Computational Statistics and Data Analysis 74 (2014) 181–197

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Learning algorithms may perform worse with increasing training set size: Algorithm–data incompatibility

Waleed A. Yousef [a,*], Subrata Kundu [b]

[a] Human Computer Interaction Laboratory (HCILAB), Computer Science Department, Faculty of Computers and Information, Helwan University, Helwan 11795, Egypt

[b] Department of Statistics, The George Washington University, Washington, DC 20052, USA

ARTICLE INFO

Article history:
Received 22 July 2012
Received in revised form 22 May 2013
Accepted 23 May 2013
Available online 29 May 2013

Keywords:
Machine learning
Pattern recognition
Statistical learning
Stable distribution
Convergence
Stochastic concentration

ABSTRACT

In machine learning problems a learning algorithm tries to learn the input–output dependency (relationship) of a system from a training dataset. This input–output relationship is usually deformed by a random noise. From experience, simulations, and special case theories, most practitioners believe that increasing the size of the training set improves the performance of the learning algorithm. It is shown that this phenomenon is not true in general for any pair of a learning algorithm and a data distribution. In particular, it is proven that for certain distributions and learning algorithms, increasing the training set size may result in a worse performance and increasing the training set size infinitely may result in the worst performance—even when there is no model misspecification for the input–output relationship. Simulation results and analysis of real datasets are provided to support the mathematical argument.

1. Introduction

"*Learning is the process of estimating an unknown input–output dependency, or structure of a system, using a limited number of observations.*" (Cherkassky and Mulier, 1998). The basic two building blocks of this process are: data collected from that system, which we call training set, and a learning algorithm that processes this data and "learns" from it. "Learning" is ubiquitous; the following is a few examples of diverse fields: Automatic Target Recognition (ATR) in military applications, Computer Aided Diagnosis (CAD) in medical imaging, DNA microarrays in Genomics, Optical Character Recognition (OCR), Speech Recognition (SR), spam email filtering, and stock market prediction.

One problem that concerns practitioners in all of the applications of learning is the size of the training dataset. Some theoretical results, simulations, and experience in the majority of applications indicate that the performance of a learning method improves as the training sample size increases (Yousef et al., 2004; Chan, 2003; Chan et al., 1998, 1999, 1997; Raudys, 1997; Raudys and Pikelis, 1980; Raudys and Jain, 1991; Takeshita and Toriwaki, 1995; Dobbin and Simon, 2007). Qualitatively speaking, performance is measured as the capability of the learning method to predict an unknown output from a new but known input. The early work by Hughes (Hughes, 1968) supports this belief. The analytical framework by Fukunaga (Fukunaga and Hayes, 1989; Fukunaga, 1990), presented under the normality assumption, was another push towards developing the same belief. These citations are a few examples from a huge literature that report the same findings.

The main objective of the current article is to show that this phenomenon is not universal and does depend on the pair of the learning algorithm and the data distribution. We prove rigorously that increasing training sample size may worsen the performance, even if the true model that describes the input–output relationship is used.

* Corresponding author. Tel.: +20 1 202 470 0909; fax: +20 2 022 554 7975.
E-mail addresses: wyousef@aucegypt.edu, wyousef@gwu.edu (W.A. Yousef).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.021

This is a somewhat surprising result that runs against both intuition and experience in the field. Proving this result is interesting for its own theoretical sake. It balances the practice of drawing generalized conclusions from mere experience, simulations, or even special case theories. More importantly, it raises an alert flag for practitioners in areas of applications where such a phenomenon may happen; (more on that in the following sections). This is a counterintuitive result for many, if not all, of those who are working in the field of pattern recognition and machine learning. On the other hand, this result may not come as a surprise to the theoreticians in the field of Probability and Statistics. We hope this article continues building the bridge between the two parties.

Before delving into the rest of the paper we would like to give a roadmap that prepares the reader and helps building the intuition. Prior knowledge of the true model that represents the input–output dependency (up to a parameter) and the availability of unlimited number of observations are not enough to ensure good performance of a given learning method (equivalently of the estimators of the model parameters). Moreover, in some cases increasing the number of observations will decrease the performance! It still depends on the underlying distribution and the "appropriateness" of the chosen learning method for this distribution. Apart from being theoretically\mathematically interesting on its own, the relation between the sample size and the performance discussed in the present article will have some impact on the way the practitioners use standard results. They should recognize the limited applicability of these standard results under certain conditions and the need for finding different algorithms. In retrospect, the relation between the sample size and the performance discussed in the present article may be of interest to those areas of application besides the mathematical interest in its own sake.

The rest of the paper is organized as follows. Section 2 sets the mathematical framework of the article and introduces some preliminaries and technical background that are necessary for the rest of the article. Section 3 provides the analysis of the main proposition of the article, along with discussion, intuition, and explanation. Section 4 reports simulation and real data studies to illustrate the mathematical results. Section 5 contains conclusion, discussions, and advice for practitioners. In the Appendix we present all the proofs.

2. Background and prerequisites

2.1. The learning problem

In engineering terminology, X is the input to a system and Y is the available measured output, which is the intrinsic system output deformed by some random noise interfering with the system. One can be interested in learning the relationship between the input X and that intrinsic system output. This is only available through observing pairs of observations from X and Y.

The learning problem is formalized as follows. We have a random variable (r.v.) Y called the response, and a set of random variables $X_1, X_2, \ldots, X_p$ called the predictors; these predictors can be considered as the components of the random vector X. Our goal is to learn how Y and X are related using the available training set $\mathbf{t} = \{t_i = (x_i, y_i) : i = 1, \ldots, n\}$, of size n. If the response Y is categorical, e.g., *diseased* or *nondiseased* in medical applications, the problem is called classification. If the response Y is quantitative, e.g., $Y \in \mathbb{R}$, the problem is called regression. The objective of the learning process is to use the set $\mathbf{t}$ to find a function $\eta_{\mathbf{t}}$ so that for any new observation $(x_0, y_0) \notin \mathbf{t}$, with known x_0 and unknown y_0, we will be able to predict y_0 by $\widehat{y}_0 = \eta_{\mathbf{t}}(x_0)$, "well". In other words, the predicted value $\widehat{y}_0$ should be in some sense "close" to y_0. The "closeness" should be defined objectively in terms of a loss function $L(y, \widehat{y})$ that gives a distance between y and $\widehat{y}$. Many loss functions can be defined. For the case of classification, the zero–one loss function is a popular one

$$L(y, \widehat{y}) = \begin{cases} 0 & y = \widehat{y} \\ 1 & y \neq \widehat{y} \end{cases}. \tag{1}$$

For the case of regression the quadratic loss function

$$L(y, \widehat{y}) = (y - \widehat{y})^2, \tag{2}$$

is very popular because of its nice mathematical properties. Since $\widehat{y} = \eta_{\mathbf{t}}(x)$, it is obvious that $\widehat{y}$ is a random variable whose randomness comes from the training set $\mathbf{t}$, along with the new observation x. The loss $L(y, \eta_{\mathbf{t}}(x))$ is a random variable, as well, whose randomness is coming from both the training set $\mathbf{t}$ and the new observation (x, y). Therefore, one should consider the average performance of the learning algorithm, averaging over the relevant populations. A common performance measure is the risk, which is the mean loss. If we denote the training set $\mathbf{t}$ by $(\mathbf{X}, \mathbf{Y})$, and the new observation by (X, Y) then the risk function is defined as

$$R(Y, \widehat{Y}) = \mathrm{E}\, L(Y, \widehat{Y}) = \mathrm{E}_{\mathbf{X}} \mathrm{E}_{\mathbf{Y}} \mathrm{E}_X \mathrm{E}_Y L(Y, \eta_{\mathbf{t}}(X)). \tag{3}$$

If the expectation of the loss is taken only over the population of new observation (X, Y) then it is the conditional risk $R_{\mathbf{t}} = \mathrm{E}_X \mathrm{E}_Y L(Y, \widehat{Y})$; this is the risk of the learning algorithm conditional on this particular training set. Then, if we carry out the expectation $\mathrm{E}_{\mathbf{X}} \mathrm{E}_{\mathbf{Y}} R_{\mathbf{t}}$, we get the unconditional risk (3) which averages the performance over the population of training datasets of the same size. This unconditional risk is a function only of the size n of the training sets.

So far, no assumptions were made about the joint distribution of X and Y. Next, we introduce the family of stable distributions and assume that the distribution of Y belongs to this family.