ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Classification of chestnuts with feature selection by noise resilient classifiers

Elena Roglia[1] Rossella Cancelliere[2]
Rosa Meo[3] *

Università di Torino - Dipartimento di Informatica
corso Svizzera 185 - Italy

Abstract.

In this paper we solve the problem of classifying chestnut plants according to their place of origin. We compare the results obtained by state of the art classifiers, among which, MLP, RBF, SVM, C4.5 decision tree and random forest. We determine which features are meaningful for the classification, the achievable classification accuracy of these classifiers families with the available features and how much the classifiers are robust to noise. Among the obtained classifiers, neural networks show the greatest robustness to noise.

1 Introduction

One of the main activities of botanic science is plants classification. As a typical problem of pattern recognition some basic issues must be addressed: (1) which attributes, called features, should be used from botanists' descriptions for classification, (2) which classifiers should be used in order to obtain, with the available features, a high classification accuracy, and finally (3) at which extent the classification accuracy degrades if the features are affected by noise. These issues are discussed in this article. We face the problem of the prediction of chestnut origin from their properties: this problem has many important industrial applications, such as production and verification of certificates of product origin. At first we worked with few features related only to fruit peculiarities. We compared the classification accuracy obtained by these features by many state of the art classifiers: a multi-layer perceptron (MLP), introduced by D. Rumelhart et al. [2], a radial basis function network (RBF), a support vector machine [8], a C4.5 decision model induced by C4.5 algorithm [3] and a random forest (RF) presented by L. Breiman [4]. The extremely poor classification performances obtained (see Section 3), suggested us to perform a initial selection of the classifiers and to seek for more informations. In fact, the initial features were supposed to be not appropriate due to the excessive variability of value from fruit to fruit. Thus, we added to the description of each chestnut instance, some features related to the entire plant with the idea that they could constitute more robust predictors. The larger data set so obtained (soon available on line) contains 1600 samples, described by 37 features taken from both chestnut plants

*Work supported by Regione Piemonte in the context of the project: Realizzazione di modelli informatici per la valorizzazione della qualità e la tracciabilità delle produzioni in specie da frutto coltivate in Piemonte, Cipe 2004.

and fruits. They are all the necessary informations to discriminate among the different classes. A choice of the best subset of these features must however be performed remembering that botanic features are extracted, collected and stored in a data set by human agents. This process is lengthy, costly and error-prone. As a consequence, the number of features should be reduced as much as possible but this reduction should not affect the classification performance.

In addition, it is very important to investigate how a classifier answers when noisy inputs are presented. We show that the selected classifiers, especially in the case of neural networks, are also robust to the presence of a noise amount consistent with the small perturbation assumption (maximum error=5% of the value of each feature).

This paper is structured as follows. In Section 2 we overview the selected classifiers: C4.5 decision trees and the random forest (MLP is a widely known method and will not be reviewed). In Section 3 we discuss the feature selection strategy and present the experimental results, in both non-noisy and noisy cases.

2 Overview of C4.5 decision trees and Random Forest

For the sake of completeness we introduce some of the basic characteristics of the adopted learning models. A complete overview can be found in [3, 4, 5].

Decision trees. A decision tree is a structure whose internal nodes answer to a test condition based on the value of some of the record attributes. Each outcome of the test condition leads either to another internal node or to a leaf node which contains a class value. That class value is the prediction of the decision tree for the set of data records that reach that final node. Thus, the outcomes of the attribute tests separate data records with different characteristics into disjoint partitions that are homogeneous for the class value.

The induction step in C4.5 follows a greedy strategy that grows a decision tree by progressively partitioning the training data into smaller partitions until each of them is homogeneous in the value of the class label. C4.5 algorithm induces the form of the decision tree, i.e., chooses the test condition at each node t of the tree by the following rule. Let denote by S the set of data records that reach node t. Given c class labels, let $p(i, S)$ denote the fraction of records in S that belong to class i. Any attribute test at node t is evaluated by entropy of the class value in S, denoted by $E(S)$:

$$E(S) \quad = \quad -\sum_{i=0}^{c-1} p(i,S) log_2[p(i,S)]$$

Entropy is a measure of impurity of the class in S. The best attribute test at node t is selected as that one that allows the higher difference between the entropy at the parent node t (before the test condition) and the entropy at the children nodes (after the test condition).

Random forest. Significant improvements in classification accuracy have resulted in a set of methods called ensemble methods. They consist in the generation of multiple, base classifiers from training data and successively combining the predictions of each of them in test. Random forest is a special ensemble learner, which is also suitable for problems involving a large number of features. In a random forest a large number of decision trees is grown where each tree depends on the values of a random vector, sampled independently and with the same distribution for all trees. Random vectors are generated using an ensemble method (called bagging) which randomly selects N features, with replacement from the original training set. Each tree in a random forest is grown at least partially at random in one of the following ways: (1) randomness is injected by growing each tree on a different random subsample of the training data; (2) randomness is injected into the attribute test selection process so that the test condition at any node is determined partially at random. When multiple trees are generated, their predictions are usually combined so that the most popular class among them is predicted. The technique of majority voting is usually adopted (where majority is eventually weighted by giving more weight to the more correct trees).

3 Experimental results

In this section we describe in more details the feature selection method, the generation of the training and test sets and the results obtained for the task of classifying chestnuts according to eight places of origin. The initial data set was made by 19 features describing 1600 samples taken from fruits and was analysed using a cross validation methods with 10 folds. Classification results from a MLP, a RBF, a binary decision tree (C4.5), a random forest (RF) and John Platt's sequential minimal optimization algorithm for training a support vector classifier (SMO) are compared. We used the default settings of Weka classification tools [8].

MLP	RBF	C4.5	RF	SMO
58.12%	47.94%	49.81%	55.06%	52.50%

Table 1: Percentage of instances correctly classified.

Table 1 shows that classification accuracy was extremely poor. After some attempts to optimize the models, we decided to add to the dataset more descriptive features related to the entire plant. We had the hope that the high number of features obtained (37) would be afterwards reduced by feature selection. The initial results also suggested us to reduce the number of classifiers in the further investigations: MLP, chosen as the best of neural methods, C4.5 (the classical and largely used decision tree method) and random forest (because of its robustness to noise and scalability).

3.1 Feature selection

In feature selection, the goal is to find a subset of significant attributes able to correctly predict unseen data and to reduce both human measurement errors and the cost activity of data extraction from plants and fruits. Ranking of features is possible since a large number of feature evaluation measures is available (see, for instance [6, 7] for a survey on some of them).

In our experiments, we tested on the whole dataset some commonly used methods for classification purposes, available in Weka. Some of them are *filter* methods, that select features on the basis of measures of feature predictivity and redundancy, like: Symmetrical Uncertainty (SU), Chi-square statistic, Gain Ratio and Information Gain. Others are wrapper methods, based on the accuracy that some learner is able to reach on the data with the selected set of features, like: attribute selector based on instance-based learners, attribute subset evaluator based on any learner, and oneR methodology based on simple rule-based classifiers. We verified that all of these methods agree on the selection of a unique core of relevant features, determined applying the above cited feature selection methods as ranking methods for the overall set of features. Selected core is made of features which are present, in the first 10 positions of the rankings. We noticed that it is exactly the feature set selected by entropy-based information gain criterion. This criterion is commonly used by decision tree algorithms when they select which attribute will become a test attribute in a given branch of the tree. Thus, information gain criterion was finally used to select the core of 6 relevant features among the 37 initial ones. They are: number of chestnuts/kg, diameter of the trunk, number of female inflorescence/ament, ament length, length of the leaf limb and height of the plant. We verified through comparison of classification performances that no information content was lost in this process; on the contrary classification performances improved because of redundancy reduction in instance description.

3.2 Classification performances in non-noisy datasets

The training set is a list of $T = 1120$ instances randomly chosen from original data set (70% of the overall data set). The test set includes the remaining 480 instances. Training set has been used to optimize the three classifiers. Random forest has been built by a forest of decision trees built on the 6 features and trained each on a different training set, built by random selection of samples with replacement (first option).

MLP classifier has 6, 12 and 8 neurons (one for each geographic zone) respectively in input, hidden and output layers. We optimized the number of hidden neurons and the most relevant parameters with respect to Weka default because performances compared with those obtained with C4.5 and RF were lower. The training phase required 100 iterations.

Decision tree and random forest produce a correct classification of all input instances while neural network correctly classifies 97.91 % of the instances.

3.3 Classification performances in noisy datasets

We also evaluate the sensitivity of the three models to noise. A noisy test set was created perturbing separately each attribute of every instance according to the following equation:

$$i'[A] = i[A] + 0.05 \cdot \eta \cdot i[A]$$

where $i[A]$ is the value of the attribute A in the i-th instance and η is a random value ($-1 \leq \eta \leq 1$). The three classifiers were finally run on the perturbed data set i.e. on the noisy version of the test set.

	Without Noise	With Noise
MLP	97.91	93.12
C4.5	100	82.29
RF	100	90.62

Fig. 1: Accuracies on test set

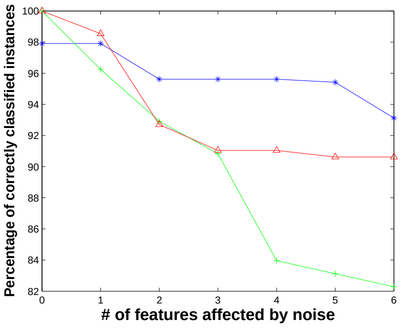

Fig. 2: Accuracy decrease.

Figure 1 shows the classification accuracy obtained by the different classifiers on non-noisy and noisy versions of the test data. It is clear that without noise the decision tree and the random forest reach a slightly higher accuracy rate than the neural network. On the contrary, on noisy data, the neural network maintains its good performance while the decision tree and the random forest degrade seriously the previously obtained results.

We also performed a paired, one-tail $t-test$ on the statistical significance of the difference in accuracy of the classifiers, conducted on the 480 test samples. The null hypothesis is that one classifier has a classification accuracy lower than the other one (mean difference$\leq$0). The observed differences between MLP and RF lead to a critical value $t_c = 1.435$ so that the null hypothesis is rejected with a p-value of 0.75%. For the difference in accuracy between MLP and C4.5 the critical value is $t_c = 5.185$ while it is $t_c = 4.969$ for the difference between RF and C4.5, both corresponding to a p-value of 10^{-5}%.

We also examined closely the classifier behavior w.r.t. an increasing number of noisy features in the data set. Figure 2 shows the performance decrease when the number of noisy features increases from 0 to 6 for decision tree (+), random forest ($\triangle$) and multilayer perceptron ($*$).

Results remark that neural networks are quite stable because class prediction results marginally affected by the presence of noise. On the contrary, decision tree and random forest are more sensitive. Although decision tree and random forest reach higher accuracy in a clean test data, classification accuracies result proportionally more affected by an increasing presence of noise. In conclusion, when decision trees and random forests are used as predictive models in the context of this peculiar domain, they are less robust with respect to neural networks to the presence of noise. This is an important issue for a learner, employed in a real environment, in which commonly some features are affected by noise or human error. The interested reader can find more experimental results in [1].

4 Conclusions

In this paper we compare the accuracy of classification of chestnuts according to their place of origin. We used state of the art learners: decision trees, random forests, multilayer perceptrons, radial basis functions networks and support vector machines. The results, in the context of this peculiar domain, confirm the robustness of neural network classification techniques and their reliability for treating noisy data. Even though decision trees and random forests reach higher accuracy rates on clean and safe test data, when noise is present, they result less robust and stable.

In this study we have also experimented the importance of feature selection for classification of botanic species. We applied several feature ranking methods. All of them agree on the selection of a core of 6 features (only 16%) as the most predictive and least redundant ones that still allow to obtain comparable classification results.

5 Acknowledgements

The authors thank IPLA and A.Ferrara, E.Viotto and F.Tagliaferro for the dataset collected for INTERREG Project funded by Regione Piemonte.

References

[1] Elena Roglia, Rossella Cancelliere, and Rosa Meo. Classification of chestnuts with experiments on feature selection and noise. *Technical Report 100-2007 - available from http://www.di.unito.it/˜ meo/Pubblist/pubbListEng.html*, November 2007.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. *Parallel distributed processing: explorations in the microstructure of cognition*, Volume 1: foundations:318–362, 1986.

[3] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1992.

[4] Leo Breiman. Random forest. *Machine Learning*, 45:5–32, 2001.

[5] P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining.* Addison-Wesley, 2005.

[6] M. Dash and H. Liu. Feature selection for classification. *Intell. Data Analysis*, 1(3), 1997.

[7] Ken McGarry. A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.*, 20(1):39–61, 2005.

[8] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2nd edition, 2005.