Neural Networks 33 (2012) 58–66

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

A comparative analysis of support vector machines and extreme learning machines☆

Xueyi Liu [a,b], Chuanhou Gao [c,*], Ping Li [a,d]

[a] School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310027, China
[b] Department of Mathematics, China Jiliang University, Hangzhou 310018, China
[c] Department of Mathematics, Zhejiang University, Hangzhou 310027, China
[d] Institute of Industrial Process Control, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 11 April 2011
Received in revised form 12 March 2012
Accepted 4 April 2012

Keywords:
SVM
ELM
VC dimension
Generalization ability
Computational complexity

ABSTRACT

The theory of extreme learning machines (ELMs) has recently become increasingly popular. As a new learning algorithm for single-hidden-layer feed-forward neural networks, an ELM offers the advantages of low computational cost, good generalization ability, and ease of implementation. Hence the comparison and model selection between ELMs and other kinds of state-of-the-art machine learning approaches has become significant and has attracted many research efforts. This paper performs a comparative analysis of the basic ELMs and support vector machines (SVMs) from two viewpoints that are different from previous works: one is the Vapnik–Chervonenkis (VC) dimension, and the other is their performance under different training sample sizes. It is shown that the VC dimension of an ELM is equal to the number of hidden nodes of the ELM with probability one. Additionally, their generalization ability and computational complexity are exhibited with changing training sample size. ELMs have weaker generalization ability than SVMs for small sample but can generalize as well as SVMs for large sample. Remarkably, great superiority in computational speed especially for large-scale sample problems is found in ELMs. The results obtained can provide insight into the essential relationship between them, and can also serve as complementary knowledge for their past experimental and theoretical comparisons.

1. Introduction

As one of the standard tools for machine learning and data mining, support vector machines (SVMs) (Cortes & Vapnik, 1995; Vapnik, 1995) have yielded many real-world applications due to their good generalization performance, even for small samples (Jonsson, Kittler, Li, & Matas, 2002), and due to their superiority to the traditional empirical risk minimization principle employed by most neural networks (Lu, Plataniotis, & Ventesanopoulos, 2001). Similarly, extreme learning machines (ELMs), which were originally proposed as novel learning algorithms for single-hidden-layer feed-forward neural networks (SLFNs) (Huang, Chen, & Siew, 2006a; Huang, Wang, & Lan, 2011; Huang, Zhu, & Siew, 2004, 2006b) and then extended to generalized SLFNs where the hidden layer neurons may not be neuron like (Huang & Chen, 2007, 2008), have also attracted a lot of research interests (Feng, Huang, Lin, & Gay, 2009; Liang, Huang, Saratchandran, & Sundararajan, 2006; Rong, Huang, Saratchandran, & Sundararajan, 2009; Wang, Cao, & Yuan, 2011; Zhu, Qin, Suganthan, & Huang, 2005) due to their better generalization performance and faster learning speed than traditional gradient-based learning algorithms. To put it briefly, these two methods are both frequently used intelligent algorithms, and they may be used interchangeably in many practical cases. However, each of them has individual merits and shortcomings which lead to their being not completely equivalent. Knowing the relationships and differences between these two methods may provide more possibility to select appropriate algorithms from them for a certain specific problem and may also be helpful to develop more effective intelligent algorithms for some practical purposes. It is not a novel issue to perform a comparative analysis between the SVM and ELM algorithms. For example, Huang et al. (2006b) pointed out that ELMs can obtain similar generalization ability to SVMs but with much less training time through simulation experiments on a few artificial and real benchmark function approximation and classification problems; Wei, Li, and Feng (2006) presented that the ELM algorithm is much faster and has better generalization performance than the SVM in a real Tennessee Eastman process; Liu, Loh, and Tor (2005),

☆ This work was supported by the National Natural Science Foundation of China under Grant Nos. 10901139, 60911130510, 61101239, and the Public Benefit Technologies R & D Program of Science and Technology Department of Zhejiang Province under Grant No. 2011C21020.

* Corresponding author.
E-mail addresses: gaochou@zju.edu.cn, gaochou@ieee.org (C. Gao).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.04.002

however, indicated that SVMs still outperform ELMs in some text classification problems; Cheng, Cai, and Pan (2009) obtained that the ELM algorithm has similar accuracy compared with the SVM but has obvious advantages in parameter selection and learning speed by addressing a reservoir permeability problem, etc. Recently, Huang, Ding, and Zhou (2010) have made a significant contribution showing the relationship between ELMs and SVMs in the framework of classification, where the following consistencies between them were found: (1) an SVM's maximal separating margin of two different classes is consistent with the minimal norm of output weights in an ELM; (2) just as an SVM does, an ELM also minimizes the training errors as well as maximizing the separating margin. Further, Huang, Zhou, Ding, and Zhang (2011) made a more in-depth exploration of their relationship, and compared the performance of ELMs, SVMs, and least-squares SVMs (LSSVMs) over 36 wide types of data sets, and drew the following important conclusions: (1) the ELM provides a unified learning platform to different applications, such as regression, binary, and multiclass classifications for the LSSVM, proximal SVM (PSVM), and other regularization algorithms; (2) because of the lack of the bias term b, from the optimization method point of view, the ELM algorithm has milder optimization constraints compared with the LSSVM and PSVM algorithms, which indicates that, in theory, LSSVMs and PSVMs can only achieve suboptimal solutions in comparison with ELMs; (3) the ELM algorithm tends to achieve similar or better generalization performance at much faster learning speed than the SVM and LSSVM algorithms.

All the above studies (i.e., (Cheng et al., 2009; Huang et al., 2010, 2011; Liu et al., 2005; Wei et al., 2006)) give hints for distinguishing SVMs and ELMs both theoretically and practically. However, there also remain several aspects needing further consideration. For example, the experimental investigations are mainly focused on the comparisons of SVMs and ELMs applied to a variety of examples, and the information is still unknown in some specific cases, such as, for the same training sample, the performance comparison of SVMs and ELMs with changing sample size. Knowing such information may provide more insight into the SVM and ELM algorithms because the former is based on the structural risk minimization principle which is especially suited for learning small samples, while the latter is based on the inductive principle known as empirical risk minimization. Besides, a comparison of the Vapnik–Chervonenkis (VC) dimensions of SVMs and ELMs has not been conducted yet, which is also important and interesting since VC theory can offer helpful theoretical insights into the nature of the learning methods and provide potential practical applicability for model complexity control in learning problems (Cherkassky, Shao, Mulier, & Vapnik, 1999; Vapnik, 1995). For these reasons, the purpose of this paper is to make comparisons of SVMs and the basic ELMs from the theoretical viewpoint of their VC dimensions, and also to perform experimental comparisons between them, including comparisons of their generalization ability under different sizes of training sample and of their computational complexity. The results can strengthen the understanding on the essential relationship between SVMs and ELMs, and can also serve as complementary knowledge for the past experimental and theoretical comparisons between them.

The rest of this paper is organized as follows. Section 2 briefly reviews the concepts of SVMs and ELMs. This is followed by the theoretical comparisons between SVMs and ELMs in Section 3. Section 4 makes the experimental comparisons between them. Finally, Section 5 concludes this paper.

2. Brief review of SVMs and ELMs

2.1. SVMs

Consider a nonlinear multiple-input single-output system

$$y = f(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$. For simplicity, but without loss of generality, the SVM is used here only to address the regression problem of the studied system. The main idea of an SVM is first to map the n-dimensional input data into a high-dimensional feature space, denoted as $\mathcal{F}$, through a nonlinear project $\Phi : \mathbb{R}^n \to \mathcal{F}$; then, a linear algorithm is performed in this feature space to approximate the underlying dynamics according to

$$y = f(\mathbf{x}) = \langle \boldsymbol{w}, \Phi(\mathbf{x}) \rangle + b, \tag{2}$$

where $\boldsymbol{w} \in \mathcal{F}$ represents the weight vector, $b \in \mathbb{R}$ is the bias term, and $\langle \cdot, \cdot \rangle$ denotes the scalar product. This process is implicitly implemented by specifying a kernel function $K(\mathbf{x}, \mathbf{x}')$, which in turn determines the high-dimensional project Φ by $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ (Aizerman, Braverman, & Rozonoer, 1964). The kernel is called a positive kernel if it satisfies Mercer's condition (Vapnik, 1995). Here, it should be noted that the dimension of $\mathcal{F}$ is usually unknown, being very high and even infinite. For a given kernel, the high-dimensional feature space with minimal dimension is called "minimal feature space". In actual operation, for a given training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, an SVM formulates the learning problem of estimating f as a variational problem that minimizes the regularized risk functional

$$\begin{aligned} R_{reg}[f] &= R_{emp}[f] + \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &= \gamma \sum_{i=1}^{N} e(f(\mathbf{x}_i), y_i) + \frac{1}{2}\|\boldsymbol{w}\|^2. \end{aligned} \tag{3}$$

Here, $R_{emp}[f]$ is the empirical risk function, $\|\boldsymbol{w}\|^2$ denotes the model complexity, γ is the regularization parameter, and $e(f(\mathbf{x}_i), y_i)$ is an ε-insensitive loss function, such as $e(f(\mathbf{x}_i), y_i) = |f(\mathbf{x}_i) - y_i|_\varepsilon^2$. Usually, the quadratic programming problem, Eq. (3), is changed to its dual form to obtain the high-dimensional vector w, i.e., the following optimization problem:

$$\begin{aligned} \max R_D = &\sum_{i=1}^{N}(\mu_i - \nu_i)y_i - \varepsilon \sum_{i=1}^{N}(\mu_i + \nu_i) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\mu_i - \nu_i)(\mu_j - \nu_j)K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{4}$$

subject to

$$\sum_{i=1}^{N}(\mu_i - \nu_i) = 0, \tag{5}$$

$$\mu_i, \nu_i \in [0, \gamma], \quad i = 1, 2, \ldots, N. \tag{6}$$

Denote the points corresponding to nonzero μ_i or ν_i to be support vectors $\mathbf{x}_k$; then $\boldsymbol{w}$ and b may be obtained by

$$\boldsymbol{w} = \sum_{k \in SV}(\mu_k - \nu_k)\Phi(\mathbf{x}_k) \tag{7}$$

$$b = y_s - \langle w, \Phi(\mathbf{x}^s) \rangle, \tag{8}$$

where $SV \subset \{1, 2, \ldots, N\}$ is the index set of the support vectors and $\mathbf{x}^s$ is one of the support vectors. Submitting Eq. (7) into Eq. (2) will give the optimal f:

$$f(\mathbf{x}) = \langle \boldsymbol{w}, \Phi(\mathbf{x}) \rangle + b = \sum_{i=1}^{N}(\mu_i - \nu_i)K(\mathbf{x}_i, \mathbf{x}) + b. \tag{9}$$