The Focused Multi-Criteria Ranking Approach to Machine Learning Algorithm Selection - An Incremental Meta Learning Assistant for Data Mining Tasks

Iain Paterson[1], Helmut Berrer[2] and Jörg Keller[2]

[1] Institute for Advanced Studies (IHS),
Department of Economics and Finance, Stumpergasse 56, A-1060 Vienna, Austria. Tel.+43 1 59991/152 - fax 163
paterson@ihs.ac.at

[2] DaimlerChrysler AG, Research & Technology, FT3/AD, P.O.-Box 2360, D-89013 Ulm, Germany
{helmut.berrer, joerg.keller}@daimlerchrysler.com

Abstract. The main goal of the ESPRIT METAL project is to use meta-learning to develop an incrementally adaptable assistant system to provide user-support in machine learning and data mining. Meta data consists of performance outcomes of ML algorithms on known datasets. Using new models of data envelopment analysis to deal with multiple criteria, an ordered ranking of algorithms is obtained for each dataset. In order to suggest a specific machine learning solution for a given application, characteristics of datasets are compared. Meta knowledge of predictive reliability combined with a nearest neighbour approach provides a means for incremental learning that is focused on the most relevant information. The meta-knowledge base is extended dynamically and can hence adapt. User experiences can be taken into account by feedback of the quality of suggested solutions into the focusing process. Results from training and testing trials are presented and informational gains are evaluated.

1 Introduction

The automated assistant system for knowledge-discovery and data-mining, which is currently being developed in the international multi-partner ESPRIT METAL project[1] (http://www.metal-kdd.org/) provides advice as guidance for choosing the appropriate machine-learning algorithms to accomplish the data-mining process. To have

[1] ESPRIT METAL Project (26.357):
Title: A Meta-Learning Assistant for Providing User Support in Machine Learning and Data Mining
Goals: To develop a Knowledge Discovery Assistant that recommends a combination of pre-processing, classification and regression techniques, is adaptive and learns from new data mining tasks, handles resource/quality trade-off, and gives explanations for its recommendations

model(s) selected as being best suited is a great boon for the inexperienced user with little or no experience of commercial software packages for data-mining.

In order to provide advice to potential users of data mining techniques as to which machine learning classification and regression algorithms may be best suited to their particular task in hand, the assistant uses, on the one hand, accumulated knowledge of the performance of various algorithms on known data sets, and characteristics of the new (target) data set to be analysed on the other. In its basic form the advice provided is a ranking of algorithms, in most-preferred order. Meta data attributes like, e.g. algorithm accuracy, and computer storage and total runtime (training and/or testing) may be taken into account in the ranking.. The ranking concept draws on the concept of *focussing* on the most suitable neighbouring datasets, which are assumed (and where verifiable, empirically shown to be) reasonably similar with respect to the performance of certain data mining algorithms on them. 'Suitability' is a combination of reliability of ranking advice– which in turn was shown to be related to similarity of data characteristics -and user-assessed quality (feedback from previous advice). Section 2 describes the levels of meta data needed for ranking and focusing while section 3 details how the modelling concept is applied dynamically, by inclusion of new datasets whose characteristics are known, but without prior knowledge of the performance of machine learners on them.

The focussing process is incremental: with each new dataset the knowledge base for proffering advice is extended. The first stage of validation is, however, to choose the parameters for the most effective ranking model for the 'static' case of known datasets (documented research on the METAL repository), given 100% reliability and quality of this data. In this case the focused reference datasets are the nearest neighbours of the target dataset. Results are detailed in section 4. Finally, a description of the DEA ranker is given in section 5. This represents a novel use of data envelopment analysis for multi-criteria ranking, for which the models were specially designed. From a data mining perspective, however, the method of ranking is less important than its success, in terms of the quality of advice it provides. Extensive assessment of the DEA ranker has been reported previously (Holzer 1999, Berrer et. al. 2000, Keller et al. 2000).

2. Meta-level Data used in the Focussing and Ranking Process

A dataset (DS) can have three associated meta data files concerning its description - MD1, the results of the experiments with the ML algorithms - MD2, and the rankings of algorithms- MD3. Metadata are defined as follows:

MD1: Statistical and information characteristics - properties such as skewness, kurtosis etc (for the classification task, regression datasets yield less measures)., as well as other measures of information content, and basic features such as the number and kind of attributes.

MD2: Performance results of ML Algorithms – for each of the datasets in the repository details of the following variables have been documented:

- *Accuracy* rate (1 - error rate) – average for the learner on the dataset
- *Hypsize* - the average model size for learning algorithm

– *Traintime* - the average CPU time needed for training a model for learning algorithm
– *Testtime* - the average CPU time needed to evaluate a model on the test sets for learning algorithm

Some or all of these result variables may be used to determine a ranking.

MD3: Ranking result - an ordered list of algorithms, with scalar 'score' attached. A ranking that is obtained using full knowledge of the MD2 for the dataset is called an *ideal ranking*.[2] In real application of the assistant by a user, as much accumulated knowledge as is available of past results (i.e. MD1, MD2, and MD3) will be used to make a *recommended ranking* for a target dataset.

Here we present an integrated concept that addresses these three meta data levels (Fig.1). The aim of our ranking procedure is to indicate which particular (classification or regression). algorithm is - likely to be in terms of previous knowledge - best suited for the particular target dataset. The basic idea is that the new dataset exhibits particular characteristics, which have already been described at meta level MD1. Based on the premise that 'similar' datasets will be similarly tractable when classified by ML algorithms, we wish to identify a set of such similar datasets to the target dataset. (The results obtained in EC ESPRIT Project *StatLog* by Michie et al (1994). indicated that particular algorithms are best suited to particular types of datasets.)

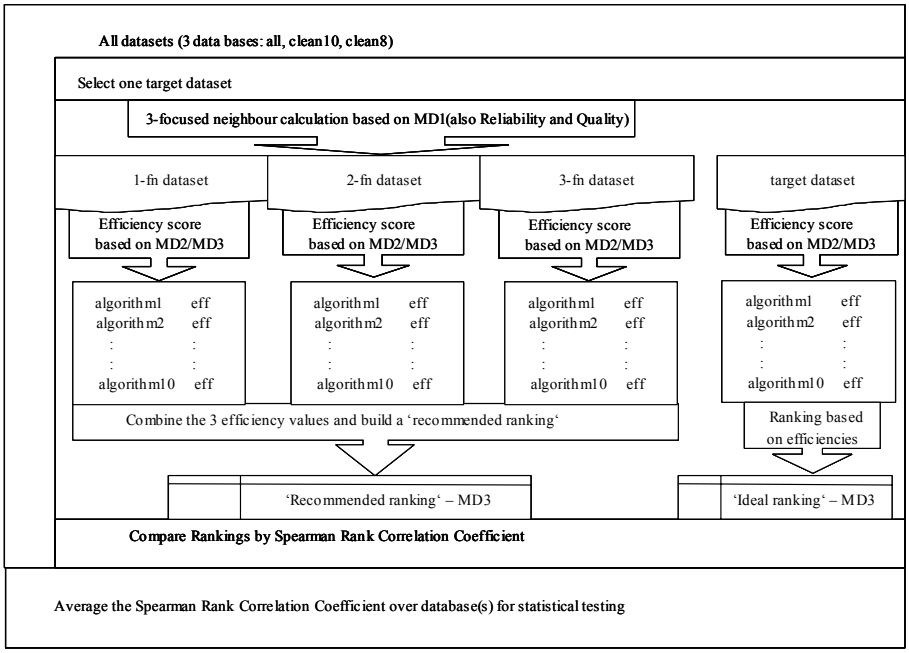

Fig. 1. An integrated concept for multi-criteria ranking of data mining-algorithms

[2] Note that 'ideal' refers to the ranking obtained by a particular model using MD2 information; the question as to which ideal ranking is best among alternative models is open.

One particular technique for identifying similar datasets in terms of their meta-characteristics is k-nearest neighbours. Although this technique is itself one of the classification ML algorithms there is no obvious reason to suppose that its use as a 'comparator' would cause a systematic bias for or against this algorithm in the ranking phase. Usually 3-nn has been used, within variations of the following generic distance function:

$$dist(x,y) = \left(\sum_{i=1}^{m} weight\_parameter(i) * \left| \frac{x(i) - y(i)}{Stdev(i)} \right|^{exponent(i)} \right)^{\frac{1}{2}} \quad \textbf{(1)}$$

The *focused distance* is a modification of 3-nn distance that takes reliability and quality factors into account (see next section), so that focused neighbours are not necessarily nearest to the target dataset. Now we can use the meta knowledge which has been acquired in METAL regarding the performance of ML algorithms on the three focused-upon datasets. If a focused neighbour dataset belongs to the repository its MD2 data is used to obtain an ideal ranking; if meta knowledge of the focused neighbour dataset has been accumulated during real time use of the system, we can not draw on MD2 but instead use the MD3 information. In this case the previously recommended ranking of a focused-upon dataset is used as a quasi-ideal ranking. The 'cost' of using this *ersatz* information is expressed by the reliability measure. User feedback on a previously recommended ranking may be incorporated by means of a quality measure. The new recommended ranking for the target dataset is obtained by a combination of the ranking scores of the 3 focused neighbours - tests have shown weighted averaging to be generally superior to simple averaging. This completes the focused ranking process for one iteration (i.e. target dataset) and augments the cumulative MD3 meta knowledge base.

3. Model Description of Incremental Focused Ranking

At the beginning of the process the set *C* of repository datasets (or a subset of these) is selected as being *canonical* i.e. they typify established meta knowledge. For these datasets ideal rankings exist and they are usually assigned both a full reliability and a quality value of 1. Subsequent calculation of rankings for target datasets *C+1, C+2,* etc. are based on previous knowledge, so that the recommended ranking of dataset *C+t* is derived under prior knowledge of *C+t*-1 datasets.

The main difference between the concept for ranking canonical datasets with the real application lies in the fact that any new dataset is previously unseen, and there exists no MD2 metadata for it. We are however able to obtain its MD1 meta data by applying the data characteristics tool (DCT) as usual, so that we can assess its nearest neighbours among all datasets, whether these are canonical or recent real-application additions, for which MD3 meta data exists (but no MD2). By investigating the relationship between k-nn distance and degree of correlation between recommended and ideal rankings (for canonical datasets) we are able to construct a reliability

function in order to characterise the a-priori quality of results using the quasi-ideal ranking approach.

In order to avoid future recommended rankings being of low quality, due to successive estrangement from the set of canonical datasets, we build in to the process a feedback loop. Specifically, we calculate a (0-1) *reliability index* based on the focused distance (defined below) from other datasets. This is an objective measure although it is essentially, as is k-nn distance, a construct. We may, however, combine this measure with a subjective measure of the user's experience with the recommended ranking, by introducing a (0-1) *quality index*. We are now interested in finding the *most-suitable* or *focused* datasets, denoted (k-fn), which may not always be identical with the k-nn nearest neighbours.

The k-nn distances of dataset $C+t$ from other datasets is given by

$$dist(j) \quad \forall j = 1,\ldots,C+t-1 \tag{2}$$

The *focused distances* are obtained by modifying these distances by previously known values of reliability (above a certain threshold with parameter p) and quality measures thus

$$\tilde{d}ist(j) = \begin{cases} \dfrac{dist(j)}{Rel(j) * Qual(j)} & \forall Rel(j) \geq Rel_P > 0, Qual(j) > 0 \\ \infty & \forall Rel(j) < Rel_P \text{ or } Qual(j) = 0 \end{cases} \tag{3}$$

This implies that datasets with reliability below the threshold or that have been judged by the user as zero quality will never be used as references, i.e. will not be focused upon. Assuming now that the indices i are arranged in order of increasing focused distance, then a weighted average for the *k focused neighbours* is obtained from

$$\tilde{w}eight(i) = \frac{\dfrac{1}{\tilde{d}ist(i)}}{\sum_1^k \dfrac{1}{\tilde{d}ist(i)}} = \frac{\dfrac{Rel(i) * Qual(i)}{dist(i)}}{\sum_1^k \dfrac{Rel(i) * Qual(i)}{dist(i)}} \tag{4}$$

$$\forall i = 1,\ldots,k$$

and the vector (length = # of algorithms) of efficiency scores is calculated from equation 5, where ideal rankings are denoted by $Eff(DS_i)$ and recommended rankings are denoted by $\tilde{E}ff(DS_i)$ respectively.

$$\tilde{E}ff\left(DS_{C+t}\right)=\sum_{i=1}^{k}\left(\tilde{w}eight(i)*\hat{E}ff\left(DS_i\right)\right) \quad (5)$$

where

$$\hat{E}ff\left(DS_i\right)=\begin{cases} Eff\left(DS_i\right) & \text{if} \quad DS_i\in\left\{DS_l : l=1,\ldots,C\right\} \\ \tilde{E}ff\left(DS_i\right) & \text{otherwise} \end{cases}$$

Rank order of ML algorithms is determined directly by sorting the elements of $\tilde{E}ff\left(DS_{C+t}\right)$ by decreasing value.

The reliability function is in general set to

$$Rel(n)=\begin{cases} 1 & n=1,\ldots,C \\ \alpha+\sum_{1}^{k}\beta_i*\tilde{d}ist(i) & \text{otherwise} \end{cases} \quad (6)$$

for chosen parameters $0<\alpha\le 1$ and $\beta_i<0, i=1,\ldots,k$. The justification for this reliability function is empirical – the correlation between recommended and ideal rankings was shown to decrease with distance from the target dataset.

Further, quality may be set for each dataset by the user to a value

$$0\le Qual\le 1 \quad \text{with default } 1 \quad (7)$$

At each stage of the incremental process, a recommended ranking of ML algorithms is thus achieved for a target dataset by focusing on its 'most-suited neighbour' datasets, and a reliability for this prediction is estimated. The user may now assign a quality index to this recommendation of the assistant which will be taken into account in all subsequent recommendations for new datasets.

4. Validation Experiments - Summary of Results

Tests of the focusing process have been carried out under a series of experimental designs in order to validate its various components step-by-step. The methodology for assessing validation is outlined below. In the first stage results from different DEA ranking specifications (radial, additive, variable returns-to-scale, constant returns-to-scale etc. – c.f. section 5) were obtained for all appropriate datasets of the repository. These results deliver the ideal rankings of algorithms for each dataset. Secondly the setting of parameters for the k-nn procedure was investigated. With this knowledge,

preferences for choice of ranker specification were established. Overall results for the performance of algorithms on the entire repository of datasets were obtained.

The steps thus far carried out pertain to a 'static' or fixed set of datasets. Since the user is ultimately concerned with the quality of a recommended ranking for a new dataset outwith the repository, such a situation is to be simulated. Target datasets (from the repository) are selected in blind tests, using a reduced repository subset to 'train' the incremental learning/prediction. To mimic real usage of the meta learning assistant, some meta knowledge (concerning MD1 and MD3) can be added to the initial repository-related meta knowledge in order to facilitate further new predictions of algorithm performance. In this dynamic phase of prediction the incremental learning aspects of focused ranking, involving reliability estimates and quality feedback can be tested by simulation in comprehensive trials.

Experimental Data

The repository build up by the METAL consortium partners consists of 53 datasets, These are referred to as the canonical datasets, because they are assumed to be representative of the universe of datasets. The MD2 database comprises performance information of 10 ML algorithms for each dataset: *c50boost*, *c50rules*, *c50tree*, *clemMLP*, *clemRBFN*, *lindiscr*, *ltree*, *mlcib1*, *mlcnb* and *ripper*. Unfortunately there some occurrences of missing values in the MD2 database. For this reason two alternative characterizations of the database are in addition examined: *MD2_clean10* possesses no missing values for any of the 10 algorithms, but the number of datasets reduces to 34. Due to the fact that two algorithms are very susceptible to missing values, viz. *clemMLP* and *clemRBFN*, and to raise the number of cases, *MD2_clean8* was established without these. It consists of 48 datasets, but now with the restriction of treating only 8 different algorithms.

Stage 1: Ideal Rankings

The experimental design involves the calculation of efficiency scores $Eff(DS_i)$ for all algorithms applied to each the repository dataset DS_i for a variety of model specifications.

If it were the case that there existed *n* equivalently performing ML algorithms, the

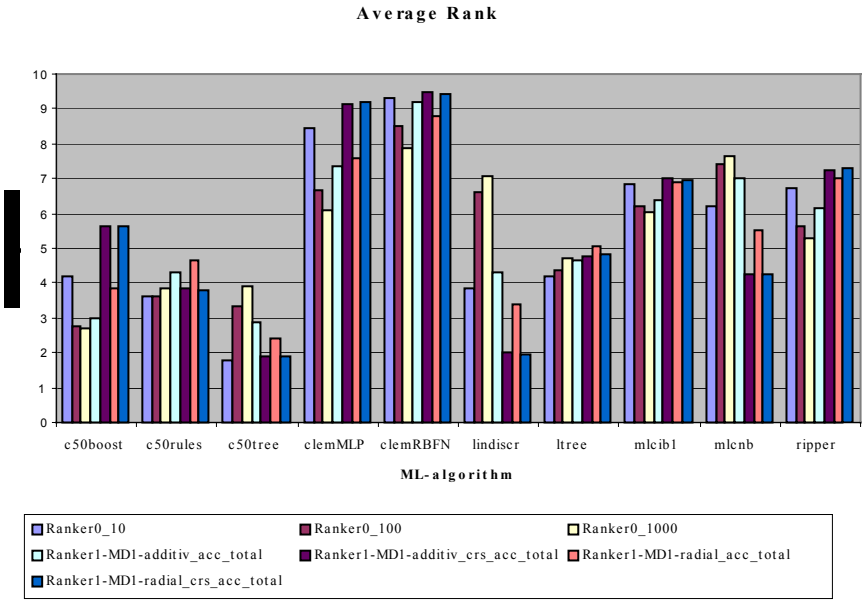

Fig. 1. Performance of Data Mining Algorithms

average of ranks obtained in ideal rankings for each would be spread closely around *(n+1)/2*. Looking at Figure 2, which shows in condensed form the performance of all machine learning algorithms as measured by a) variations of the DEA ranker and b) another ranker designed by the METAL partner LIACC (here referred to as Ranker0), one can see that there are overall both more (*c50rules* and *c50tree*) and less (*clemMLP* and *clemRBFN*) promising algorithms. The two last named algorithms only perform well on a small subset of datasets, or even on only one outlier dataset.

Stage 2: Experimental Design for the k-nn Calibration

To find the best appropriate k-nn setting a trial and test procedure was used. All items concern every dataset in the repository and all rankers and their variants.

- First an ideal ranking is determined for each dataset.
- Different recommended rankings – based on different k-nn settings - are calculated. Recommended rankings $\widetilde{Eff}(DS_i)$ are based on all datasets *except* DS_i. At this stage, all meta knowledge is assumed to be perfectly reliable and of highest quality (i.e. = 1), so that the focusing process reduces to the selection of k nearest neighbours.
- The Spearman rank correlation coefficient between the ideal and various recommended rankings is computed for each dataset.
- The most promising this k-nn setting is identified, that which tends to provide the highest correlation irrespective of the ranker and variant used.

Having determined all ideal rankings of the datasets in the repository, the nearest neighbours of each dataset are identified for each k-nn variant. The basic surmise is, that with increasing distances the Spearman rank correlation coefficient will decrease[3]. This is another factor, besides the overall level of correlation, which determines the merit of the k-nn parameter setting. Although the occurring variation in rank for each particular ML algorithm is limited by its overall effectiveness, it is possible to show at least a small effect on the rank correlation coefficient.

Table 1 shows the average Spearman rank correlations coefficients[4] between the ideal and the recommended rankings. The recommended rankings have been derived using efficiency scores from the 3 nearest neighbour datasets for each of five different choices of MD1 variables, in the case of the column 'leave-one-out' all efficiency values except those from the target dataset are used . The leave-one-out scenario acts here as a null hypothesis., which we would expect to be outperformed by the k-nn method. For the 3-nn version 002 all available MD1 variables are used in calculating distances. The low average Spearman rank correlation coefficients expose this as being a poor choice, but nevertheless it is mostly within range of the leave-one-out threshold. Tests with varying MD1 variable subsets, and varying weight_parameters and exponents in the k-nn formula, have shown the best as yet identified k-nn setting to be the basic version 003, which achieves the highest average (over all 53 datasets) Spearman rank correlation coefficients in 9 out of 12 ranker variant cases (rows of the

[3]Calculate correlations between the distances and the Spearman rank correlation coefficient!

[4] Highest Spearman rank correlation coefficients are marked in bold style.

table). This most promising k-nn parameter set uses only the data characteristics *Nr_sym_attributes*, *Nr_num_attributes*, *Nr_examples* and *Nr_classes*.

		Average Spearman Rank correlation	3-nn weighted (focusing) Variants				
Ranker Variants	database	leave-one-out	3nn-001	3nn-002	3nn-003	3nn-004	3nn-005
ADDITIVEVRS	All	**0.6465**	0.6278	0.5889	0.6152	0.6003	0.5514
ADDITIVECRS	All	0.7334	0.7712	0.7389	**0.7893**	0.7712	0.7225
RADIALVRS	All	0.5546	0.6046	0.5720	**0.6248**	0.5834	0.5264
RADIALCRS	All	0.7824	0.8907	0.8736	**0.8947**	0.8905	0.8751
ADDITIVEVRS	clean8	0.5382	0.5556	0.6171	**0.6190**	0.5724	0.5253
ADDITIVECRS	clean8	0.5625	0.6265	0.6171	**0.6607**	0.6166	0.5804
RADIALVRS	clean8	**0.6507**	0.6107	0.5584	0.6450	0.6107	0.5475
RADIALCRS	clean8	0.8016	0.8586	0.8690	**0.8938**	0.8557	0.8636
ADDITIVEVRS	clean10	0.6444	0.6606	0.5706	**0.6665**	0.6514	0.5295
ADDITIVECRS	clean10	0.7403	0.7789	0.7388	**0.7877**	0.7741	0.7051
RADIALVRS	clean10	0.6510	0.6743	0.6366	**0.7158**	0.6600	0.6260
RADIALCRS	clean10	0.9126	**0.9522**	0.9456	0.9515	0.9486	0.9394

Table1. Average Spearman rank correlation coefficients between ideal and recommended rankings for various DEA ranking specifications and k-nn parameters.

Stage 3: Ranker Variant Preference

After deciding upon a k-nn parameter setting, we can examine the different ranker variants in detail. The same level of focusing - i.e. k-nn and efficiency calculations with a fixed number of datasets – applies as in stage 2. Table 2 gives an extract of all 53 Spearman rank correlation coefficients between the ideal and the recommended ranking for the selected k-nn version 003. The table additionally shows results for unfocused ranker versions that use simple averaging of efficiency scores from k-nn datasets (similar to the 'zooming' process used by LIACC).

In the extended focusing concept (incremental learning) we will need recommended rankings that are 'good' enough to be used quasi-ideal rankings, thus high rank correlation coefficients between the recommended and ideal rankings in this test phase are desirable. Statistical tests (Wilcoxon signed rank test) identified *Ranker1-fradial_crs* (i.e. the focused radial version) with an overall average Spearman rank correlation of 0.8947 as best, followed by *Ranker1-radial_crs* (the zoomed radial version). In general the constant return to scale formulations of the Universal Models possess higher 'robustness'– correlation between ideal and recommended ranking - and the focused variants are generally more effective than the unfocused variants due to the proportionately higher weight given to the scores of 'nearer' datasets in focusing.

Dataset	Ranker1-additiv	Ranker1-radial	Ranker1-additiv_crs	Ranker1-radial_crs	Ranker1-fadditiv	Ranker1-fradial	Ranker1-fadditiv_crs	Ranker1-fradial_crs
DS1_abalone	0.188	0.200	0.552	0.927	0.188	0.200	0.564	0.927
DS2_acetylation	0.188	0.212	0.818	0.939	0.200	0.212	0.818	0.927
DS3_adult	0.467	0.430	0.758	0.952	0.467	0.430	0.758	0.952
DS4_allbp	0.806	0.976	0.976	0.976	0.806	0.976	0.976	0.976
DS5_allhyper	0.976	0.988	0.988	0.976	0.976	1.000	0.988	0.988
DS6_allhypo	0.976	0.988	0.988	0.988	0.976	0.988	0.988	1.000
DS7_allrep	0.939	0.879	0.952	1.000	0.939	0.879	0.952	1.000
DS8_ann	0.600	0.842	0.636	0.867	0.600	0.842	0.636	0.867
DS9_Byzantine	0.055	0.382	0.770	0.939	0.055	0.297	0.770	0.939
DS10_c_class	0.333	0.091	0.976	0.964	0.430	0.115	0.976	0.964
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
DS53_yeast	0.600	0.745	0.309	0.952	0.382	0.842	0.309	0.939

Table2. Spearman rank correlation coefficient for different variants of the ranker.

'Robustness'– the correlation between ideal and recommended ranking - is one way to screen different rankers and their variants for their effectiveness. It has to be recognized that this procedure measures principally the internal consistency of each model. While this is in itself a very desirable property for a ranker, it is not to be regarded as a sole arbiter of success. There is no single criterion for assessing the performance of multi-criteria rankers, however. If it did exist, then a ranker could probably be designed to achieve optimal performance. However, in view of the *No Free Lunch* Theorem, we do not expect to be able to show optimality, because there are always datasets to be found that will yield contrary results when one particular ranker is applied. However, the balance of empirical evidence may be employed to favour one ranker or another. One approach towards independently evaluating the range of effectiveness of 2 dimensional rankers based on accuracy and time, in order to discriminate between them, is the methodology of Discounted Information Flow (Berrer, Paterson and Keller, 2000).

Stage 4: Experimental Design for Testing the Incremental Focusing Concept

After setting appropriate k-nn parameters and choosing a promising ranker variant, the next development is to examine the incremental aspects of the focusing concept. The purpose of validation in this stage is to assess real application of the ranking assistant by a user. Since only repository datasets are available at present, a 'training and testing' experimental design is used to simulate real application

The whole dataset repository is separated into *k* canonical datasets and *(53-k)* target datasets. The choice of the canonical datasets is random in every run and also the number *k* is varied. For the canonical datasets, the ideal ranking based on the known MD2 information is used and for the target datasets the MD3 information is built up in successive increments according to the focusing concept described in section 3. After each run the quasi-ideal rankings (i.e. recommended rankings) of the target datasets are compared to the true known ideal rankings in a similar fashion to stage 3. However the appropriate null hypotheses for this stage are $ideal_k$ rankings based only

on the corresponding *k* canonical datasets, not on the 'full information' of 52 datasets of the leave-one-out procedure.

Other parameters, of reliability and user quality are then introduced successively in the simulation experiments in order to validate their usefulness. Parameters to be tuned at this stage include the coefficients of the reliability function and the reliability lower threshold. User feedback can be simulated by including randomly generated estimates of quality. In this stage, as in others, justification of a model component depends on empirical results tested against an appropriate null hypothesis. Currently scripts for this stage are being tested, and the full results will be presented at ECML/PKDD 2001 in Freiburg.

5. The DEA Approach to Multi-criteria Ranking

The possibility of using DEA for ranking algorithms was first suggested Nakhaeizadeh et al. (1997) A partial ranking is achieved in DEA by comparing the (in)efficiency of the decision making units (DMUs). In the terminology of METAL, each DMU is a data-mining algorithm, which has positive - i.e. 'more' means 'better' - or negative - i.e. 'more' means 'worse' - attributes (like accuracy, time etc.). These variables (unrestricted in the number of each) are the output and input components, respectively.

Some DEA approaches to productivity measurement and ranking of units depend on measuring the 'distance' of a DMU (Decision Making Unit) to the convex hull spanned by *other* DMUs; calculating Malmquist productivity indices is one case, 'superefficiency' measurement (Andersen and Petersen 1993) is another. Super-efficiency provides a methodology for ranking efficient DMUs. Standard input- or output-oriented DEA models exhibit deficiencies when applied to this task. Superefficiency is, however, only well-defined for CRS (constant returns-to-scales) specifications in the traditional DEA approaches - but not for variable returns-to-scales, and such models are units invariant but not translation invariant.

Novel DEA models have, however, been developed by Paterson (2000) which extend the concept of measuring superefficiency to the useful case of variable returns-to-scales (VRS), thus enabling the complete ranking of efficient as well as inefficient units The invariance properties of the objective function introduced by Lovell and Pastor (1995) are utilized. Two of these models also have strong invariance properties.

1. *Universal Radial*: This model is units and translation invariant (also for slacks) for the VRS specification: input or output data may thus assume negative or zero values. It is units invariant (also for slacks) for non-negative data in CRS.
2. *Universal Additive*: This model is units and translation invariant (slacks being included in the efficiency score) for the VRS specification: input or output data may thus assume negative or zero values. It is units invariant for non-negative data in CRS. Unlike the universal radial model, it exhibits discontinuity in the (super)efficiency values along the weak efficient boundary, so the former model may be preferred for ranking purposes.

For both of these models a method for normalising the efficiency scores has been devised, so that inefficient units obtain an 'efficiency' value less than one, weak efficient units obtain a value of one, and superefficient units obtain a value greater than one. In addition to the script used in METAL, these models have been implemented in a windows software package DCRanker. and applied to the ranking of units in an industrial context. In this paper the Universal Radial Model is presented in more detail.

The basic terminology is as follows:

Y is the k×n matrix for k outputs and n DMUs

Y_{-d} is the k×(n-1) matrix for k outputs and n-1 DMUs (i.e. excluding DMU d)

Y^d is the k×1 output vector for DMU d being evaluated

X is the m×n matrix for m inputs and n DMUs

X_{-d} is the m×(n-1) matrix for m inputs and n-1 DMUs

X^d is the m×1 input vector for DMU d being evaluated

$\underline{\lambda}$ is a vector of length n-1; $\overline{\sigma_i^{-1}}$ and $\overline{\sigma_o^{-1}}$ are row vectors of length m, and k, respectively

$\bar{1}$ is a (n-1) row vector of 1's; $\underline{0}$ denotes zero vectors of appropriate length

$-\langle\rightarrow|+\rangle$ and $+\langle\rightarrow|-\rangle$ are 'algorithmic operators'. They imply that the sign may change once, in the direction shown, during the algorithm, iff no feasible initial solution is found. An asterisk * denotes pointwise vector multiplication.

The Universal Radial Model

First we introduce the *Reference Base* point

$$r^* = \{ X_i^*, Y_o^* \} \text{ for } i = 1,..., m \text{ and } o = 1,..., k \quad \textbf{(8)}$$

as an 'artificial DMU', defined as follows

$$X_i^* = \max_{j=1,...,n} X_i^j \text{ and } Y_o^* = \min_{j=1,...,n} Y_o^j \quad \forall \text{ inputs } i, \text{ and } \forall \text{ outputs } o. \quad \textbf{(9)}$$

The 'DMU' r^* is now added to the set of DMUs so that there are n+1 DMUs altogether and X and Y are $m \times (n+1)$ and $k \times (n+1)$ matrices. Likewise X_{-d} and Y_{-d} are $m \times n$ and $k \times n$ matrices now, etc.

Likewise, standard deviation vectors $\underline{\sigma_i}, \underline{\sigma_o}$ for each input and output variable are calculated after adding the reference base-point, so that. $\underline{\sigma_i} = \sigma(X_i^j)$ for j = 1,…,n+1 DMUs, for example.

The model is formulated as follows:

For DMU *d* inside or outside the hull, inefficiency I_{UR} is calculated by the linear program

$$I_{UR} = \underset{\zeta \geq 0, \underline{\lambda} \geq \underline{0},}{\boldsymbol{max}} \langle \rightarrow | - \rangle \zeta \tag{10}$$

s.t.

$$X_{-d}\,\underline{\lambda} + \langle \rightarrow | - \rangle \underline{\sigma_i}\zeta \leq X^d \quad \forall i$$

$$Y_{-d}\,\underline{\lambda} - \langle \rightarrow | + \rangle \underline{\sigma_o}\zeta \geq Y^d \quad \forall o$$

$$\bar{1}\underline{\lambda} = 1$$

$$\zeta \geq 0, \underline{\lambda} \geq \underline{0},$$

where ζ is a scalar.

This model is called radial because the projection is always along all input and output dimensions, which is not necessarily the case for the additive model, for example. It is 'equi-radial' because the distance to the frontier is the same in each input and output, after the normalisation in terms of standard deviations is taken into account.

Just as is the case for the Universal Additive model, the Universal Radial model is also completely translation and units invariant. In other words, no matter what zeros or negative values are contained in the data, the solution will be found, and is identical for any affine translation. There is no discontinuity in efficiency measurement in this model. It is truly a universal model which can be used to measure efficiency from inside or outside the hull for the specification of variable returns-to-scale. A constant returns-to-scale option has also been implemented in the DEA ranker.

Normalised Universal Radial Efficiency

The Normalised Universal Radial Efficiency for a DMU *d* is defined as

$$E^*_{UR}(d) = 1 - \frac{I_{UR}(d)}{I_{UR}(r^*)} \tag{11}$$

Thus $E^*_{UR} = 1$ for DMUs on the strong *and weak* efficient boundaries, $E^*_{UR} \geq 1$ for DMUs outside the hull , $E^*_{UR}(r^*) = 0$ and in general: $E^*_{UR} \geq 0$.

In the focused ranking procedure of section 3 $Eff(DS_i)$ represents a vector of efficiency scores $E^*_{UR}(d)$ for all named algorithms d applied to dataset DS_i

6. Conclusions

Thus far, the focused ranking approach (by DEA) for providing advice to potential users of machine learning algorithms, the core activity of an automated assistant, has proved to be promising. Extensive trials have been carried out in order to determine the most appropriate DEA specification, settings for k-nn parameters, and the focusing procedure. Repositories for both classification and regression algorithms that have been produced by the METAL project have been taken as the basis for the assistant's knowledge base. Various ideas for evaluating the appropriateness of the assistant's recommendations have been developed and tested both on the results obtained our focusing approach and from the complementary work on 2-dimensional ranking carried out by the METAL partner LIACC in Porto (Berrer et al. 2000). Patently, the eventual usefulness of the assistant will depend as much on how representative for the tasks of users the datasets and algorithms selected for investigation in METAL are, and their inherently affording a basis for extrapolating meta knowledge of performance, as it depends on the procedures defined by the focused ranking or other approaches.

Incremental focused ranking, currently being validated, has the potential of building up the meta knowledge base dynamically while indicating the reliability of advice offered, and for the 'personalisation' of this knowledge by the individual user by focusing on the quality of the assistance.

References

Andersen P, Petersen N C; A Procedure for Ranking Efficient Units in Data Envelopment Analysis, Management Science, Vol. 39, No. 10, October 1993, page 1261–1264.

Berrer, H., Paterson, I., and Keller, J.: Evaluation of Machine-Learning Algorithm Ranking Advisors, paper presented at the DDMI 2000 Workshop, at PKDD-2000, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, Lyon, September 2000.

Keller, J.; Paterson, I. and Berrer, H.: An Integrated Concept for Multi-Criteria- Ranking of Data-Mining Algorithms. 11th European Conference on Machine Learning WS: Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination Barcelona, Catalonia 2000

Holzer, I. (1999): Einsatz der Data-Envelopment-Analysis zur Optimierung von Drall und Durchflußzahl (in German, English title: Application of Data-Envelopment-Analysis for Swirl and Discharge Optimization), diploma thesis, 1999, DaimlerChrysler AG, Stuttgart, Germany

Lovell K, Pastor J; Units Invariant and Translation Invariant DEA Models, Operations Research Letters 18, 1995, page 147–151.

Michie, D.; Spiegelhalter, D.J. and Taylor, C.C. (1994): Machine Learning, Neural and Statistical Classification, EC ESPRIT project StatLog, Ellis Horwood Series in Artificial Intelligence

Nakhaeizadeh, G. and Schnabel, A.: Development of Multi-Criteria Metrics for Evaluation of data-Mining Algorithms, Third International Conference on Knowledge Discovery and Data-Mining, Proceedings, Newport Beach, California, August 14-17, (1997), pp.37-42

Paterson, I.: New Models for Data Envelopment Analysis, Measuring Efficiency Outwith the VRS Frontier, Economics Series No. 84, July 2000, Institute for Advanced Studies, Vienna.