

# Evaluating Model Selection Abilities of Performance Measures

Jin Huang and Charles X. Ling

Department of Computer Science  
The University of Western Ontario  
{jhuang, cling}@csd.uwo.ca

## Abstract

Model selection is an important task in machine learning and data mining. When using the holdout testing method to do model selection, a consensus in the machine learning community is that the same model selection goal should be used to identify the best model based on available data. However, following the preliminary work of (Rosset 2004), we show that this is, in general, not true under highly uncertain situations where only very limited data are available. We thoroughly investigate model selection abilities of different measures under highly uncertain situations as we vary model selection goals, learning algorithms and class distributions. The experimental results show that a measure's model selection ability is relatively stable to the model selection goals and class distributions. However, different learning algorithms call for different measures for model selection. For learning algorithms of SVM and KNN, generally the measures of RMS, SAUC, MXE perform the best. For learning algorithms of decision trees and naive Bayes, generally the measures of RMS, SAUC, MXE, AUC, APR have the best performance.

## Introduction

Some machine learning and data mining tasks, such as facial and hand writing recognitions, usually need to train a highly robust and accurate learning model. In these cases a learning model trained with the default or arbitrary parameter settings is not enough because it usually cannot achieve the best performance. To satisfy these requirements we vary the parameter settings to train more than one learning models and then select the best one as the desired model. Instances of selecting learning model include choosing the optimal number of hidden nodes in neural networks, choosing the optimal parameter settings of Support Vector Machines, and determining the suitable amount of pruning in building decision trees. This arises the *model selection* problem, which is an important task in statistical estimation, machine learning, and scientific inquiry (Vapnik 1982; Linhart & Zucchini). *Model selection* attempts to select the model with best future performance from alternate models measured with a model selection criterion. Traditional model selection tasks usually use accuracy as model selection criterion. However, some data mining applications often call for other measures as criteria. For example, ranking is an important task in machine

learning. If we want to select a model with best future ranking performance, then AUC (Area Under the ROC Curve), instead of accuracy, should be used as the model selection criterion. A model selection criterion is called "model selection goal". Holdout testing method is a primary approach to perform model selection. It uses a holdout data to estimate a model's future performance: repeatedly using a subset of data to train the model and using the rest for testing. In the testing process we may choose other measures to evaluate a model's performance. These measures are called "model evaluation measures". A common consensus in the machine learning community is that the model selection goal measure and the model evaluation measure should be same.

In practice we often encounter situations where resources are severely limited, or fast training and testing are required. We only have very limited data for model training and for future performance evaluation, which is called the highly uncertain situations. Naturally one may ask whether the common consensus that the model selection goal measure and the model evaluation measure should be same is also true under the highly uncertain conditions. Rosset (Rosset 2004) performed an initial research on this question with two special measures: accuracy and AUC. He compared the performance of model evaluation measures AUC and accuracy when the model selection goal is accuracy. He showed that AUC can more reliably identify the better model compared with accuracy for Naive Bayes and k-Nearest Neighbor models, even when the model selection goal is accuracy. However, his work has several limitations. First, he only chose very limited data (one synthetic dataset and one real world dataset) to perform the experiment. Second, he did not study model selection with different goals (other than accuracy) using different evaluation measures (other than AUC and accuracy), as learning algorithms and class distributions vary.

In this paper we thoroughly investigate the problem of model selection under highly uncertain conditions. We analyze the performance of nine different model evaluation measures under three different model selection goals, four different learning algorithms, on a variety of real world datasets with a wide range of class distributions.

We have obtained some surprising and interesting results. First, we show that the common consensus mentioned above is generally not true under the highly uncertain conditions. With the model selection goals of accuracy, AUC or lift,

many measures may perform better than these measures themselves. Second, we show that a measure’s model selection ability is relatively stable to different model selection goals and class distributions. Third, different learning algorithms call for different measures for model selection.

### Evaluation Measures

We review eight commonly used evaluation measures, Accuracy (acc), AUC, F-score (FSC), Average Precision (APR), Break Even Point (BEP), Lift, Root Mean Square Error (RMS), Mean Cross Entropy (MXE). Details of these measures can be found in (Caruana & Niculescu-Mizil 2004).

(Caruana & Niculescu-Mizil 2004) categorizes different machine learning measures into three groups: threshold measures, ranking measures, and probability-based measures. Accuracy, F-score, lift and Break Even Point are called threshold measures because they all use thresholds in their definitions. AUC and Average Precision have the common characteristic that they measure the quality of ranking: how well each positive instance is ranked compared with each negative instance. Thus they are called ranking measures as they only consider the ordinal relations of instances. RMS and MXE, however, depend on the predicted probabilities. This kind of measures are called probability-based measures. For RMS and MXE, the closer the predicted probabilities to the true probabilities, the smaller the values.

However, the ranking measures and probability-based measures both have some weaknesses. Ranking measures completely ignore the predicted probabilities, while probability-based measures need the true probabilities, which is usually not available in the real world applications. To overcome these weaknesses, a new measure, SAUC (Softened Area Under the ROC Curve), is proposed. Suppose that there are  $m$  positive examples and  $n$  negative examples. If we use  $p_i^+$ ,  $p_j^-$  to represent the predicted probabilities of being positive for the  $i$ th positive example and the  $j$ th negative example, respectively, then

$$SAUC = \frac{\sum_{i=1}^m \sum_{j=1}^n (p_i^+ - p_j^-) I(p_i^+ - p_j^-)}{mn} \quad (1)$$

where

$$I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

Clearly, SAUC is in the range of [0,1]. The closer the predicted probabilities to the true probabilities, the larger the SAUC. SAUC and AUC have the common point in that they both measure how each positive instance is ranked compared with each negative instance. However, AUC only cares whether each positive instance is ranked higher or lower than each negative instance, while SAUC also considers the probability differences in the ranking. In addition, SAUC also reflects how well the positive instances are separated from the negative instances according to their predicted probabilities. Thus SAUC can be categorized both as a ranking and a probability-based measure. As a more refined and delicate measure than AUC, SAUC can reflect both ranking and probability predictions.

### Experiments to Evaluate Measures for Model Selection

We perform experiments to simulate model selection tasks under highly uncertain conditions. The goal of these experiments is to study the model selection abilities of measures under different model selection goals, learning algorithms, and class distributions.

#### Model Selection Goals

In our experiments we choose three model selection goals: accuracy, AUC and lift. Accuracy is chosen because it is the most commonly used measure in a variety of machine learning tasks. Most of previous researches adopted accuracy as the model selection goal (Schuurmans 1997; Vapnik 1982). Ranking is increasingly becoming an important task in machine learning. We choose AUC as a model selection goal because it reflects the overall ranking performance of a classifier. Actually AUC has been widely used to evaluate, train and optimize learning algorithms in terms of ranking. We also choose lift as another model selection goal because it is very useful in some data mining applications, such as market analysis.

#### Data Sets and Learning Algorithms

We select 17 large data sets, each with at least 5000 instances. 13 of them are from the UCI repository (Blake & Merz 1998) and the rest are from (Delve 2003) and (Elena 1998). The properties of these datasets are listed in Table 1. All multiclass datasets are converted to binary datasets by categorizing some classes to the positive class and the rest to the negative class. For six multiclass datasets, letter, chess, artificial character, pen digits, isolet and satimage, we also vary the class distributions to generate more than one binary datasets. For example, the letter dataset contains 26 classes. We generate 6 different binary datasets with 50%, 38.2%, 25%, 11.5%, 7.8% and 4% of the positive class by selecting the letters of A-M, A-J, A-G, A-C, A-B, A as positive class, respectively. We generate different class distributions because we will investigate whether class distributions influence a measure’s model selection ability. From the multiclass datasets we can obtain a total of 41 binary datasets for our experiment as shown in Table 1.

We choose four learning algorithms: Support Vector Machine (SVM),  $k$ -Nearest Neighbor (KNN), decision trees (C4.5) and Naive Bayes in our study. We choose four different learning algorithms because we want to investigate whether different learning algorithms affect a measure’s model selection ability. For each learning algorithm we vary certain parameter settings to generate 10 different learning models with potentially different future predictive performance. For SVM, we choose the polynomial kernel with the degree of 2 and we vary the regularization parameter  $C$  with the values of  $10^{-6}$ ,  $10^{-5}$ ,  $\dots$ , 1, 10, 50, and 100. For KNN we set  $k$  with different values of 5, 10, 20, 30, 50, 100, 150, 200, 250, and 300. For C4.5 we vary tree construction stopping parameter  $m = 2, 5, 10$  and tree pruning confidence level parameter  $c = 0.1, 0.25, 0.35$ . For Naive Bayes we vary the number of attributes of each datasets used to train different

Table 1: Properties of datasets used in experiments

Dataset	Size	Training Size	Attribute #	Class #	Positive Class Ratio
Letter	20000	2000	16	26	50%, 38.2%, 25%, 11.5%, 7.8%, 4%
Adult	30162	4000	14	2	24.8%
Artificial Char	31000	2500	6	10	50%, 30%, 20%, 10%
Chess	28060	2500	6	16	47%, 23.5%, 10%, 5%
Page blocks	5473	1000	10	5	10.2%
Pen digits	10992	1000	16	10	50%, 40%, 30%, 1.4%, 7%, 3%
Nursery	10992	1000	8	5	33.3%
Covtype	29000	2900	54	7	48.8%
Connect-4	38770	3877	42	3	65.8%
Nettalk	20000	1000	3	2	28.2%
Musk	7075	700	50(166)	2	45%
Mushroom	8124	810	22	2	48.2%
Isolet	7797	780	60(617)	26	50%, 38.2%, 25%, 11.5%, 7%, 4%
Satimage	6435	640	5	7	9.7%, 23.8%, 30.8%, 47.2%
Phoneme	5427	540	5	2	29.4%
Texture	22000	2200	40	14	36.7%
Ringnorm	7400	740	21	2	27%

learning models. We train a sequence of Naive Bayes models with an increasing number of attributes used, with the attributes of any former model is the subset of any latter model. For example, for the pen digits dataset, we choose the first 1, 2, 4, 6, 8, 10, 12, 14, 15, 16 attributes in training 10 different Naive Bayes models. We use WEKA (Witten & Frank 2000) implementations for these algorithms.

## Experiment Process

We use the holdout testing method to perform model selection. Our approach is different from the standard cross validation or bootstrap methods. Here only a small sample of the original dataset is used to train learning models, and lots of small test sets are used to simulate the small future unseen data. This is a simple approach to simulate model selection in highly uncertain conditions (Rosset 2004). Given a model selection goal  $f$ , a model evaluation measure  $g$ , a learning algorithm and a binary dataset, we use the following experimental process to test the model selection ability of  $g$ .

The binary dataset is stratified<sup>1</sup> into 10 equal subsets. One subset is used to train different learning models and the rest are stratified into 100 small equal-sized test sets. We train 10 different learning models of the learning algorithm on the same training subset. For each model we evaluate it on the 100 small test sets. For two models  $X$  and  $Y$ ,  $X$  is better than  $Y$  iff  $E(f(X)) > E(f(Y))$ , where  $E(f(X))$  is the mean  $f$  score measured on  $X$ 's 100 testing results.  $g$  is used to measure  $X$  and  $Y$ 's testing results on each of the 100 testing sets and compare them to see whether or not they agree with  $E(f(X))$  and  $E(f(Y))$ . If  $f$  agrees with  $g$  then  $g$  selects the correct model; otherwise  $g$  selects the wrong model. We count in how many cases (among 100) that  $g$  selects the correct model. This leads a percentage (or probability) that

<sup>1</sup>“stratify” means to partition a dataset into some equal-sized subsets with the same class distribution.

$g$  can choose the better model between  $X$  and  $Y$ , representing how well a measure can do in selecting model. When all pairs of learning models are considered, we use the measure  $MSA$  to reflect the overall model selection ability of  $g$ . It is defined as

$$MSA(g) = \frac{2}{N(N-1)} \sum_{i < j} p_{ij}$$

where  $N$  is the number of learning models ( $N = 10$ ),  $p_{ij}$  is the probability that measure  $g$  can correctly identify the better one from models  $i$  and  $j$ .

We repeat the above process 10 times by choosing a different subset for training each time. We use the average  $MSA(g)$  to measure the model selection ability of  $g$ .

## Experimental Results Analysis

We use the  $MSA$  measure as the criterion to explore two issues from the experimental results. First, we will compare the  $MSA$  of the goal measure with other measures. This will tell us whether it is true that we should always use the model selection goal as the evaluation measure to do model selection. Second, we will explore whether different model selection goals, class distributions and learning algorithms influence a measure's model selection ability.

To clearly explore the above two issues, we need to directly present and analyze the  $MSA$  of all the measures in all cases. If a model selection task with a specific model selection goal, dataset, and learning algorithm is called a “model selection case”, there are a large number of such model selection cases. One direct approach to clearly show the  $MSA$  of different measures is to use a figure to depict the  $MSA$  performance for each model selection case.

However, the major problem of this approach is that there are too many such figures to be presented. Since in our experiments we use 41 binary datasets, 4 learning algorithms

and 3 model selection goals, there are totally  $41 \times 4 \times 3 = 492$  figures. If these figures are categorized according to different model selection goals, there are 164 figures for each model selection goal category. On the other hand, it is also difficult to choose the representative and diverse figures for different model selection cases.

To overcome this difficulty, we use a statistical method to evaluate a measure's MSA. To compare a measure's MSA with that of a model selection goal, we categorize the model selection cases according to different model selection goals. For each model selection case, there is a measure that achieves the best MSA. We compute the percentage of the cases in which one measure can reach the maximum MSA within a varying  $x\%$  tolerance range, to the total cases. This percentage indicates the success rate that one measure can reach the maximum within an  $x\%$  range. The success rates of different measures can be depicted in a figure, in which each curve line represents the success rate of a measure.

**Comparing a Measure's MSA with Goal Measure** Figure 1(a) depicts the success rates of different measures when we choose accuracy as the model selection goal, while varying the tolerance ranges from 1% to 5%. We can see that the measures SAUC, RMS, MXE, AUC, APR statistically perform better than accuracy for different learning algorithms and datasets. The measures lift and BEP, however, are constantly worse than accuracy.

In Figure 1(b) AUC is used as the model selection goal. Only SAUC, RMS and MXE perform better than AUC in most of the sub figures. All other measures are inferior to AUC.

In Figure 1(c) lift is used as model selection goal. We can see that except for BEP all measures are constantly better than lift. Furthermore, by comparing Figure 1(c) with Figure 1(b) and Figure 1(a), we can see that the differences of success rates between SAUC, RMS, MXE, AUC, APR with lift are much more than their corresponding differences with accuracy and AUC in Figures 1(a) and 1(b).

The above discussion shows that under the highly uncertain condition, in general, we should not use the model selection goal measure to perform model selection. This result extends the preliminary work of (Rosset 2004) to more general situations.

**The Stability of a Measure's MSA** We next discuss whether one measure's MSA is stable under different model selection goals, class distributions, and learning algorithms.

#### (i) Model Selection Goals

From the analysis of the previous subsection, we can see that a measure's absolute ability (MSA) is stable to the model selection goals.

#### (ii) Class Distributions

To explore whether class distributions influence a measure's MSA, we analyze the experimental results according to the datasets with different class distributions. The experimental results are categorized into three groups according to the datasets with class distributions of 40%-50%, 25%-30%, 1.4%-10%, respectively. Each group includes the experimental results with all model selection goals and learning models. The success rates of measures are depicted in Figure

2. If we rank measures according to their MSA, we can see that generally this ranking is stable to class distributions.

#### (iii) Learning Algorithms

We explore how a measure's MSA is influenced by different learning algorithms. We first discuss how different measures perform for the learning models of SVM and KNN. Here we fix the learning algorithms and vary the datasets and model selection goals. The success rates of measures are depicted in Figure 3(a) and Figure 3(b) for SVM and KNN, respectively.

As shown in Figure 3(a) and 3(b), the measures can be categorized into three different groups according to their performance.

The probability-based measures, including SAUC, RMS and MXE, achieve the best performance. MXE and RMS perform very similarly in most situations. The second group of measures, including AUC and APR, are inferior to the first group measures (SAUC, RMS and MXE). The third group includes the measures of accuracy, F-score, BEP, and lift. This group measures are inferior to the second group measures. F-score is generally competitive with accuracy. Lift and BEP are the two measures always with the worst performance.

Surprisingly, the above three groups of measures match the categories of probability-based measures, ranking measures and threshold measures. Therefore it seems that there is a strong correlation between a measure's category with its model selection ability. An appropriate explanation lies in two aspects. First, the outstanding performance of probability-based measures (RMS, MXE) is partly due to the high quality probability predictions of SVM and KNN learning algorithms. Second, the discriminatory power of the measures also plays an important role. The discriminatory power of a measure reflects how well this measure can discriminate different objects when it is used to evaluate them. Generally a measure's discriminatory power is proportional to the different possible values it can reach. As an example, for a ranked list with  $n_0$  positive instances and  $n_1$  negative instances, accuracy and lift can only reach  $n_1 + n_0$  and  $(n_0 + n_1)/4$  different values (if we use a fixed 25% percentage for lift). The ranking measure AUC can reach  $n_0 n_1$  different values. The probability-based measure RMS, however, can have infinitely many different values. Thus these measures can be ranked according to their discriminatory power (from high to low) as RMS, AUC, accuracy, lift. This discriminatory power ranking matches with the model selection performance sequence. Therefore we can claim that a measure's model selection ability is closely correlated with its discriminatory power for the SVM and KNN learning algorithms. The possible reason is that a measure with high discriminatory power usually uses more information in evaluating objects and thus is more robust and reliable. Probability-based measures use the predicted probability information, and thus they are more accurate than ranking measures which only use the relative ranking position information. Similarly, ranking measures also use more information than accuracy or lift, which only considers the classification correctness in the part or whole dataset ranges.

However, compared with SVM and KNN learning al-

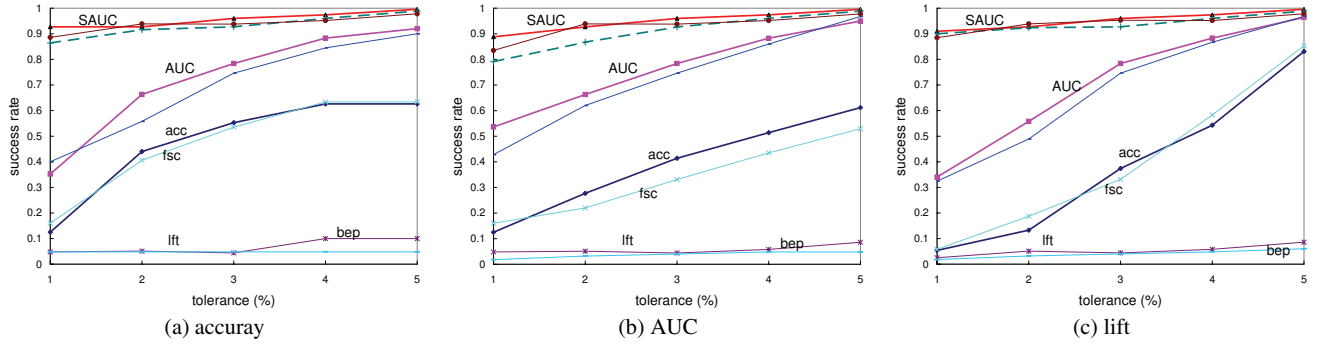


Figure 1: Ratio of datasets on which each measure's MSA is within x% tolerance of maximum MSA, using accuracy, AUC and lift as model selection goals.

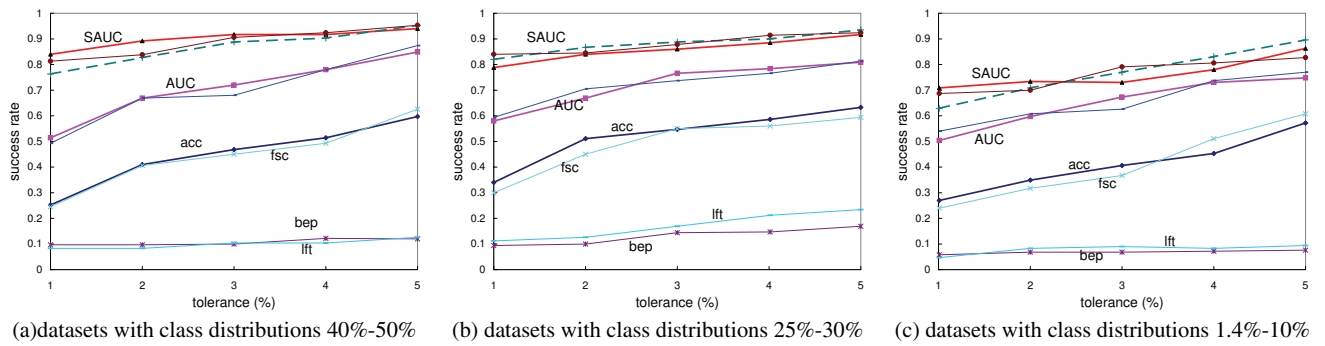


Figure 2: Ratio of datasets on which each measure's MSA is within x% tolerance of maximum MSA, for datasets with varied class distributions.

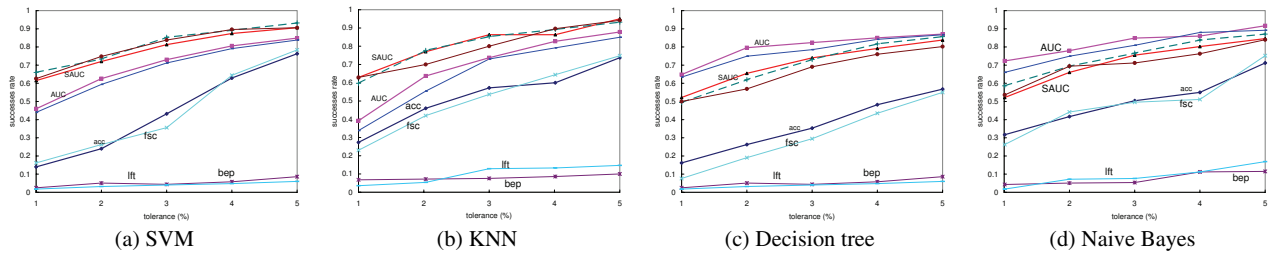


Figure 3: Ratio of datasets on which each measure's MSA is within x% tolerance of maximum MSA, with SVM, KNN, Decision tree and Naive Bayes algorithms.

gorithms, measures perform differently for decision trees (C4.5) and Naive Bayes. The success rate graphs are shown in Figure 3(c) and Figure 3(d) for Naive Bayes and decision trees. We can see that probability-based measures do not always perform better than ranking measures. This indicates that they might be unstable for some datasets and model selection goals. By comparing ranking measures with threshold measures, however, we can see that these two kinds of measures are less influenced by learning algorithms. We can conclude that generally the measures of RMS, SAUC, MXE, AUC, APR have the best performance for decision trees (C4.5) and Naive Bayes algorithms.

(Domingos & Pazzani 1997; Provost, Fawcett, & Kohavi 1998) have shown that learning algorithms of C4.5 and Naive Bayes usually produce poor probability estimations. The poor probability estimations directly degrade the performance of SAUC, RMS and MXE when they are used to rank learning models. This explains why the probability-based measures perform unstably for C4.5 and Naive Bayes models. Although the poor probability estimations also influence the ranking measures of AUC and APR, these influences are not so strong. This also explains why the ranking measures relatively perform stably.

In summary, from the above discussions we can draw the following conclusions.

1. For model selection tasks under the highly uncertain conditions, the common consensus that the goal measure should be used to do model selection is not true.
2. A measure's model selection performance is relatively stable to the selection goals and class distributions.
3. Different learning algorithms need to choose different measures for model selection tasks. For learning algorithms with good quality of probability predictions (such as SVM and KNN) a measure's model selection ability is closely correlated with its discriminatory power. The probability-based measures (SVM, SAUC, MXE) perform best, followed by ranking measures (AUC, APR), followed by threshold measures (Accuracy, FSC, BEP, lift). For learning algorithms with poor probability predictions (such as C4.5 and Naive Bayes), the probability-based measures such as SVM, SAUC and MXE perform quite unstable. AUC and Average Precision become robust and well performed measures.

### Conclusions and Future Work

Model selection is a significant task in machine learning and data mining. In this paper we perform a thorough empirical study to investigate how different measures perform in model selection under highly uncertain conditions, with varying learning algorithm, model selection goals and dataset class distributions. We show that a measure's model selection performance is relatively stable by model selection goals and class distributions. However, different learning algorithms call for different measures for model selection.

For our future work, we plan to investigate model selection tasks under other uncertain conditions. We also plan to devise new model selection measures that are specialized under different conditions.

### References

- Blake, C., and Merz, C. 1998. UCI repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>. University of California, Irvine, Dept. of Information and Computer Sciences.
- Caruana, R., and Niculescu-Mizil, A. 2004. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the 10th ACM SIGKDD conference*.
- Delve. 2003. Delve project: Data for evaluating learning in valid experiments. <http://www.cs.toronto.edu/delve/>.
- Domingos, P., and Pazzani, M. 1997. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning* 29:103–130.
- Elena. 1998. Elena datasets. <ftp://ftp.dice.ucl.ac.be/pub/neural-nets/ELENA/databases>.
- Linhart, H., and Zucchini, W. *Model Selection*. New York:Wiley.
- Provost, F.; Fawcett, T.; and Kohavi, R. 1998. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann. 445–453.
- Rosset, S. 2004. Model selection via the AUC. In *Proceedings of the 21st International Conference on Machine Learning*.
- Schuermans, D. 1997. A new metric-based approach to model selection. In *Proceedings of National Conference on Artificial Intelligence(AAI-97)*.
- Vapnik, V. 1982. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag NY.
- Witten, I. H., and Frank, E. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco.