

Meta-Analysis: From Data Characterisation for Meta-Learning to Meta-Regression

Christian Köpf¹, Charles Taylor², and Jörg Keller¹

¹ DaimlerChrysler AG, Research & Technology, FT3/AD, P.O.-Box 2360,
D-89013 Ulm, Germany,

{christian.koepf,joerg.keller}@daimlerchrysler.com

² University of Leeds, Department of Statistics, Leeds, LS2 9JT, U.K.,
charles@amsta.leeds.ac.uk

Abstract. An extended Meta-Analysis fertilizes a Meta-Learning, which is applied to support the user with an automated guidance in model selection and data-transformation. Two major application fields were selected in METAL (Meta-Learning assistant, ESPRIT project 26.357): classification and regression learning. In phase 1 of the project, the data characteristics, measures and tests have been evaluated for an automated use of classification algorithms. For regression learning, the statistics, information theoretical measures and tests had to be proved. This paper works out necessary statistics and tests for regression learning. The new approach of this paper is to use a Meta-Regression: A regression learning on the meta level for Meta-Learning. In comparison to a classification of error rates, calculated for cross-validation tests, our new approach could improve the accuracy for Meta-Learning.

1 Introduction

Nowadays, Knowledge Discovery in Databases (KDD) presents many challenges with popular commercial tools, e.g. (1) selection of a suitable model and (2) the combination of methods. Most non-expert users must either resort to trial-and-error or consult an expert. In public, industrial or even in research application fields, almost all applicants require an automatic and systematic guidance for preprocessing and thereafter use of Machine Learning (ML) and Data Mining (DM) methods. A brief overview can be found in [10].

In the past decade, a number of research projects have examined Meta-Learning. In Europe, the most prominent ones include the ESPRIT **StatLog** (1991-1994) [9] and METAL (1998-2001) [8] projects. Several international workshops were dedicated to this subject, e.g. ECML'98, ICML'99, ECML'2000 and MSL'2000.

2 Meta-Learning and Meta-Analysis

Meta-learning is closely related to meta-analysis defined by the National (US) Library of Medicine as “a quantitative method of combining the results of independent studies (usually drawn from the published literature) and synthesizing summaries and conclusions which may be used to evaluate therapeutic effectiveness, plan new studies, etc., with application chiefly in the areas of research and medicine.” Some of the first attempts [5] at meta-analysis were focussed on combining p -values from various studies, but the name “meta-analysis” was not coined until 1971. Early examples were taken from psychology, but the recent rapid growth has been mainly drawn from and analysis of clinical studies. Two central aspects of meta-analysis are methods for gathering meta-data (for example whether certain results should be omitted) and how to summarize the findings. In particular, methods have been developed for detecting selection bias and building models to appropriately adjust statements of inference. More generally, consideration of sampling errors indicate that the meta-data observations should be weighted according to sample size. Clearly, there is much that can be learned from the theory which has built up around meta-analysis. But there are several distinctive features for which meta-learning deserves special attention. These include: the possibility to re-analyze and augment the data; the feasibility of simulating useful data; the fact that the values to be predicted are multivariate, and possibly qualitative; and the absence of a requirement for domain-specific knowledge. A final difference is that there are often known distributional results for some of the predictor variables, and this can lead to suggestions for optimal transformations and combinations of variables to improve predictability of the response.

3 Statistics for Dataset Characterisation

Data characteristics from the given data are the second knowledge source for the selection of pre-processing and classification methods. In the following sections, we define a vocabulary of measurements which can be computed on the datasets. We distinguish between simple or standard characteristics, discriminating measures and information theoretical measurements. The data that is available possesses certain properties, which should be cataloged at first. This means that meta-data, i.e. information theoretic measurements describing properties of the data, need to be collected on the data. Meta-data can then be used for several purposes:

- Support for the selection of model generation.
- Support for the data preparation phase of a KDD process [3]. A special gain is to be expected from the application of meta-data in defining the data transformation stage of the model generation phase.
- Guide to understanding the algorithm performance characteristics and more general behaviour

Data characteristics are divided into several groups, including standard statistics and enhanced statistics. Standard statistics describe the properties of the dataset as well as properties that the variables in the dataset possess. On the other hand, enhanced statistics are divided into two main parts, description of properties of the numerical subspace in which the domain is described and description of properties of the nominal properties of the data space.

A different approach in order to perform any kind of supervised meta-learning is landmarking [2]. With landmarking, the performance of simple and efficient learners on a task are calculated. The learner’s space is divided into *areas of expertise* which are classes of tasks on which it performs specially well. So, the performance of the simple learners on a special task might give an indication on the performance of other learners [12]. We did not use any landmarkers for this work. However, this serves as an idea and motivation for future work.

3.1 Simple Measures

These simple measures include general information of the dataset such as the total number of observations, denoted by n , the total number of classes, denoted by q , the number of observations for each class, denoted by n_i for class i , and the number of attributes, denoted by p of which p_{sym} are symbolic (categorical).

3.2 Information Theoretic Measures

When analysing symbolic attributes with the help of information theoretic measures, only the distribution of the attributes is of importance. There are however χ^2 -squared measures of association [16]. In contrast to numerical attributes, it is not straightforward to generate relationships between the attributes using measurements. Therefore, each measure is appropriate for one single attribute. Since each data tuple belongs to exactly one of q classes, the class affiliation can be regarded as an additional attribute, the class attribute C . It is of interest to compare the distribution of the target attribute with the distribution of one of the other attributes for each class in order to find possible connections. The following entropy measures can be regarded as the qualitative analogy of a dispersion measure for numerical attributes. The entropy itself is a measure of information of the distribution of a single numerical attribute from the equipartition and traces back to [13]. The basic idea is to interpret the entropy of a probability distribution, for example for an attribute, as the average number of yes/no-questions that are necessary to determine the concrete value of an attribute. For the following measures, let B denote a symbolic attribute with k characteristics, and let C denote the target variable. The attribute entropy of B (as a realisation of a discrete random variable X where q_i is the probability of X taking the i th value) is defined as

$$H_B = H(B) = - \sum_{i=1}^p q_i \log_2 q_i$$

whereas the class entropy of C is

$$H_C = H(C) = - \sum_{i=1}^q \pi_i \log_2 \pi_i$$

indicating how much information is necessary to specify one class. In the latter case, π_i is the prior probability for class A_i .

For each group we suppose the variables are x_1, x_2, \dots, x_p . Let $y_i = x_i/SD(x_i)$, $i = 1, \dots, p$ and compute a kernel density estimate as $\hat{f}_i(y)$ with an automatic choice of kernel bandwidth. Then the entropy measure for each variable is defined by

$$E_i = - \int \hat{f}_i(y) \log(\hat{f}_i(y)) dy, i = 1, \dots, p$$

Then, for group $j = 1, 2$ we have

$$IM_j = \frac{1}{p} \sum_{i=1}^p E_i \quad (1)$$

and, to allow for highly correlated values we also compute

$$IS_j = E^T (X^T X)^{-1} E \quad (2)$$

where $E = (E_1, E_2, \dots, E_p)^T$.

IM and IS are similarly defined, but now $\hat{f}_i(y)$ is calculated for *all* the data (regardless of group membership).

3.3 Statistical Measures

In discriminant analysis, several covariance matrices are calculated in order to be able to determine eigenvalues and eigenvectors. They are calculated by solving the characteristic polynomial $\det(A - \lambda I) = 0$ where A is the covariance matrix of interest and I is the identity matrix. The solutions $\lambda_1, \lambda_2, \dots, \lambda_n$ of this equation are called the eigenvalues.

$$Frac1 = \lambda_1 / \sum_{i=1}^r \lambda_i \quad (3)$$

denotes the relative importance of the largest eigenvalue, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$, and r is the rank of the covariance matrix of the attributes. Note that these eigenvalues are non-negative, so the sum is strictly positive. The above mentioned covariance matrices are fundamental in the theory of linear and quadratic discrimination. When to apply one and not the other is dependent on the homogeneity or otherwise of the covariances [9]. The hypothesis that all populations have a common covariance structure can be tested via Box's M test statistic [7]. When discriminating between two multinormal populations,

the two-sample Hotelling's T^2 test can be used. Then, under the null hypothesis ($\mu_1 = \mu_2$, given $\Sigma_1 = \Sigma_2$)

$$\left\{ \frac{n_1 n_2 (n-2)}{n} \right\} d^T W^{-1} d \sim T^2(p, n-2) = \left\{ \frac{(n-2)p}{n-p-1} \right\} F_{p, n-p-1} \quad (4)$$

and the null hypothesis is rejected for large values of this statistic [7]. Here, $d = \bar{x}_1 - \bar{x}_2$ and $W = \sum n_i S_i$

Tests of Normality To examine if the unknown distribution function $F(x)$ matches a hypothetical distribution function $F_0(x)$, the Kolmogorov-Smirnov goodness-of-fit test is appropriate. The major assumption is that the distribution $F_0(x)$ is a continuous function. If this assumption is not met, the test is conservative, i.e. adheres to the hypothesis of equality longer than appropriate. For "small" sample sizes, this test is more appropriate than the χ^2 -square test since the latter is only working approximately [14]. The hypothesis

$$\begin{aligned} H_0 : F(x) &= F_0(x) \quad \forall x \\ \text{vs. } H_1 : F(x) &\neq F_0(x) \text{ for some } x \end{aligned}$$

is tested by comparing $S_n(x)$, the empirical distribution function of the n observations ($S_n(x) = I(x_i < x)/n$), with $F_0(x)$. The test statistic is given by $\sqrt{n}D_n$, where $D_n = \sup_x |F_0(x) - S_n(x)|$ specifies the largest vertical difference between the hypothetical and empirical distribution function. Critical values can be found in [6] or [11].

In the StatLog project [9], the normal distribution was tested by using skewness and kurtosis statistics. These statistics measure how far the present distribution is different from the normal distribution in respect of asymmetry (measured by skewness) and tail thickness (measured by kurtosis), respectively. The classical measure of skewness for the distribution of a real-valued random variable X is defined as

$$\gamma_1 = E \left(X - E(X) / \sqrt{\text{var}(X)} \right)^3 . \quad (5)$$

If the skewness is larger than zero, one is dealing with a so-called right-skewed distribution, if it is less than zero with a left-skewed distribution. For a symmetrical distribution such as the normal distribution, it is equal to zero. The kurtosis of the distribution of a real-valued random variable X is defined as

$$\gamma_2 = E \left(X - E(X) / \sqrt{\text{var}(X)} \right)^4 . \quad (6)$$

If the kurtosis assumes values larger than 3, the kurtosis of the present distribution is above that of the normal distribution, in case it is less than 3 below that of the normal distribution [1].

3.4 Choice of Measures

Many of our dataset measures have been frequently used in other studies in meta-learning [9] [8]. The selection of measures was based partly on the fact that the simulated data was primarily real-valued (rather than qualitative), and also because of what was readily available - or easily programmable - within the statistical package R (with which the meta-dataset was constructed) [4]. The measure T^2 (based on Hotelling's statistic) was chosen since it was hoped this would provide a measure of the overall noise in the data. More specifically, if two classes have the same covariance (as determined by Box's M-statistic), then Hotelling's T^2 test is essentially a multivariate t-test and so is used to decide if the means of the two groups are (un)equal. If the measure T^2 is small (and hence the means are very similar), then it suggests a large amount of noise in the data, and so we would expect to obtain high error rates from any classifier. Note that, if Box's M statistic is large (meaning that the covariance matrices are unequal), or there are a large number of binary attributes (which means that the data are not normally distributed) then the reliability of the T^2 measure to predict noise is reduced. In this paper, we mainly follow the **StatLog** approach regarding the choice of measures [9]. We found that some measures were helpful for building most of the models. In particular, Box's M statistic, Hotelling's T^2 (Equation 4), and the functions of the univariate entropy measures (as described in Equations 1 and 2) proved to be useful.

3.5 Regression Measures

There are some important decisions concerning how to assert the generalisation error of the methods. While in classification the 0/1 loss function is very common (except in medical diagnoses, and credit-scoring applications), within the regression setting things are not so clear. Errors in regression are metric, meaning that the amplitude of the difference between the true target variable value and its predicted value is relevant [15]. As one error measurement, we used the normalised error/residual mean square

$$NMSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (7)$$

where the numerator displays the squared distances between the actual and the predicted values, and the denominator describes the squared differences between the actual values and the mean of the observations. The $NMSE$ measure can be regarded as a performance ratio between the regression algorithm and the simplest model that would be to always predict the average training set target variable value. Good scores should be between 0 and 1 (closer to 0 is better), while values above 1 would basically mean that the regression algorithm is performing even worse than the simple average.

The other error measurement we used was the normalised measure for absolute differences, the normalised mean absolute deviation

$$NMAD = \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{\sum_{i=1}^n |Y_i - \nu_Y|} \quad (8)$$

where ν_Y is the median target variable value. The use of the median is justified by the fact that this is the constant that minimises the absolute deviation. The advantage of MAD, when compared to MSE, is that it is expressed in the same unit as the target variable, thus having a more intuitive meaning. As a matter of fact, both measures are incomparable in a sense that one method might be preferred to another in terms of MSE and vice versa when using MAD [15].

4 Classification versus Regression for Meta-Learning

Some meta-data were created in a large simulation study using datasets with two classes. The meta-data comprised one observation for each dataset. The observations consisted of the number of variables (both total and symbolic), the number of elements in each group, their skewness (Equation 5) and kurtosis (Equation 6), the proportion of the first eigenvalue for each group as well as for the pooled data (3). It also consisted of Box’s M-statistic and Hotelling’s T^2 statistic (as described in Equation (4)), the Kolmogorov-Smirnov statistic to examine normality as well as of functions of univariate entropy measures. Additionally, for each simulated dataset error rates for a leave-one-out cross validation were determined for the following classification algorithms: linear discriminant (LDA), quadratic discriminant (QDA), and 1-NN.

In total, the dataset consisted of 5450 observations with 25 attributes of continuous nature and one class attribute. The sample sizes for each individual observation ranged from 110 to 2000, the number of variables from two to ten of which a maximum of five were symbolic. In 56% of the cases, QDA produced the lowest error rate compared to 42% of LDA and 2% of 1-NN. The dataset is available by anonymous ftp from “amsta.leeds.ac.uk” in the directory “pub/users/charles/metal”.

Figure 1 gives a graphical overview of how the problem was approached and also shows the different meta-levels of learning. For each dataset, error rates for various algorithms are obtained leading to meta-errors for both regression and classification errors. Note that for each dataset a certain amount of data characteristics, say k , are calculated. So, each of the m sets of data characteristics on the first meta level consists of k values of which each is corresponding to a certain data characteristic.

For simplicity, suppose that the performance of an algorithm is measured by a single variable (say the error rate e_{ij} which corresponds to algorithm i on dataset j). In prediction, we would seek to predict the error rate of an algorithm using all the available measures. In classification, we would seek to predict the best algorithm using all the available measures. The classification approach will

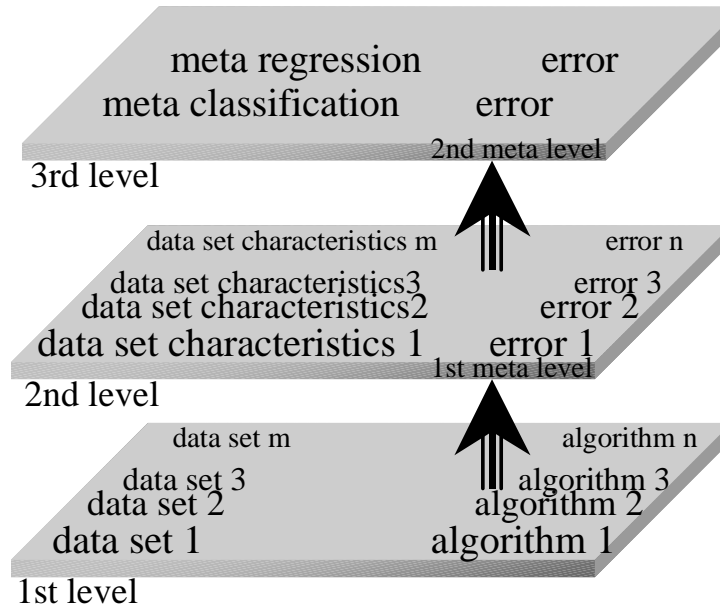


Fig. 1. From Data Characterisation for Meta-Learning to Meta-Regression

overlook the fact that, in some cases, the performance of two methods will be very similar, and this information will be lost by merely recording which is best. To illustrate this point, we offer the following example:

In a more comprehensive study we would also measure other performance indicators such as storage costs, CPU time, comprehensibility, but for simplicity suppose that the performance on a dataset is linearly related to a measure, say x (which could be a landmarker or data measure). Algorithm i may give an error rate (as a percentage) of the form $e_1 = 1 + x$ and $e_2 = 0.5 + 2x$ where we suppose that x lies on a range from 0 to 1. Additionally, we suppose that both of these performance error rates are observed with observation error. A classification meta-dataset will consist of pairs $(x, class)$ (where class takes the value 1 or 2 according as $e_1 < e_2$), whereas a regression meta-dataset will consist of observations (x, e_1, e_2) . Given a new dataset with data measure x_0 we can use a classification rule (linear discriminant) which has been learned from the classification meta-dataset. Alternatively, fitting a separate regression function for each of the error rates, then selecting the algorithm which gives the smallest predicted error using measure x_0 gives another means of classification.

In order to give an indication of the difference between these two approaches, we simulated 100 replicates of sample size 100 from the above models, and for each replicate we simulated a further 100 values as the test data on which we calculated the error rate. In order to check the robustness of the methods to outliers, we carried out experiments when the residual errors were normally

distributed, and for errors following a t-distribution with 20 and 2 degrees of freedom. We also investigated violations in the model assumptions by simulating data from a non-linear model with $e_1 = 1 + x + 2(x - 1/2)^2$, and $e_2 = 0.5 + 2x + 2(x - 1/2)^2$, and normally distributed errors, but continue to carry out predictions using a linear regression.

	Classification	Prediction	p -value for paired test
normal errors	21.1	15.5	0.002
t_{20} errors	16.9	16.1	0.567
t_2 errors	29.9	34.5	0.116
quadratic model	18.3	14.2	0.007

Table 1. Percentage of error rates

As can be seen from Table 1, prediction is superior when the model is correct and the errors are normally distributed. However, when there are outliers created by a thick-tailed distribution, then the classification approach is better. In general, we believe that when the model is specified correctly and the errors are reasonably normally distributed, then the prediction approach will perform better than that of classification. However, in the case of real meta-data, when the model is unknown and will generally be highly non-linear, and the errors are probably asymmetric, then classification should be better, unless suitable transformations are applied.

5 Empirical Results

Three regression models - one for each error rate of LDA, QDA, and 1NN - and one classification model - choosing the algorithm with the lowest error rate- were created using M6 and C5.0, respectively. Additionally, another regression model was built using the lowest error rate of the algorithms. By doing so, a certain comparability of the results was given. The goal was to assess error rates to each algorithm for predicting its own error rates. The calculated error measurements were MAD, MSE, NMAD, and NMSE, respectively. The results of the error rates for each of the models were obtained by ten-fold cross-validation and are presented in Table 2. Note however that each value does not represent the error rate for the corresponding algorithm. What it does represent is actually the average error rate in predicting the error rate for the actual algorithm.

Table 2 shows that LDA has the highest error rate in predicting the actual error rate for the algorithm itself for both MAD and MSE whereas QDA and 1NN don't seem to differ significantly. Note, however, that 1NN only provided in 2% of the meta-data the lowest error rate for classification.

For classification of the algorithm with the lowest error rate using just the lowest error rate, an error rate of 0.123 was obtained whereas the regression

	MAD	MSE	NMAD	NMSE
LDA	0.097	0.0098	1.008	0.631
QDA	0.066	0.0044	0.876	0.400
1NN	0.053	0.0055	0.579	0.397

Table 2. Error rates for LDA, QDA, and 1NN

approach for predicting the lowest error rate delivered 0.032 for absolute and 0.002 for squared deviations. That basically means that the regression approach delivers a more exact idea of the error measurement. When trying to predict the class that gives the lowest error rate by using all three predicted error rates of LDA, QDA, and 1NN, the then obtained error rate was 0.084. By using informative error rates, we got better results than just by comparing algorithms since the latter always means that information is “thrown away”.

The following figure displays the general approach and shows how different regression and classification models can be obtained. Note that in this context each regression problem can be transformed into a classification problem and vice versa just by connecting the lowest error rate of a data characteristics to the algorithm that produced it.

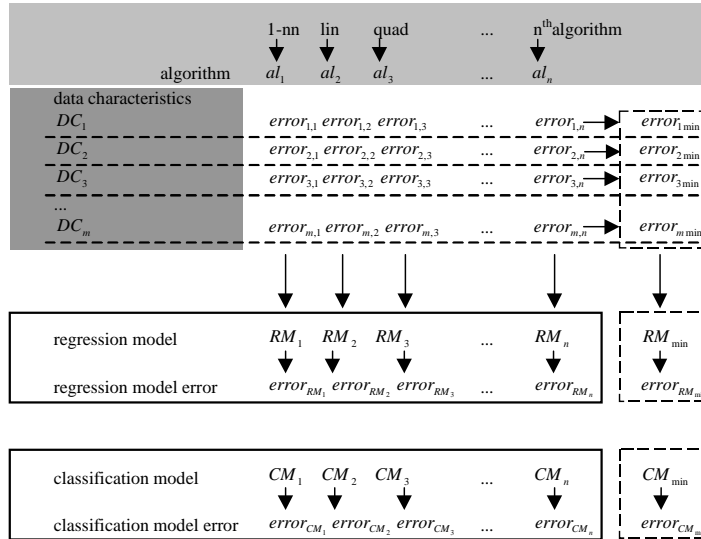


Fig. 2. General Approach to Meta-Regression

6 Conclusions and Future Work

In this paper, we address the problem of connecting data characterisation for meta-learning to regression. Several measurements were introduced that built the basic of our data characteristics as well as the reason why we chose these measurements was explained. Part of our aim (and any future study's aim) was to determine which variables were useful. A very helpful tool in order to improve on estimates might be to take combinations of variables and transformations. This is surely a challenge for future work.

Assessing the quality of any method by a 0,1 loss function implies a lack of information when compared to the regression settings.

The use of simulated data may not sufficiently cover the space of a (previously unseen) real dataset. In practice, simulations will have to be used since the dimensionality of the data characteristics is large, and there are simply not enough real datasets at present. It would be interesting to consider how to simulate datasets which are "similar to" a given real dataset in the sense that they encompass the space around.

The right mixture of landmarks and information theoretic measures will be the subject of further investigations as well.

Acknowledgements

The authors would like to thank Reza Nakhaeizadeh and Luis Torgo for fruitful discussions. The research was supported financially by EC METAL project (ESPRIT # 26.357) and DaimlerChrysler.

References

1. Becker, B. (1993). Statistik, Oldenbourg, Munich, Germany.
2. Bensusan, H. and Giraud-Carrier, C. (2000). Casa Batló is in Passeig de Gràcia or Landmarking the Expertise Space, ECML-2000 Workshop on Meta-Learning.
3. CRoss-Industry Standard Process for Data Mining, <http://www.crisp-dm.org>.
4. Ihaka, R. & Gentleman, R. (1996). R: A Language for DataAnalysis and Graphics, *Journal of Computational and Graphical Statistics*, 5: 299-314.
5. Jones, L.V. and Fiske, D.W. (1953). Models for testing the significance of combined results, *Psychological Bulletin*, (50)5:375-382.
6. Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association*, Vol.62, pp.399-402.
7. Mardia K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press, London, UK.
8. EC ESPRIT MetaL project # 26.357 (1998-2001). Partners: Universities of Bristol, Geneva, Vienna, and Porto, DaimlerChrysler, and Integral Solutions, Ltd., <http://www.cs.bris.ac.uk/cgc/METAL/>.
9. Michie, D., Taylor, C., and Spiegelhalter, D. (eds.) (1994). *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, London, UK.

10. Nakhaeizadeh, G., Reinartz, T. and Wirth, R. (1998). Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick, in *Data Mining*, Nakhaeizadeh, G. (edt.) , Physica, Heidelberg, Germany.
11. Pearson, E.S. and Hartley, H.O. (1972). Biometrika tables for Statisticians II, Cambridge University Press, London, UK.
12. Pfahringer, B., Bensusan, H. and Giraud-Carrier, C. (2000). Meta-Learning by Landmarking Various Learning Algorithms, to appear in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML-2000.
13. Shannon, C.E. (1948). The Mathematical Theory of Communication, The Bell Systems Technical Journal (27):379-423.
14. Slakter, M.J. (1965). A comparison of the Pearson Chi-square and Kolmogorov Goodness-of-fit-tests with respect to validity, Journal of the American Statistical Association, Vol.60, pp.854-858.
15. Torgo, L. (1999). Inductive Learning of Tree-Based Regression Models, PhDthesis, University of Porto, Porto, Portugal.
16. Yang, Y. and Pedersen, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization, In *Proceedings of the 14th International Conference on Machine Learning*, pp.412-420, Morgan Kaufman.