HOW TO ALLOCATE A RESTRICTED BUDGET OF LEAVE-ONE-OUT ASSESSMENTS FOR EFFECTIVE MODEL SELECTION IN MACHINE LEARNING: A COMPARISON OF STATE-OF-ART TECHNIQUES

Olivier Caelen[1] Gianluca Bontempi

ULB Machine Learning Group, Computer Science Department,
Université Libre de Bruxelles, 1050 Brussels - Belgium
http://www.ulb.ac.be/di/mlg/

Abstract

The problem of selecting the best among several alternatives in a stochastic context has been the object of research in several domains: stochastic optimization, discrete-event stochastic simulation, experimental design. A particular instance of this problem is of particular relevance in machine learning where the search of the model which could best represent a finite set of data asks for comparing several alternatives on the basis of a finite set of noisy data.

This paper aims to bridge a gap between these different communities by comparing experimentally the effectiveness of techniques proposed in the simulation and in the stochastic dynamic programming community in performing a model selection task. In particular, we will consider here a model selection task in regression where the alternatives are represented by a finite set of K-nearest neighbors models with different values of the structural parameter K. The techniques we compare are i) a two-stage selection technique proposed in the stochastic simulation community, ii) a stochastic dynamic programming approach conceived to address the multi-armed bandit problem, iii) a racing method, iv) a greedy approach, v) a round-search technique.

1 Introduction

A common practice in machine learning consists in adopting *cross-validation*, and most specifically its leave-one-out version, to assess different models (e.g. a neural network vs. a support vector machine or two RBFs with a different number of basis functions) and to select the best one. Leave-one-out cross-validation is known to be an *almost* unbiased estimator of the generalisation error [5]. However, for a dataset of N examples and a number S of alternative models to be assessed, this technique asks to carry out $N \times S$ parametric identifications and $N \times S$ predictions. This means that, when S or N is very large and/or when the time constraints are tight, this technique may require too much computation to be affordable.

In this context, it is crucial to study techniques that, once a restricted number of cross-validation assessments is allowed, be able to exploit the fixed budget of possible estimations in an optimal way in order to detect the best structure among a set of alternatives. The issue here is that, given a set of S alternative model structures and their relative assessments based on cross-validation, there is always a non zero probability that the ranking returned by a cross-validation be not compatible with the unkwnown true ranking of their generalization accuracies. An interesting answer to this problem may come from two disciplines which, together with machine learning, focuses on the issue of selecting the best system among a large number of alternatives on the basis of a finite number of

[1]The work of the researcher was supported by the the F.I.R.S.T. project "Data mining prédictif en anesthésie intraveineuse informatise en vue de certification" (project number EP1A320501R059F/415717) of the Région Wallonne, Belgium.

noisy samples. This is the case of discrete-event stochastic simulation [9] and stochastic dynamic programming [2].

In stochastic simulation a relevant issue is how to develop procedures that can efficiently select the best among a set of competing system designs, where best is defined by the maximum or minimum expected simulation output. A well known procedure is the ranking and selection procedure (R&S) first proposed in [6]. This is a two-stage procedure for selecting the best design or a design which is very close to the best one. In the first stage, all the designs are assessed with a fixed number of replications. Based on the results of the first stage, the second one determines the number of additional replications required to attain a specified confidence in the selection.

Stochastic dynamic programming is another discipline which focuses on the problem of optimal choice in a stochastic context. A well-known example is the k-armed bandit problem where the goal is to find the strategy or the sequential design that maximises the total expected sum of outcomes coming from k stochastic processes. An optimal and tractable solution to this problem in the normal and independent case relies on the adoption of the dynamic allocation indices (also known as Gittins indices [7]), which depend on the number of times that the process has been sampled and the resulting outcome. If we intepret the k stochastic processes as the alternative model structures to be assessed and their outcome as their estimated generalization error, it appears evident the importance of a bandit process strategy when the selection of the best model on the basis of a reduced number of assessments is at stake. The bandit interpretation of the model selection problem raises the well known exploration-exploitation issue where the challenge is how to trade off leave-one-out assessments that will help knowing more about the different alternatives versus assessments intended to improve the accuracy of the model structure currently observed as the best.

This paper advocates the need to import the above mentioned techniques in the machine learning arena in order to assess their capacity to deal with problems characterized by a finite set of noisy data and where the goal is to search for the model with the best generalization capability. If at our knowledge, no application of simulation optimization techniques to machine learning tasks exists, the closest research work in terms of adoption of the k-armed bandit algorithm for a learning task is the paper of Schnedier and Moore where the bandit algorithm was used in a context of active learning for experimental design.

The two main contributions of the paper to this new research perspective are: a reformulation of these techniques in the machine learning formalism and an experimental evaluation of these techniques in a regression model selection problem. In particular, we consider here a set of S competing *K-Nearest-Neighbours* (KNN) models, characterized by different values of K. The KNN models are local models. When a prediction is asked on an input query point q, the KNN model searches the K nearest neighbours of q in the learning set D_N and returns as output the average of the K nearest neighbours.

Throughout all the paper we assume a fixed budget made of L leave-one-out train and test assessments which is made avaialble to all the techniques in order to select the best model structure among a set of S alternatives. The experimental session will then evaluate the different strategies on the basis of their ability of sequentially choosing the next candidate to be assessed, by taking into account the total constrained number of estimations and the need of returning at the end the structure with the lowest generalization error.

In order to build up a reliable benchmark, the simulation R&S technique and the bandit strategy [7] are compared on a set of 26 regression datasets to three yardstick search strategies : (i) a *round-search* [10] approach where the budget of L assessments is uniformly shared by all the competing models, (ii) a *greedy-search* strategy [10] which sequentially spends the budget on the most promising model structure on the basis of the previous assessments (iii) a racing strategy for model selection, proposed first by Maron and Moore [12, 13] and developed furtherly in [3, 4].

2 The model selection strategies

Consider a supervised regression problem where the training set $D_N = \{z_i\} = \{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_N, y_N\rangle\}$ is made of N pairs $z_i = \langle x_i, y_i\rangle \in \mathcal{X} \times \mathcal{Y}$ i.i.d. distributed according to the joint distribution $P(\langle X, y\rangle) = P(y|X)P(X)$. Let us define a *learning machine* by the following components: (i) a class Λ of *hypothesis* functions $h(\cdot, \alpha)$, $\alpha \in \Lambda$, which can be represented as a nested

sequence of model structures $\Lambda_1 \subset \Lambda_2 \subset \ldots \subset \Lambda_s \subset \ldots \subset \Lambda_S$, (ii) a *quadratic cost* function $C(y,h) = (y-h)^2$, (iii) an *algorithm* of parametric identification that for a given structure Λ_s and a given training set D_N returns a hypothesis function $h(\cdot, \alpha^s_{D_N})$ with $\alpha^s_{D_N} \in \Lambda^s$ such that $\sum_{\langle x,y\rangle \in D_N} C\left(y, h(x, \alpha^s_{D_N})\right) \leq \sum_{\langle x,y\rangle \in D_N} C\left(y, h(x, \alpha^s)\right)$ for all $\alpha^s \in \Lambda^s$. The problem of model selection consists in finding the class of hypothesis which is optimal in terms of generalization accuracy. In formal terms this corresponds to search for the class Λ_{s^*} such that

$$s^* = \arg\min_{s \in S} \mathrm{MISE}(s) = \arg\min_{s \in S} \left\{ E_{D_N} \left[\int_{\mathcal{X}} \int_{\mathcal{Y}} (y - h(x, \alpha^s_{\mathbf{D}_N}))^2 dP(y|x) dP(x) \right] \right\} \tag{1}$$

where the mean integrated square error $\mathrm{MISE}(s)$ is a quantitative measure of the generalization accuracy of the class s of models. Unfortunately, in practice, the quantity $\mathrm{MISE}(s)$ is not observed directly but can only be estimated, for instance by cross-validation.

The common model selection practice to minimize the generalisation error MISE consists in: (i) computing an estimate $\widehat{\mathrm{MISE}}(s)$ of the unknown $\mathrm{MISE}(s)$ on the basis of the whole dataset D_N, (ii) implementing a search procedure to search for the *structure index* s that minimizes $\widehat{\mathrm{MISE}}s$.

In this paper, we focus on a model selection task which has to be accomplished on the basis of an overall budget of L Monte-Carlo leave-one-out (l-o-o) cross-validation assessments. The resulting procedure is then a sequence of L steps (or trials) where each step consists in:

1. choosing a model among the S alternatives according to a specific strategy,
2. selecting randomly a sample $z_i = \langle x_i, y_i \rangle$ from the learning set D_N,
3. performing the parametric identification step (e.g. by empirical risk minimization) on the dataset $D_{(i)}$, obtained by removing the sample z_i from the learning set D_N,
4. assessing the prediction error of the chosen model $e_i^{(i)} = \left(y_i - h(x_i, \alpha^s_{D_{(i)}})\right)^2$.

Each of the strategies presented in the following uses a different sequential policy to choose the model structure and to allocate the L assessments. Let l_s $(s = 1, \ldots, S)$ be the number of assessments allocated to the model structure indexed by s, after l $(l = 1, \ldots, L)$ steps. At the $l+1$th step of the procedure, each technique will rely on the S estimations $\widehat{\mathrm{MISE}}(s, l_s)$ of the generalisation error performed so far, where

$$\widehat{\mathrm{MISE}}(s, l_s) = \frac{1}{l_s} \sum_{j=1}^{l_s} \left(y_i - h\left(x_i, \alpha_{D_{(i)}}\right)\right)^2 \tag{2}$$

is the estimation of the generalisation error of the model s on the basis of l_s l-o-o assessments, and $\sum_{s=1}^{S} l_s = l$.

In Sections 2.1 and 2.2 we will discuss how the R&S and the bandit technique, borrowed from simulation and stochastic dynamic programming respectively, can be applied to a model selection task. Section 2.3 will rapidly sketch the three reference methods we use to perform the benchmarking session.

2.1 The R&S search

The ranking-and-selection procedure aims to select the best (e.g. without loss of generality the lowest quantity) among S alternatives by controlling the probability that the selected alternative is really the best one. Suppose we have S estimations $\hat{\mu}(s)$, $s = 1, \ldots, S$ of S unknown expected value $\mu(s)$, $s = 1, \ldots, S$. Let $\mu(s_j)$ be the jth smallest of the $\mu(s)$'s so that $\mu(s^*) = \mu(s_1) \leq \mu(s_2) \leq \cdots \leq \mu(s_S)$. Suppose we want to select on the basis of the $\hat{\mu}(s), s = 1, \ldots, S$, the value s^* such that $s^* = \arg\min \mu(s)$. Given the random nature of the estimators $\mu(s)$, we can never be absolutely sure that we shall make the optimal selection. However, we could accept as correct also some choice $\tilde{s}$ such that $|\mu(\tilde{s}) - \mu(s^*)| < \delta$ for a given δ (also known as the *indifference zone parameter*).

Let us denote by CS (for "correct selection") the event related to the choice of such a $\tilde{s}$. The R&S algorithm [6] defines a procedure for ensuring that, for a given δ the probability of having

accomplished the correct selection be greater than a certain value P^* (also known as the *probability of correct selection*).

Let us now reformulate the problem in the model selection terminology. Given S structures, let s_j be the index of the jth best structure:

$$\text{MISE}(s^*) = \text{MISE}(s_1) \leq \text{MISE}(s_2) \leq \ldots \leq \text{MISE}(s_j) \leq \ldots \leq \text{MISE}(s_S)$$

We would like to select the structure Λ_{s_1} returning the smallest generalisation error but we are ready to accept as correct the structure indexed by $\tilde{s}$ where, with probability P^*, the difference between $\text{MISE}(\tilde{s})$ and $\text{MISE}(s_1)$ is smaller than δ. P^* and δ are two parameters of the method fixed by the designer.

The statistical procedure for solving this problem involves two stages of sampling on each of the S alternative structures. In the first-stage sampling, we make $N_0 \geq 2$ tests on each of the S structures. The l-o-o error means and variances after the first N_0 assessments on the S structures are:

$$\widehat{\text{MISE}}(s, N_0)^{(1)} = \frac{\sum_{i=1}^{N_0} \left(y_i - h(x_i, \alpha^s_{D_{(i)}}) \right)^2}{N_0} \tag{3}$$

and

$$\sigma^2_{\widehat{\text{MISE}}(s,N_0)} = \frac{\sum_{i=1}^{N_0} \left[\left(y_i - h(x_i, \alpha_{D_{(i)}}) \right)^2 - \widehat{\text{MISE}}(s, N_0)^{(1)} \right]^2}{N_0 - 1} \tag{4}$$

where the estimetion of MISE is restricted to a subset of the training set. Then we compute $N_1(s)$, the number of assessments needed to attain the given accuracy:

$$N_1(s) = \max \left\{ N_0 + 1, \left\lceil \frac{\left(h^S_{P^*}(N_0) \right)^2 . \sigma^2_{\widehat{\text{MISE}}(s,N_0)}}{(\delta)^2} \right\rceil \right\}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x and the value $h^S_{P^*}(N_0)$ (dependent on S, P^* and N_0) can be obtained from the tables in [9][2]. Next we make $N_1(s) - N_0$ new tests on each structure and obtain the second-stage test means:

$$\widehat{\text{MISE}}(s, N_1(s) - N_0)^{(2)} = \frac{\sum_{i=N_0+1}^{N_1(s)} \left(y_i - h(x_i, \alpha_{D_{(i)}}) \right)^2}{N_1(s) - N_0} \tag{5}$$

Then we define the weights W^s_1 and W^s_2:

$$W^s_1 = \frac{N_0}{N_1(s)} \left[1 + \sqrt{1 - \frac{N_1(s)}{N_0} \left(1 - \frac{(N_1(s) - N_0)(\delta)^2}{h^S_{P^*}(N_0) \sigma^2_{\widehat{\text{MISE}}(s,N_0)}} \right)} \right]$$

where $W^s_2 = 1 - W^s_1$. Finally, according to [6], the estimation of the generalisation error of the sth structure is:

$$\widehat{\text{MISE}}(s) = W^s_1 \cdot \widehat{\text{MISE}}(s, N_0)^{(1)} + W^s_2 \cdot \widehat{\text{MISE}}(s, N_1(s) - N_0)^{(2)}$$

Note that although the method assumes that the leave-one-out errors are normally distributed, we need to assume that neither the values of the variance are known, nor they are the same for all s. At the same time, the main limitation of the technique is that it considers complete independence both for observations within individual alternatives and across alternatives, that does not evidently hold in a cross-validation case. A blocking version of the algorithm that removes this assumption can be found in [1].

For our experimental session we set P^* to 0.9 and δ to the average of the differences between the quantities $\widehat{\text{MISE}}^{(1)}$ of each pair of structures after the first stage.

[2]For a more extended description of the meaning of h we refer the reader to [6, 1]

2.2 The bandit-search

The *bandit-search* is a method for model selection inspired to the solution of the *k-armed bandit problems* based on the adoption of the Gittins' indices [7]. In the k-armed bandit problem setting, we consider a casino player who has to find the optimal strategy to play with a random k-armed machine. When one arm is pulled, the bandit machine generates a random reward based on a normal distribution specific to that arm with unknown mean and variance. The player must find the sequence of actions on the bandit machines which maximizes the rewards of the machines.

An optimal *sequence strategy* must then solve the *exploration versus exploitation trade-off*. Since the parameters of the normal distribution are unknown, the strategy would require a maximum of actions on each arm (*exploration*) in order to improve the estimation of its reward. On the other hand, since the goal is to maximize the rewards of the machines, the strategy should privilege the actions on the best observed arm (*exploitation*). Gittins, in his book [7], computes a set of indices to solve this allocation trade-off problem in an optimal manner.

Back to our model selection problem, the arms can be considered as the alternative models and the expected rewards as the expected generalisation error of each of them. The goal here is not to maximize the reward but to minimize the generalization error on the basis of the L assessment trials made available.

The indices of Gittins return the optimal solution in the case of an infinite *temporal discount factor* problem. Our model selection problem is not time infinite, but is based on a finite number of trials. To adapt this approach to our setting, we have recourse to the solution proposed in [14]. If a system with a discount factor γ ($0 < \gamma < 1$) receives, at every step, a reward R, then the total reward will be : $R + \gamma R + \gamma^2 R + \gamma^3 R + \ldots = R \sum_{i=0}^{\infty} \gamma^i = R/(1-\gamma)$. Thus its total reward will be the same as if there is no infinite temporal and it stops after $1/(1-\gamma)$ loops. This given a heuristic for converting the number of loops in an effective γ : $\gamma = \frac{L-1}{L}$.

Our implementation of the *bandit-search* method begins by performing in a round-robin fashion $L/2$ assessments overall. Then, the method carries out the remaining $(L - L/2)$ steps by at each step (i) computing the Gittins indices for all the model structures, (ii) allocating the assessment to the structure with the lowest Gittins index. Finally, when the budget runs out, the *bandit-search method* returns the structure with the lowest generalisation error. The values of the *Gittins indices* at the lth step are given by

$$\widehat{\text{MISE}}(s, l_s) - \sigma_{\widehat{\text{MISE}}_{(s,l_s)}} v(l, \gamma) \tag{6}$$

where $\widehat{\text{MISE}}(s, l_s)$ and $\sigma_{\widehat{\text{MISE}}_{(s,l_s)}}$ are respectively the average mean and standard deviation of the l-o-o errors, and the value $v(l, \gamma)$ is obtained by the tables of [7]. Note that a simple linear interpolation is used to estimate the values of $v(l, \gamma)$ not included in the tables provided by the Gittins' book.

2.3 The yardstick techniques

We will consider three yardstick techniques to benchmark the quality of the methods discussed in the two previous sections:

Round-search algorithm: this simple algoritmh allocates the l-o-o assessments in a round fashion [10]. This means that at the lth step it assesses the structure of index $s = ((l-1) \bmod (S)) + 1$.

Greedy-search algorithm: after an initial assessment of all the S structures, this algorithm selects at the lth ($l = s+1, \ldots, L$) step always the current best structure according to the observed $\widehat{\text{MISE}}(s, l_s)$ [10].

F-race-search algorithm: The idea of racing consists in assessing a large number of models by performing cross-validation only on a reduced test set. On the basis of well-known statistical results, it is possible to show that families of good feature subsets can be rapidly found by quickly discarding the bad subsets and concentrating the computational effort on the better ones. This model selection technique was called the *Hoeffding race* by Maron and Moore [11], with reference to Hoeffding's formula which puts a bound on the accuracy of a sampled mean of l_s observations as an estimator of the expected value. Let Λ^* be a set with S structures

Name	ABALONE	AILERONS	BANK-32FH	BANK-32FM	BANK-32NH
N	4177	10000	8192	8192	8192
n	10	40	32	32	32
Name	BANK-32NM	BANK-8FH	BANK-8FM	BANK-8NH	BANK-8NM
N	8192	8192	8192	8192	8192
n	32	8	8	8	8
Name	BUPA	COVTYPE	ELEVATORS	HOUSING	KIN32FH
N	345	10000	10000	506	8192
n	6	54	18	13	32
Name	KIN32FM	KIN32NH	KIN32NM	KIN8FH	KIN8FM
N	8192	8192	8192	8192	8192
n	32	32	32	8	8
Name	KIN8NH	KIN8NM	MPG	OZONE	POL
N	8192	8192	392	330	10000
n	8	8	7	8	48
Name	STOCK				
N	950				
n	9				

Table 1: The datasets use to compare the search methods. N is the number of samples and n is the number of covariates.

M_1 vs M_2	num. victories M_1	num. victories M_2	num. equalities
R&S vs *bandit*	5	11	10
R&S vs *F-race*	9	9	8
R&S vs *round*	5	10	11
R&S vs *greedy*	5	9	12
bandit vs *F-race*	8	2	16
bandit vs *round*	5	3	18
bandit vs *greedy*	9	6	11
F-race vs *round*	5	8	13
F-race vs *greedy*	6	9	11
round vs *greedy*	5	4	17

Table 2: Comparison of the number of victories (*paired t-test* with confidence = 90%) for each couple of search methods where the R&S-search defines the number L. The first colomn is the number of victories for the first search method, the second column is the number of victories of the second search method and the last column is the number of equalities between the two methods.

of model. To find the best index of structure in $[1, \ldots, S]$, the *race-search* [11] tests all the structures in parallel (it makes a *race*). At each loop of the race-search, when a structure is significantly worse than the current best structure, this structure is removed of the race and the method can focus the computational power to distinguish the best structure. In this paper, we focus on an enhanced version of the racing algorithm, the F-Race, introduced by [3] for comparing metaheuristics for combinatorial optimization problems and applied to feature selection problems in [4]. In the experimental session we use the implemented version of F-Race made available by Mauro Birattari in the R-package `race`[3].

3 The experimental results

This section compares the two techniques discussed in Section 2.1 and 2.2 respectively, with the three yardstick methods sketched in Section 2.3. The experimental session uses 26 well-known regression datasets reported in table 1. For the purpose of the experiments, the datasets are randomly split in two parts; a training set (1/3 of the samples) and a test set (2/3 of the samples). The comparative assessment is performed by computing a set of paired t-test (confidence 90%) on

[3]http://cran.r-project.org/src/contrib/Descriptions/race.html

	$L = 300$			$L = 400$			$L = 500$			$L = 600$		
bandit vs *F-race*	13	8	5	10	9	7	11	5	10	8	5	13
bandit vs *round*	14	5	7	6	6	14	8	8	10	9	7	10
bandit vs *greedy*	9	7	10	7	6	13	10	6	10	7	5	14
F-race vs *round*	11	10	5	8	6	12	5	12	9	8	6	12
F-race vs *greedy*	7	11	8	9	6	11	8	13	5	7	9	10
round vs *greedy*	4	11	11	7	7	12	8	6	12	7	9	10

	$L = 900$			$L = 1200$		
bandit vs *F-race*	7	8	11	7	10	9
bandit vs *round*	10	5	11	8	5	13
bandit vs *greedy*	7	5	14	10	6	10
F-race vs *round*	12	4	10	9	4	13
F-race vs *greedy*	9	5	12	8	6	12
round vs *greedy*	4	8	14	5	8	13

Table 3: Comparison of the number of victories (*paired t-test* with confidence = 90%) for each couples of search methods with $L = \{300, 400, 500, 600, 900, 1200\}$. For each value of L we have three columns, the first one is the number of victories for the first search method, the second column is the number of victories of the second search method and the last column is the number of ties between the two methods.

the vectors of test squared errors. In order to take into account the peculiarities of the R&S method we need to perform two experiments. In the first one the budget of l-o-o assessments is fixed by the two-stage R&S procedure. This is the only way to have a fair assessment of the R&S procedure with respect to the competitors. In this case we assess $S = 10$ different KNN model structures: $K = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$. Table 2 reports for all the pairs of considered methods, the number of times that a method significantly outperforms the others, the number of times that it is significantly outperformed and the number of ties.

In the second experiment we assess $S = 30$ different KNN models structures: $K = [1, \ldots, 30]$. Since the R&S algorithm is not taken into account we can range over six prespecified different L values ($L = \{300, 400, 500, 600, 900, 1200\}$). Table 3 reports for all the the pairs of considered methods, the number of times that a method significantly outperforms the others, the number of times that it is significantly outperformed and the number of ties. Two are the main results coming from these preliminary results. The R&S technique does not seems to be competitive neither with the bandit strategy nor with the simplest assessment strategies. This is probably due to the strong assumptions made on the independence of the samples. On the other hand good performance of the bandit strategy is very promising. This is particularly evident in the case of the comparison with R&S and in the second experiment for the very reduced budget $L = 300$ cases (that corresponds on average to 10 l-o-o assessments per model). The superiority of the bandit search becomes less striking when the number L increases. For example, for $L = 1200$ the F-race technique seems to obtain the best accuracy. This is probably due to the fact that the multiple paired statistical tests used by the racing technique become more effective when the number of l-o-o samples is sufficiently high. As a general conclusion, it seems that the bandit strategy is more aggressive than racing for small number of samples and tends to differentiate less when the budget of assessments begins to become sufficiently high. A possible reason could be that the bandit criteria does not rely on paired tests as this is the case of F-race.

4 Conclusion

The issue of selecting the best system is of crucial relevance in machine learning but is of equal importance in other disciplines. This paper aims to present some preliminary results on the usefulness of an improved cross-fertilization between different research communities that address similar topics. Future work will focus on the extension of the model selection task to more complicated settings (e.g. feature selection) and the adoption of enhanced version of the simulation-based and the bandit algorithms. Interesting techniques to test are the sequential versions of the two-stage R&S algorithm discussed in [8] and the use of methods that remove the assumption of independence.

References

[1] R. E. Bechhofer, T. J. Santer, and D. M. Goldsman. *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. Wiley, 1995.

[2] D. Berry and B. Fristedt. *Bandit problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

[3] M. Birattari, T. Stützle, L. Paquete, and K. Varrentrapp. A racing algorithm for configuring metaheuristics. In W. B. Langdon, editor, *GECCO 2002*, pages 11–18. Morgan Kaufmann, 2002.

[4] G. Bontempi, M. Birattari, and P.E. Meyer. Combining lazy learning, racing and subsampling for effective feature selection. In *Adaptive and Natural Computing Algorithms: Proceedings of the International Conference ICANNGA05*, pages 393–396. Springer Computer Science, 2005.

[5] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

[6] E. J. Dudewicz and S. R. Dalal. Allocations of observations in ranking and selection with unequal variances. *Sanhkya*, 37:28–78, 1975.

[7] J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley, 1989.

[8] K. Inoue, S. E. Chick, and C-H Chen. An empirical evaluation of several methods to select the best system. *ACM Transactions on Modeling and Computer Simulation*, 9(4):381–407, 1999.

[9] A.M. Law and W.D. Kelton. *Simulation Modeling & analysis*. McGraw-Hill International, second edition, 1991.

[10] Omid Madani, Daniel J.Lizotte, and Russell Greiner. Active model selection. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 357–365, 2004.

[11] O. Maron and A. Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11(1–5):193–225, 1997.

[12] O. Maron and A. W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 59–66. Morgan Kaufmann Publishers, Inc., 1994.

[13] Oden Maron and Andrew W. Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11(1-5):193–225, 1997.

[14] Jeff Schneider and Andrew Moore. Active learning in discrete input spaces. In *Proceedings of the 34th Interface Symposium*, 2002.