Neurocomputing 174 (2016) 250–264

Contents lists available at ScienceDirect

Neurocomputing

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

An unsupervised discriminative extreme learning machine and its applications to data clustering

Yong Peng [a], Wei-Long Zheng [a], Bao-Liang Lu [a,b,*]

[a] Center for Brain-like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China

[b] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 18 September 2014
Received in revised form
18 November 2014
Accepted 25 November 2014
Available online 18 August 2015

Keywords:
Extreme learning machine (ELM)
Unsupervised learning
Manifold information
Discriminative information
Image clustering
EEG

ABSTRACT

Extreme Learning Machine (ELM), which was initially proposed for training single-layer feed-forward networks (SLFNs), provides us a unified efficient and effective framework for regression and multiclass classification. Though various ELM variants were proposed in recent years, most of them focused on the supervised learning scenario while little effort was made to extend it into unsupervised learning paradigm. Therefore, it is of great significance to put ELM into learning tasks with only unlabeled data. One popular approach for mining knowledge from unlabeled data is based on the manifold assumption, which exploits the geometrical structure of data by assuming that nearby points will also be close to each other in transformation space. However, considering the manifold information only is insufficient for discriminative tasks. In this paper, we propose an improved unsupervised discriminative ELM (UDELM) model, whose main advantage is to combine the local manifold learning with global discriminative learning together. UDELM can be efficiently optimized by solving a generalized eigenvalue decomposition problem. Extensive comparisons over several state-of-the-art models on clustering image and emotional EEG data demonstrate the efficacy of UDELM.

1. Introduction

ELM as an emerging learning technique provides an efficient unified solution to generalized feed-forward networks such as SLFNs. The main merit of ELM is that the network input weights are randomly assigned and independent from specific applications [1,2], which makes the analytical solution of network output weights be efficiently obtained by solving a least square formula. Despite the fact that the determination of the network hidden layer outputs is based on randomly generated network input weights, it has been proven that SLFNs trained based on ELM algorithm still have the global approximation ability [3,4]. ELM is a unified framework for regression and multiclass classification [5]. Due to its effectiveness and fast learning process in comparison with gradient descend-based optimization, ELM has been adopted in many applications such as face recognition [6], action recognition [7,8], gesture recognition [9], security assessment [10], EEG signal processing [11], data privacy [12], image quality assessment [13,14] and remote sensing [15].

Though many ELM variants were proposed in the last few years [16–21,8], the extension on ELM research focused mainly on the supervised learning tasks. This greatly limits the applicability of ELM in utilizing unlabeled data. Moreover, in many real world applications, labeled data is usually expensive to obtain but unlabeled data is relatively easy to collect, which drives us to extend ELM into unsupervised learning by properly harnessing the unlabeled data. On the basis of manifold regularization, Huang and his colleagues proposed two ELM variants, semi-supervised ELM and unsupervised ELM (USELM) [22]. He et al. proposed to do clustering in the ELM hidden layer space in view of the good properties of its random feature mapping, which generates better results than clustering in the original data space [23]. The part from hidden layer to output layer of ELM was discarded and the hidden layer representation was used for clustering. The rationality of ELM feature mapping was also analyzed in [23]. A new ELM clustering technique was presented by Akusok et al. [24] by incorporating some prior knowledge into clustering. This method utilizes the prior knowledge of the exact number of points in each cluster; however, this requirement is usually hard to satisfy. We are sometimes provided with imbalance data sets which have

* Corresponding author at: Center for Brain-like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China.
E-mail addresses: stany.peng@gmail.com (Y. Peng), bllu@sjtu.edu.cn (B.-L. Lu).

http://dx.doi.org/10.1016/j.neucom.2014.11.097
0925-2312/

different number of points in different clusters. These unsupervised models greatly enlarge the applicability of ELM.

In this paper, we aim to make improvements on the basis of USELM [22] for the reasons that (1) we want to retain the whole architecture of ELM network, from input layer, hidden layer to output layer; (2) we need not to know the exact number of points in each cluster before clustering. USELM, which was designed to exploit the underlying structure of data, shows excellent performance in clustering when comparing with several state-of-the-art unsupervised algorithms [22]. However, it pays only attention to the local structure of data and ignores the discriminative information of different classes. Various studies have shown that both structure and discriminative information are important in dealing with discriminative tasks such as classification [25–27] and clustering [28,29]. Specifically, Guan et al. introduced the manifold regularization and the margin maximization into non-negative matrix factorization and presented the manifold regularized discriminative non-negative matrix factorization [25]. In [26], similar technique was incorporated into the ELM framework for EEG-based emotion recognition. Shu et al. included a graph regularization term into the discriminative analysis based on spectral regression [27]. The formulated LocLDA method covers both local and global structure information, which is more effective for face recognition. In [28], Yang et al. proposed to exploit the discriminative information in each local data clique based on constructing an elaborate local graph Laplacian and then globally integrating the local models of all the local cliques. The formulated model, local discriminant models and global integration (LDMGI), was put into spectral clustering and promising results were demonstrated in comparison with ordinary normalized cut [30]. In [29], both local manifold learning and global discriminative learning are incorporated into non-negative matrix factorization to learn effective data representation.

Inspired by existing studies, we propose a novel unsupervised ELM model, unsupervised discriminative ELM, to utilize both the local structure and global discriminative information of data. Our goal is to learn a well-structured data representation for clustering. On the one hand, the learned data representation can preserve the intrinsic structure as much as possible through efficiently exploiting the local manifold information; on the other hand, the global discriminative information is utilized to make the learned representation achieve discriminative power, e.g., differentiating samples from different clusters.

The main contributions of this paper can be summarized as follows:

(1) We propose the *unsupervised discriminative ELM* to derive better data representations for clustering. UDELM utilizes both the local structure and global discriminative information of data.
(2) Different from USELM, which needs to tune the number of output neurons, UDELM defines such value as the number of the clusters. This exactly coincides with the original ELM definition.
(3) Extensive experiments are conducted to evaluate the clustering performance of UDELM by comparing with several state-of-the-art algorithms. Results on five widely used image data sets and one emotional EEG data set demonstrate the efficacy of UDELM.

The remainder of this paper is organized as follows. Section 2 provides a brief review of ordinary ELM and USELM [22]. Section 3 proposes the model formulation and optimization method of UDELM. Experimental studies to evaluate the performance of UDELM are given in Section 4. Section 5 concludes the paper.

2. Preliminaries

2.1. Extreme learning machine

Denote $\{\mathbf{x}_i, c_i\}_{i=1,\dots,N}$ a set of N raw feature vectors $\mathbf{x}_i \in \mathbb{R}^D$ and the corresponding class labels $c_i \in \{1,\dots,C\}$. The task is to train a SLFN with $\{\mathbf{x}_i, c_i\}_{i=1,\dots,N}$. Such a network consists of D input (the dimensionality of $\mathbf{x}_i$), L hidden and C output (the number of classes) neurons. In ELM, the number of hidden neurons is usually set to be larger than the number of classes to ensure the global approximation ability [5], i.e., $L \gg C$. For each training vector $\mathbf{x}_i$, the corresponding network target vector is $\mathbf{t}_i = [t_{i1},\dots,t_{iC}]$. Generally, when $\mathbf{x}_i$ belongs to class k, that is $c_i = k$, we have $t_{ij} = 1$ if $j = k$ and $t_{ij} = -1$ otherwise. In ELM, the network input weights $\mathbf{W} \in \mathbb{R}^{L\times D}$ and the hidden layer biases $\mathbf{b} \in \mathbb{R}^L$ are randomly generated, which leads to the analytical calculation of the network output weights $\boldsymbol{\beta} \in \mathbb{R}^{L\times C}$.

Based on the above settings, the network response $\mathbf{o}_i = [o_{i1},\dots,o_{iC}]$ corresponding to $\mathbf{x}_i$ can be calculated by

$$o_{ik} = \sum_{j=1}^{L} \beta_{jk} h_j(\mathbf{x}_i), \quad k = 1,\dots,C \tag{1}$$

where $\mathbf{h}(\mathbf{x}_i) = [h_1(\mathbf{x}_i),\dots,h_L(\mathbf{x}_i)] \in \mathbb{R}^{1\times L}$ is the output row vector of the hidden layer corresponding to the input $\mathbf{x}_i$. $\mathbf{h}(\mathbf{x}_i)$ actually maps the sample $\mathbf{x}_i$ from the D-dimensional input space $\mathcal{X}$ to the L-dimensional ELM feature space $\mathcal{H}$. By storing the network hidden layer outputs for all the training vectors in one matrix, we have

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_1) \\ \mathbf{h}(\mathbf{x}_2) \\ \vdots \\ \mathbf{h}(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} h_1(\mathbf{x}_1) & h_2(\mathbf{x}_1) & \cdots & h_L(\mathbf{x}_1) \\ h_1(\mathbf{x}_2) & h_2(\mathbf{x}_2) & \cdots & h_L(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \vdots \\ h_1(\mathbf{x}_N) & h_2(\mathbf{x}_N) & \cdots & h_L(\mathbf{x}_N) \end{bmatrix}.$$

We can rewrite (1) in a compact form as

$$\mathbf{O} = \mathbf{H}\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{O} \in \mathbb{R}^{N\times C}$ is a matrix containing the network responses for all training samples $\mathbf{x}_i, i = 1,2,\dots,N$.

The original ELM assumes that $\mathbf{o}_i = \mathbf{t}_i$, $i = 1,\dots,N$ (or $\mathbf{O} = \mathbf{T}$ in matrix form), where $\mathbf{T} = [\mathbf{t}_1;\dots;\mathbf{t}_N]$ is a matrix containing the network target vectors. By using (2), the closed form of the network output weights is

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{T}, \tag{3}$$

where $\mathbf{H}^{\dagger}$ is the Moore–Penrose generalized inverse of $\mathbf{H}$. If $\mathbf{H}^T\mathbf{H}$ is nonsingular, $\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$; or when $\mathbf{H}\mathbf{H}^T$ is nonsingular, $\mathbf{H}^{\dagger} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}$ [5]. Once the network output weights are obtained, the network response for an unseen vector $\mathbf{x}_{new}$ is given by

$$\mathbf{o}_{new} = \mathbf{h}(\mathbf{x}_{new})\boldsymbol{\beta}. \tag{4}$$

To avoid the singularity problem when calculating the inverse of $\mathbf{H}^T\mathbf{H}$, a regularization term is introduced to minimize the norm of the network output weights, which results in the following objective of regularized ELM as

$$\arg\min_{\boldsymbol{\beta}} \mathcal{J}_{RELM} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{\lambda}{2}\sum_{i=1}^{N}\|\xi_i\|_2^2,$$
$$\text{s.t.}\xi_i = \mathbf{t}_i - \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta}, \quad i = 1,\dots,N \tag{5}$$

where $\xi_i \in \mathbb{R}^{1\times C}$ is the error vector corresponding to $\mathbf{x}_i$ and $\lambda > 0$ is a regularization parameter. Therefore, the network output weights in regularized ELM can be estimated as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^T\mathbf{H} + \frac{\mathbf{I}}{\lambda}\right)^{-1}\mathbf{H}^T\mathbf{T}. \tag{6}$$