IMAGE CLASSIFICATION USING NON-PARAMETRIC CLASSIFIERS AND CONTEXTUAL INFORMATION *

F.J. Cortijo and N. Pérez de la Blanca
Depto. Ciencias de la Computación e I. A. (DECSAI)
E.T.S. Ingeniería Informática
Universidad de Granada
18071 Granada, Spain
cb@robinson.ugr.es

Commision III, Working Group 2

KEY WORDS: Classification, Learning, Algorithms, Combination, Accuracy, Pattern Recognition, Contextual Classification

ABSTRACT

This paper shows some combinations of classifiers that achieve high accuracy classifications. Traditionally it is used the maximum likelihood classification as the initial classification for the contextual correction. We will show that using non-parametric spectral classifiers to obtain the initial classification we can improve the accuracy of the classification significatively with a reasonable computational cost. More specifically we propose to apply the contextual correction performed by the ICM algorithm to some non-parametric spectral classifications.

1 INTRODUCTION

Supervised classifiers assume the existence of a training set $\mathcal{T}$ composed by n labeled training samples, where the labels represent informational classes (labels). This information is used for learning -construction of the classifier- and usually for testing too. We will note by $\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\}$ to the set of informational classes and by X to the samples used for learning and classifying. We assume they are d-dimensional random variables.

Spectral classifiers use only the spectral information related to the pixel to be classified. The thematic map they give as output has the overall impression of a "noisy" classification. This effect is more evident when there is overlapping among the training sets in the spectral space [Cortijo et al., 1995]. In this case it is necessary a post-processing over the initial classification because it is expected to find homogeneous regions in the map as they can be found in the Nature. The straightforward solution consists in incorporating additional information into the classifier related to the spatial neighborhood -its context- of the pixel to classify. That information may be the spectral values of the spatial-neighbors pixels, their labels or both kinds of information combined in some way. When this kind of information is used for classification the classifier is known as a *contextual classifier*.

From a general point of view a contextual classifier can be seen as a smoothing process over an initial image of labels. This map is obtained usually by a spectral classifier. It is well known that some contextual classifiers achieve a local optimum [Besag, 1986] determined by the initial classification. It is used traditionally the maximum likelihood (ML) classification as the starting point for the smoothing process. We have shown [Cortijo & Pérez de la Blanca, 1996a] that the ML classifier is not the best choice when the training sets are high-overlapped. In this work we propose the use of different spectral classifications as initial classifications to a contextual classifier in order to obtain high-accuracy classifications with a reasonable computational cost.

In order to achieve a higher accuracy it looks reasonable to adopt a high accuracy spectral classification as starting point to the contextual classifier, given that contextual classifiers assure convergence to a local maximum. Our proposal consists in adopting different spectral classifications as starting points with the aim of improving the accuracy of the conventional methodology consisting in contextually correcting the ML classification. Many others spectral classifiers improve significatively the results obtained by the ML classifier and the classifications obtained by them are good candidates to be the initial classifications for contextual classifiers. Finally, we must considere the required computational effort to perform the global process: spectral classification followed by the contextual classification. For a particular contextual classifier it is obvious that the contextual classification effort is the same for any initial classification, thus the global computational effort is determined by the spectral classification computing demands.

This paper is organized as follows: In section 2 we describe the methodology we have adopted in this work together with a brief description of the classifiers we have used. In section 3 we describe the datasets used in this paper and in section 4 we show the results obtained. Finally, the main conclusions we have achieved are summarized in section 5.

2 METHODOLOGY

Our objective in this work is to show some combinations of classifiers that achieve high-accuracy classifications. In order to determine some interesting combinations of classifiers for Remote Sensing image classification we have tested a wide number of families of spectral and contextual classifiers.

2.1 Spectral Classifiers

Spectral classifiers are partitioned in two main categories: a) parametric classifiers, if they assume the existence of an underlying probability distribution of the data and b) non-parametric classifiers, if they do not assume anything about the probability distribution.

The structure of the Bayes classifier is determined, basically, by the probability density functions (*pdf's*) $p(X|\omega_i)$. The objective in the construction of a supervised parametric classification rule is to characterize the pattern of each class in

*This work has been supported by the Spanish "Dirección General de Ciencia y Tecnología" (DGCYT) under grant PB-92-0925-C02-01

terms of its *pdf*. It is assumed that it is known the form of that function, that is, it is only needed to know some parameters (estimated from the training set) in order to characterize each class. The *pdf's* are usually multivariate Gaussian functions so it is only needed to estimate two sets of parameters for each class: the mean vector, μ_i and the covariance matrix, Σ_i.

The maximum likelihood classifier (**ML classifier**) imposes quadratic decision boundaries between the clusters of samples in the representation space. If it is assumed a common covariance matrix, $\Sigma_i = \Sigma$ for $i = 1, 2, \ldots, J$, then the quadratic classifier yields to a linear classifier in which the decision boundaries have a linear form. The quadratic classifier is more sensitive to the violation of the Gaussian assumption of the *pdf's* than the linear classifier [Lachenbruch, 1979] and the training set size required by a quadratic classifier is higher than the training set size required by a linear classifier. It is well-know that the Hughes effect [Hughes, 1968] arises in high dimensionality data when the training set size is not large enough to estimate properly the covariance matrices.

When adopting a quadratic or a linear classifier we are imposing an extreme degree of adjustment of the decision boundaries to the training samples. When the training samples are highly overlapped in the representation space they are not good choices and it is plausible to allow a wider degree of adjustment, that is, a wider set of possible decision boundaries. This can be achieved using the *regularized discriminant analysis* classifier (**RDA classifier**) proposed by Friedman [Friedman, 1989].

RDA allows a wide family of parametric classifiers including the quadratic ML classifier and the linear classifiers as particular cases. The estimation of the covariance matrices is performed by a regularization process determined by two parameters, the values of which are customized to individual situations by jointly minimizing a sample-based estimate of future misclassification risk. The joint optimization of these parameters minimize the cross-validation error so this technique is more convenient for problems in which the training set size is small.

In most classification applications the assumption that the forms of the underlying density functions are known is unrealistic. The common parametric forms rarely fit the densities actually encountered in practice. For instance, the parametric models manage unimodal densities whereas many practical problems present multimodal densities [Duda & Hart, 1973]. The only available information is the training set and the classification rules must be built just from it with no additional assumptions.

We can find a wide variety of spectral non-parametric classifiers. They can be summarized in three main categories. The first approximation consists in non-parametric statistical techniques to estimate $p(X|\omega_i)$ (*nearest neighbor estimation* techniques and *kernel estimation* techniques [Duda & Hart, 1973], [Parzen, 1962], [Devijver & Kittler, 1982]). The second consists in estimate directly the a posterior probability $p(\omega_i|X)$ (*nearest neighbor classification* rules [Devijver & Kittler, 1982]) and the third consists in splitting recursively the representation space by means of binaries questions related to the values of the variables involved (*classification trees*). In this work we have adopted two different approaches due to their wide use, accuracy and knowledge: we have used **CART** [Breiman et al., 1984] as classification tree based technique and the nearest neighbor classification rules [Devijver & Kittler, 1982] applied to many different reference sets which have been built by many different learning algorithms, as we will see below.

One of the most popular and widely used non-parametric classification rule is the k *nearest neighbor* rule (k-NN). In the simplest formulation, given a training set $\mathcal{T}$, a metric δ on the representation space and a sample to classify X, the k-NN searches the k nearest neighbors to X in $\mathcal{T}$ and assigns to X the most populated class in the selected neighbors. If $k = 1$ the k-NN rule is known as the nearest-neighbor-rule or 1-NN rule. In this work we will note by 1-NN to the 1-NN classifier that uses the complete (as given by experts) training set to search the nearest neighbor.

The requirement of a large training set to assure the convergence of the k-NN rule [Devijver & Kittler, 1982] is the main drawback of the nearest neighbor rules in practical problems. Moreover there are two additional drawbacks in the application of these rules: firstly, they are very influenced by incorrectly labeled training samples ("noisy" samples or outliers) and secondly, the computational complexity associated to the search of the nearest neighbor(s) in $\mathcal{T}$ can be $O(n^2)$ or higher. To circumvent these problems it is possible to obtain a *reduced* and *representative* reference set, $\mathcal{R}$, from $\mathcal{T}$ with the objective of searching the nearest neighbor(s) in $\mathcal{R}$ with an acceptable trade-off between the accuracy of the classification and the required computational effort ([Devijver & Kittler, 1982],[Kohonen, 1990], [Geva & Sitte, 1991] among others). This can be done in two ways: a) by *editing-condensing* techniques or b) by *adaptative learning* techniques. In the first case $\mathcal{R} \subseteq \mathcal{T}$ whereas in the second case there is not an explicit relation between both sets.

The aim of *editing-condensing* techniques is two-fold: improving the accuracy of the classification by removing samples located in overlapping acceptation surfaces (*editing techniques*) and decreasing the computational effort required to find the nearest neighbor(s) (*condensing techniques*). Editing techniques take as input the original training set and give as output a subset of the original training set. Condensing techniques take usually as input the edited training set and give as output a subset of the previously edited set. $\mathcal{R}$ is a reduced (sometimes drastically) and representative version of $\mathcal{T}$. The joint application of these techniques improve the trade-off between the accuracy of the 1-NN classification and the computational effort required for that classification. A different approach consists in *adaptative learning* techniques. Adaptative learning is a powerful alternative to classical editing-condensing techniques as it allows to fix the reference set size. Now the reference set is not usually a subset of the training set. Adaptative learning algorithms can be tuned by means of a set of parameters in such a way that it is possible to directly supervise the learning process. The training samples are used to tune a fixed number of *codebooks* or *prototypes* and the reference set is called the *codebooks set* or the *prototypes set*. Adaptative learning is performed in two sequential phases: *initialization* and *learning*. The prototypes set is initially a subset of the training set and the values of the prototypes are updated in a iterative learning process.

As editing algorithm we have chosen the *multiedit algorithm* [Devijver & Kittler, 1982] and as condensing algorithm we have adopted the Hart's condensing al-

gorithm [Hart, 1968]. As adaptative learning algorithms we have chosen *DSM* [Geva & Sitte, 1991] -*Decision Surface Mapping*- and *LVQ-1* [Kohonen, 1990] -*Learning Vector Quantization*, version 1-. The values of the parameters involved in LVQ-1 learning have been estimated by using two algorithms proposed by the authors [Cortijo & Pérez de la Blanca, 1996b].

Now we can apply the 1-NN classifier using the reference set learned by these algorithms. We will note by **1-NN ($\mathcal{T}_M$)** to the 1-NN classifier that uses $\mathcal{T}_M$ as reference set, that is, the multiedited training set. Following this notation, if $\mathcal{T}_{MC}$ is the multiedited-condensed training set, then **1-NN ($\mathcal{T}_{MC}$)** is the 1-NN classifier that uses $\mathcal{T}_{MC}$ as reference set. To apply DSM learning it is required that the training set to be previously edited [Geva & Sitte, 1991]. We have used $\mathcal{T}_M$ as initial set for DSM learning. Now if $\mathcal{T}_{DSM}$ is the reference set after DSM learning, **1-NN ($\mathcal{T}_{DSM}$)** is the 1-NN classifier that uses $\mathcal{T}_{DSM}$ as reference set. Finally, if $\mathcal{T}_{LVQ-1}$ is the reference set after LVQ-1 learning, **1-NN ($\mathcal{T}_{LVQ-1}$)** is the 1-NN classifier that uses $\mathcal{T}_{LVQ-1}$ as reference set. More details about these algorithms can be found in [Cortijo & Pérez de la Blanca, 1996a].

2.2 Contextual Classifiers

The contextual classifiers we have tested are based in the assumption of a Markov random field to model the prior distribution of the labels in the image. Stochastic models and *random fields* (RF) in particular represent accurately information a priori on the map. This information can be used in such a way that the Bayes decision theory can be applied. A random field is a joint probability distribution imposed on a set of M random variables $L = \{L_1, \ldots, L_M\}$ representing objects of interests that imposes statistical dependence in a spatially meaningful way. In contextual classification each $L_i \in \Omega$. The spatial dependence can be specified by a global model such as the Gibbs random field (GRF). A GRF describes the global propertied of an image in terms of the joint distribution of labels for all pixels [Dubes & Jain, 1989]. A *Markov random field* (MRF) is defined in terms of local properties. It is needed to fix a neighborhood system in which the spatial dependence is relevant. Two neighborhood systems are mainly used, the first order neighborhood which includes the four-nearest-spatial-neighbors, and the second order neighborhood which includes the eight-nearest-spatial-neighbors.

Given a set of observations, $X = x$, and the contextual information modeled as a MRF, $P(L = l)$, in a Bayesian context the objective is to find the estimator $\hat{l}$ which maximizes equation 1, that is, the a posteriori probability of $L = \hat{l}$, given $X = x$.

$$P(L = l \mid X = x) = \frac{P(X = x \mid L = l)\, P(L = l)}{P(X = x)} \qquad (1)$$

This is known as the MAP (maximum a posteriori) method. The model relating observation x to labeling l is chosen to ensure that the posterior distribution of L, given $X = x$, is also a MRF. If we require conditional independence of the observed random variables. given the true labels, it is enough to ensure that the posterior distribution is also a MRF. Thus we assume that

$$P(X = x \mid L = l) = \prod_{i=1}^{M} P(X_i = x_i \mid L_i = l_i) \qquad (2)$$

If both $P(X = x \mid L = l)$ and $P(L = l)$ are known we can compute L which maximizes the MAP by applying equation 1. In the practice it is clear that even if M and J are low it is not possible to calculate directly the MAP as given in equation 1. To circumvent this problem some alternatives are available to estimate the MAP [Dubes & Jain, 1989]. The first approximation consists in the *simulated annealing* algorithm [Geman & Geman, 1984] which find MAP estimates for all pixels simultaneously. As the computational demands of this algorithm are considerable there are two computationally feasible approximations to the MAP estimate: a) the *ICM* algorithm (*iterated conditional modes*) and b) the *MPM* algorithm (*maximizer of posterior marginals*). A detailed discussion on these methods can be found in [Dubes & Jain, 1989] and references therein. We will center our interest in the **ICM algorithm** [Besag, 1986] which has been demonstrate to have an excellent trade-off between the accuracy of the contextual correction and the required computational effort [Cortijo, 1995].

Another approximation to contextual correction using a MRF consists in *point-to-point contextual correction* methods. They are based in complex conditioned-probability models which are extensions of the MAP expression given in equation 1 by adding an additional term, the *contextual correction factor*, into the denominator of the MAP expression [Sæbø et al., 1985]. Assuming conditional independence of the feature vectors (observations) in a spatial neighborhood two models can be adopted [Sæbø et al., 1985]: a) the **Welch and Salter, Haslett's model** and b) the **Owen and Switzer's model**. We have tested both models in this work.

Contextual classifiers accept as input the classifications obtained by the 8 spectral classifiers described in section 2.1, so we have performed 24 additional classifications for each problem.

3 DATA

The data used to test the performance of the classifiers are two LANDSAT images, landscapes from Greenland, Denmark[1]. The first image is a LANDSAT-2 MSS image of the Igaliko region. The second is a LANDSAT-5 TM image of the Ymer Ø region. Both images are 512×512 pixels in size. The training sets have been selected by expert geologists [Conradsen et al., 1987] and their spectral distribution represent different problematics.

In Igaliko we have five classes to discriminate, the training set size is 42796 samples and there is a slight overlapping in the the spectral distribution of the training samples. In Ymer Ø we have twenty classes to discriminate, the training set size is 12574 samples and there is a high overlapping in the spectral distribution of the training samples. See [Conradsen et al., 1987] for more details.

In this work we have adopted the *test sample estimation* to measure the accuracy of the classifications. The training set, $\mathcal{T}$ is splited into two disjoint sets: $\mathcal{T}^l$ (learning set) and $\mathcal{T}^t$ (test set). $\mathcal{T}^l$ has been built by selecting randomly 2/3 of the available training samples; the remainder are placed into $\mathcal{T}^t$. We use the learning set to construct the classifier and the test set for testing. In tables 1 and 2 we show the learning and test set sizes for each dataset.

[1]We must thank to the **IMM** (Denmak University of Technology, Lyngby, Denmark) for providing the LANDSAT images used in this work.

Class	$\mathcal{T}^l$	$\mathcal{T}^t$	Sum
1	3806	1919	5725
2	7542	3830	11372
3	5463	2768	8231
4	2796	1395	4191
5	8834	4443	13277
Total	28441	14355	42796

Table 1: Learning and test set size. Igaliko.

Class	$\mathcal{T}^l$	$\mathcal{T}^t$	Sum
1	2464	1234	3698
2	843	392	1235
3	413	194	6071
4	196	83	279
5	480	234	714
6	476	233	709
7	178	77	255
8	344	149	493
9	52	21	73
10	187	79	266
11	94	33	127
12	656	313	969
13	144	64	208
14	369	167	536
15	227	96	323
16	192	81	273
17	274	119	393
18	453	220	673
19	271	118	389
20	247	107	354
Total	8560	4014	12574

Table 2: Learning and test set size. Ymer Ø.

4 EXPERIMENTAL RESULTS

In table 3 we show the accuracy of the classifications performed on the Igaliko image and in table 4 we show the accuracy of the classifications performed on the Ymer Ø image. We show in the first column the name of the spectral classifier used to get the initial map, and the accuracy of that classification, in the second column. The remainder columns show the accuracies of the contextual corrections made over the initial map by using the three models adopted in this paper.

5 DISCUSSION AND CONCLUDING REMARKS

From tables 3 and 4 we must note that the accuracy of the spectral classifications can be improved -sometimes drastically- if they are used as input to a contextual classifier independently of the nature of the spectral classifier. This is true for the three contextual classifiers tested in this work.

We can conclude that among the contextual classifiers ICM gives the best results and we must note that the required computational effort is lower than the others. As the ICM computational effort is identical for every initial classification, the global computational cost is determined by the spectral classification cost.

We must note that in both problems the accuracies got with the combinations:

Spectral Classifier	Orig.	ICM	Welch	Owen
ML	73.51	81.33	79.83	80.21
RDA	78.97	89.37	85.46	85.68
CART	80.66	92.30	86.66	86.55
1-NN ($\mathcal{T}$)	74.61	86.94	85.87	85.70
1-NN ($\mathcal{T}_M$)	77.76	83.02	84.63	84.66
1-NN ($\mathcal{T}_{MC}$)	77.08	82.83	84.83	84.85
1-NN ($\mathcal{T}_{DSM}$)	77.50	85.32	84.12	84.52
1-NN ($\mathcal{T}_{LVQ-1}$)	79.07	90.80	86.42	86.44

Table 3: Accuracy of the classifications. Igaliko.

Spectral Classifier	Orig.	ICM	Welch	Owen
ML	61.92	91.37	85.11	85.33
RDA	64.29	85.55	69.36	69.57
CART	62.35	95.58	86.73	87.16
1-NN ($\mathcal{T}$)	78.50	97.98	86.50	87.08
1-NN ($\mathcal{T}_M$)	65.67	90.07	82.96	83.60
1-NN ($\mathcal{T}_{MC}$)	63.23	81.09	70.12	70.35
1-NN ($\mathcal{T}_{DSM}$)	64.55	80.97	72.66	73.22
1-NN ($\mathcal{T}_{LVQ-1}$)	68.18	93.64	85.55	86.41

Table 4: Accuracy of the classifications. Ymer Ø.

a) **CART + ICM**, and

b) **1-NN ($\mathcal{T}_{LVQ-1}$) + ICM**

are very high. The computational effort associated to CART is mainly influenced by the learning step (a function of the training set size) but we must note that it is a relatively low cost step. LVQ-1 learning is a quick process and as a additional advantage we can select the training set size and the parameters involved [Kohonen, 1990]. As an additional advantage the values of the parameters involved in the LVQ-1 learning have been automatically estimated by using two algorithms proposed by the authors.

These combinations have also been tested on synthetic very-high-spectral images [Cortijo, 1995] and the results obtained do extend these shown here.

REFERENCES

[Besag, 1986] Besag, J., 1986. On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society. Ser. B, 48(3), pp. 259–302.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R. and Stone, C., 1984. Classification and Regresion Trees. Wadsworth International Group.

[Conradsen et al., 1987] Conradsen, K., Nielsen, A.A., Nielsen, B.K., Pedersen, J.L. and Thyrsted, T., 1987. The Use of Structural and Spectral Enhancement of Remote Sensing Data in Ore Prospecting - East Greenland Case Study. Technical Report, IMM, The Technical University of Denmark, Lyngby, Denmark.

[Cortijo et al., 1995] Cortijo, F.J., Pérez de la Blanca, N., Molina, R. and Abad, J., 1995. On the combination of non-parametric nearest neighbor classification and contextual correction. In: Pattern Recognition and Image Analysis,

Proccedings of the VI Spanish Symposium on Pattern Recognition and Image Analysis held in Córdoba, Spain, on 3-7 April, 1995. Edited by A. Calvo and R. Medina. pp. 503–510.

[Cortijo, 1995] Cortijo, F.J., 1995. Un Estudio Comparativo de Métodos de Clasificación de Imágenes Multibanda. Ph.D. Thesis, DECSAI, Universidad de Granada, Spain *(In spanish)*.

[Cortijo & Pérez de la Blanca, 1996a] A comparative study of some non-parametric spectral classifiers. Applications to problems with high-overlapping training sets. Technical Report. #DECSAI-96-03-08. Department of Computer Science and Artificial Intelligence, University of Granada, Spain. `http://decsai.ugr.es/diata/tech_rep.html`, electronic edition.

[Cortijo & Pérez de la Blanca, 1996b] Cortijo, F.J. and Pérez de la Blanca, N. (1996). Image classification using adaptative-learning techniques. Automatic estimation of the LVQ-1 parameters. Technical Report. #DECSAI-96-03-10. Department of Computer Science and Artificial Intelligence, University of Granada, Spain. `http://decsai.ugr.es/diata/tech_rep.html`, electronic edition.

[Devijver & Kittler, 1982] , Devijver, P.A. and Kittler, J.V., 1982. Pattern Recognition. A Statistical Approach. Prentice Hall, Englewood Cliffs.

[Dubes & Jain, 1989] , Dubes, R.D. and Jain, A.K., 1989. Random Field Models in Image Analysis. Journal of Applied Statistics. 16, pp. 131–164.

[Duda & Hart, 1973] , Duda, R.O. and Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Wiley & Sons.

[Ferri, 1995] Ferri, F.J., 1993. Selecció de Referències en Reconeixment de Formes. Aplicacions en Visió Artificial. Ph.D. Thesis, Universitat de València, Spain.

[Friedman, 1989] Friedman, J.H., 1989. Regularized Discriminant Analysis. Journal of the American Statistical Association, 84(405), pp. 165–175.

[Geman & Geman, 1984] Geman, S. and Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, pp. 721–741.

[Geva & Sitte, 1991] Geva, S. and Sitte, J., 1991. Adaptative Nearest Neighbor Pattern Classification. IEEE Transactions on Neural Networks, 2(2), pp. 318–322.

[Hart, 1968] Hart, P.E., 1968. The Condensed Nearest Neigbour Rule. IEEE Transactions on Information Theory, 14, pp. 515–516.

[Hughes, 1968] Hughes, G. F., 1968. On the mean accuray of statistical pattern recognizers. IEEE Transactions on Information Theory, 14, pp. 55–63.

[Kohonen, 1990] Kohonen, T., 1990. The Self-Organizing Map. Proc. of the IEEE, vol. 78. pp. 1464–1480.

[Lachenbruch, 1979] Lachenbruch, P. A., 1979. Note on initial missclassification effects on the quadratic discriminant function. Technometrics, 21, pp. 129–132.

[Parzen, 1962] Parzen, E., 1962. On estimation of probability density functions and mode. Annals on Mathematics and Statistics, 33, pp. 1065–1076.

[Sæbø et al., 1985] Sæbø , H.V., Bråten, K., Hjort, N.L., Llewellyn, B. and Mohn, E., 1985. Contextual Classification of Remotely Sensed Data: Statistical Methods and Development of a System. Technical Report 768, Norwegian Computer Center, Oslo, Norway.