THEMATIC MAP COMPILATION USING NONPARAMETRIC CLASSIFICATION METHODS.

Vladimir Cervenka

Institut of Surveying and Mapping, Prague
Czechoslovakia

Commission No.: III/3

ABSTRACT:

A method combining unsupervised clustering and supervised nonparametric classification of multispectral image data will be described. The creation of sufficiently representative training sets for supervised classification may be a serious problem - it is difficult to find training samples, which cover the whole feature space. Therefore results of unsupervised classification are used for completion of terrestrial investigation. Then the training data are verified using generalized entropy measure and mutual information. Finally the principles of nonparametric Bayesian decision based on Parzen windows are applied. Nonparametric methods have been shown to yield excellent results in applications other than remote sensing for the present. These methods are suitable especially when there is a poor knowledge about real probability densities or about their functional form. Unfortunately, they require storage and computation proportional to the number of samples in the training set.

KEY WORDS: Algorithm, Artificial Intelligence, Classification, Feature Extraction, Image Interpretation, Thematic, Training

1. INTRODUCTION

Gathering of information on the land use belongs to the main goals of remote sensing methods. This task is of special importance in regions with complicated structural zoning, e.g. in urban aglomerations and their surrounding. At present, Thematic Mapper (TM) data are frequently exploited for these purposes. A great attention has also been paid to the development of their automatic interpretation (classification). There are two principal approaches to the classification: supervised and unsupervised one.

Any computer classification that will lead to a ground-cover thematic map is based on the ground truth data gathered from selected area. The choice of training samples has to be representative, but random. However, the creation of sufficiently representative training sets may be a serious problem. Satellite images cover some hundreds km^2 nevertheless it is difficult to find suitable training samples, which cover the whole feature space. Therefore results of unsupervised classification are used for completion of terrestrial investigation when significantly different spectral classes are determined. The unsupervised classification enables to reduce the extent of subsequent supervised classification to a selected subset of spectral classes.

The notion of unsupervised classification will be presented in Section 2. The interpretation of clustering results in terms of mutual information will be proposed in Section 2.1. The practical aspects of nonparametric classification methods and various approaches are discussed in Section 3.

2. UNSUPERVISED CLASSIFICATION

The clustering method ISODATA has been used to analyze satellite data (Charvat, 1987a). Using this method approximately 50 % sample of pixels in the scene is clustered. In the k-means ISODATA method the pixels are placed in k groups (clusters) according to the similarity of digital features. The cluster centres are established during the iterative clustering. Then all pixels are mapped onto the original spatial domains using the nearest neighbour classifier. To avoid the excessive CPU time requirements, a threedimensional histogram is used when all samples in the feature space with the same feature values are represented with a specific histogram cell. The clustering process is realized in the reduced feature space only. A feature reduction technique is necessary for this reason - usually three new synthetic features (images) are computed.

2.1 Feature reduction

There are two basic reasons for incorporating the feature reduction procedure into the classification process. The ISODATA method uses threedimensional histogram, so the maximal number of features is three. A color composite production is the second reason for transformation of all disposable spectral bands into the three ones. The color composite created on the basis of the three uncorrelated features preserves great deal of spectral information from all original spectral bands. The method used improves the contrast of the color composite significantly. The color composites seems to be a useful tool for collection and verification of training samples as well as for the visual verification of classification results.

The use of "Tasseled Cap" transformation (Crist, 1984a) or the principal component method for this purpose has been described. A method based on neural networks can be utilized successfully (Charvat, 1990) when the back propagation algorithm (Hinton, 1987) is used. The neural net proposed consists of three layers. Input and output layer has the same number of nodes (neurons) equal to the number of spectral bands, the middle layer has three nodes in our case. Each node in the middle layer is connected with all nodes in preceding and succeeded layer. The neural net can be described by a unidirectional graph where nodes (neurons) bear some value. A certain weight is assigned to every connection. In the course of adaptation the feature values of selected samples are assigned to the nodes in the input layer and the values x_i in the middle and output layer are computed according to the expression

$$x_i = S\left(\sum_{j \in J} w_{ij} \cdot x_j \right) , \qquad (1)$$

where J is a set of neurons from previous layer and S is usually a sigmoid function. The weights of connections are changed using the back propagation algorithm until all node values in the input and output layer are approximately the same. Then the corresponding values in the middle layer can be considered as the effective comprimation of the original information. Finally the new features are computed for all pixels when the original feature values are introduced to the input layer of "instructed" network. The synthetic images computed may be used as R, G, B components of additive color composite. The technical details and description of adaptation process are beyond the scope of this paper and has been discussed by several authors (Fahlman, 1988).

2.2 Interpretation of clustering results

The ISODATA method produces clusters that can be bounded by a hypersphere or by a hyperellipsoid. Therefore it is necessary to group the data into more clusters than is the number of spectral classes. Some of the classes are broken into a several clusters. Higher number of clusters brings problems in subsequent interpretation of classification results. The theory of information gives us an efficient tool for solution of this problem (Charvat, 1990).

If P(x) denotes probability distribution of random variable X in some discrete space, the Shannons entropy H(Px) is defined as follows:

$$H(P_X) = - \sum_x P(x) \cdot \log P(x). \qquad (2)$$

When P(x,y) is a probability distribution of a composed variable (X,Y) and P(x), P(y) are the marginal distributions, then mutual information between variables X and Y is defined as follows:

$$I(X,Y) = \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x) \cdot P(y)} . \qquad (3)$$

The mutual information I(X,Y) can be considered as a general dependency measure between the variables X and Y.

The result of unsupervised classification may be interpereted easily when using the mutual information. Number of resulting clusters even after removing of nonsignificant ones is usually high. It is necessary to join several classes in the resulting image. Let $P(\omega_i, \omega_j)$ is a probability that the classes ω_i and ω_j occur in the neighbouring pixels and $P(\omega_i)$, $P(\omega_j)$ are aposteriori probabilities of classes (areal extents). Then the spatial dependency between individual classes may be described using the mutual information by the expression:

$$\sum_{i,j} P(\omega_i, \omega_j) \cdot \log \frac{P(\omega_i, \omega_j)}{P(\omega_i).P(\omega_j)} . \qquad (4)$$

It is the mutual information computed in the image space. For every two classes the value of loss of this mutual information is computed if they are joined. The system recommends to join such two classes for which this loss of information is minimized. The procedure is repeated until a satisfactory result is reached.

A preliminary unsupervised classification and interpretation yields the approximate areal extents of the cover classes. They can be used as estimates of apriori class probabilities when supervised classification is applied. The resulting cluster domains and color composite map are used to route the terrestrial investigations when main landcover classes are delineated.

3. SUPERVISED CLASSIFICATION

3.1 Verification of training samples

When the supervised classification is used for satellite image data interpretation, the gathering of suitable training samples creates the main problem. It is necessary to test the separability of classes and verify labeling of training polygons. Some methods solving these problems for normally distributed data have been already investigated (Charvat, 1987b). They are based on the statistical comparisons of mean vectors and covariance matrices.

The mutual information can characterize the separability between classes. Let the training sets are collected for every class $\omega_j \in \Omega$, for $j = 1,\ldots,M$ (M is a number of classes), x will be a random vector of feature space $\mathcal{X}$ which represents the multispectral image. The probability distributions $P(\omega_i)$, $P(x)$ and $P(x,\omega_i)$ can be estimated on the ground of training samples. In the case of absolutely separable classes the mutual information $I(X,\Omega)$ and entropy $H(P_\Omega)$ of probability distribution of classes are equal. It follows:

$$I(X,\Omega)/ H(P_\Omega) = 1 , \qquad (5)$$

where

$$I(X,\Omega) = \sum_{x,i} P(x,\omega_i) \cdot \log \frac{P(x,\omega_i)}{P(x) \cdot P(\omega_i)} \qquad (6)$$

and

$$H(P_\Omega) = - \sum_i P(\omega_i) \cdot \log P(\omega_i). \qquad (7)$$

The algorithm for verification of training samples is based on this idea:

1) Class identifiers are assigned to every training polygon - every polygon is considered as a temporary spectral class.

2) The mutual information $I(X,\Omega)$ and entropy $H(P_\Omega)$ are computed using the estimates of $P(x)$, $P(\omega_i)$, $P(x, \omega_i)$. The method of Parzen windows which will be described is used for this purpose.

3) If $1 - I(X,)/H(P) < \varepsilon$ then the algorithm stops.

4) For every two temporary classes the loss of $I(X,\Omega)$ is computed if they are joined.

5) Such two classes for which is the loss minimal are found and joined. The current set of features cannot be probably used for discrimination of these classes. The algorithm goes back to the step 2).

Usually a number of well separable spectral cathegories is received. It is possible to compare the results with the real identifiers of target classes (considering the order in which temporary classes are joined) and to correct the unperfectly labeled training polygons.

3.2 Nonparametric supervised classification

The principles of Bayesian decision are frequently applied to the classification of remotely sensed multispectral data. The use of Bayesian strategy supposes the identification of the probability density function of each spectral class in order to determine the decision function that minimizes the probability of missclassification.

Classified pixels (vectors x) are assigned to one of the classes according to $P(\omega_j|x)$ - probability that class ω_j occurs at vector x. Using the Bayes theorem we find ω_i so that

$$P(\omega_i) \cdot P(x|\omega_i) = \max_j P(\omega_j) \cdot P(x|\omega_j) , \quad (8)$$

where $P(\omega_i)$ is an apriori probability of class ω_i (extent of the class in the image) and $P(x|\omega_i)$ is a conditional probability density. A set $\Gamma_j = \{x_1^j, \ldots, x_{n^j}^j\}$ of n_j observations of x for every class ω_j is available. Let $\Gamma = \cup \Gamma_j$ be the set of all training samples. To estimate the density $P(x|\omega_j)$ at random vector x the analyst can use the parametric or nonparametric methods.

Parametric classifiers are based on the assumption that all vectors come from a certain type of statistical distribution. The assumption of Gaussian distribution is very often used. Then the mean vectors and covariance matrices may be estimated from the sets Γ_j. The parametric methods are very time consuming for the application on large area. There are some improvements possible (Feiveson, 1983). But - what is more important the data do not fulfil the presumption of normality. The landuse classes have usually complex decision boundary, especially in high-dimensional feature space. However, the classifier decision is influenced most heavily by the samples close to the decision boundary.

That is why many authors suggest nonparametric density estimations (Skidmore, 1988), (Cervenka, 1990). Nonparametric classifiers make no assumption about the shape of the data distribution. Such techniques are based on the concept, that the value of density $P(x|\omega_j)$ at the point x can be estimated using the training samples located within a small region around the point x. The Parzen windows (Parzen, 1962) are very often used for the density estimations :

$$P(x|\omega_j) = 1/n_j \sum_{k=1}^{n_j} h_n^{-N} \cdot F((x - x_k^j)/h_n) , \quad (9)$$

$N = \dim(X)$,

where the function F(Y) is widely used in this functional form (so called uniform kernel):

$$F(Y/h_n) = \begin{cases} 2^{-N} & \text{if } |Y_1| / h_n \le 1 \quad 1 = 1,\ldots,N \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Usually, $h_{n_j} = n_j^{-C/N}$, where C will be within the interval (0,1). In fact, numbers of samples from Γ_j within a hypercube centered at x are computed in practical applications. Such a function can be evaluated easily - individual features can be tested one by one, and many samples can be eliminated quickly. Then the classification using (8) can be applied.

The nonparametric methods require a large number of computations. Common classification problems consist in the classification of millions vectors into 10 - 20 classes using 3 - 7 features. However, the decision of nonparametric classifiers is based on the small subregion of the total feature space. Several authors propose efficient solutions of this problem. Fukunaga (Fukunaga, 1975) suggests a decomposition of the original training set into hierarchically arranged subsets. Then the whole training set is represented by a tree structure, where the succeeded nodes create a decomposition of the preceding node. The root corresponds to the whole set Γ. The clustering method is used for the decomposition of the training samples. The cluster analysis is subsequently applied on the individual nodes. Following information is recorded for every node p: mean vector M_p of all samples in the node p (this set is denoted S_p), minimal and maximal values of individual features and the value r_p of maximum distance between M_p and $x_i \in S_p$. The distance of all samples to the corresponding sample mean are saved at the final tree level. The classification of any vector x corresponds to the tree search. All vectors sufficiently near to the vector x are sought. With the help of informations which are saved in the tree nodes most of the nodes can be eliminated. The given tree structure can be used for nonparametric classification methods.

When using the Parzen windows with uniform kernel, the test is performed at every tree level, if there is an intersection between the window and the parallepiped which contents all samples from the tested node. Minimal and maximal feature values in the node are exploited in such a case. These tests are repeated for succeeded nodes in an affirmative case only. At final level individual training samples are tested if they fall within the window centered at the classified sample. The features can be checked one by one. In most cases (when the training sample fall outside the window), it is not necessary to check all features. From this point of view it is advantageous to check the features with greater entropy at first.

3.3 Nearest neighbours method

The k nearest neighbours (k-NN) method is based on similar (local) principles as nonparametric Bayesian classifiers. They find k nearest neighbours to given sample x in the training set Γ. The sample x is assigned to the class ω_j, if the majority of its k - nearest neghbours belongs to that class ω_j. Ties may be broken arbitrarily using some heuristics. These classification methods have been proposed by many authors (Cover, 1967), (Tomek, 1976). One of the most important theoretical results is, that these methods have good behaviour in the asymptotical sense (Wilson, 1972). For large values of n_j the expected probability of error P^e is bounded as follows:

$$P^* \le P^e \le A \cdot P^* , \quad (11)$$

(A=2 for k=1).

P* is the Bayes probability of error, i. e. the probability of a possible error when the true apriori probabilities and density functions are known. If the value of k rises the coefficient A becomes much smaller. Wilson has shown, that editing of sample set improves the performance of nearest neighbours rule.

Now, the process of finding the nearest neighbour to the vector x will be described. The procedure can be simply extended for k - nearest neigbours method. The tree is searched through and tested, if the nearest neighbour could occur in any node or in its successors. The branch and bound algorithm is applied during this process. Two basic rules are used (B is the distance to the current nearest neighbour of x). If the inequality

$$B + r_p < d(x,M_p) \qquad (12)$$

(d(x,y) is a distance of vectors x, y)

holds, then (from the triangle inequality) follows that any nearest neighbour of x cannot be situated in the node p. The search for corresponding branch can be cut in such a case. Second rule concerns the individual samples x_i at final level of the tree. If the inequality

$$B + d(x_i,M_p) < d(x,M_p) \qquad (13)$$

holds, then the sample x_i cannot be the nearest neighbour of x. The distances $d(x_i,M_p)$ are saved during the creation of the tree, so that most of time consuming computation can be eliminated. During the search the distances $d(x,M_p)$ are computed and stored for the nodes p at every level. Then the inequality (12) is tested. If it does not hold the node p is placed into the list of active nodes at the relevant level. There is a possibility at present that this node contents the nearest neighbour of x. If the active list is not empty, the nearest node to x is chosen and the procedure is repeated for its successors. After the final level is reached, the nearest node q is chosen again. Then, all samples x_i in the node q are tested in accordance with the rule (13). If this inequality holds, the tested sample cannot be the nearest neghbour. Otherwise the distance $d(x,x_i)$ to the tested sample x_i is computed. If $d(x,x_i) < B$, the value of B is updated and the new nearest neighbour is saved.

When all samples in q are processed, remaining nodes in the active list at final level are searched (the test (12) is applied at first). If the search through all possible nodes at some level is finished, the procedure backtracks to the higher levels and tests remaining nodes in active lists (with updated - usually much smaller value of B).

3.4 Modifications of basic algorithms

This contribution suggests another improvement which speeds up the seeking of nearest neighbours as well as the nonparametric density estimates. At every tree level only one node (the nearest node to x) is chosen. The search is realized exclusively within the range of successors of the node having been chosen at preceding level. At the final level the nearest node q is determined (so called terminal node). The nearest neighbours to x or the samples within the Parzen window are found in the terminal node as well as in its neighbouring nodes. These nodes are determined in the course of creating the tree structure. All nodes at final tree level, to which the samples from q are classified using the nearest neighbour method, are included to the neighbourhood of node q. Thus, for every sample from the node q the nearest neighbour among other terminal nodes (being represented with their mean vectors) is found.

If the nearest neighbours to the sample x are found, the rules (12), (13) are applied during the search in the neighbourhood of node q. When using the nonparametric density estimates, the same tests, as used in basic algorithm, are performed. However, the volume of computation is much smaller.

This procedure uses certain heuristic. It does not guarantee, that all nearest neighbour of sample x or all training samples from corresponding Parzen window are precisely found. However, this procedure works successfully in the majority of reasonable cases.

3.5 Supervised classification using lookup table

A serious disadvantage of classification algorithm mentioned is the fact, that many samples with absolutely identical feature values are individually classified many times. This problem is solved by means of lookup tables where all previous classifications (the feature values and corresponding classification result) are recorded. A certain key is computed for every sample at first. This key defines a position in the lookup table. If this position is occupied, all feature values are compared. If there is an agreement in all dimensions, further computations are unnecessary. The sample is classified into the class having been found out in the table. If the value of any feature does not agree, various solutions may be used. If such a situation occurs rarely, it is possible to classify these samples in a usual way.

The choice of the key (hashing function) seems to be a critical problem. This function should be economic - a huge number of possible keys cannot be implemented easily on small computers. On the other hand, the set of possible keys cannot be too small, because of key collisions with different samples. An application of any orthogonal linear transformation (dimension reduction) seems to be a suitable solution. The method of principal component (Karhunen - Loeve's transformation) or "Tasseled Cap" transformation preserves a great deal of data variability. When using three principal components (the intrinsic dimensionality of TM data is aproximately three (Crist, 1984b)), it is easy to assign the key to every sample. The sample has the same position as in the three-dimensional array where coordinates correspond to principal components. The computation of the principal components and table position can be combined effectively. Considering the TM data the range of feature values in every principal component is over one hundred. The lookup table must have 1.5 - 2 millions of entries from this reason.

But, even the reduced feature space is not filled completely. There are some regions where only few samples are located. However, there is a method enabling to reduce the storage requirements which uses the proposed hierarchic decomposition of a training set. A certain part of the lookup table is assigned to every terminal node q of the tree. It

is a space saving information about some small area of feature space which cover the sample set S_q. Every part of the table corresponds to the three-dimensional array again. But the range in every dimension is much smaller. During the classification of certain sample the appropriate terminal node is found and the key is computed. But this key serves as an entry to the "local" lookup table. It is necessary to estimate the average space needed for every part of main lookup table. This space can be estimated after the creation of a tree structure. All training samples from the corresponding terminal node are used and the range in every principal component is determined.

4. CONCLUSION

The methods described have been succesfully applied at the Earth Remote Sensing Centre (Institut of Surveying and Mapping) in Prague, especially for the classification of TM data. Several thematic maps from the Northeast Bohemia and Prague region have been produced.

The CPU times requirements depend on the number of spectral classes as well as on the number of the training samples. In fact, the performance of the proposed algorithms depends upon the spatial configuration of the data set in the feature space. The influence of the number of classified pixels can be substantially reduced when using the method described in 3.5.

The average number of pixels was approximately 400 for every landuse cathegory in Prague region. The number of spectral cathegories was 12. Then the time requirements on IBM PC 386 personal computer for 10^6 pixels when using the 6 TM bands (the thermal band has not been included) were:

- nearest neighbour (1-NN) classification - 15 min.
- 5 - nearest neighbours classification - 60 min. (both nearest neigbours classifiers were implemented according to the part 3.4)

- Bayesian nonparametric classification (basic algorithm) - 88 min.
- Bayesian nonparametric classification (algorithm according to the part 3.4.) - 51 min.
- Bayesian nonparametric classification using lookup table - 46 min.

The probability estimates of correct classification were similar for all methods (using the resubstitution method) and exceed 90 %.

In certain situations it is desirable not to classify samples which cannot be assigned with sufficient certainty. Such samples are marked as nonclassified. This "reject option" can be implemented with the classifiers using the nonparametric probability estimates easily. If the extent of nonclassified areas is too large, it is necessary to complete the training sets and to repeate the classification process.

5. REFERENCES

Cervenka, V., Charvat, K., 1990. Digital processing of pictorial image data in remote sensing (in Czech). Tech. Rep. No 33/1990, Geodetic and cartographic enterprise, Prague - Czechoslovakia.

Charvat, K., Cervenka, V., 1987a. The Development of Classification and Segmentation System (in Czech). In: Digital image processing '87, CSVTS TESLA A.S.Popova, Prague - Czechoslovakia.

Charvat, K., Cervenka, V., Soukup, P., 1987b. Using Statistical Tests for Computation of the Classification Parameters in Remote Sensing of Earth (in Czech). In: Application of Artificial Intelligence AI '87, ÚISK, Prague - Czechoslovakia.

Charvat, K., Cervenka, V. 1990. Pseudocolor photomaps production using neural networks. In: International symposium on thematic mapping from satellite imagery., Paris - France, Bulletin of French Committee for Cartography, No 127 - 128, pp. 75-78.

Cover, T. M., Hart, P. E., 1967. Nearest Neighbour Pattern Classification. IEEE Transactions on Information Theory, Vol. IT-13, January, pp. 21-27.

Crist, F. P., Cicone, R. C., 1984a. A Physically Based Transformation of TM Data - the Tasseled Cap. IEEE Transaction on Geoscience and Remote Sensing, No.3.

Crist, F. P., Cicone, R. C., 1984b. Comparisons of the Dimensionality and Features on Simulated LANDSAT 4 MSS and TM Data. Remote Sensing of Environment, Vol 15.

Fahlman, S. E., 1988. An Empirical Study of Learning Speed in Back Propagation Networks. tech. Rep. CMU-CS-88-162.

Feiveson, A. H., 1983. Classification by thresholding. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5, January, pp. 48-53.

Fukunaga, K., Narendra, P.M., 1975. A Branch and Bound Algorithm for Computing k-Nearest Neighbours. IEEE Transactions, Vol. C-24, July, pp. 751 - 753.

Hinton, G. E., 1987. Connectionist Learning Procedures, Tech. Rep., CMU-CS-87-195.

Parzen, E., 1962. on Estimation of Probability Density Function and Mode. Ann. Math. Statist., Vol. 33, pp. 1065-1076.

Skidmore, A. K., Turner, B. J. 1988. Forest Mapping Accuracies Are Improved Using a Supervised Nonparametric Classifier with SPOT Data. Photogrammetric Engineering and Remote Sensing, Vol. 54, October, pp. 1415-1421.

Tomek, I., 1976. A Generalization of the k-NN Rule. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-6, February, pp. 121-126.

Wilson, D., 1972. Asymptotic Properties of Nearest Neighbour Rules Using Edited Data. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-2, pp. 408 - 421.