A Meta-Analysis of Classification Methods

Adrien Jamain
August 2006

1. Aim of the project
Supervised classification has a substantial history, but has recently progressed dramatically as a result of the advent of the computer, and is now considered as one of the main components of the relatively new field of data-mining. Numerous methods have been designed in the past twenty years or so, but by different communities who had different approaches and interests. Comparative studies have thus been carried out to assess the relative merits of the methods. Perhaps paradoxically, however, there is still little understanding of the circumstances under which certain methods will perform better than others. Providing some elements of such an understanding was our initial and ideal goal.

2. Basic idea
Instead of creating yet another comparative study, our idea was to try to re-use the vast quantity of results which has already been published in the literature. Hence the novelty of the project: this kind of second-level analysis (or meta-analysis) had not been done in a systematic fashion in the classification area before. Although it is a common and accepted approach in medicine and the social sciences, it appears that the challenges involved in meta-analysis here are quite different, and that the task is undoubtedly very hard.

3. Approach
Our first task was to collect results from the literature and to store them in a purposely built database. The literature search itself is not complete in the strict sense of the term—and it cannot be, given the size and nature of the domain—but we hope to have included all the major studies, in addition to a set of smaller ones. We have then explored the database through various subsets and methods. Subsets that we have used include a small based on 6 articles, taken as a pilot study, all the results from the famous Statlog study, and results related to 9 of the most well-known classification methods. Methods that were applied comprise variations around psychometric-inspired rating, hierarchical clustering, logistic regression, principal components analysis, and others.

4. Results
We have written three articles summarising the important results of our full study [1]:

- In [2] we describe the analysis of a subset of our data related to 9 of the most famous classification methods. We show how one can use logistic regression, linear regression, and Bradley-Terry models to provide an overall assessment of classification methods.
- In [3] we show an intriguing pattern discovered in our whole database, namely that error rate seem to decrease with dataset size across different datasets. We investigate different plausible explanations for the phenomenon.
- In [4] we report the discovery of a probable mistake that we have spotted in the literature while analysing our data. We suggest that the Naïve Bayes method in one famous comparative study has been reported coming from the wrong source, and we present the evidence for our claim.

Besides the relatively limited conclusions we have obtained through the use of various models on our data, one of our most significant achievements is the constitution of the database that we have built to contain the results collected from the literature. This database is a simple Excel file, and offers a set of macros which implements a front-end interface for the user (links below). The database will hopefully be of major usefulness for all the researchers interested in designing or comparing classification methods, along with being a potential subject of research by itself.

5. Discussion
One problem that we had to face is that the literature is very heterogeneous, and therefore it is absolutely impossible to account for all the variability in the published results. Two simple examples of this heterogeneity concern the datasets on the one hand, and the methods on the other. Datasets are indeed very rarely if ever used in the same way in two different studies, and methods are similarly modified at will by the authors of each study. Hence we have come to the belief that published results reflect other factors, like the time spent by the authors of each study “per result” (which is roughly inversely proportional to the size of the study) or the authors’ initial intention when they designed the study (which is also linked to its size). In [1] we have also criticized a few biasing aspects of the UCI repository, which is the largest source of datasets for classification on the Internet, and propose some bases to improve its adequateness for the future of research in data mining. In overall conclusion we give words of caution against any strong interpretation of the results published in the literature.

6. References