Characterization of Classification Algorithms

J. Gama and P. Brazdil

LIACC, University of Porto
Rua Campo Alegre 823
4150 Porto, Portugal
email: {jgama, pbrazdil}@ncc.up.pt

Abstract

This paper is concerned with the problem of characterization of classification algorithms. The aim is to determine under what circumstances a particular classification algorithm is applicable. The method used involves generation of different kinds of models. These include regression and rule models, piecewise linear models (model trees) and instance based models. These are generated automatically on the basis of dataset characteristics and given test results. The lack of data is compensated for by various types of preprocessing. The models obtained are characterized by quantifying their predictive capability and the best models are identified.

1 Introduction

Previous studies have shown that it is difficult to identify a classification algorithm that would perform well on all tasks [7]. Although some classification algorithms may perform quite well in general, they may be easily surpassed by others. It is then desirable to take the characteristics of the task into account when attempting to identify a suitable algorithm. We are interested to find models that permit us to make reasonably reliable predictions concerning applicability. If we use Shaffer's [11] terminology our aim is to characterize the space in which the individual classification algorithms achieve positive generalization performance.

Shaffer [10] described a method which involves testing competing models using cross-validation. It was demonstrated that the truly best model is selected with high probability. The disadvantage of this approach is that we have to do quite a lot of testing before a decision can be made. As Brodley [4] has shown the decision as to which is the 'best' model can be guided by rules and can thus be potentially quicker. The rules used in [4] incorporate the knowledge of domain experts and hence are not easy to update when new algorithms, or test results, become available.

Various attempts have been made to automate the generation of such rules ([2], [3]). The method in [3] incorporated the process of learning, using decision tree/rule learning system C4.5 [8]. The advantage of this approach is that the rules characterizing the applicability can not only be generated with relative ease, but also updated whenever new results become available.
algorithms are expressed in terms of this error margin (i.e. how many EM's it is above the default error rate).

Normalization Method 3: This method consists of first assembling the test results of all classification algorithms on one particular dataset. These are used to calculate the mean and the standard deviation. All error rates are normalized by subtracting the mean value and by dividing the result by the standard deviation. For the Segment dataset, for instance, the mean error rate of the classification algorithms considered is 8.9%. The standard deviation in this case is 9.8%. If we use these figures to normalize the error rate of, say, ALLOC80, we will get the normalized error rate $ERN_3$ of -0.605.

The advantage of this method is that the values have rather clear interpretation. Higher negative values (i.e. $<-0.5$) indicate that the error rate is rather low. Values around 0 show that the error rate is not far from the average. Positive values may be interpreted as worse than average performance. High positive values (such as 10) suggest that the algorithm may have failed to run and hence was attributed maximum error rate.

2.2 Normalization of Dataset Characteristics

Datasets are characterized using certain number of measures which are similar to those in [3] or in [7]. These include simple measures, statistical measures and information based measures. The simple measures include:
- number of examples (N),
- number of attributes (p),
- number of classes (k),
- proportion of binary attributes (Bin.att),
- errors quantified by costs (Cost).

The statistical measures include:
- standard deviation ratio (SDratio),
- mean value of correlation (Correl),
- canonical correlation for the best single combination of attributes (Cancor1),
- the first normalized eigenvalues of canonical discriminant matrix (Fract1),
- skewness,
- kurtosis.

The information based measures include:
- entropy of class (Hc),
- entropy of attributes (Ha),
- mean mutual information of class and attributes (Mca),
- noise-signal ratio (NSratio) equal to $(Ha-Mca)/Mca$.

In the experiments with regression models the dataset measures were also normalized. The process is similar to normalization 3 of error rates. For each measure mean and standard deviation is calculated across all datasets. Each measure is normalized by subtracting the mean and by dividing the result by the standard deviation. Here all measures normalized in this way are marked by '. For instance, $Kurtosis'$ represents the normalized version of $Kurtosis$ measure.