Knowledge-Based Systems 37 (2013) 394–414

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

A competitive ensemble pruning approach based on cross-validation technique

Qun Dai *

Institute of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 3 March 2012
Received in revised form 25 August 2012
Accepted 26 August 2012
Available online 3 October 2012

Keywords:
Neural networks ensemble
Ensemble pruning
Concept-drifting data
Cross-validation
Competitive learning

ABSTRACT

Ensemble pruning is crucial for the consideration of both efficiency and predictive accuracy of an ensemble system. This paper proposes a new Competitive technique for Ensemble Pruning based on Cross-Validation (CEPCV). The data to be learnt by neural computing models are mostly drifting with time and environment, therefore a dynamic ensemble pruning method is indispensable for practical applications, while the proposed CEPCV method is just the kind of dynamic ensemble pruning method, which can realize on-line ensemble pruning and take full advantage of potentially valuable information. The algorithm naturally inherits the predominance of cross-validation technique, which implies that those networks regarded as winners in selective competitions and chosen into the pruned ensemble have the "strongest" generalization capability. It is essentially based on the strategy of "divide and rule, collect the wisdom", and might alleviate the local minima problem of many conventional ensemble pruning approaches only at the cost of a little greater computational cost, which is acceptable to most applications of ensemble learning. The comparative experiments among the four ensemble pruning algorithms, including: CEPCV and the state-of-the-art Directed Hill Climbing Ensemble Pruning (DHCEP) algorithm and two baseline methods, i.e. BSM, which chooses the Best Single Model in the initial ensemble based on their performances on the pruning set, and ALL, which reserves all network members of the initial ensemble, on ten benchmark classification tasks, demonstrate the effectiveness and validity of CEPCV.

1. Introduction

Ensemble learning is an important topic of interest in the research communities of pattern recognition and machine learning for its desirable generalization capability [1,2]. It refers to training a collection of base predictors for a given classification or regression task and then combining their outputs with a combinational strategy [3]. It is also termed multiple classifier systems [4], expert committee[5], decision forest [6,7], etc. Remarkable improvement in generalization performance has been observed from ensemble learning in a broad scope of application fields, for example: face recognition [8], optical character recognition [9], scientific image analysis [10,11], medical diagnosis [12,13], financial time series prediction [10], military purposes. [14], intrusion detection [15], etc.

Typically, ensemble learning algorithms consist of two main stages: the generation of multiple predictive models and their fusion [2]. Recently, a so-called ensemble pruning stage has been considered as an additional intermediate stage which deals with the selection of the appropriate ensemble members prior to combination [16–24]. It is also termed ensemble pruning, selective ensemble, ensemble thinning or ensemble selection.

Ensemble pruning is important and necessary for the consideration of two factors: *efficiency* and *predictive accuracy* [2]. Firstly, an ensemble system with large size will lead to heavy computational burdens. In certain applications, such as stream data mining, it is especially important to minimize the running time expenses. And when models are distributed over a network, a large number of constituent models will certainly lead to another serious problem, i.e. a large amount of communication costs [2]. Secondly, the other factor of *predictive accuracy* is equally influential. An ensemble may comprise constituent models with either high or low predictive accuracy. Those ensemble members with low predictive accuracy will negatively affect the overall predictive performance of the whole ensemble. Pruning these models while still maintaining a rather high diversity among the reserved ones is typically considered a proper method for the construction of an efficient and effective ensemble system [2].

The problem of ensemble pruning has been proven to be an NP-complete problem [25,26]. Enumerative algorithm for searching the best subset of classifiers is not easily worked for ensembles that contain a large number of constituent models. Greedy algorithms, however, possess high speed, since they only consider a very small subspace among all the possible combinations [16–18,21,27]. But this characteristic may result in suboptimal solutions of the ensemble pruning problem [25]. A compact review

* Tel.: +86 25 84593038; fax: +86 25 84498069.
E-mail address: daiqun@nuaa.edu.cn

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.08.024

about the related works on ensemble pruning is given in Section 2.2 of this paper.

This work, however, studies the problem of ensemble pruning from a perspective of competitive learning. For the research purpose, the *n*-Bits Binary Coding ICBP Ensemble System (*n*BBC-ICBP-ES) proposed in our previous work [28] is employed as the basic ensemble. A brief introduction about *n*BBC-ICBP-ES is given in Section 5.2. The reason why *n*BBC-ICBP-ES is adopted as the initial ensemble system for this work is very natural and intuitive. Because *n*BBC-ICBP-ES is successfully implemented in our previous work. It is simple but efficient and effective, and its effectiveness has been verified through experiments on several Benchmark classification tasks. And it is anticipated that the investigation about CEPCV Algorithm would improve the classification performance and generalization capability of the initial *n*BBC-ICBP-ES further, so that a desirable selective ensemble could be achieved, which is an original neural network system completely resulted from our own research works.

After the basic *n*BBC-ICBP-ES has been generated, the proposed Competitive Ensemble Pruning Algorithm Based on Cross-Validation (CEPCV) is started up for the purpose of ensemble pruning, wherein the final pruned ensemble is dynamically constructed with the help of cross-validation technique. Explicitly, for the specific test instance $\boldsymbol{t}$ under consideration, we calculate its squared Euclidean distance from every validation sample $\boldsymbol{v}_i$. After that, all the validation samples are arranged according to their above calculated squared distances values from $\boldsymbol{t}$. Then, the first VS_n validation instances in the arranged array of validation set, i.e. the VS_n nearest neighbors of test sample $\boldsymbol{t}$ in the validation set, are picked out to form the dynamic validation subset associated with the specific test instance $\boldsymbol{t}$. Each constituent ICBP model in the basic *n*BBC-ICBP-ES is then employed to provide its classification results to the above selected VS_n dynamic validation instances. Those ICBP components which correctly classify at least τ dynamic validation instances are declared the winners in the competition and selected into the dynamically pruned NNE associated with test instance $\boldsymbol{t}$. Finally, the classification decision for test sample $\boldsymbol{t}$ is made based upon the dynamically pruned NNE using the method of majority voting.

Our motivations for the development of CEPCV algorithm mainly consist of: First of all, the data needed to be learnt by the neural computing models are usually drifting and changing along with time and environment [2]. However, a majority of the typical ensemble pruning strategies imply that the component models selected to comprise the pruned ensemble are changeless once decided. They are incapable to realize ensemble pruning flexibly and changeably. This kind of defect will inevitably lead to neglect of valuable heuristic information in the data. In contrast, the proposed CEPCV method can actualize ensemble pruning dynamically, the pruning decisions of which are alterable and sensitive to each different testing sample under processing. This characteristic constitutes a remarkable novelty of CEPCV algorithm, which makes it evidently different from other typical ensemble pruning methods, i.e. it realizes pruning operation at the same time with the test procedure of the ensemble system, resulting in a final pruned ensemble with significantly higher accuracy and reliability.

Secondly, CEPCV algorithm naturally inherits the competence of the technique of cross-validation. The cross-validation technique is a standard tool in statistics which provides an appealing guiding principle to choose, within a set of candidate model structures, the "best" one according to a certain criterion [29]. The hope here in CEPCV is that the networks regarded as winners and selected into the pruned ensemble have the "best" generalization capability.

Thirdly, CEPCV algorithm boosts up the holistic predictive performance of selected models, while maintaining a high diversity among them. Constitutionally, the basic thinking behind CEPCV algorithm is the divide-and-conquer strategy [30], which is a significant research strategy of ensemble learning. And the unique strategy of CEPCV algorithm itself can be explained as "rout the enemy forces one by one". It might alleviate the local minimum problem of many traditional ensemble pruning approaches at the cost of a little greater computational cost, which is generally acceptable to the requirements of most applications.

The notion of *diversity* here is a rather broad sense of concept. It means that those specific selective subensembles are diversified among each other, which is associated with each different test instance $\boldsymbol{t}$. In this sense, it could be considered that, the selected subensembles maintain a high diversity among each other. Kuncheva and Whitaker have studied several statistics which can measure diversity among binary classifier outputs in their published work [31]. However, most of these diversity measures are not applicable to dynamically pruned ensembles, such as those resulted from CEPCV algorithm. Therefore, it would be our future work to investigate some diversity measures that could be applied to the scenario of dynamic ensemble pruning.

The remains of this work are structured as follows: Section 2 presents the method of ensemble pruning, including a theoretical analysis and a compact review about its related works in reference papers. Section 3 briefly reviews the technique of cross-validation for neural network optimization. Section 4 presents the proposed Competitive neural network Ensemble Pruning algorithm based on Cross-Validation technique (CEPCV). Section 5 reports the results of experimental study. From these experimental results, the final conclusions are drawn in Section 6.

2. Ensemble pruning method

2.1. Theoretical analysis on ensemble pruning

It was about 22 years ago when Hansen and Salamon proposed Neural Network Ensemble (NNE) [29]. They claim that ensembling a group of neural networks can improve the generalization capability of each individual component network significantly. This technology has recently become a very hot topic in both neural networks and machine learning communities for its remarkably desirable performance. However, it should be noticed that the law of "the more, the better" is not always true for all occasions [32]. It is necessary to use an appropriate method to select some individual members from those overproduced multiple ensemble members for the goal of successful ensemble prediction [32]. Ensemble pruning is important for the requirement of both efficiency and predictive performance. The following is the theoretical analysis on the generalization capability of neural network ensemble pruning [32].

Suppose the task is to use an ensemble comprising *N*component neural networks to approximate a function $f:\boldsymbol{R}^m \to C$, where C is the set of class labels and the classification results of the component networks are aggregated by the approach of majority voting. For simplicity and convenience of discussion, we assume that C includes only two class labels, i.e. the function to be approximated is $f:\boldsymbol{R}^m \to \{-1,1\}$.

Suppose there are *m*samples, the desired output, i.e. $\boldsymbol{D} = [d_1, d_2, \ldots, d_m]^T$, where d_j denotes the desired output on the jth sample, and the actual output of the ith component neural network, i.e. $\boldsymbol{f}_i$, on those samples is $[f_{i1}, f_{i2}, \ldots, f_{im}]^T$, where f_{ij} denotes the actual output of the ith component network on the jth sample. $\boldsymbol{D}$ and $\boldsymbol{f}_i$ satisfy that $d_j \in \{-1,1\}$ $(j = 1, 2, \ldots, m)$ and $f_{ij} \in \{-1,1\}$ $(i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, m)$, respectively. Then the generalization error of the ith component network on those m samples is:

$$E_i = \frac{1}{m}\sum_{j=1}^{m} I(f_{ij} \neq d_j) \tag{1}$$

where $I(\cdot)$ is an indicator function ($I(true) = 1$ and $I(false) = 0$).