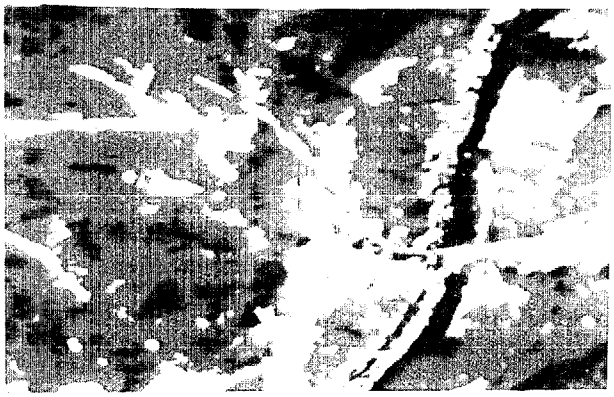

image 3 - résultat à partir de la combinaison (1 3 7) des canaux originaux.

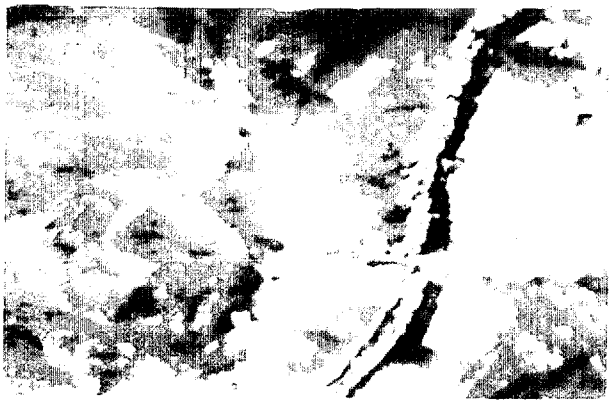

image 4 - résultat à partir de la combinaison (1 5 7) des canaux originaux.