

Comparison of Learning Methods: Pitfalls and Challenges

Vladimir Cherkassky and Wuyang Dai

EXTENDED ABSTRACT

The growing importance of discovering regularities in observed data in many applications has led to a number of diverse data-analytic methodologies, such as Pattern Recognition, Machine Learning, Data Mining, Artificial Neural Networks etc. Most algorithms developed in these fields pursue the goal of estimating (learning) predictive models from available data. Such a predictive model is then used for prediction with new (test) inputs. The prediction (or generalization) performance of a model can be objectively evaluated using an independent test set. Over the past 10 years, hundreds or maybe thousands of new learning algorithms have been proposed. Typically, introduction of a new algorithm includes empirical comparisons suggesting that the proposed method is ‘better’ (in terms of generalization performance) than already existing methods. Remarkably, the process of inventing new algorithms continues at a rapid pace, even though it is logically inconsistent to have hundreds of ‘best’ algorithms. A closer inspection of empirical comparisons used to justify new learning methods suggests that:

- (a) Often the experimental procedure (used for comparisons) is poorly designed (or not described at all).
- (b) The authors of a proposed learning method put special effort in tuning its parameters.
- (c) Sometimes comparisons fail to differentiate between resampling error (used for model complexity control) and true prediction error.
- (d) Comparisons fail to account for the fact that generalization performance always depends on statistical characteristics of the data (such as sample size, amount of noise etc.).

This paper describes issues arising in empirical comparison of learning methods, and illustrates potential pitfalls via simple examples.

Inductive Learning Setting:

Our discussion starts with a brief review of inductive learning setting underlying most learning algorithms. Standard inductive learning [3,4,5,6] attempts to estimate a model or function f which maps an input vector $\mathbf{x} \in \mathbf{X}$ to

an output $y \in \mathbf{Y}$. This model is selected from a set of possible models $f(\mathbf{x}, w)$ parameterized by a general set of parameters w . Estimation (or learning) is performed using finite training samples that are identically and independently generated from an unknown probability distribution $P(\mathbf{x}, y)$. The goal is to find the best function f such that the expected loss

$$R(w) = \int L(f(\mathbf{x}, w), y)P(x, y)d\mathbf{x}dy$$

is minimized. Here $L(f(\mathbf{x}, w), y)$ denotes a loss function appropriate for a given application (i.e., classification error, squared loss etc.). This standard inductive learning setting implies that:

- (a) the model is estimated using *only* finite training set.
- (b) Prediction accuracy is estimated using *large* test set.

In practice, test set is finite (at best) or not available (at worst). In the latter case, prediction accuracy is usually estimated using resampling of available training data. As a result, estimated prediction error always depends on selected resampling procedure and pure luck. Remarkably, many empirical comparisons do not give any details of the resampling procedure.

Further, prediction performance of a method is strongly affected by implementation of model selection, i.e. tuning of method’s parameters, such as the value of k in k -nearest neighbors, the number of hidden units in MLP networks, regularization parameter C in SVM etc. These parameters are typically selected via resampling. However, different implementations of resampling, i.e. leave-one-out vs 5-fold cross-validation may yield different model complexity. Unfortunately, detailed description of the experimental procedure for model selection is (almost) never reported.

High-Dimensional Data:

Particular issues and challenges arise with application of learning methods to sparse high-dimensional data. Such a data is common in biomedical applications, imaging, text categorization etc. These applications usually favor simple and robust methods such as k -nearest neighbors, linear SVM and linear discriminant analysis. Due to sparseness of such data, clear and detailed description of resampling procedure used for comparisons becomes especially important.

Non-Standard Learning Settings:

Many recent powerful algorithms do not follow standard inductive setting. For example, transduction and semi-supervised learning incorporate the knowledge of x -values of test data into learning. Comparisons involving such non-

Vladimir Cherkassky and Wuyang Dai are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. (e-mail: cherk001@umn.edu).

standard learning formulations are especially challenging because:

- (a) one has to clearly understand underlying assumptions of these non-standard formulations vs assumptions used in inductive learning.
- (b) Such new formulations have more tuning parameters, which makes model selection more difficult. Hence, performance estimates become especially sensitive to resampling procedure used in comparisons.

We discuss in detail several examples of non-standard learning setting where the training data includes additional (group) information. This leads to new learning settings known as Multi-Task Learning [1,2,8] and Learning with Structured Data (aka SVM+) [7], as discussed next.

Suppose that training data can be represented as a union of t related groups, i.e. each group $r \in [1, 2, \dots, t]$ contains n_r samples independently and identically generated from a distribution P_r on $\mathbf{X} \times \mathbf{Y}$. Therefore, available data is a union of $t > 1$ groups:

$$\{\{X_r, Y_r\}, r = 1, \dots, t\}, \{X_r, Y_r\} = \{\{\mathbf{x}_{r_1}, \mathbf{y}_{r_1}\}, \dots, \{\mathbf{x}_{r_{n_r}}, \mathbf{y}_{r_{n_r}}\}\}$$

and can be thought as samples identically and independently generated from the distribution $P = \cup_{r=1, \dots, t} P_r$.

If the group labels of future test samples are not given, the problem is “Learning With Structured Data (LWSD)” formulation [7]. In this formulation, the goal is to find one best mapping function f such that the expected loss

$$R(w) = \int L(f(\mathbf{x}, w), y) P(\mathbf{x}, y) d\mathbf{x} dy$$

is minimized. Note that even though the expected loss is in the same form as in the supervised learning setting, the difference is that in supervised learning setting P is unknown, while in LWSD, P is a union of t sub-distributions.

On the other hand, if the group labels of future test samples are given, the problem is Multi-Task Learning (MTL) problem [1,2,8]. The goal in multi-task learning is to find t mapping functions $\{f_1, f_2, \dots, f_t\}$ such that the sum of expected losses for each task

$$R(w) = \sum_{r=1}^t \left(\int L(f_r(\mathbf{x}, w), y) P_r(\mathbf{x}, y) d\mathbf{x} dy \right)$$

is minimized. Figure 1 illustrates that standard supervised learning, multi-task learning and learning with structured data handle training and test data in different ways.

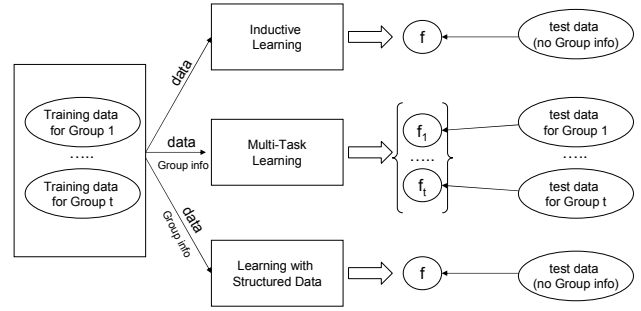


Figure 1: Inductive learning, Multi-task learning, and Learning with structured data use different ways to handle training and test data.

“Learning with structured data” formulation and multi-task learning formulation are similar in the sense that they all try to exploit the group information. However, there are several important differences: (1) LWSD comes out one model, while MTL comes out t models; (2) LWSD does not require group membership of new testing data, but MTL does require that. Let’s consider two realistic application problems that distinguish the two formulations. One example is handwritten digit recognition, where the training data originates from t persons (each person provides labeled examples of all 10 digits). Then goal 1 (LWSD) is to find a classifier that can generalize well for other (previously unseen) samples written by these people (we don’t know who writes this test samples). In contrast, goal 2 (MTL) is improved generalization for each person who contributed to training data (ie. group membership for future samples is known). Another application example is fMRI data analysis or more generally, medical diagnosis. Here you try to estimate a predictive model (predict/diagnose a disease) from the training samples from t patients. Then goal 1 (LWSD) is to find a predictive model that has good generalization for other (new) samples from these patients, whereas the goal 2 (MTL).

We show empirical comparisons between different learning approaches for utilizing group information in the data. In particular, we compare:

- multiple SVM approach where a separate SVM classifier is estimated for each group
- SVM+ approach implementing LWSD setting
- SVM+MTL implementing multi-task learning using SVM+ methodology [9].

Comparisons are performed using synthetic data. Comparison results indicate that there is no single winner, and that relative performance strongly depends on the size of training data set.

Acknowledgement: This work was supported, in part, by NSF grant ECCS 0802056, and by the [A. Richard Newton Breakthrough Research Award](#) from [Microsoft Research](#)

REFERENCES

- [1] Ando, R. and Zhang, T. A Framework for Learning predictive structures from multiple tasks and unlabeled data, *Journal of Machine Learning Research*, 2005.
- [2] Ben-David, S., Gehrke, J. and Schuller, R. A theoretical framework for learning from a pool of disparate data sources. *ACM KDD*, 2002.
- [3] Cherkassky, V. and Mulier, F. *Learning from Data*, John Wiley & Sons, New York, second edition, 2007.
- [4] Hastie, T. , Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, 2001.
- [5] Vapnik, V. *Estimation of Dependences Based on Empirical Data*, Springer Verlag, New York, 1982.
- [6] Vapnik, V. *Statistical Learning Theory*, Wiley, New York, 1998.
- [7] Vapnik, V. *Empirical Inference Science Afterword of 2006*, Springer, 2006.
- [8] Evgeniou, T. and Pontil, M.. Regularized multi-task learning. In *Proc. 17th SIGKDD Conf. on Knowledge Discovery and Data Mining*, 2004.
- [9] Liang, L. and V. Cherkassky, Connection between SVM+ and multi-task learning, *Proc. IJCNN-2008*, Hong Kong, China.