Predicting and Optimizing Classifier Utility with the Power Law

Mark Last
Ben-Gurion University of the Negev, Israel
mlast@bgu.ac.il

Abstract

When data collection is costly and/or takes a significant amount of time, an early prediction of the classifier performance is extremely important for the design of the data mining process. Power law has been shown in the past to be a good predictor of decision-tree error rates as a function of the sample size. In this paper, we show that the optimal training set size for a given dataset can be computed from a learning curve characterized by a power law. Such a curve can be approximated using a small subset of potentially available data and then used to estimate the expected trade-off between the error rate and the amount of additional observations. The proposed approach to projected optimization of classifier utility is demonstrated and evaluated on several benchmark datasets.

1. Introduction

In real-world data mining projects, the amount of available data that can be used for inducing a data mining model is often restricted by economic, timing, and other constraints. Examples of such constraints include a limited number of medical records that can be provided by a hospital over time, high costs of manufacturing records representing results of an engineering experiment, low frequency of obtaining new seasonal records in an agricultural database, etc. All these situations require a careful consideration of the added value of new examples (e.g., in terms of the improved predictive accuracy) vs. the costs and/or the times involved in acquiring those examples.

According to [13], the ultimate goal of *utility-based data mining* is to maximize the utility of the entire data mining process by taking into account all utility considerations. In the case of a classification task, this means considering the trade-off between an increase in predictive accuracy and the cost of acquiring new data. As shown in [13], for each data set and learner there is an optimal training set size that maximizes the overall utility of the classifier. Given a ratio between the data acquisition cost and the error cost, a nearly optimal sample size can be found iteratively by one of progressive sampling schemes presented in [13]. A clear disadvantage of the progressive sampling approach, beyond the potential overhead associated with each sampling increment, is its inability to estimate in advance such parameters as the minimal achievable error rate subject to the budgeting constraints, the number of examples required to reach a particular error rate, or the optimal training set size that is expected to maximize the overall utility.

Frey and Fisher [2] have conducted an extensive set of experiments showing that the power law is the best fit for modeling the error rates of the C4.5 decision-tree algorithm [10]. The percentage of explained variation (r^2) of the power law was compared to r^2 of linear, logarithmic, and exponential functions across 14 benchmark datasets of relatively small size (having less than 1,000 instances on average). The power law has produced the highest value of r^2 in 12 datasets out of 14 resulting in the best models predicting diminishing returns in the error rate for increasing the amount of training data. In their further experiments, the authors of [2] have also shown that a power law model derived from a small portion of data (15%) can be used to reliably estimate the error rate for decision trees learned on the remaining amount of data as a function of the training set size.

A more recent study by Singh [12] argues that the power law is only second best to the logarithmic regression for a variety of classification algorithms, namely ID3, k-Nearest Neighbors, Support Vector Machines, and Artificial Neural Networks. His results are based on four datasets from the UCI Repository [9] with the number of instances varying between 101 and 1,728. The C4.5 decision-tree algorithm used by Frey and Fisher in [2] was not included in Singh's experiments.

This paper is organized as follows. In Section 2 we present the calculation of the optimal training set size for a given dataset based on the total utility measure of [13] and a learning curve characterized by a power law. The proposed optimization methodology is applicable to *any* classification algorithm and sampling technique provided that the classifier performance can

0-7695-3019-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDMW.2007.31

be accurately approximated by this type of a learning curve. Section 3 evaluates the proposed approach to projected optimization of classifier utility on six benchmark datasets using an oblivious decision tree algorithm (Information Network [6]), which is shown to fit the power law better than other performance models (linear, logarithmic, and exponential). The paper is concluded by Section 4, which summarizes the results and outlines directions for future research.

2. Optimizing the Training Set Size

Weiss and Tian [13] have defined the total utility of the classification process as a sum of the two terms: Data Cost and Error Cost. Based on the common model assessment approach (see [4]), the authors of [13] assume the data to be divided into three parts: the *training set* of n examples used to induce the model, the *test set* used to estimate the model accuracy, and the *score set* S of future examples to be classified by the model. The Error Cost in [13] is proportional to the number of errors made when classifying the score data set S. Without loss of generality, a fixed size of S (e.g., 100) can be assumed. The Data Cost is just the cost of data acquisition. Two unit costs involved include C_{tr} for acquiring each new training example and C_{err} for each misclassified example from the score set. The error rate measured on the test set is denoted by *err*. Consequently the total cost of a classifier can be calculated by the following expression:

$$\text{Total Cost} = n \cdot C_{tr} + err \cdot |S| \cdot C_{err} \quad (1)$$

If the error rate is characterized by the power law, it can be projected using the following equation (based on [2]):

$$err = a \cdot n^{-b} \quad (2)$$

where a and b are two non-negative coefficients, which can be easily calculated from the available training data by applying the linear regression model to the logarithmic transformations of n and *err*. This approach can be verified by taking a *log* or a *ln* of both sides in Eq. (2):

$$\log err = \log a - b \cdot \log n \quad (3)$$

Substituting the power function (2) into Eq. (1) results in the following expression for the Total Cost as a function of the training set size n:

$$\text{Total Cost}(n) = n \cdot C_{tr} + |S| \cdot C_{err} \cdot a n^{-b} \quad (4)$$

Due to the fact that the Data Cost in Eq. (4) is a non-decreasing function of n while the Error Cost, based on the power law, is a non-increasing function of n, an optimal trade-off between these two costs should exist. To find the optimal training set size n^* that minimizes the above cost function we need to compute the first derivative of Total Cost (n) and then set it to zero:

$$(n \cdot C_{tr} + |S| \cdot C_{err} \cdot a \cdot n^{-b})' = 0 \quad (5)$$

$$C_{tr} - |S| \cdot C_{err} \cdot a \cdot b \cdot n^{-b-1} = 0 \quad (6)$$

$$n^* = \left(\frac{|S| \cdot C_{err} \cdot a \cdot b}{C_{tr}} \right)^{\frac{1}{1+b}} \quad (7)$$

Since the optimal training set size in Equation (7) depends only on the ratio of the costs, we can follow another simplification of [13] by assuming C_{tr} to be equal to 1. Consequently we obtain the following expression for n^*:

$$n^* = \left(|S| \cdot C_{err} \cdot a \cdot b \right)^{\frac{1}{1+b}} \quad (8)$$

It can be easily shown that n^* is the *global minimum* of Total Cost (n). Since a and b can take only non-negative values, the expression n^{-b-1}, or $1/n^{b+1}$, is a monotone non-increasing function of n and, consequently, the first derivative of Total Cost (see Eq. 6) is a monotone non-decreasing function of n. This implies that Total Cost (n) is a convex function and thus, according to [1], the point n^*, where Total Cost$'(n) = 0$ is a global minimum.

To make the Equation (8) useful for estimating the optimal amount of training examples required for inducing a classification model from a real-world dataset, the following conditions should hold:

1) The learning curve of the classification algorithm fits the power law.
2) A small portion of the available data is sufficient for a reliable projection of the learning curve on examples to be collected in the future.

The usefulness of the optimal training set size calculated by Equation (8) can also be evaluated directly by comparing the actual minimal cost (found by progressive sampling) to the cost associated with n^*. In the next section, we apply the proposed methodology to several benchmark datasets using one of decision-tree classifiers.

3. Empirical Results

3.1. Design of Experiments

Decision trees are considered one of the most popular classification methods [3]. Since the learning curves of C4.5 and ID3 algorithms have been studied elsewhere (see [2] and [12] respectively), this paper focuses on a different decision-tree algorithm called Information Network (IN), which was shown in [6] to produce more compact and stable decision-tree models than C4.5, while preserving nearly the same level of predictive accuracy. Information network is an oblivious read-once decision graph built by a top-down information-theoretic algorithm, which uses the likelihood-ratio test as a pre-pruning criterion. More details on IN can be found in [6] and [7].

Similarly to [13], each dataset was randomly partitioned into approximately 25% of test examples and approximately 75% of examples potentially available for training. The actual percentage of training examples (out of the examples remained for training) was varied between 1% and 99% in increments of 1% at a time. To increase the statistical significance of our results, the error rate of each percentage is based on averages over 50 random partitions of the training set. The equation of the learning curve used to find the optimal amount of training examples was computed from the first 15 points (i.e., only 15% of the training data). The same amount of data was used in [2] to build projected learning curves. The projected optimum was evaluated on the remaining 84 points.

We have used the entire training sets to compare the percentage of explained variation (r^2) of the power law to r^2 of linear, logarithmic, and exponential functions shown by equations (9-11) respectively, where n stands for the amount of training data and a, b are the specific parameters of each function:

$$err = an + b \qquad (9)$$

$$err = a\log n + b \qquad (10)$$

$$err = ab^n \qquad (11)$$

Though previous studies have already demonstrated the non-linear behavior of most learning curves, we use the linear fit as a baseline in this paper.

In our experiments, we have analyzed 6 datasets from the UCI Machine Learning Repository [9]. The datasets are described in Table 1. Three datasets have between 600 and 900 records, while the other three contain several thousands of records. We believe that such "medium-size" datasets are most appropriate for studying the learning curves of classification algorithms, since small datasets may not extend at all into the "plateau" region of the curve, where the optimal sample size is usually located, while in the large datasets a majority of records may be completely redundant due to their negligible contribution to a decrease in the error rate.

Table 1. Datasets description

Dataset	Total Size	Potentially Training Examples	Test Examples
Breast Cancer	699	523	176
Chess	3196	2384	812
Hypothyroid	3163	2379	784
Mushroom	8124	6213	1911
Soybean large	683	510	173
Vehicle	846	623	223

3.2 Modeling the Learning Curves

The learning curves of the Information Network algorithm on the three largest datasets out of six are shown in Figures 1-3. In general, all curves are characterized by diminishing returns in error rate for increasing the training set size. However, there are some minor differences between the datasets, which are worth noting. Thus, in the Hypothyroid curve, a long plateau of nearly a constant error rate is followed by an inflection point, which is very close to the maximum amount of available training records. When the training set is slightly increased beyond this inflection point, an additional though small decrease in the error rate is observed. Though the Hypothyroid dataset preserves the non-increasing behavior of the error rate as a function of n, the last few points of its learning curve may not fit the same function as the other points. Another dataset (Chess) exhibits a slight increase of the error rate for the last few points, which appears to be a result of overfitting. Obviously, extending the training set into this range should be counter-productive disregarding the value of the unit error cost All these small departures from the "ideal" diminishing returns behavior (e.g., shown by Mushroom dataset) are quite marginal and they do not affect the vast majority of points in each learning curve.

Figure 1. Learning curve for Chess dataset

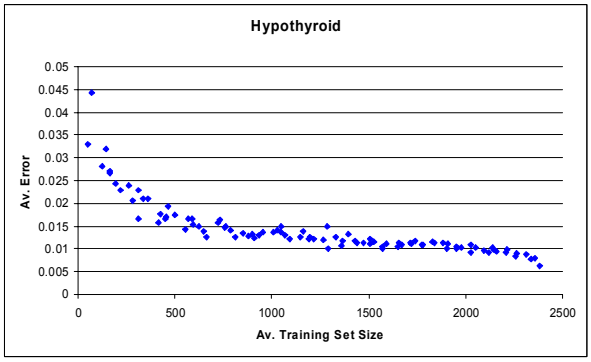

Figure 2. Learning curve for Hypothyroid dataset

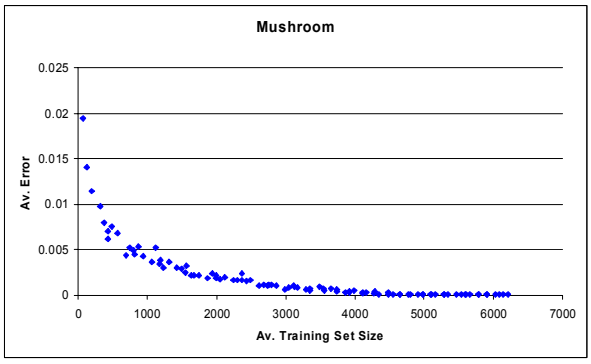

Figure 3. Learning curve for Mushroom dataset

In Table 2, we compare the percentage of explained variation (r^2) of the power law, linear, logarithmic, and exponential functions that are fit to the average error rates of the six datasets across the training set size varied from 1% to 99% of available examples. The largest values of r^2 per each dataset are shown in bold. The power law is the best fit of the learning curve for five datasets out of six, with its r^2 values ranging between 0.84 and 0.93. The only exception is the Mushroom dataset, where the exponential fit is much better than the power law (r^2 = 0.96 vs. 0.80). It is worth noting that all correlation coefficients in Table 2, including the lowest ones representing the linear fit, were found statistically significant using the *t*-test based on [8].

Table 2. Function fits to the learning curves (r^2)

Dataset	Power	Linear	Log	Exp
Breast Cancer	**0.8419**	0.3058	0.6529	0.5417
Chess	**0.8913**	0.2118	0.5380	0.5424
Hypo-thyroid	**0.9344**	0.6204	0.8871	0.7866
Mushroom	0.7974	0.5769	0.9191	**0.9639**
Soybean large	**0.9095**	0.5723	0.8582	0.7175
Vehicle	**0.9280**	0.5747	0.8895	0.6596

3.3 Projected Optimization of the Training Set Size

Our primary goal is to utilize the power law for projecting the optimal training set size for a given dataset from a small subset of potentially available data. This goal is based on the following assumptions:

- The learning curve induced from a subset of the data fits the error rate of the future data.
- The actual utility curve of the entire dataset has a unique global optimum associated with the minimum cost of the classification process.
- The actual cost of the optimal training set size n^* projected by the power law (see Eq. 8) is not much higher than the minimum cost.

We proceed with the experimental testing of the above assumptions on six benchmark datasets.

3.3.1 Fitting Projected Error Rates

In our experiments, which are based on the experimental settings of [2], the power law curves were induced from the first 15% of the available data. For each percentage between 1% and 15%, the entire set of available data was randomly partitioned 50 times and the 15 averages of 50 runs were used as the data points for finding the parameters of the regression equation. The resulting power law function was used to calculate the error rates for the remaining 84 points (varying between 16% and 99% of the available data). Following the experimental methodology of [2], the goodness of fit of the projected error rates was evaluated by two parameters: the value of the chi-square test [8], which represents the probability that the

projected and the observed error rates are generated by the same distribution, and the mean value of the absolute differences between the observed and projected error rates. These parameters are shown in Table 3, along with the r^2 values of the 15-point regression and the induced power law equations.

Table 3. Projected error rates fits to the learning curves

Dataset	R^2 / Equation	Chi-square test value	Aver. Abs. Diff.
Breast Cancer	0.8323 $err = 0.523n^{-0.452}$	1.00	0.0148
Chess	0.9372 $err = 3.445n^{-0.725}$	1.00	0.0156
Hypo-thyroid	0.7498 $err = 0.163n^{-0.357}$	1.00	0.0010
Mushroom	0.9471 $err = 0.221n^{-0.567}$	1.00	0.0013
Soybean large	0.5236 $err = 1.213n^{-0.209}$	1.00	0.1679
Vehicle	0.7353 $err = 1.081n^{-0.187}$	1.00	0.0122

The results in Table 3 show that the power law functions induced from 15% of the observed data do not differ significantly from the predicted 84% of the error rate data for all six datasets. The average absolute differences between the observed and projected error rates are usually much lower than the minimal error rates achievable with the maximum amount of available data (see Figures 1-3).

3.3.2 Utility Curves

Using the learning curves of the six experimental datasets and Equation (1), we have computed the total classification cost vs. the training set size for nine different cost ratios between 1 and 50,000 (based on the ratios used in [13]). As indicated above, the cost ratio is equivalent to C_{err} given that $C_{tr} = 1$. The size of the score set |S| was fixed at 100 in all calculations. The plots of the normalized costs, obtained by dividing the total cost by the maximum total cost for a given cost ratio, are shown in Figures 4-6 for the three largest datasets. While for Chess and Mushroom, the optimal training set size is lower than the maximum number of available training examples, in Hypothyroid, starting with C_{err} = 5,000, the optimum strategy is to use all of the training data.

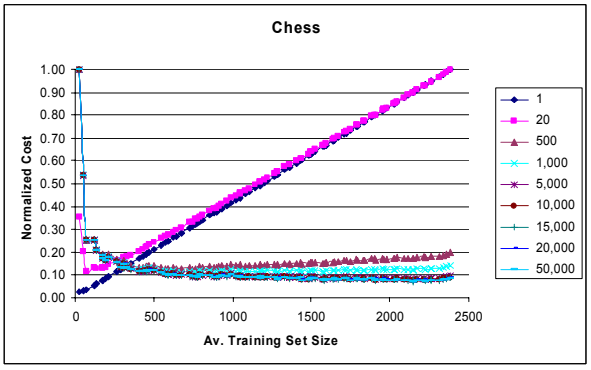

Figure 4. Normalized utility curves for Chess dataset

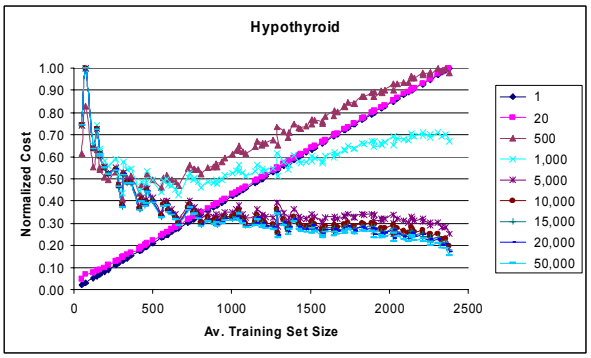

Figure 5. Normalized utility curves for Hypothyroid dataset

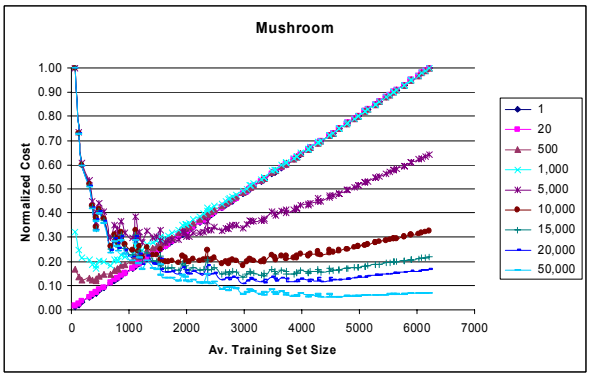

Figure 6. Normalized utility curves for Mushroom dataset

Table 4. Average cost difference between projected and actual optima

Dataset	Average Cost Difference
Breast Cancer	0.0%
Chess	18.2%
Hypothyroid	6.4%
Mushroom	14.8%
Soybean large	0.2%
Vehicle	4.7%
Overall	**7.4%**

3.3.3 Projected Optimum

For each cost ratio, we have calculated the optimal training set size n^* using Eq. (8) and the power law functions obtained from 15% of data (see Table 3). The observed classification cost with n^* records was compared to the minimum cost. If the projected optimal size of the training set was less than 15% of all available data, the projected optimum was ignored, since in that case the exact optimum can be found without projection. Also, whenever the projected optimal size exceeded the maximum amount of training data, the minimum cost was compared to the cost of using all available records. Table 4 presents the average difference (in percent) between the costs of the projected optimum and the actual optimum over various cost ratios. In two datasets (Breast and Soybean), the difference is almost zero, in two other datasets (Hypothyroid and Vehicle) the difference is less than 10%, and only in the remaining two (Chess and Mushroom) it exceeds 10%. Overall, we believe that these initial results demonstrate the usefulness of the power law for projecting the optimal amount of training records.

4. Conclusions

This paper presents a new methodology for projecting the optimal amount of training data for a given dataset using the power law function induced from a small subset of potentially available data. Such a method is beneficial whenever there is a known trade-off between the number of collected records and the accuracy of the resulting classification model. The proposed methodology is successfully evaluated on several benchmark datasets using an oblivious decision-tree classifier and random sampling. Future research may include extending this technique to additional classification algorithms, exploring the effect of the number of data points on the accuracy of the projected learning curve, and experimentation with real-world datasets, where actual cost data is readily available. Modeling learning curves obtained with non-random sampling techniques, such as active sampling [11], is another important task.

5. References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[2] L. J. Frey and D. H. Fisher, "Modeling Decision Tree Performance with the Power Law", In Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics, 1999, pp. 59-65.

[3] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann, 2006.

[4] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag, 2003.

[5] G. John and P. Langley, "Static versus dynamic sampling for data mining", In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1996, pp. 367-370.

[6] M. Last and O. Maimon, "A Compact and Accurate Model for Classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 2, pp. 203-215, February 2004.

[7] O. Maimon and M. Last, *Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology*, Kluwer Academic Publishers, Massive Computing, Boston, December 2000.

[8] E.W. Minium, R.C. Clarke, and T. Coladarci, *Elements of Statistical Reasoning*, New York: John Wiley & Sons, Inc. 2nd Ed., 1999.

[9] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, *UCI Repository of Machine Learning Databases*, Irvine, CA: University of California, Department of Information and Computer Science, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html].

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[11] M. Saar-Tsechansky and F. Provost, "Active Sampling for Class Probability Estimation and Ranking", *Machine Learning,* Vol. 54, No. 2, Feb. 2004, pp. 153-178.

[12] S. Singh, "Modeling Performance of Different Classification Methods: Deviation from the Power Law", Project Report, Department of Computer Science, Vanderbilt University, USA, April 2005.

[13] G. M. Weiss and Y. Tian, "Maximizing Classifier Utility when Training Data is Costly", *SIGKDD Explorations*, Volume 8, Issue 2, December 2006, pp. 31-38.