Comparing Prequential Model Selection Criteria in Supervised Learning of Mixture Models

Petri Kontkanen, Petri Myllymäki, Henry Tirri
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science, FIN-00014 University of Helsinki, Finland
http://www.cs.Helsinki.FI/research/cosco/, cosco@cs.Helsinki.FI

Abstract

In this paper we study prequential model selection criteria in supervised learning domains. The main problem with this approach is the fact that the criterion is sensitive to the ordering the data is processed with. We discuss several approaches for addressing the ordering problem, and compare empirically their performance in real-world supervised model selection tasks. The empirical results demonstrate that with the prequential approach it is quite easy to find predictive models that are significantly more accurate classifiers than the models found by the standard unsupervised marginal likelihood criterion. The results also suggest that averaging over random orderings may be a more sensible strategy for solving the ordering problem than trying to find the ordering optimizing the prequential model selection criterion.

1 Introduction

In this paper we are concerned with the problem of defining practical model selection criteria for learning predictive models from sample data — in other words we wish to find computationally feasible scoring functions that can be used for distinguishing accurate predictive models from poor models in machine learning contexts. It should be noted that as we define the quality of a model in terms of predictive accuracy, this definition is dependent on how we measure the accuracy of a predictive distribution, i.e., on the loss function used. In the unsupervised setting the loss function is defined in terms of a *joint* distribution on the domain variables. In contrast to this, in the following we consider *supervised* situations where the domain variables can be partitioned into two separate sets, and we know a priori that *all* future prediction tasks involve predicting the values of variables in the second set, given the values of variables in the first set. In particular, in this paper we focus on a special case of such problems, the *classification problem*, where the second set consists of a single (class) variable.

The standard Bayesian approach for solving the model selection problem is to view the possible models as values of a random variable, and to choose the model maximizing the posterior probability, given the sample data. Assuming all the models to be equally probable a priori, this leads to choosing the model maximizing the *marginal likelihood* or the *evidence* of the data. However, as discussed in [13], the maximal evidence model represents well the joint distribution of the domain variables, and is hence a solution for unsupervised model selection tasks. Nevertheless, this approach is frequently used also for supervised model selection tasks, such as the classification problem at hand. This issue is discussed in more detail in Section 2.

In our earlier work [15] we demonstrated empirically that marginal likelihood can be in practice a poor model selection criterion for classification domains, and that model selection criteria based on *prequential (predictive sequential) approaches* [5, 6, 7, 19] or *cross-validation* [23, 9] lead to more accurate predictive models. In this paper we extend and elaborate our previous work in two ways. First, instead of constraining ourselves to simple variants of the Naive Bayes model, here we change the model family to consist of more complex *finite mixture models*, where the joint probability distribution is obtained by a weighted sum of component distributions. The second extension concerns the use of the prequential approach in supervised model selection. Namely, we assume that there is no natural ordering in the data, but wish to treat the data as an unordered list. However, as already pointed out in [15], in this case the value of the prequential model selection criterion depends on the order of which the data is processed. In Section 3 we discuss several approaches for addressing the ordering problem.

The empirical results obtained support the observations reported in [15], and demonstrate similar behavior with the mixture models as with the Naive Bayes model: supervised model selection criteria clearly outperform the unsupervised marginal likelihood criterion also in this case. The results also suggest that the greedy heuristic suggested in [19, 20] for handling the ordering problem, or the simple variants considered here, do not yield satisfactory results in practice, but more efficient solutions are needed. In this set of experiments, better results were obtained by averaging the prequential criterion over a number of random orderings. The results are summarized in Section 4.

2 The Supervised Model Selection Problem

Let $\mathcal{D} = \mathbf{x}^N$ denote the *training data*, a matrix of N vectors each consisting values of n random variables $X_1, \ldots, X_n$. For simplicity, in the sequel we will assume the random variables X_i to be discrete. By a *model* M we mean here a parametric model form so that each parameterized instance (M, θ) of the model produces a probability distribution $P(X_1, \ldots, X_n | M, \theta)$ on the space of possible data vectors $\mathbf{x}$. Although it is intuitively appealing (and in many cases conceptually convenient) to think of the data $\mathcal{D}$ as a random sample from some "true" but unknown probability distribution, it should be pointed out that the model selection problem can also be formalized without such an assumption, as demonstrated in, e.g., [5, 20, 18].

Given a set $\mathcal{F} = \{M_1, \ldots, M_m\}$ of possible models, and a data sample $\mathcal{D}$, in the (unsupervised) model selection problem, the task is to choose a model $M \in \mathcal{F}$ so that the resulting predictive distribution

$$
\begin{aligned}
&P(X_1, \ldots, X_n | \mathcal{D}, M) \\
&\quad = \int P(X_1, \ldots, X_n | \mathcal{D}, M, \theta) P(\theta | \mathcal{D}, M) d\theta \qquad (1)
\end{aligned}
$$

yields more accurate predictions in the future than any of the predictive distributions defined by the other models. Consequently, in this paper we do not consider the problem of choosing the model parameters, but use in each case the predictive distribution (1), and assume that the models M are such that this type of marginalization can be done in closed form. We also do not address here the important problem of how to find good sets of models, but concentrate on model validation, and assume the set $\mathcal{F}$ to be given.

In the Bayesian approach, the model selection problem is typically solved by regarding $\mathcal{F}$ as a random variable (with possible values $M_1, \ldots, M_m$), and by choosing the model maximizing the posterior probability $P(M_i | \mathcal{D})$. Assuming all the models to be equally probable a priori, this leads to choosing the model M^* maximizing the *marginal likelihood* or the *evidence* of the data $\mathcal{D}$:

$$
\begin{aligned}
M^* &= \arg\max_M P(M | \mathcal{D}) = \arg\max_M P(\mathcal{D} | M) \\
&= \arg\max_M \int P(\mathcal{D} | M, \theta) P(\theta | M) d\theta. \qquad (2)
\end{aligned}
$$

We see that the marginal likelihood measure depends on the prior distribution $P(\theta | M)$ defined on the model parameters. This prior can either be regarded as a formalization of our prior domain knowledge, in which case we are faced with the question of compatibility and consistency between different priors [12, 3], or only as a technical parameter representing no such information. In the latter case, it can be shown that a certain prior known as Jeffreys' prior [14, 1] can be given strong theoretical justification from the predictive performance point of view with respect to the so called minimax loss formulation [21, 10]. Some empirical results concerning the effect of Jeffreys' prior on predictive accuracy can be found in [16, 17, 11]. In the remainder of this paper we do not address the important problem of choosing the prior distributions, but simply use uniform non-informative priors for the model parameters θ as well as for the models M.

In the supervised classification framework considered in this paper, the goal in the model selection is to choose from $\mathcal{F}$ the model M which yields the most accurate classifications with respect to the loss function used, and the classification predictive distribution $P(V | \mathbf{u}, M)$, where V denotes the class variable, the value of which is to be predicted, and $\mathbf{u}$ denotes the values of the other variables, which are assumed to be given. It is now important to realize that although the joint probability distribution $P(v, \mathbf{u} | M)$ can be used for producing the required classification probability distribution by marginalization,

$$
P(v | \mathbf{u}, M) = \frac{P(v, \mathbf{u} | M)}{P(\mathbf{u} | M)} = \frac{P(v, \mathbf{u} | M)}{\sum_{v'} P(v', \mathbf{u} | M)},
$$

the model M producing the most accurate predictive distribution in the joint probability estimation sense does not necessarily have to produce the most accurate classification probability distribution, *unless the joint distribution $P(v, \mathbf{u} | M)$ represents the "true" domain probability distribution exactly.* As we can safely say that in reality this assumption is never true, we can conjecture that proper supervised model selection criteria may favor different models than unsupervised model selection criteria.

3 Supervised Prequential Model Selection

As discussed in [15, 13], there are many alternative approaches for constructing theoretically valid model selection criteria for the supervised framework discussed in the previous section. In the following we concentrate on *prequential approaches* where the model selection criteria are typically computed predictively and sequentially ("prequentially"). Theoretical frameworks for prequential model selection can be found in [5, 6, 7, 19, 20, 24]. It is noteworthy that although these frameworks are motivated by various different considerations, all the suggested approaches lead to quite similar results if the predictive accuracy is measured by using the logarithmic loss function.

As prequential model selection principles are usually described in the unsupervised model selection domain, the approach has to be modified accordingly for our supervised classification case. In [15] we followed the suggestion given in [6], based on the observation that the marginal likelihood can be factorized into two products as follows:

$$
\begin{aligned}
&P(\mathcal{D}|M) = P(\mathbf{v}^N, \mathbf{u}^N|M) \\
&= \prod_{i=1}^{N} P(v_i, u_i|\mathbf{v}^{i-1}, \mathbf{u}^{i-1}, M) \\
&= \prod_{i=1}^{N} P(v_i|\mathbf{v}^{i-1}, \mathbf{u}^{i}, M) \prod_{i=1}^{N} P(u_i|\mathbf{v}^{i-1}, \mathbf{u}^{i-1}, M). \qquad (3)
\end{aligned}
$$

Of these two products, the first one was called the *partial (marginal) likelihood* in [4] and *conditional node monitor* in [22].

We now see that if we use the partial marginal likelihood as a basis for a prequential scoring function, this results in a sequential process where at time i, the classification predictive distribution

$$P(V_i|\mathbf{v}^{i-1}, \mathbf{u}^{i}, M) = P(V_i|\mathbf{v}^{i-1}, \mathbf{u}^{i-1}, u_i, M) \qquad (4)$$

is computed by using the information preceding v_i in the matrix $\mathcal{D}$ (assuming that the values of V are stored in the last column of $\mathcal{D}$). Consequently, assuming the logarithmic loss function, this approach suggests that one should select the model M minimizing the following prequential model selection criterion

$$S(\mathbf{v}^N, \mathbf{u}^N|M) = \sum_{i=1}^{N} -\log P(\mathbf{v}_i \mid \mathbf{v}^{i-1}, \mathbf{u}^{i}, M). \qquad (5)$$

It is now important to notice that unlike in the unsupervised case, where the prequential log-loss score is equivalent to the marginal likelihood criterion and hence order-independent, the value of the partial marginal log-likelihood (5) depends on the ordering of the data. The ordering is of course irrelevant asymptotically, but this raises the question of whether the ordering is relevant with small sample sizes, and if it is, how should we then select the data ordering?

We can now distinguish two alternative approaches for addressing the ordering problem. First of all, if we think of our data as an unordered list (of vectors) instead of a vector (of vectors), this suggests that we should marginalize over different orderings; in other words, we should sum the partial marginal likelihood score (5) over all the permutations of the data. This marginalization is obviously computationally infeasible in practice, which leaves us with approximative methods. The simplest solution is to generate a number of random data orderings, and average the results over the individual prequential scores obtained.

An alternative viewpoint was taken in [19], where it was suggested that instead of summing over data orderings, given a model M, one should try to find the ordering minimizing the prequential score (5). Nevertheless, as in performing the marginalization, this minimization is again of course computationally infeasible in practice. For this reason, Rissanen suggested in [19, 20] a simple greedy procedure, where the data is ordered so that at each step the data vector yielding (locally) the largest gain in the prequential score is processed next. In the sequel we call the resulting model selection score the *best-first prequential score.*

The greedy best-first search is in most cases probably a poor optimization method that is prone to get stuck in a local optimum. Nevertheless, even this simple heuristic requires $O(N^2)$ time to run, which can be computationally demanding with large data sets. For this reason, in this paper we do not consider more elaborate optimization methods, but focus on using simple heuristics.

One alternative to the best-first approach is to use the worst-last procedure, where the ordering is determined from the last vector to the first vector in a greedy fashion, but so that at each stage the data vector yielding the smallest local gain in the prequential score (5) is chosen. This approach can be motivated by argumenting that the cases that are the most difficult to predict should be given as much history data as possible, hence those vectors should be processed last. Two obvious counterparts of the best-first and worst-last procedures are given by the best-last and worst-first heuristics.

4 Empirical Results

4.1 The Setup

In *finite mixture models* the classification predictive distribution (4) is obtained by a weighted sum of component distributions,

$$P(V_{i+1}|\mathbf{v}^i, \mathbf{u}^i, \mathbf{u}_{i+1}) = \sum_z P(V_{i+1}|\mathbf{v}^i, \mathbf{u}^{i+1}, Z = z)P(Z = z|\mathbf{v}^i, \mathbf{u}^{i+1}), \quad (6)$$

where Z denotes a (hidden) latent variable indexing the component distributions of the mixture. In the following setup we consider only finite mixture models where the variables X_i are assumed to be independent of each other, given the value of the latent variable Z.

One way to look at finite mixture models is to treat the latent variable Z as a clustering variable, the hidden values of which index the data source (a probability distribution representing a cluster of cases) where a data vector "originates" from. However, assuming this type of a latent variable is in contradiction with the assumption made in Section 2, where we assumed that the models used are such that the predictive distributions can be obtained by integrating over the parameters. For this reason, in the following we simplify the setup and assume that the values of the latent variable corresponding to each of the training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are set to some fixed values $\mathbf{z} = (z_1, \ldots, z_N)$. We do not address the problem of how to find the clustering $\mathbf{z}$, but assume it to be given. In this case the predictive distribution (6) is fully determined by $\mathbf{z}$, and we can regard the clusterings $\mathbf{z}$ as our models M.

The prequential model selection criterion alternatives discussed in Section 3 were empirically validated by using 14 classification data sets from the UCI data repository [2]. A single model selection experiment was performed in the following way. The data was first partitioned into two equal size sets, the training data and the test data. A pool of 40 candidate models (clusterings) $\mathbf{z}_1, \ldots, \mathbf{z}_{100}$ was then produced by running the K-means clustering algorithm (see, e.g., [8]) 40 times with the training data, starting from random initial points. The number of mixture components (the number of clusters, i.e., the number of possible values of Z) varied randomly between 3 and 20.

With each model selection criterion, all the 40 candidate models were then evaluated by using the criterion with the training data, and with each criterion, the model with the best score was selected. After that, all the selected models were evaluated by using the previously unseen test data, by computing both the log-score and 0/1-score for each of the test vectors. The score obtained by the candidate model selected by a model selection criterion was recorded as the prediction accuracy associated with this criterion.

One should observe that the test vectors were treated as independent classification tasks, not as a sequence, and with each model selection criterion, the average of the resulting N individual classification prediction scores was stored as the predictive accuracy obtained by using the criterion with this training data and test data. This whole procedure was then repeated 15 times by splitting the full data set randomly into training data and test data, and the same was repeated with all the 14 classification data sets. It should be emphasized that the experiment is completely fair in the sense that at no time before the actual classification task had the model selection criteria access to the test data.

4.2 The Results

The model selection scoring methods used in the experiments are listed in Table 1, and the results of the experiments are summarized in Figures 1 and 2. The predictive scores obtained with each model selection criterion are scaled with respect to the score obtained by the marginal likelihood model selection criterion, so that a score of 0,0% means the equivalent result as with the marginal likelihood criterion, and a score of, say +5.0%, means that the corresponding classification score was on the average 5.0% better than the score obtained with the marginal likelihood criterion.

Table 1: The methods used in the experiments.

ABBREVIATION	**EXPLANATION**
L-O-O	Leave-one-out crossvalidation.
PREQ-RAND(N)	The prequential score (5) averaged over N random permutations of the data.
PREQ-MAX(N)	The prequential score (5) optimized over N random permutations of the data.
BEST-FIRST	The prequential score (5) with the data ordering determined by a greedy best-first optimization.
WORST-LAST	The prequential score (5) with the data ordering determined by a greedy worst-last optimization.

From the figures we can see that all the relative scores are positive, which means that the supervised model selection criteria clearly outperformed the "unsupervised" marginal likelihood in the classification domains used in the experiments. Actually, all the supervised model selection scores tested gave almost always a positive relative score, and the score was in many cases

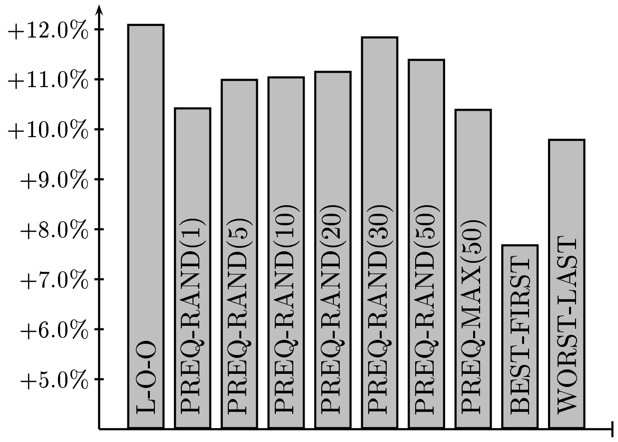

Figure 1: Average relative prediction gains with the logarithmic loss.

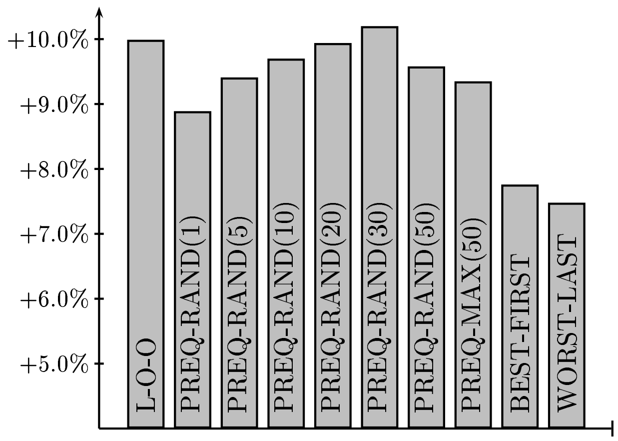

Figure 2: Average relative prediction gains with the 0/1-loss.

over 50% with the 0/1-score and over 30% with the log-score. It should be noted that in the log-score case the scale is logarithmic, which means that the differences in the relative score are in this case much more significant.

The results obtained with the PREQ-RAND(N) sampling method first improve as N increases, but decrease slightly after N reaches 30. However, the results are consistently better than those obtained with the greedy heuristics, or with the PREQ-MAX(N) method. These results suggest that averaging over the orderings gives better results than optimizing. An additional interesting observation is that in the log-score case the worst-last heuristic gave on the average better results than the best-first method suggested. The best-last and worst-first heuristics performed clearly worse than best-first and worst-last and were excluded from figures.

5 Conclusions and Future Work

We studied the model selection problem in supervised classification domains, and demonstrated empirically that the inherently unsupervised marginal likelihood model selection criterion can be outperformed in practice by the prequential approach and by crossvalidation, which were both designed for the supervised model selection problem at hand. The models used in this study consisted of finite mixture models with the typical assumption of independence between the domain variables, given the value of the clustering variable.

For addressing the ordering problem inherent to the prequential method used, we considered two alternative approaches: the sampling approach and the optimization approach. In the sampling approach the effect of data ordering was smoothed out by averaging the score over a number of random data permutations. In the optimization approach, motivated by the predictive MDL approach advocated by Rissanen [19, 20], the prequential score was determined by using the single data ordering optimizing the prequential score.

The results show that the prequential score with the sampling approach can lead to better results than crossvalidation. However, in the experiments reported here, the results do not seem to improve monotonically with the number of random orderings used, after a certain sample size is reached. We believe that this is probably a random effect caused by the relatively small number of repetitions used in the experiments. On the other hand, it is also possible that this result indicates that the limits of the naive random sampling have been reached, and in order to get better results with the sampling approach, one needs to use more elaborate sampling methods. This question will be studied in more detail in our future work.

The prequential score with the minimization approach did not produce as good results as the sampling approach, although the results were significantly better than with the "vanilla" marginal likelihood approach. Whether this means that the greedy search methods used were just overly naive, or that the sampling approach is the more proper solution for the ordering problem, remains as an open question that will also be studied in the future.

Acknowledgments

This research has been supported by the National Technology Agency and the Academy of Finland.

References

[1] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[2] C. Blake, E. Keogh, and C. Merz. UCI repository of machine learning databases, 1998. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[3] R. Cowell. On compatible priors for Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):901–911, September 1992.

[4] D.R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[5] A.P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, 147:278–292, 1984.

[6] A.P. Dawid. Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society B*, 53(1):79–109, 1991.

[7] A.P. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 109–125. Oxford University Press, 1992.

[8] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis.* John Wiley, 1973.

[9] S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, June 1975.

[10] P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty.* PhD thesis, CWI, ILLC Dissertation Series 1998-03, 1998.

[11] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In G. Cooper and S. Moral, editors, *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192, Madison, WI, July 1998. Morgan Kaufmann Publishers, San Francisco, CA.

[12] D. Heckerman and D. Geiger. Likelihoods and parameter priors for Bayesian networks. Technical Report MSR-TR-95-54, Microsoft Research, 1995.

[13] D. Heckerman and C. Meek. Models and selection criteria for regression and classification. In D. Geiger and P. Shenoy, editors, *Uncertainty in Arificial Intelligence 13*, pages 223–228. Morgan Kaufmann Publishers, San Mateo, CA, 1997.

[14] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, 186:453–461, 1946.

[15] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of Bayesian networks. In K. Laskey and H. Prade, editors, *Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pages 334–342. Morgan Kaufmann Publishers, 1999.

[16] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. *Statistics and Computing*, 10:39–54, 2000.

[17] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and K. Valtonen. Exploring the robustness of Bayesian and information-theoretic methods for predictive inference. In D. Heckerman and J. Whittaker, editors, *Proceedings of Uncertainty'99: The Seventh International Workshop on Artificial Intelligence and Statistics*, pages 231–236. Morgan Kaufmann Publishers, 1999.

[18] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, October 1998.

[19] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14(3):1080–1100, September 1986.

[20] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, New Jersey, 1989.

[21] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.

[22] D. Spiegelhalter, P. Dawid, S. Lauritzen, and R. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

[23] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)*, 36:111–147, 1974.

[24] Vovk. V. Competitive on-line statistics. Manuscript, submitted for publication.