Hierarchical Feature Selection: A Decision Tree based Approach.

F. J. Ferri*, J. Albert
Dept. Informàtica i Electrònica
Universitat de València
46100 Burjassot (VALÈNCIA) Spain

I. Gracia, F. Pla
Dept. Informàtica
Universitat Jaume I
12071 Castelló Spain

P. Pudil, J. Novovičová
Inst. Inform. Theory and Automation
Academy of Sciences
182 08 Prague 8 Czech Republic

(*) email: ferri@uv.es, tel+fax: +34 6 386 4768

1 Introduction

The problem of selecting a subset of representative features or variables for discrimination purposes has been studied from very different points of view over the few past decades. As in other combinatorial application-dependent problems, no optimal and generally applicable solution exists. Consequently, most researchers have developped their own solution to particular problems with more or less degree of applicability to other domains.

In Statistical Pattern Recognition, several families of FS techniques exist, each one with their own properties [2]. In these methods, two separate components can be identified: the criterion used to assess the feature subset significance or discriminant power, and the method used to effectively search the subset space. The *criterion function* assigns to a particular subset of features a measure of its ability to perform the desired task. In particular, when a classification system is to be built, the criterion has to be related with its expected accuracy. On the other hand, the *search method* usually constructs feature subsets and computes the value of the criterion function. This eventually leads to a particular subset or family of subsets with different sizes that are near to the optimal solution. In particular, there are optimal algorithms that give the best solution for a given subset size if the criterion function satisfies the monotonicity condition [4]. Apart from the fact that the concept of best solution is not clearly defined, these optimal solution are not applicable in practice due to the high computational burden.

From the fact that specific classification problems may require more accurate feature selection, there is a recent interest in developping feature selection methods for certain interesting types or families of classifiers [7, 3]. The motivation is two-fold. First, it constitutes an improvement on the plain classifier from the point of view of their accuracy; and second, it makes the final solution more feasible from the point of view of their implementation.

In this work, several recent attempts to apply feature selection methodology in the context of decision tree classifiers are briefly described. Some results from preliminary experiments are reported and discussed. Finally, the possibility of applying this approach and further extensions for general problems in Pattern Recognition is considered and evaluated.

2 Feature Selection and Decision Trees

Decision Trees (DT) are a particular and interesting type of classifier which allows dealing with many different types of features, and even with problems where the feature set is not fixed (uncertainty in some measurements). Training samples are hierarchically split into smaller and homogeneous groups (usually) taking into account only one feature at a time. Classification is made by dropping the patterns through a binary tree while performing simple tests on a feature at each step [1, 5].

DT construction implies some kind of selection among the features initially available. Instead of assigning higher weights to highly discriminant features as in the case of feedforward neural nets [9], discriminant features are usually used in the first levels of the tree. Nevertheless, because of particularities of the tree construction methods, some non discriminant features tend to be selected at the lower levels of the tree which gives rise to overtraining. These effects are partially overcome by the pruning methods [8] but the information gained in the early steps about the discriminant power of certain features is never used in the process.

Partially supported by grants SERC GR/E 97549, Czech Acad.Sci. A275107 and CICYT TIC95-676-C02-01

3 Estimating Feature Significance

An alternative to prunning methods for tree construction are descending methods. Instead of growing large trees that overfit the data, a more careful error estimation process can be applied when adding new nodes to the tree. A descending method using this ideas was shown to perform as good as the most used ascending (prunning) methods [6].

Moreover, descending methods have the advantage that they can be used to retrieve convenient information about the discrimination power of each feature. In particular, an algorithm for estimating feature discriminant power while growing a decision tree has been recently proposed [3]. In this algorithm, the same information used to add a new node using a particular feature is used to update the relative significance of this feature with respect to the overall performance of the tree.

As a result, a ranking among features is obtained. This ranking can be used either to rebuild the tree according to the selected features or to use them in a different classifier.

4 Experiments and Conclusions

Experiments using synthetic and real data have been performed. The ranking of features has been used to construct feature subsets with an increasing number of features. Decision trees and other type of classifiers have been built with these feature subsets. The initial results were encouraging but they put forward the dependency of the final classifier with respect to the criterion used for selecting features. In this sense, the feature selection method presented is specially suitable when a final tree classifier is considered. In this case, the number of features used can be significantly reduced if they are used to construct a decision tree without degrading the performance.

A number of possible extensions to the work described can be considered. First, the way of estimating discriminant power could be generalised to include not only ranking of features but construction of different subsets of features with different cardinalities. In this way the well-known problem of dependence among some particular features could be removed. On the other hand, the method could be applied directly as the tree is being built. This could lead to a combined method for tree construction and feature selection that would necessarily imply some kind of backtracking or prunning of some nodes of the tree at early steps.

References

[1] Breiman L. et. al., *Classification and Regression Trees*, Chapman & Hall, 1984

[2] Devijver, P.A. and Kittler, J. *Pattern Recognition: a Statistical Approach*, Prentice-Hall International, 1982

[3] Gracia, I; Pla, F.;Ferri, F.J.;Garcia,P. "Estimating Feature Discriminant Power in Decision Tree Classifiers" in Proc. of 6th Intl. Conf. on Computer Analysis of Images and Patterns, Prague. *Lecture Notes in Computer Science, 970*, pp. 612-617, 1995.

[4] Kittler, J. "Feature Selection and Extraction", *Handbook of Pattern Recognition and Image Processing*, 1986

[5] Murthy, S.K.; Kasif, S. and Salzberg, S.; "A System for Induction of Oblique Decision Trees", Journal of Artificial Intelligence Research, **2**, 1994, pp. 1-32.

[6] Pla, F. *Estudios de Técnicas de Análisis de Imagen en un Sistema de Visión para la Recolección Robotizada de Cítricos*, (in Spanish) Ph.D. Thesis, Universitat de València, 1993.

[7] Pudil, P.; Novoviçová, J; Ferri, F. and Kittler J. "Advances in the Statistical Methodology for the Selection of Image Descriptors for Visual Pattern Representation and Classification", in Proc. of 6th Intl. Conf. on Computer Analysis of Images and Patterns, Prague. *Lecture Notes in Computer Science, 970*, pp. 832-837, 1995.

[8] Quinlan, J.R. "Simplifying decision trees", International Journal of Man-Machine Studies **27**, pp. 221-234, 1987.

[9] Raudys, S. "Accuracy of Feature Selection and Extraction in Statistical and Neural Net Pattern Classification", in Proc. of the 11th Intl. Conf. on Pattern Recognition, The Hague, pp. 62-70, 1992.