Deriving Consensus Rankings from Benchmarking Experiments

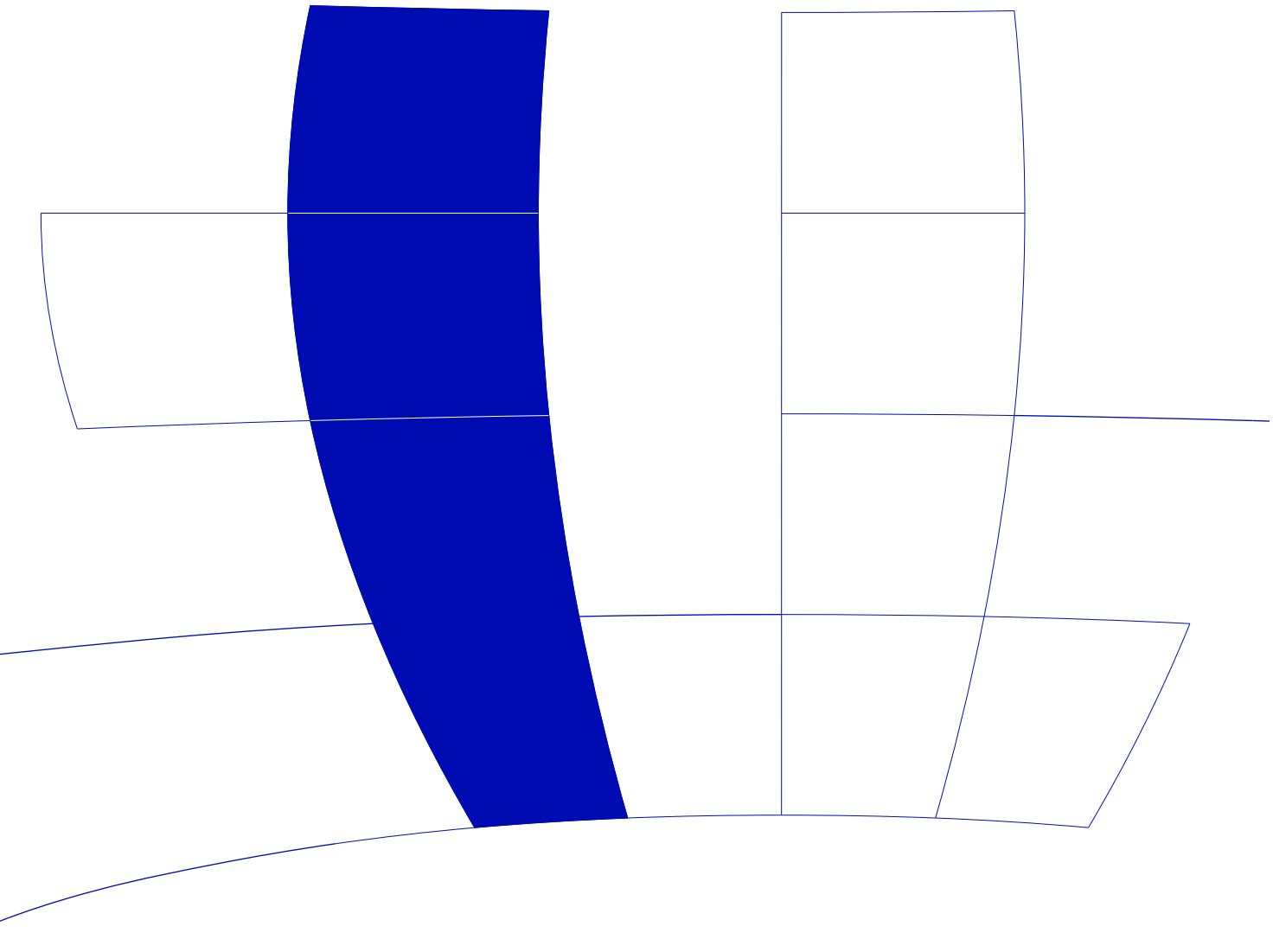
Kurt Hornik, David Meyer

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 33
May 2006

http://statmath.wu-wien.ac.at/



Deriving Consensus Rankings from Benchmarking Experiments

Kurt Hornik **David Meyer**
Wirtschaftsuniversität Wien

Abstract

Whereas benchmarking experiments are very frequently used to investigate the performance of statistical or machine learning algorithms for supervised and unsupervised learning tasks, overall analyses of such experiments are typically only carried out on a heuristic basis, if at all. We suggest to determine winners, and more generally, to derive a consensus ranking of the algorithms, as the linear order on the algorithms which minimizes average symmetric distance (Kemeny-Snell distance) to the performance relations on the individual benchmark data sets. This leads to binary programming problems which can typically be solved reasonably efficiently. We apply the approach to a medium-scale benchmarking experiment to assess the performance of Support Vector Machines in regression and classification problems, and compare the obtained consensus ranking with rankings obtained by simple scoring and Bradley-Terry modeling.

Keywords: benchmark experiments, consenus rankings, Borda, Condorcet, symmetric difference, linear order, poset, linear programming.

1. Introduction

The past decades have featured an immense proliferation of available statistical or machine learning algorithms for supervised and unsupervised learning tasks, including decision trees, neural networks, support vector machines, and resampling methods such as bagging or boosting. With theoretical analyses of the properties of such algorithms becoming ever more challenging, detailed experiments based on suitable combinations of artificial and real-world data sets are employed to study these algorithms. In particular, performance is typically investigated using benchmarking experiments where several competing algorithms are used on a collection of data sets (e.g., from the UCI Machine Learning repository (Blake and Merz 1998)).

Quite surprisingly, solid methodological frameworks for the *analysis* of the results of such benchmarking experiments are typically lacking. Often, p-values reported for assessing significant difference in the performance of algorithms are rather incorrect (e.g., necessary independence assumptions cannot be guaranteed in commonly employed experimental designs) or potentially misleading (e.g., by solely focusing on the means of performance distributions which can be considerably skewed). Hothorn, Leisch, Zeileis, and Hornik (2005) provide a framework which allows the comparison of algorithms on *single* data sets based on classical statistical inference procedures, making it possible to test one-sided hypotheses ("Does algorithm A_i perform significantly better than algorithm A_j on data set D_b?") as well as the hypothesis of non-equivalence.

An overall analysis of the benchmarking experiment would suitably aggregate the performance "measurements" on the individual data set, resulting, e.g., in the determination of a "winner", or more generally a *consensus ranking* which orders the algorithms according to their overall performance. Clearly, conclusions drawn from such an analysis should be taken with the appropriate grain of salt: the results depend on the specific collection $\mathcal{D}$ of data sets employed and hence are primarily *conditional* on the data. They can only be "representative" across learning tasks in as much as $\mathcal{D}$ can serve this purpose. With no algorithm being able to uniformly outperform all

others for all possible data sets, it is clearly impossible to use benchmark experiments to determine whether a certain algorithm is "generally" the best. Still, a chosen $\mathcal{D}$ might be reasonably representative of the needs of a group of researchers or practitioners, and there is an obvious need for a well-founded group decision based on the benchmarking results (e.g., which algorithm to deploy in a specific application).

In this paper, we indicate how consensus rankings can naturally be obtained from paired performance comparisons on the benchmark data sets. The underlying theory and computational issues are discussed in Section 2. An application to a medium-scale benchmarking experiment to assess the performance of Support Vector Machines in regression and classification problems (Meyer, Leisch, and Hornik 2003) is given in Section 3. The obtained rankings are also compared to those provided by a simple scoring approach and a Bradley-Terry model.

2. Consensus Rankings

Consider a benchmarking experiment featuring n learning algorithms $\mathcal{A} = \{A_1, \ldots, A_n\}$ and B data sets $\mathcal{D} = \{D_1, \ldots, D_B\}$, and suppose that it is possible to "rank" the algorithms according to their performance on each data set D_b. Such rankings could for example be obtained based on the means or (quite surprisingly, far less popular) median performances obtained from several runs of the algorithms on suitable bootstrap samples from the data set. Note that distributions of performance measures typically exhibit considerable skewness: hence, whereas means or medians may be employed to investigate differences in location, aggregation should not be based on the "raw" values of the performance measures (but could, e.g., use the ranks or a related scoring method instead). In any case, we feel that it is both more natural and preferable to derive rankings based on the *comparisons* of performances only, in particular, basing these on a notion of one algorithm A_i performing *significantly better* than another algorithm A_j, symbolically, $A_i > A_j$. Using the experimental designs of Hothorn *et al.* (2005), "classical" hypothesis tests can be employed for assessing significant deviations in performance.

The collection of paired comparisons for a data set D_b induces a *relation* (more precisely, endorelation) R_b on the set of algorithms $\mathcal{A}$ which expresses either the strict preference relation as indicated above or its dual, or a "$\leq$" relation taking ties (indicating equivalent performance) into account. The collection of benchmark data sets thus induces a *profile* (ensemble) of relations $\mathcal{R} = \{R_1, \ldots, R_B\}$. A consensus ranking is a suitable aggregation of the relation profile into a relation R. Hereafter, we will assume that a *linear order* is sought, i.e., that the consensus relation be an endorelation on $\mathcal{A}$ which is reflexive, asymmetric, and transitive.

There is a huge literature on consensus methods for relations, starting in the late 18th century with the approaches of Borda (1781) and Condorcet (1785) to aggregate the preferences of voters. In Borda's approach, the objects are ranked according to the so-called Borda marks, the overall numbers of "wins" in the paired comparisons. As this may result in one object being ranked above another in the consensus relation R even though it was consistenly ranked below the other in the invidual relations, Condorcet suggested to base R on a "majority" rule which ranks an object i above object j iff the number of individual wins of i over j exceeds the number of losses. This rule may result in intransitivities ("Effet Condorcet") even when aggregating strict preference relations; if not, it agrees with the Borda solution.

The Borda and Condorcet approaches are examples of so-called *constructive* consensus methods, which simply specify a way to obtain a consensus relation. In the *axiomatic* approach (e.g., Day and McMorris 2003), emphasis is on the investigation of existence and uniqueness of consensus relations characterized axiomatically. The *optimization* approach formalizes the natural idea of describing consensus relations as the ones which "optimally represent the profile" by providing a criterion to be optimized over a suitable set $\mathcal{C}$ of possible consensus relations. This approach goes back to Régnier (1965), who suggested to determine R by solving (a non-weighted variant of) the problem

$$\sum_{b=1}^{B} w_b d(R, R_b) \Rightarrow \min_{R \in \mathcal{C}},$$

where d is a suitable dissimilarity (distance) measure. Such a relation R has also been termed the *median* (more precisely, the $\mathcal{C}$-median) of the profile (Barthélemy and Monjardet 1981). For order relations, Kemeny and Snell (1962) have shown that there is a unique d satisfying a few natural axioms (basically, metricity and betweenness). This so-called Kemeny-Snell distance d_{KS} in fact coincides with the *symmetric difference distance* d_Δ between relations, i.e., the cardinality of the symmetric difference of the relations, or equivalently, the number of pairs of objects being in exactly one of the two relations. This is also the minimal path length distance d_{MPL} between the relations: in the lattice obtained by equipping the set of endorelations with its natural (pointwise incidence) order, d_{MPL} is the minimal number of moves for transforming one relation into the other along the edges of the covering graph (Hasse diagram) of the poset (Monjardet 1981). Both characterizations suggest that d_Δ is the most natural way to measure distance between relations, and to use for the optimization-based consensus approach.

Median linear orders based on d_Δ can be computed by integer linear programming (e.g. Marcotorchino and Michaud 1982). Write $r_{ij}(b)$ and r_{ij} for the incidences of relations R_b and R, respectively. Noting that $u = u^2$ for $u \in \{0, 1\}$ and hence $|u - v| = u + v - 2uv$ for $u, v \in \{0, 1\}$, we have

$$\begin{aligned}
\sum_{b=1}^{B} w_b d(R, R_b) &= \sum_b w_b \sum_{i,j} |r_{ij}(b) - r_{ij}| \\
&= \sum_b w_b \sum_{i,j} (r_{ij}(b) + r_{ij} - 2 r_{ij}(b) r_{ij}) \\
&= \text{const} - \sum_{ij} \left(\sum_b (2 w_b r_{ij}(b) - 1) \right) r_{ij}
\end{aligned}$$

so that, letting $c_{ij} = \sum_b (2 w_b r_{ij}(b) - 1)$, the median linear order R can be obtained by solving

$$\sum_{i \neq j} c_{ij} r_{ij} \Rightarrow \max$$

with the constraints that the r_{ij} be the incidences of a linear order, i.e.,

$$\begin{array}{rcll}
r_{ij} & \in & \{0, 1\} \quad i \neq j & \text{(binarity)} \\
r_{ij} + r_{ji} & = & 1 \quad i \neq j & \text{(asymmetry)} \\
r_{ij} + r_{jk} - r_{ik} & \leq & 1 \quad i \neq j \neq k & \text{(transitivity)}
\end{array}$$

We note that this is a "very hard" combinatorial optimization problem (in fact, NP complete), see Wakabayashi (1998). Its space complexity is related to the number of variables and constraints which are of the orders n^2 and n^3, respectively. In fact, the asymmetry conditions imply that we can, e.g., work only with the upper diagonal part of R, i.e., $r_{ij}, i < j$, and use $r_{ij} = 1 - r_{ji}$ for $i > j$. For each triple of distinct i, j, k the 6 transitivity conditions reduce to 2 non-redundant ones for $i < j < k$. The worst case time complexity is at most of the order 2^n. Quite often, solutions can be found efficiently via Lagrangian relaxation (Marcotorchino and Michaud 1982), i.e., by replacing the binarity constraints $r_{ij} \in \{0, 1\}$ by $0 \leq r_{ij} \leq 1$, $i \neq j$, and iteratively adding "cutting planes" selectively enforcing binarity to the relaxation (Grötschel and Wakabayashi 1989). One can also use state of the art general-purpose integer programming software, such as the open source lp_solve (Berkelaar, Eikland, and Notebaert 2006) or GLPK (Makhorin 2006).

If the explicit asymmetry and transitivity conditions are dropped, the corresponding consensus relation can be determined immediately: obviously, the maximum is obtained by taking $r_{ij} = 1$ if $c_{ij} > 0$ and $r_{ij} = 0$ if $c_{ij} < 0$. This is exactly the Condorcet solution, as for preference relations the r_{ij} are the incidences of the wins and $\sum_b (2 r_{ij}(b) - 1) > 0$ iff $\sum_b r_{ij} > B/2$, i.e., i wins over j in more than half of the comparisons in the profile. Thus, the Concordet approach can be given an optimization ("metric") characterization as yielding the (unconstrained) median endorelation when employing symmetric difference distance.

Determining the median linear order can also be interpreted as finding the maximum likelihood paired comparison ranking (deCani 1969). More generally, constructive consensus approaches could be based on the intrinsic or extrinsic worths (Brunk 1960) obtained by probabilistic modeling of the paired comparison data. The Bradley-Terry model (Bradley and Terry 1952) is the most prominent such model, representing the odds that i wins over j as α_i/α_j using worths ("abilities") α_i, or, in an equivalent logit-linear formulation, $\text{logit}(\Pr(i \text{ beats } j)) = \lambda_i - \lambda_j$ with $\lambda_i = \log(\alpha_i)$. Ordering objects according to their fitted abilities yields another simple constructive consensus approach.

3. Application: Benchmarking Support Vector Machines

Meyer *et al.* (2003) report the results of a benchmark experiment of popular classification and regression methods on both real and artificial data sets. Its main purpose was to compare the performance of Support Vector Machines to other well-known methods both from the field of machine learning (such as neural networks, random forests, and bagging) and "classical" statistics (such as linear/quadratic discriminant analysis and generalized linear models). Most data sets originate from the the UCI Machine Learning repository (Blake and Merz 1998) and are standard in benchmarking. The size and structure of the data sets cover a wide range of problems: The numbers of cases vary from 106 to 3,196, and the numbers of variables range from 2 to 166, involving a mix of dichotomous, polytomous, and metric variables. Both real and artificial data sets were employed. In total, the study involved $n_c = 17$ methods on $B_c = 21$ datasets for classification, and $n_r = 9$ methods on $B_r = 12$ datasets for regression.

All methods were repeatedly (10 times) trained and tested on all data sets, resulting in $n_c \times B_c = 357$ performance measure *distributions* for classification (misclassification rates) and 108 for regression (root mean squared errors). The error distributions were summarized by three statistics: mean, median, and interquartile range, and reported by means of 8 tables. Even using state-of-the-art visualization methods such as parallel boxplots in a trellis-layout for all data sets, it is hard to compare the performance of one method across several data sets, and to come to an overall assessment.

The method of consensus rankings provides a simple clue to further analysis: for each data set D_b, we computed two-sample t tests on the error distributions of all method pairs (A_i, A_j) to assess whether method A_i performed significantly better than A_j on data set D_b (significance level: 5%). The B relations induced by these paired comparisons were then aggregated by means of three consensus ranking methods described above (Median linear order, Borda, and the Bradley/Terry model). The resulting rankings are compared in Table 1 for classification and Table 2 for regression. Interestingly, for classification, all three methods agree at least for the top 5 methods, whereas the top rankings differ for regression. The space and time complexities for the median linear order consensus on the benchmark experiment results are summarized in Tables 3 and 4. For both the classification and regression experiments, the results were immediate on a machine with a Pentium M processor with 1.6 GHz and 1 GB of memory, using the **lpSolve** interface (Buttrey 2005) to R (R Development Core Team 2005) for solving the integer linear programming problem. The corresponding values of the criterion function $\Phi(R) = \sum_{b=1}^{B} d(R_b, R)$ are 1,902 (median linear order), 1,916 (Borda), and 1,938 (Bradley-Terry) for the classification and 331, 355, and 333 for the regression datasets, respectively.

4. Outlook

Median linear orders are only fully interpretable provided that they uniquely solve the corresponding optimization problem. This suggests employing solvers which yield *all* solutions of the underlying binary program (e.g., Branch and Bound methods), as well as considering other types of consensus relations (e.g., preorders allowing for ties, or equivalence relations giving classes of algorithms which perform "equally well"). We are currently exploring these issues, along with the

	Median	**Borda**	**Bradley-Terry**
1	svm	svm	svm
2	dbagging	dbagging	dbagging
3	randomForest	randomForest	randomForest
4	bagging	bagging	bagging
5	nnet	nnet	nnet
6	fda.mars	mart	mart
7	mart	fda.mars	fda.mars
8	multinom	multinom	multinom
9	glm	glm	glm
10	mda.mars	lda	lda
11	lda	mda.mars	mda.mars
12	rpart	knn	mda.bruto
13	lvq	rpart	fda.bruto
14	qda	lvq	knn
15	knn	mda.bruto	qda
16	mda.bruto	qda	rpart
17	fda.bruto	fda.bruto	lvq

Table 1: Comparison of three consensus rankings for the classification data sets. The abbreviations are the same as in Meyer *et al.* (2003).

	Median	**Borda**	**Bradley-Terry**
1	randomForest	nnet	randomForest
2	ppr	randomForest	nnet
3	nnet	ppr	ppr
4	svm	svm	svm
5	bruto	mart	bruto
6	mart	bagging	mart
7	mars	lm	mars
8	bagging	rpart	bagging
9	rpart	bruto	lm
10	lm	mars	rpart

Table 2: Comparison of three consensus rankings for the regression data sets. The abbreviations are the same as in Meyer *et al.* (2003).

n	# variables	# constraints
9	36	240
17	136	1,632

Table 3: Space complexity for the regression/classification benchmark experiments in terms of number of variables and constraints.

n	2^n
9	512
17	131,072

Table 4: Worst case time complexity for the regression/classification benchmark experiments. Given n data sets, it is at most of the order 2^n.

development of an R package which offers computational infrastructure for relations, and methods for computing consensus rankings.

References

Barthélemy JP, Monjardet B (1981). "The median procedure in cluster analysis and social choice theory." *Mathematical Social Sciences*, **1**, 235–267.

Berkelaar M, Eikland K, Notebaert P (2006). "`lp_solve`." Version 5.5.0.7, URL http://groups.yahoo.com/group/lp_solve.

Blake CL, Merz CJ (1998). "UCI Repository of Machine Learning Databases." URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Borda JC (1781). "Mémoire sur les élections au scrutin." Histoire de l'Académie Royale des Sciences.

Bradley RA, Terry ME (1952). "Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons." *Biometrika*, **39**, 324–245.

Brunk HD (1960). "Mathematical Models for Ranking from Paired Comparison." *Journal of the American Statistical Association*, **55**(291), 503–520.

Buttrey SE (2005). "Calling the `lp_solve` Linear Program Software from R, S-PLUS and Excel." *Journal of Statistical Software*, **14**(4). ISSN 1548-7660. URL http://www.jstatsoft.org/v14/i04/.

Condorcet MJA (1785). "Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix." Paris.

Day WHE, McMorris FR (2003). *Axiomatic Choice Theory in Group Choice and Bioconsensus*. SIAM, Philadelphia.

deCani JS (1969). "Maximum Likelihood Paired Comparison Ranking by Linear Programming." *Biometrika*, **56**(3), 537–545.

Grötschel M, Wakabayashi Y (1989). "A Cutting Plane Algorithm for a Clustering Problem." *Mathematical Programming*, **45**, 59–96.

Hothorn T, Leisch F, Zeileis A, Hornik K (2005). "The Design and Analysis of Benchmark Experiments." *Journal of Computational and Graphical Statistics*, **14**(3), 675–699.

Kemeny JG, Snell JL (1962). *Mathematical Models in the Social Sciences*, chapter Preference Rankings: An Axiomatic Approach. MIT Press, Cambridge.

Makhorin A (2006). *GNU Linear Programming Kit (GLPK)*. Version 4.9, URL http://www.gnu.org/software/glpk/glpk.html.

Marcotorchino F, Michaud P (1982). "Agregation de similarites en classification automatique." *Revue de Statistique Appliquée*, **XXX**, 21–44.

Meyer D, Leisch F, Hornik K (2003). "The Support Vector Machine under Test." *Neurocomputing*, **55**, 169–186.

Monjardet B (1981). "Metrics on Partially Ordered Set: A Survey." *Discrete Mathematics*, **35**, 173–184.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Régnier S (1965). "Sur quelques aspects mathématiques des problèmes de classification automatique." *ICC Bulletin*, pp. 175–191.

Wakabayashi Y (1998). "The Complexity of Computing Medians of Relations." *Resenhas*, **3**(3), 323–349.