Multisource Data Classification With Dependence Trees

Mihai Datcu, Farid Melgani, Andrea Piardi, and Sebastiano B. Serpico, *Senior Member, IEEE*

***Abstract*—In order to apply a statistical approach to the classification of multisource remote-sensing data, one of the main problems to face lies in the estimation of probability distribution functions. This problem arises out of the difficulty of defining a common statistical model for such heterogeneous data. A possible solution is to adopt nonparametric approaches, which rely on the availability of training samples without any assumption about the related statistical distributions. The purpose of this paper is to investigate the suitability of the concept of dependence trees for the integration of multisource information through estimation of probability distributions. First, this concept, introduced by Chow and Liu, is used to provide an approximation of a probability distribution defined in an N-dimensional space by a product of $N-1$ probability distributions defined in two-dimensional (2-D) spaces; this approximation corresponds, in terms of graph theoretical interpretation, to a tree of dependence. For each land cover class, a dependence tree is generated by minimizing an appropriate closeness measure. Then, a nonparametric estimation of the second-order probability distributions is carried out through the Parzen window approach, based on the implementation of 2-D Gaussian kernels. In this way, it is possible to reduce the complexity of the estimation, while capturing a significant part of the interdependence among variables. A comparison with other multisource data fusion methods, namely, the multilayer perceptron (MLP) method, the k-nearest neighbor (k-NN) method, and a Bayesian hierarchical classifier (BHC), is made. Experimental results obtained on multisensor [airborne thematic mapper (ATM) and synthetic aperture radar (SAR)] and multisource (experimental synthetic aperture radar (E-SAR) and a textural feature) data sets show that the proposed fusion method based on dependence trees is able to provide a classification accuracy similar to those of the other methods considered, but with the advantage of a reduced computational load.**

***Index Terms*—Dependence trees, discrete probability distribution approximation, multisource fusion, Parzen window approach, remote sensing image classification.**

I. Introduction

In the classification scheme of a remote-sensing image, an important step is to translate the observations concerning a study area into an adequate mathematical model. One of the most usual approaches to characterizing a source of observations is to develop statistical models through the estimation of probability distributions (for discrete observations) or probability density functions (pdfs) (for continuous observations). However, when one wants to apply this statistical approach to the classification of multisource remote-sensing data, problems arise out of the difficulty of defining a common statistical model for such heterogeneous data. A possible solution is to adopt a nonparametric approach, which does not rely on a specific distributional assumption and allows the kind of distribution to be derived entirely from data.

Histogram methods can be regarded as the earliest nonparametric estimators of a density function. On the basis of a partition of the feature space into cells of fixed or variable size, histogram methods involve the problem of their application to high-dimensional spaces, where the scarcity of available data makes it likely to find empty cells [1]. Finding the k-nearest neighbors (k-NNs) represents another way to compute a density function estimate. It provides a simple estimation of the density, yielded by the volume of the hypersphere containing the k-NNs and centered on the point where the density function has to be computed [2]. However, this method involves many distance computations and requires the application of sophisticated search algorithms (e.g., the Kd-Tree [3]) to speed up the search process. Another more recent nonparametric approach consists in the modeling of the class distributions by means of artificial neural networks, among which the multilayer perceptron (MLP) can be regarded as the best known. It has been proven that an MLP with as few as one hidden layer of neurons and a specific type of activation functions (including the sigmoidal function and the hyperbolic tangent) can approximate any functional relation arbitrarily well, provided that enough hidden neurons are available [4]. In particular, their powerful capability to model class posterior probabilities makes them an interesting solution to the classification problem [5], [6]. Their main drawback is the lack of a general criterion to determine the optimal architecture and training parameters, in addition to long training times. Another interesting nonparametric approach consists in deriving the density function from the superposition of kernel functions centered on data samples and acting as smoothing operators [7], [8]. Gaussian, triangular, and rectangular kernels represent typical examples of kernel functions. Despite the computational load they involve, kernel estimators offer rather a powerful tool for approximating the true density. In addition to the above nonparametric approaches, there exist many others, like expansion by basis functions, adaptive kernel estimators and projection pursuit density estimation. More detailed descriptions can be found in [9] and [10].

In the context of remote-sensing data from different sources, one should take into account that the larger the number of features, the larger the amount of training samples required to avoid the well-known Hughes effect [11], which would otherwise involve a limitation on the classification accuracy. Unfortunately,

Manuscript received May 18, 2001; revised October 12, 2001.

M. Datcu is with the German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen, D-82234 Weßling, Germany (e-mail: mihai.datcu@dlr.de).

F. Melgani, A. Piardi, and S. B. Serpico are with the Department of Biophysical and Electronic Engineering, University of Genoa, I-16145 Genova, Italy (e-mail: vulcano@dibe.unige.it).

Publisher Item Identifier S 0196-2892(02)03268-0.

the availability of training samples is usually very limited, due to the fact that the collection of the ground truth is a difficult and expensive task. One possible solution is to express the true class-conditional distributions through class-conditional distributions of reduced dimensions, for which the problem of few training samples is less critical.

The problem of approximating an Nth order binary distribution by a product of several of its lower-order component distributions was considered by Lewis [12] and Brown [13]. On the one hand, it is evident that component distributions of increasingly high order allow one to capture much more information about the stochastic dependence among the features and, consequently, to get a better estimate. On the other hand, practical considerations push to use low-order component distributions to estimate the class-conditional distributions. A first answer was given by Chow and Liu [14]: they proposed a class of product approximations derived from the "chain rule," which approximate the N th-order distribution as a product of $(N-1)$ second-order conditional distributions. It is shown that this approximation can be related to the so-called "dependence tree," whose branches represent the relationships among the nodes (features). Ku and Kullback [15] presented an iterative approach, in which the restriction to a tree-type dependence model is discarded to get closer product approximations by using component distributions whose order may be higher than two. This approach proposes a general estimation procedure, for which the class of second-order product approximations introduced by Chow and Liu is a particular case. It can yield to better results since it handles higher-order product approximations that, during the iterative computation, may exploit lower-order ones.

As only approximations can be obtained, it is necessary to have a tool capable to quantify the goodness of an approximation by adopting a measure of distance between the true and estimated probability distributions. One such measure is the Kullback–Leibler divergence measure [16], appreciated for its simplicity when applied to product expansions. In order to minimize this theoretical information measure, the well-known mutual information measure (MIM) was used by Chow and Liu [14] as a practical tool for measuring the degree of dependence between all pairs of features and, consequently, for selecting the most dominant dependences, which determine the dependence tree.

In this paper, the suitability of the concept of dependence tree for the integration of multisource remote sensing data through estimation of probability distributions is investigated. In particular, the technique described in the following represents a tool for reducing the complexity of the original class-conditional distributions to a combination of second-order class-conditional distributions, for which the problem of the small number of training samples is less crucial. Then, a nonparametric method based on Gaussian kernels is used to estimate the second-order class conditional distributions. In the experiments involving multisensor (ATM + SAR) and multisource data sets, it has been found that the fusion process based on the concept of dependence tree is able to provide similar classification accuracies but with the advantage of a reduced computational load, as compared with other classifiers, like the MLP, the k-NN classifier, and the Bayesian hierarchical classifier (BHC) [17].

II. Approximating Probability Distributions With Dependence Trees

A. Problem Formulation

We consider a multisource data set consisting of images showing the same ground area and taken by S sensors. Let us assume that each image contains N_s channels (with $s = 1, \ldots, S$) and that each related pixel is represented by the discrete feature vector X_s. We regard the global N-dimensional feature vector X as the concatenation of the feature vectors X_s (each of dimension N_s) provided by the S sensors. It is assumed that the scene includes C thematic classes represented by the set of labels $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$, with which each image pixel should be associated. Let us adopt a pixel-based (noncontextual) classification scheme by applying the Bayes rule for minimum error (BRME), which assigns the optimal label $\omega^* \epsilon\, \Omega$ to X, such that

$$P(\omega^*|X) = \underset{\omega_k \in \Omega}{\text{Max}} \{P(\omega_k|X)\} \tag{1}$$

where the posterior probability $P(\omega_k|X)$ can be expressed as a function of the class-conditional distributions by means of the Bayes theorem, that is

$$P(\omega_k|X) = \frac{P(X|\omega_k) \cdot P(\omega_k)}{\sum_{i=1}^{C} P(X|\omega_i) \cdot P(\omega_i)} \tag{2}$$

and $P(\omega_k)$ represents the prior probability of the class ω_k. In order to apply the BMRE, the problem becomes one of estimating the class-conditional distributions (also known as class-conditional probability mass functions) $P(X|\omega_i)$ for all classes ω_i in Ω.

B. A Particular Class of Product Approximations

The "chain rule" [18] allows one to expand the class-conditional distribution $P(X|\omega_k)$ as a product of conditional distributions of increasing order, such as in the following:

$$\begin{aligned} P(X|\omega_k) &= P(x_1|\omega_k) \cdot P(x_2|x_1, \omega_k) \\ &\quad \cdot P(x_3|x_1, x_2, \omega_k) \ldots . P(x_N|x_1, x_2, \ldots x_{N-1}, \omega_k) \end{aligned} \tag{3}$$

where $X = [x_1, x_2, \ldots x_N]$. In this expansion, the feature x_1 is called the "root of the chain." Obviously, (3) can be reformulated by considering any other feature x_i as the root. The basic idea of the approximation based on the dependence tree is to reduce the conditioning on multiple features to a conditioning on only one feature of the form

$$P(X|\omega_k) \cong \hat{P}(X|\omega_k) = \prod_{i=1}^{N} P\left(x_{mi}|x_{mj(i)}, \omega_k\right) \tag{4}$$

where $\{m_1, m_2, \ldots, m_N\}$ is an unknown permutation of the integers $\{1, 2, \ldots, N\}$ and $j(i)$ represents the corresponding mapping such that $0 \leq j(i) < i$. Equation (4) is the translation of a particular class of product approximations of (3) in which only second-order probabilistic relationships are used to point out the interfeature dependence.

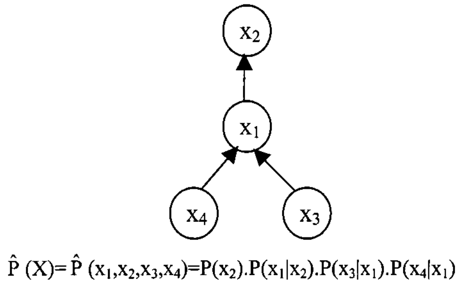


Fig. 1. Example of dependence tree for a four-dimensional distribution.

C. Notion of Dependence Tree

If we reason in terms of graph-theoretical interpretation, the unknown mapping $j(i)$ makes up a graph whose nodes are the features $x_n (n = 1, 2, \ldots, N)$, the edges represent the pairs of probabilistic relationships $(x_{mi}, x_{mj(i)})$, and the root node guides the edge orientation. This graph is called the first-order dependence tree of the distribution [14]. Fig. 1 illustrates an example of the (first-order) dependence tree. Note that the root is the only feature x_r for which $j(r) = 0$. For the sake of simplicity, the notation x_{mi} is dropped and replaced with the notation x_i. The product approximation expressed in (4) becomes

$$\hat{P}(X|\omega_k) = \prod_{i=1}^{N} P\left(x_i | x_{j(i)}, \omega_k\right). \tag{5}$$

D. Dependence Tree Optimality

As (5) is only an approximation, it involves a loss of information, as compared with the true class-conditional distribution (3). The problem becomes one of minimizing this information loss, i.e., of finding the optimal approximation [in the context of the class of approximations defined by (5)] for which the distance between the true and approximate class-conditional distributions is the minimum one. As mentioned above, one of the criteria used to quantify the closeness of the approximation is the Kullback–Leibler divergence measure, defined as follows:

$$I\left(P, \hat{P}\right) = \sum_{X} P(X|\omega_k) \cdot \log\left(\frac{P(X|\omega_k)}{\hat{P}(X|\omega_k)}\right). \tag{6}$$

$I(P, \hat{P})$ can be seen as the difference between the information contained in $P(X|\omega_k)$ and that contained in $\hat{P}(X|\omega_k)$ about $P(X|\omega_k)$. It is characterized by the following property:

$$I\left(P, \hat{P}\right) \geq 0 \tag{7}$$

and is equal to zero only if $\hat{P}(X|\omega_k)$ and $P(X|\omega_k)$ are identical. Finding the optimal approximation $\hat{P}^*$ is equivalent to finding the optimal dependence tree, or the optimal mapping $j^*()$ over the set M of all possible mappings $j()$, such that

$$I\left(P, \hat{P}_{j^*}\right) = \underset{j \in M}{\text{Min}} \left\{ I\left(P, \hat{P}_j\right) \right\} \tag{8}$$

where $\hat{P}_j$ stands for the approximate class-conditional distribution defined by the mapping $j()$. In the case where the approximation model is a dependence tree, it can be shown [14] that the measure of closeness defined in (6) can be expressed as a function of the MIMs among the features (nodes) of the dependence tree, that is

$$I\left(P, \hat{P}\right) = Q - \sum_{i=1}^{N} \text{MIM}\left(x_i, x_{j(i)} | \omega_k\right) \tag{9}$$

where Q is a constant independent of the approximation. The MIM plays the role of weight of each branch $(x_i, x_{j(i)})$ of the tree and is given by

$$\text{MIM}\left(x_i, x_{j(i)} | \omega_k\right) = \sum_{x_i, x_{j(i)}} P\left(x_i, x_{j(i)} | \omega_k\right) \cdot \log\left(\frac{P\left(x_i, x_{j(i)} | \omega_k\right)}{P(x_i|\omega_k) \cdot P\left(x_{j(i)} | \omega_k\right)}\right). \tag{10}$$

The problem of minimizing the Kullback-Leibler divergence measure (6) is transformed into the problem of finding a dependence tree with a maximum total branch weight or, in other words, the optimal mapping $j^*()$ such that

$$\sum_{i=1}^{N} \text{MIM}\left(x_i, x_{j^*(i)} | \omega_k\right) = \underset{j \in M}{\text{Max}} \left\{ \sum_{i=1}^{N} \text{MIM}\left(x_i, x_{j(i)} | \omega_k\right) \right\}. \tag{11}$$

E. Optimal Dependence Tree Construction

The algorithm introduced by Kruskal [19] can be used to construct the tree with a maximum total branch weight, branch by branch, thanks to the additive property of the branch weights. First, the weights of all possible branches are computed, that is, since the MIM is symmetric, $N(N-1)/2$ branches from which only $(N-1)$ will be selected to build the optimal dependence tree. The procedure for selecting the best branches (i.e., those capturing the most important probabilistic interdependence among the features) starts by sorting, in decreasing order, the branch weights MIM$(x_u, x_v | \omega_k)$, with $u = 1, 2, \ldots, N$ and $u < v \leq N$. The first branch chosen by the branch-selection process is that for which the weight is the maximum one. Then, the list is scanned and a branch is added if it exhibits the acyclicity property and shows a link with one of the previously selected branches. In this case, it is removed from the sorted list and, in the next step, the search restarts from the beginning of the list. Otherwise, the branch is rejected and one continues the search by considering the next branch on the list. This process goes on until it gathers the $N-1$ branches required by the dependence tree to approximate the N-dimensional class-conditional distribution. It is worth recalling that, as each dependence tree, in the present paper, is constructed to approximate a distribution conditioned by a particular class, the above procedure must be applied several times to obtain a number of dependence trees equal to that of the classes available.

III. Second-Order Distribution Estimation

A. Use of Second-Order Class-Conditional Relative Frequencies

The previously discussed approximation for an N-dimensional class-conditional distribution involves second-order probabilistic relationships still difficult to estimate. Furthermore, as seen above, the search for the optimal dependence tree requires the estimation of $N(N-1)/2$ second-order class-conditional distributions $P(x_u, x_v|\omega_k)$ in order to compute all possible branch weights MIM$(x_u, x_v|\omega_k)$ $(u = 1, 2, \ldots, N$ and $u < v \leq N)$. The application of a nonparametric estimation method at this stage would be very expensive from a computational point of view. As a tradeoff, one can approximate the second-order class-conditional distributions by using the simple second-order class-conditional relative frequencies derived from the training samples, as follows:

$$\hat{P}(x_u = \alpha, x_v = \beta|\omega_k) = \frac{F(x_u = \alpha, x_v = \beta|\omega_k)}{T_k} \tag{12}$$

where $F(x_u = \alpha, x_v = \beta|\omega_k)$ denotes the number of training samples belonging to the class ω_k whose features are equal to α in the uth channel and to β in the vth channel. T_k is the total number of training samples belonging to the class ω_k. It is evident that the simplicity of the estimation based on the relative frequencies allows one to decrease the computational burden of the generation of the optimal dependence tree. Once this stage has been completed, one can perform a more accurate, even though more complex, estimation of the $(N-1)$ second-order class-conditional distributions related to the product approximation in (5). An interesting solution is to adopt the powerful nonparametric estimation method based on Gaussian kernels.

B. Estimation Based on Gaussian Kernels

The density estimation based on kernel functions (also called the Parzen window approach) is a well-known nonparametric approach that has been shown to be able to provide an asymptotic, unbiased, consistent estimate of the true distribution [9]. The density estimate is obtained by the superposition of kernel functions K, each centered on one of the available samples drawn from the true density [9]. In our case, we apply this approach to estimate two-dimensional (2-D) class-conditional distributions by adopting kernel functions of the Gaussian type, as defined by the following equation:

$$\hat{P}(x_u, x_v|\omega_k) = \frac{1}{T_k} \cdot \sum_{t=1}^{T_k} \frac{1}{2\pi h^2} \exp\left(-\frac{\left(x_u - x_u^{t,k}\right)^2 + \left(x_v - x_v^{t,k}\right)^2}{2h^2}\right) \tag{13}$$

where T_k stands for the total number of training samples associated with the class ω_k; $x_u^{t,k}$ and $x_v^{t,k}$ refer to the uth and vth features of the tth training sample of the class ω_k, respectively; and h represents the width (standard deviation) of the Gaussian kernel. The performances of kernel estimators depend strongly on the width parameter h, which plays the role of a smoothing parameter [20]. If h is too large, the density approximation will be affected by the problem of over-smoothing (too low resolution), whereas, if h is too small, the density will exhibit spurious discontinuities. In our case, as the optimization of the width parameter h is not the focus of the present work, its value will be obtained by a simple validation approach based on the subdivision of the labeled samples into two sets, namely, the training and validation sets. The best width parameter is chosen to optimize the overall classification accuracy on the validation set.

IV. Experimental Results

Two experiments are described in order to assess the accuracy and performance of the classifier based on dependence trees, in comparison with other classifiers. In Section IV-A, a multisensor (radar and optical sensors) remote-sensing data set is considered; a comparison with the k-NN and MLP neural network approaches is made. In Section IV-B, the second experiment refers to a multisource data set including information of the textural type. The dependence tree classifier (DTC) is compared in terms of classification accuracy with another powerful classifier, namely, the BHC described in [17]. In both cases, in order to assess the performances of the different classifiers, two accuracy measures are utilized: the overall accuracy (OA), that is, the percentage of correctly classified pixels among all the considered pixels (independently of the classes to which they belong) and the average accuracy (AA), that is, the average over the classification accuracies obtained for the different classes.

A. Multisensor Data Set

The multisensor data set, on which the proposed method was tested, was extracted from two images showing the same ground area. The first image was taken in July 1989 by a Daedalus 1268 airborne thematic mapper (ATM) scanner of which six channels were considered, that is, those corresponding to all the channels of the Landsat TM sensor, except for the thermal infrared channel. The second image was acquired in August 1989 PLC-band, fully polarimetric, NASA/JPL synthetic aperture radar (SAR) airborne sensor, which yielded nine different channels corresponding to all possible combinations of bands (P, L, or C) and polarizations (HH, HV, or VV). The selected sections of the ATM and SAR images, separate in time by only a few days, were registered by using the SAR image as the reference image (see Fig. 2). The size of such sections was 250 pixels × 350 pixels; the portion of the scene they represented corresponded to an agricultural area near Feltwell, U.K., in which five land cover types are dominant, namely, sugar beets, stubble, bare soil, potatoes, and carrots. Table I gives the numbers of training and validation samples used to train the classifiers and assess their accuracy values, respectively.

MLP and k-NN Classifiers: The results obtained by two other nonparametric data classification methods and published in [21] are used as reference values to evaluate the performances of the DTC. The MLP classifier used for this experiment was a feedforward neural network with one hidden layer with eight nodes, whereas the input and output layers consisted of 15 nodes and five nodes, respectively. The MLP was trained by the error backpropagation learning procedure at a learning rate $\eta = 0.01$. Concerning the k-NN classifier, results were

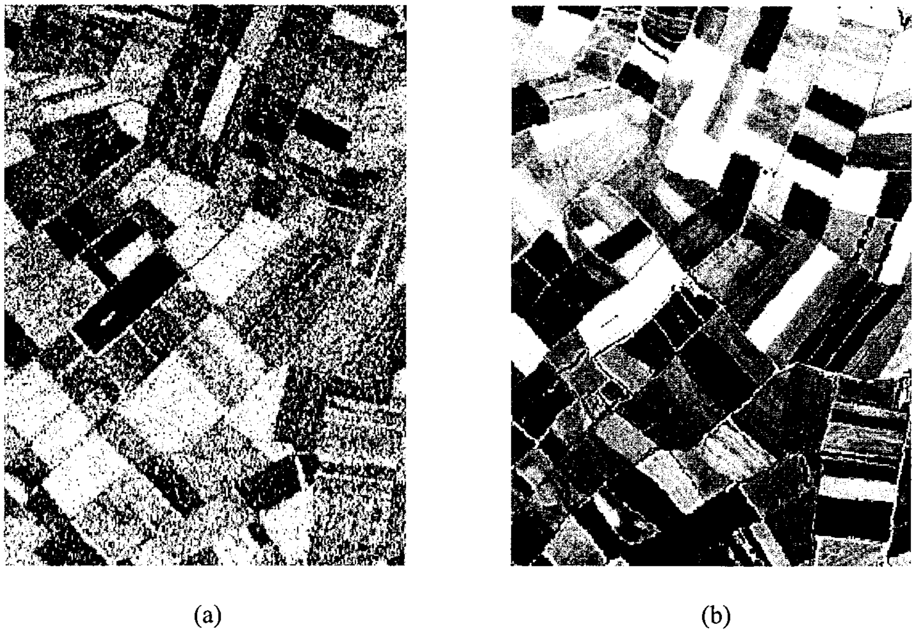


Fig. 2. View of the section of the multisensor data set utilized for the experiments: (a) SAR image (band L, polarization HV) and (b) ATM image (band 9).

TABLE I
NUMBERS OF TRAINING AND VALIDATION SAMPLES FOR THE MULTISENSOR DATA SET

CLASS	TRAINING	VALIDATION
Sugar beets	1488	2043
Stubble	1070	1371
Bare soil	341	555
Potatoes	1411	884
Carrots	814	967
TOTAL	5124	5820

TABLE II
BRANCH WEIGHT VALUES OF THE OPTIMAL DEPENDENCE TREE FOUND FOR THE CLASS "CARROTS" (CORRESPONDING TO ω_5)

Feature n	Feature m	MIM(x_n,x_m/ω_5)
7	8	3.16
8	9	3.09
8	10	2.94
8	11	2.83
8	12	2.58
8	14	2.44
8	4	2.30
8	13	2.13
8	15	2.02
4	6	1.77
6	5	1.76
5	1	1.58
5	2	1.51
2	3	1.64

obtained for a value of $K = 25$ and are shown in Table IV, together with those of the MLP classifier.

Results: For each class, a tree of dependence among the different random variables representing the 15 multisensor features (six from the ATM sensor and nine from the SAR sensor) had to be built, based on the available training samples. The first step of the process was the computation of the upper-right part of the matrix of MIMs MIM(x_u, x_v), with $u = 1, \ldots, 15$ and $u < v \leq 15$. After sorting the MIM(x_u, x_v) in decreasing order, the branch-selection procedure described in Section II was run. Table II shows, as an example, the MIM values of the branches characterizing the best dependence tree found for the class "carrots," illustrated in Fig. 3. Different MIM values and hence different dependence trees were found for the remaining classes. In order to apply the BRME, the prior probabilities of the classes were estimated from the training sets. The first classification experiment was carried out by estimating the members of the distribution product approximation through the use of the simple second-order class-conditional relative frequencies. As

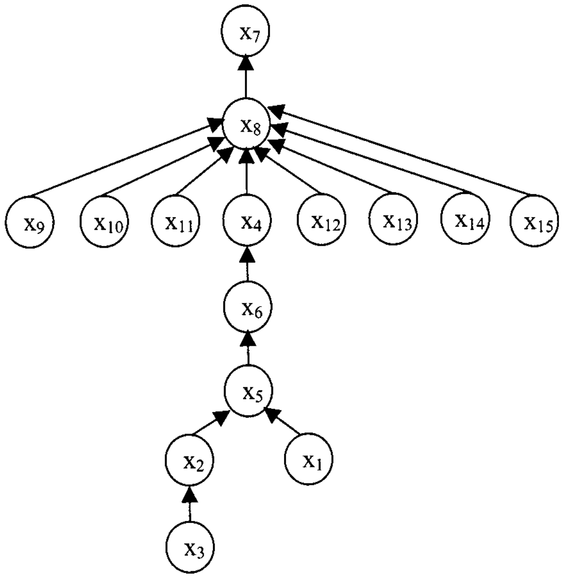


Fig. 3. Optimal dependence tree found for the class "carrots."

TABLE III
OVERALL ACCURACY VALUES OBTAINED FOR THE VALIDATION SAMPLES FOR DIFFERENT VARIANCE VALUES

Variance	OA [%]
50	85.0
150	87.8
300	88.7
350	88.8
400	88.5
700	87.7
1200	86.6

TABLE IV
COMPARISON OF THE ACCURACY VALUES YIELDED BY DIFFERENT CLASSIFIERS FOR THE MULTISENSOR DATA SET (OA: OVERALL ACCURACY; AA: AVERAGE ACCURACY)

Method	OA [%]	AA [%]	Sugar beets [%]	Stubble [%]	Bare soil [%]	Potatoes [%]	Carrots [%]
DTC	88.8	87.0	97.0	88.8	79.6	78.5	91.3
MLP	89.6	86.4	98.0	84.6	80.9	80.1	88.2
K-nn	89.8	87.0	97.4	88.4	76.0	86.4	87.1

expected, the results were very poor, with an OA of 16.6%. By contrast, the use of the nonparametric Gaussian kernel estimators turned out to be effective and allowed the results to be considerably improved. Several Gaussian kernels were generated by setting the kernel width parameter h (standard deviation) to different values. The overall classification accuracy values obtained for the validation samples by using Gaussian kernel estimators are provided in Table III. When the variance was increased from 50 to 350, the OA increased from 85.0% to 88.8%. From an empirical point of view, this value seems to correspond to the optimal kernel width ($h = 18.7$), since, for larger values of the variance (up to 1200), the OA decreases from 88.8% to 86.6%. This is due to the over-smoothing effect, which tends to increase the overlap among the densities and, consequently, the confusion among the classes. Finally, from Table IV, one can observe that the results obtained by the DTC, MLP and k-NN classifiers, are very similar, with OA values equal to 88.8%, 89.6%, and 89.8%, respectively, and with AA values equal to 87.0%, 86.4%, and 87.0%, respectively.

The analysis of the above results suggests some considerations. In our opinion, the fact that the results of the different classifiers are so similar to one another may mean that all the considered classifiers were able to fully exploit the information available from the training set. Therefore, they all arrived close to the best accuracy that can be reached by using such information, given the variability of the classes and the degree of overlapping among them. An explanation for the fact that the accuracies of the considered classifiers are not very high (only the accuracy of the class "sugar beets" is close to 100%) may be that each kind of growing appears in several agricultural fields, often with different appearances (the ground truth utilized for this data set is shown in [6]). The training samples were taken only from some fields; therefore, they are not completely representative of the statistics of the classes. In addition, a high percentage of boundary pixels were misclassified due to the presence of spurious ground coverings (trees, lanes, etc.) between neighboring fields. In the ground truth, such pixels were assigned to the class of one of the neighboring fields, whereas they should be regarded as mixed pixels or even as belonging to other classes, different from the agricultural classes considered.

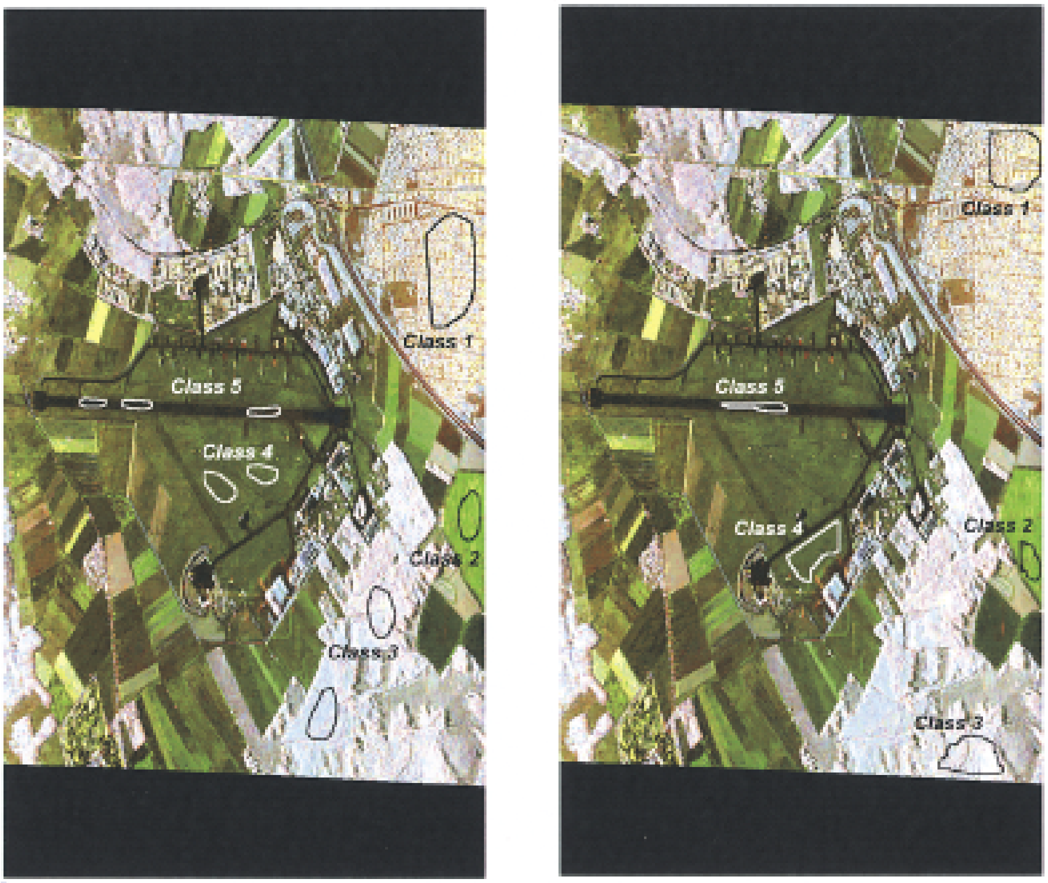
B. Multisource Data Set

The DTC and the BHC were compared in a multisource classification problem. The latter classifier is based on a hierarchical organization of extracted information. The different levels of the hierarchy are arranged according to their different degree of semantic abstraction. Elements at each level are obtained by a step of Bayesian inference from one or more lower levels. The process of information extraction is split into two parts: 1) a robust, signal-oriented description of the content of an image (which requires a significant computational load) and 2) a fast user-oriented labeling of the extracted image information content [22], [23]. The signal-oriented description is thought to have already been generated once and for all when the image is archived. Therefore, the related computation can be considered to have been performed "off-line."

The multisource data set, consisting of both radiometric and textural information, was acquired over the area of the German Aerospace Center (DLR, Oberpfaffenhofen, Germany) by the DLR airborne experimental synthetic aperture radar (E-SAR). An L-band fully polarimetric image and a texture "virtual" channel (extracted as the norm of textural features computed by the model-based algorithm described in [24]) were used for the experiments. Note that, given the difficulty with the statistical modeling of the distribution of textural information, the presence of this source highlights the usefulness of nonparametric methods, like the proposed one.

Results: Some objects and some cover types were selected as classes, like the airport runway, the city of Neughilching, the forest and some grass areas (see Fig. 4). The overall and average accuracies achieved by the two classifiers were quite similar (see Table V). The most significant difference can be observed for the class "forest," which was better recognized by the DTC (99.9% accuracy) than by the BHC (93.6% accuracy). In particular, from the confusion matrix (not shown in the paper for the sake of brevity) one can deduce that the BHC suffers more than the DTC from the overlapping between the distributions of the classes "forest" and "city."

The better results obtained by this experiment, as compared with the previous one, are mainly associated with the nature of the data set, the feature extraction and the selection of the training regions; these factors led to very accurate class models and, consequently, to very good classification results. The E-SAR data set was acquired by a "single sensor" thus the time change and the image registration did not affect the class models. The feature extraction from the E-SAR data set was performed according to [24], thus very much reducing the effect of the speckle process. Each polarimetric channel was "despeckled" to achieve uniform backscatter regions; as a consequence, a very smooth signal was obtained. Finally, the training and validation areas were selected as internal regions of single ground structures, thereby minimizing the risk of including mixed pixels.



(a) (b)

Fig. 4. Color composite of the image of the study area acquired by the L-band fully polarimetric E-SAR sensor: (a) the training area and (b) the validation area are highlighted. The considered classes are: city (class 1), grass 1 (class 2), forest (class 3), grass 2 (class 4), and airport runway (class 5).

TABLE V
COMPARISON OF THE ACCURACY VALUES ACHIEVED BY THE DTC AND BHC CLASSIFIERS FOR THE MULTISOURCE DATA SET (OA: OVERALL ACCURACY; AA: AVERAGE ACCURACY)

Method	OA [%]	AA [%]	City [%]	Grass 1 [%]	Forest [%]	Grass 2 [%]	Airport Runway [%]
DTC	99.94	99.93	99.99	99.86	99.91	99.89	100
BHC	97.53	98.20	98.21	99.50	93.62	99.65	100

TABLE VI
COMPUTATIONAL TIME REQUIRED FOR THE CLASSIFICATION OF THE MULTISENSOR DATA SET (DESCRIBED IN SECTION IV-A) BY USING A PERSONAL COMPUTER WITH A PENTIUM II PROCESSOR

Processing Step	Computational time [s]
MIM estimation and optimum DPT construction	26.2
Second-order distribution estimation	56.9
Classification	4.2

V. CONCLUSIONS

By capturing the most significant second-order probabilistic relationships, the dependence tree approach allows one to approximate the N-dimensional class-conditional distributions (representing the multisource class distributions) by products of second-order class-conditional distributions. Then the latter distributions can be estimated by use of the nonparametric Parzen window approach based on Gaussian kernels. The similar classification accuracies obtained in the experiments confirm that the DTC may compete with other powerful classifiers considered in this paper, namely, the k-NNs, the MLP, and the BHCs.

Concerning the computational aspect, we note that the DTC is fast in both the training and classification phases. Table VI gives the processing times for the experiments with the multisensor data set. In this case, the training time was about 1 min and 23 s. For the classification phase, as the processing time grows linearly with the image size, the time required to classify a typical image of 1024 pixels × 1024 pixels is about 50 s. Concerning the MLP classifier, in addition to the problems of defining its architecture and finding effective training parameters, this classifier suffers from the long training time required to optimize the weights of the neural network. The k-NN classifier requires a simple training step (to optimize the number of neighbors to be considered); however, in the classification phase, it involves the computation of a large number of distances in order to find, for each sample, the k nearest training samples. The DTC and the BHC are similar for they model the feature space by Gaussian kernels (explicitly by the DTC and implicitly by the BHC). The BHC was designed for a server-client architecture, having as a goal the computation speed in the interactive phase; thus, part of the computations is performed off-line and is "invisible" to the user. This classifier is faster than the DTC, if one considers only the on-line computations in the interactive phase; by contrast, it is much slower if all the processing involved is taken into account.

It is worth noting that the DTC has also the advantage of making the class models explicit as regards the role of the features and their relationships; therefore, other types of stochastic inference are possible (e.g., for data-mining purposes).

In conclusion, in this paper, we have proven the effectiveness of the dependence-tree method in both probability distribution modeling and classification. In our opinion, this method has been underestimated not only in the remote-sensing literature but also in the pattern-recognition one. Thanks to its capability to merge heterogeneous information sources and to its limited computational requirements, it may reveal especially useful for the analysis of multisource remote-sensing data.

Finally, work is currently in progress to deal with the issues associated with the curse of dimensionality.

References

[1] G. S. Sebestyen and J. Edie, "An algorithm for nonparametric pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-15, pp. 908–915, 1966.

[2] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 515–516, 1968.

[3] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, pp. 509–517, 1975.

[4] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[5] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data," *IEEE Trans. Geosci. Remote Sensing*, vol. 28, pp. 540–552, July 1990.

[6] S. B. Serpico and F. Roli, "Classification of multisensor remote-sensing images by structured neural networks," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 562–578, May 1995.

[7] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.

[8] T. Cacoullos, "Estimation of a multivariate density," *Ann. Inst. Stat. Math.*, vol. 18, pp. 178–189, 1966.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[10] A. J. Izenman, "Recent developments in nonparametric density estimation," *J. Amer. Stat. Assoc.*, vol. 86, pp. 205–224, 1991.

[11] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55–63, 1968.

[12] P. M. Lewis, "Approximating probability distributions to reduce storage requirement," *Inf. Contr.*, vol. 2, pp. 214–225, 1959.

[13] D. T. Brown, "A note on approximations to discrete probability distributions," *Inf. Contr.*, vol. 2, pp. 386–392, 1959.

[14] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 462–467, 1968.

[15] H. H. Ku and S. Kullback, "Approximating discrete probability distributions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 444–447, 1969.

[16] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.

[17] M. Schroder, K. Seidel, and M. Datcu, "Bayesian labeling of remote sensing image content," in *Maximum Entropy and Bayesian Methods*, W. Von der Linden, V. Dose, R. Fischer, and R. Preuss, Eds., 1999, pp. 199–206.

[18] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984, p. 181.

[19] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," in *Proc. Amer. Math. Soc.*, vol. 7, 1956, pp. 48–50.

[20] J. S. Marron and W. J. Padgett, "Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples," *Ann. Statist.*, vol. 15, pp. 1520–1535, 1987.

[21] S. B. Serpico, L. Bruzzone, and F. Roli, "An experimental comparison of neural and statistical nonparametric algorithms for supervised classification of remote-sensing images," *Pattern Recognit. Lett.*, vol. 17, pp. 1331–1341, 1996.

[22] M. Datcu and K. Seidel, "Bayesian methods: Applications in information aggregation and data mining," *Int. Archives Photogramm. Remote Sens.*, pt. 7-4-3 w6, vol. 32, pp. 68–73, 1999.

[23] M. Schroder, H. Rehrauer, K. Seidel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives," *IEEE Trans. Geosci. Remote Sensing*, vol. 38, pp. 2288–2298, 2000.

[24] M. Datcu, K. Seidel, and M. Walessa, "Spatial information retrieval from remote sensing images—Part I: Information theoretical perspective," *IEEE Trans. Geosci. Remote Sensing*, vol. 36, pp. 1431–1445, 1998.

Mihai Datcu received the Ph.D. degree in electronics and telecommunications from the University "Politehnica" of Bucharest (UPB), Romania, in 1986, and the title "Habilitation à diriger des recherches" from Université Louis Pasteur, Strasbourg, France, in 1999.

He has held a professorship in electronics and telecommunications with UPB since 1981. Since 1993, he has been a Scientist with the German Aerospace Center (DLR), Oberpfaffenhofen, Weßling, Germany. He is developing algorithms for scene understanding from SAR and interferometric SAR data, as well as model-based methods for information retrieval, and conducts research in information theoretical aspects and semantic representations in advanced communication systems. He held Visiting Professor appointments from 1991 to 1992 at the Department of Mathematics, University of Oviedo, Spain, from 2000 to 2001 at the Université Louis Pasteur, from 1992 to 2001 at the Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, and in 1994 was Guest Scientist with the Swiss Center for Scientific Computing (CSCS), Manno, Switzerland. He was teaching stochastic image analysis, fractal analysis, and image processing in medical sciences and designing and developing new concepts and systems for image information mining, realistic visualization, query by image content from very large image archives, and new algorithms for parameter estimation. Currently, he is Senior Scientist and Image Analysis group leader with the Remote Sensing Technology Institute (IMF) of DLR. His interest is in Bayesian inference, information theory, stochastic processes, model-based scene understanding, image information mining, with applications in information retrieval and understanding of high-resolution SAR and optical observations.

Farid Melgani received the State Engineer degree in electronics from the University of Batna, Algeria, in 1994, and the M.Sc. degree in electrical engineering from the University of Baghdad, Iraq, in 1999. Currently, he is pursuing the Ph.D. degree in electronic and computer engineering at the University of Genoa, Genoa, Italy.

He was with the International Computer Services, Algeria, in 1994 and 1995. Currently, he cooperates with the Signal Processing and Telecommunications group, Department of Biophysical and Electronic Engineering, University of Genoa. His current research interests are in the area of processing and pattern recognition techniques applied to remote-sensing images (classification, data fusion, and multitemporal analysis).

Mr. Melgani served on the Scientific Committee of the SPIE International Conferences on Signal and Image Processing for Remote Sensing VI (Barcelona, Spain, 2000) and VII (Toulouse, France, 2001).

Andrea Piardi received the M.S. degree (magna cum laude) in telecommunications engineering from the University of Genoa, Genoa, Italy, in 2000, and the Pianoforte Diploma from the Academy of Music of Genoa in 1998.

From August 1999 to March 2001, he cooperated with the Department of Biophysical and Electronic Engineering (DIBE), University of Genoa, and with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, within remote-sensing projects. This work was part of his dissertation focused on the development of methods for the classification of multisource remote-sensing data.

Sebastiano B. Serpico (M'87–SM'00) received the Laurea degree in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1982 and 1989, respectively.

As an Assistant Professor in the Department of Biophysical and Electronic Engineering (DIBE), University of Genoa, from 1990 to 1998, he taught pattern recognition, signal theory, telecommunication systems, and electrical communication. Since 1998, he has been an Associate Professor of Telecommunications at the Faculty of Engineering, University of Genoa, where he currently teaches signal theory and pattern recognition. Since 1982, he has cooperated with DIBE in the field of image processing and recognition. His current research interests include the application of pattern recognition (feature selection, classification, change detection, and data fusion) to remotely sensed images. From 1995 to the end of 1998, he was Head of the Signal Processing and Telecommunications research group (SP&T) at DIBE, and is currently Head of the SP&T Laboratory. He is the author or coauthor of more than 150 scientific publications, including journals and conference proceedings.

Dr. Serpico was a recipient of the "Recognition of TGARS Best Reviewers" from the IEEE Geoscience and Remote Sensing Society in 1998 and a Guest Editor of the Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of hyperspectral image data (July 2001). Since 2001, he has been an Associate Editor of TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a member of the International Association for Pattern Recognition Society (IAPR).