Published in final edited form as:

Semin Oncol. 2010 February ; 37(1): 65–68. doi:10.1053/j.seminoncol.2009.12.002.

Learning Curves in Classification with Microarray Data

Kenneth R. Hess and Caimiao Wei

Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston TX, 77030 USA

Kenneth R. Hess: khess@odin.mdacc.tmc.edu

Abstract

The performance of many repeated tasks improves with experience and practice. This improvement tends to be rapid initially and then decreases. The term “learning curve” is often used to describe the phenomenon. In supervised machine learning, the performance of classification algorithms often increases with the number of observations used to train the algorithm. We use progressively larger samples of observations to train the algorithm and then plot performance against the number of training observations. This yields the familiar negatively accelerating learning curve.

To quantify the learning curve, we fit inverse power law models to the progressively sampled data. We fit such learning curves to four large clinical cancer genomic datasets, using three classifiers (diagonal linear discriminant analysis, K-nearest-neighbor with 3 neighbors and support vector machines) and four values for the number of top genes included (5, 50, 500, 5000). The inverse power law models fit the progressively sampled data reasonably well and showed considerable diversity when multiple classifiers are applied to the same data. Some classifiers showed rapid and continued increase in performance as the number of training samples increased, while others showed little if any improvement. Assessing classifier efficiency is particularly important in genomic studies since samples are so expensive to obtain. It is important to employ an algorithm that uses the predictive information efficiently. But with a modest number of training samples (>50), learning curves can be used to assess the predictive efficiency of classification algorithms.

Introduction

The performance of many repeated tasks improves with experience and practice (Ramsay, 2001; Ritter, 2001). This improvement tends to be rapid initially and then decreases over time, ultimately reaching a plateau. The term ‘learning curve’ is often used to describe this phenomenon. Such curves have been used since the early 1900s in studies of cognitive and motor learning, (Thorstone, 1919) and since the 1930s in studies of manufacturing labor and cost (Yelle, 1979). The term ‘learning curve’ became part of the popular culture and language in the 1960s with the publication and popularization of many psychological studies of human performance (Ramsay, 2001).

There is an amazing ubiquity to these negatively accelerating learning curves. The same basic shape has been seen in individual short-term perceptual tasks such as performing mental arithmetic, in cooperative short-term tasks such as surgical procedures (Ramsay, 2001) and in cooperative long-term tasks such as airplane or ship-building (Ritter, 2001). Research in cognitive psychology and the theory of learning has been used to explain the negative acceleration phenomenon (Ritter, 2001).

In the 1980s, learning curves were applied to artificial intelligence and computational learning theory (Kadie, 1991). In machine learning, the performance of classification or rule induction algorithms often increases with the number of observations used to train the algorithm. If performance is plotted against the number of training observations, the familiar, negatively accelerating curve is often seen. With massive datasets, learning curves can be used to find the size of a smaller sample of the observations that can be used to train an algorithm, yielding similar accuracy to that achieved when training on the entire dataset (Provost, 1999). Progressively larger samples of observations are used to estimate performance and then these data are typically analyzed by fitting a mathematical model to them. A power law model is one choice to yield the usual negative acceleration, but other models are also used (Ramsay, 2001, Gu, 2001).

In genomic studies, learning curves have been used to estimate the number of samples needed to train a particular algorithm to achieve a specified accuracy (Mukherjee, 2003). After selecting a specific classifier and set of features to use and obtaining a reasonable number of training samples (e.g., 50 to 70), the learning curve is fit just as for massive datasets (i.e., with progressively larger samples). The learning curve is fit to these pilot data and then results are extrapolated to estimate the number of samples needed to achieve the desired accuracy.

Methods

Datasets

We downloaded datasets for our study from the Oncomine website (<http://www.oncomine.org/>). We sought out datasets with: large numbers of samples (> 150), use of clinical cancer samples, use of the Affymetrix platform, and binary classification problems with no fewer than 20 samples in the smaller of the two classes. The first dataset (Bowtell) consists of data from 196 ovarian cancer patients: 78 with grade 2 cancer and 118 with grade 3 cancer. The second dataset (Bittner) consists of data from 387 cancer patients: 199 with ovarian cancer and 188 endometrial cancer. The third dataset (Desmedt) consists of data from 198 breast cancer patients: 134 with ER positive tumors and 64 with ER negative tumors. The fourth dataset (Ivshina) consists of data from 247 breast cancer patients: 58 with p53 mutations and 189 without. Datasets were normalized via RMA prior to analysis.

Progressive sampling

We performed stratified (i.e., done in such a way as to preserve the class distributions) progressive sampling, with training dataset sizes $N_{tr} = 20$ to $N_{max} - 20$ by 5 (where N_{max} is the total number of samples available). The test set size was $N_{max} - N_{tr}$ (i.e., all of the

remaining samples). At each training set size, we randomly selected points for the training set. We repeated this process five times at each training set size.

Prediction performance was examined for three widely used classifiers: diagonal linear discriminant analysis (DLDA), K-nearest neighbor with 3 neighbors (kNN3) and support vector machines (SVM) with a linear kernel, on each of the datasets (Dudoit, 2002). DLDA and SVM fit hyper-planes to the data (in quite different ways) to separate classes, while the boundary between classes can be very nonlinear for kNN3. Progressive sampling, classification, and learning curve fitting were performed with the R-language for statistical computing. DLDA and kNN3 were implemented with internal scripts, SVM with the e1071 R library, downloadable from the R-Project CRAN contributed packages website: <http://cran.us.r-project.org/>.

At each repetition of progressive sampling we ranked the top genes using a between-class t-test on log-transformed expression values. For each classifier we formed models using the top 5, 50, 500 and 5000 genes and recorded the area under the ROC curve (AUROCC) (Pepe, 2003). Note that we are using the term model to refer to a given classifier used with a given number of top genes. It is also important to note that we re-ranked the genes for each repetition of sampling prior to fitting the model and estimating performance. This provides a more realistic estimate of performance akin to full cross-validation.

Learning Curve Estimation

The learning curve model we used, often called an inverse power-law model, is defined as

$$E(Y) = \alpha + \beta \times N_{tr}^{-\gamma} \quad \alpha, \beta, \gamma \geq 0 \quad (1)$$

In this representation, Y is area above the ROC curve ($AAC = 1 - AUROCC$), and N_{tr} is the training set size. $E()$ is the expectation operator. Here α is the minimum AAC that can be expected as $N_{tr} \rightarrow \infty$, γ is referred to as the learning rate, and β the scale. The speed with which an individual classifier learns is a function of both γ and β , as both parameters determine the 1st and 2nd derivatives of the learning curve. The model parameters were estimated using the `nlminb` function in R (Huet, 2004). We use AAC in order to have a performance metric that operates such that lower values are better.

RESULTS

Progressive sampling data with inverse power law learning curves for kNN3 with the top 5 genes for each dataset are shown in Figures 1(B)(a)–(d). These demonstrate that the inverse power-law model captures the trend in performance reasonably well. The classic shape of negative acceleration in performance is seen. The variability of points around the inverse power law curves indicates the uncertainty of the estimates. Similar results were seen for the other classifiers and the other numbers of top genes.

The curves for the 12 combinations of classifier and number of top genes are shown for each of the 4 datasets in Figures 2(a) – (d). For each dataset, learning curves vary considerably in initial performance, limiting performance and in the rate of transition from initial

performance to limiting performance. For all but a few models, performance on the test set improves as the size of the training set. In no dataset was the model with the best initial performance also the model with the best limiting performance. In one dataset, the model with the second worst initial performance had the best limiting performance (Desmedt, SVM, 5 top genes). Generally, the model with the best performance at 25 samples is not the model with the best performance at 250 samples.

Discussion and Conclusion

Our findings support the inverse power law relationship between model prediction performance and training set size for genomics datasets. The diversity of the 12 learning curves fit to a given dataset illustrates how different models fit to the same data behave differently as the training set size increases. Some models make efficient use of the increased number of training samples, while others do not. Learning curves can be useful for both estimating predictive efficiency for a particular classifier and for selecting among several candidate models based on their efficiency. It is important to note that in the present study the relative efficiency of a given model varied considerably from dataset to dataset. Thus it is important to assess for each dataset how candidate models perform.

The observation that machine learning performance generally improves with the number of training samples is an important consideration when designing genomic prediction studies. Determining the number of training samples and the classifier to be used are key elements in the design of such studies. Learning curves can be used to assess the predictive efficiency of classification algorithms. Assessing classifier efficiency is particularly important in genomic studies since samples are so expensive to obtain. It is important to employ an algorithm that uses the predictive information efficiently.

References

- Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc.* 2000; 97:77–87.
- Gu, B.; Feifang, H.; Liu, H. Research Paper. National University of Singapore; Republic of Singapore: 2001. Modeling classification performance for large data sets: an empirical study. (Available at <http://www.nus.edu.sg>)
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction.* Springer-Verlag; New York, NY: 2001. p. 214-221.
- Huet, S.; Bouvier, A.; Poursat, MA.; Jolivet, E. *Statistical Tools for Nonlinear Regression. A Practical Guide with S-PLUS and R Examples.* 2. Springer-Verlag; New York, NY: 2004. p. 137-139.
- Kadie, CM. Quantifying the Value of Constructive Induction, Knowledge, and Noise Filtering on Inductive Learning. *Proceeding of the 8th Machine Learning Workshop*; 1991. p. 153-157.
- Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*; Morgan Kaufman, San Francisco, CA. 1995. p. 1337-1143.
- Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, Golub TR, Mesirov JP. Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol.* 2003; 10:119–142. [PubMed: 12804087]
- Pepe, MS. *Oxford Statistical Science Series 28.* Oxford University Press; New York, NY: 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*; p. 66-129.
- Provost, F.; Jensen, D.; Oates, T. Efficient progressive sampling. *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*; 1999; 1999. p. 23-32.

- Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT. Statistical assessment of the learning curves of health technologies. *Health Technol Assess*. 2001; 5:1–79. [PubMed: 11319991]
- Ritter, FE.; Schooler, LJ. The learning curve. In: Smelser, NJ.; Baltes, PB., editors. *International Encyclopedia of the Social and Behavioral Sciences (IESBS)*. Elsevier Science Ltd; New York, NY: 2001. p. 8602-8605.
- Thorstone. The learning curve equation. *Psych Mon*. 1919; 26:51.
- Yelle LE. The learning curve: historical review and comprehensive survey. *Decision Sci*. 1979; 10:302–327.

Figure 1. Progressive sampling data with inverse power law learning curves. kNN3 using the top 5 genes.

Figure 2.
Fitted inverse power law learning curves