Expert Systems With Applications 53 (2016) 192–203

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Root-quatric mixture of experts for complex classification problems

Elham Abbasi[a], Mohammad Ebrahim Shiri[a,b,*], Mehdi Ghatee[a,c]

[a] Department of Computer Science, Amirkabir University of Technology, N. 424, Hafez Ave, Tehran, Iran
[b] Laboratory of Systems and Intelligent Agents (SINA), Amirkabir University of Technology, N. 424, Hafez Ave, Tehran, Iran
[c] Laboratory of Network Optimization Research Center (NORC), Amirkabir University of Technology, N. 424, Hafez Ave, Tehran, Iran

ARTICLE INFO

Keywords:
ME
Negative correlation learning
Neural network ensemble
Ensemble learning
Diversity
Root–quartic negative correlation learning

ABSTRACT

Mixture of experts (ME) as an ensemble method consists of several experts and a gating network to decompose the input space into some subspaces regarding to the experts specialties. To increase the diversity between experts in ME, this paper incorporates a correlation penalty function into the error function of ME. The significant of this modification is providing an occasion to encourage experts to specialize on different parts of the input space and to create decorrelated experts. The experimental results of this approach reveals that the impacts of this penalty function is extremely improved the diversity of experts and the tradeoff between the accuracy and the diversity in ME. Moreover in the implementation of this method, the experts are trained simultaneously and they can communicate by the aid of the correlation penalty function. The performance of the proposed method on ten classification benchmark datasets shows that the average of accuracy of this method improves 1.94%, 3.7%, and 3.74% compared with the mixture of negatively correlated experts, ME and the negative correlation learning, respectively. Thus the proposed method can be considered as a better classifier for healthy and medical problems and also when the great non-stationary data should be classified.

1. Preliminaries and related works

Ensemble learning directs to combine multiple experts (classifiers or regressions) that are trained on a sample problem. Their decisions are combined to obtain better generalization ability in comparison with the base models. For this aim, ensemble learning applies diverse experts and tries to minimize their mistakes. Different methods are designed to create decorrelated (diverse) experts that can be classified as explicit and implicit methods (Brown & Yao, 2001). Implicit methods indirectly effect on the learning path to encourage the experts to be divers. These methods use different strategy such as different weight initializations, different training data and different network topologies in order to create diversity (Brown & Yao, 2001). For instance, some methods in the implicit group provide for each expert different sub sets of training data. They train each expert on the whole of feature space but with a portion of input samples such as Breiman, 1996, Freund and Schapire (1997), Gaikwad and Thool (2015), Simidjievski, Todorovski, and Džeroski (2015). For example bagging (Breiman, 1996) and boosting (Freund & Schapire, 1997) implicitly create diversity in the base experts. In bagging, different training data is generated by resampling techniques. In this method, M training sets are created by resampling M times from the original training set with replacement. Also, boosting generates different training sets by resampling the original training data. But the instances that misclassified with the previous classifier have greater weights to resample for new classifier (Freund & Schapire, 1997). Other methods in this group use different subsets of features or different features as input features for each expert such as Cruz, Sabourin, Cavalcanti, and Ren (2015), Kheradpisheh, Sharifizadeh, Nowzari-Dalini, Ganjtabesh, and Ebrahimpour (2014), Li, Zou, Hu, Wu, and Yu (2013), Pedrajas and Osorio (2011), Peralta and Soto (2014), Tamponi (2015). A dynamic classifier ensemble has been proposed in Li et al. (2013) that random feature selection has been adopted to generate diverse classifiers. A subset of classifiers was selected dynamically according to the confidence of the classifiers. In Kheradpisheh et al. (2014) a 3-phase ensemble system based on ME has been proposed that an optimal subset of features were selected in the first phase then each expert was trained on a subset of features and the experts were trained with the standard ME training algorithm at the end. A regularized ME has been proposed in which L_1 regularization was applied for local feature selection in experts and gating network in Peralta and Soto (2014). In Pedrajas and Osorio (2011) a linear and non-linear supervised

* Corresponding author at: Department of Computer Science, Amirkabir University of Technology, Tehran, Iran. Tel.: +98 2166460948; fax: +98 2166497930.
E-mail addresses: e.abbasi@aut.ac.ir (E. Abbasi), shiri@aut.ac.ir (M.E. Shiri), ghatee@aut.ac.ir (M. Ghatee).
URL: http://www.aut.ac.ir/shiri (M.E. Shiri)

http://dx.doi.org/10.1016/j.eswa.2016.01.040
0957-4174/

projection has been applied to construct an accurate and diverse ensemble system. To find the projection, the misclassified instances have been applied. Cruz et al. (2015) have proposed a dynamic ensemble selection method that five different meta-features are extracted from training set to train meta-classifier to determine the level of suitability of classifier to classify the input sample. In other methods in this group, the experts are specialized in different parts of input space by applying a weighting strategy for input samples (Tamponi, 2015). In Tamponi (2015) forest of local trees has been proposed that first the input space was divided into the centralized sub spaces according to some centers selected from training set. For this aim, a strategy was applied to place the centers in the input space in the way that be far from each other. To specialize trees in specific region of input space, a center was assigned to each tree and a weighted training set was applied to encourage the trees to subspaces of the input space close to their centers. The other approaches in this group use different kinds of classifiers or different topologies for them to create diversity (Fossaceca, Mazzuchi, & Sarkani, 2015). In explicit group, the learning algorithm is manipulated for creating diversity among the experts (Brown & Yao, 2001). In this group, at the same time with the learning, the experts are encouraged to learn different subspaces of the problem by incorporating a penalty correlation term in their error function (Liu & Yao, 1999a; Masoudnia, Ebrahimpour, & Arani, 2012b; McKay & Abbass, 2001). For example in negative correlation learning (NCL) and mixture of experts (ME), a specific error function is used to reduce the correlation between the base experts. ME applies a specific error function to directly effect on the learning path and encourage base experts to learn different aspects of input spaces. It consists of a gating network that dynamically assigns the weights to the output of experts and combines them. In NCL, a regularization term is incorporated into the error function of each expert. This term is used in order to quantify the error correlation and can be minimized explicitly during the training phase.

Explicit and implicit methods have different advantages and disadvantages that are complementary each other (Islam, Yao, Nirjon, Islam, & Murase, 2008; Liu & Yao, 1999b). For example in bagging and boosting, experts are added sequentially and they are independent of each other. So, there is the loss of cooperation and interaction among the individual networks during learning in implicit approaches (Liu & Yao, 1999a). Also, there is no feedback from combination phase to the training phase of base experts. But the base experts in NCL and ME are trained in parallel and they have cooperation with each other and there is a feedback from combination phase to the training phase of base experts. Some researchers tried to combine the advantages of implicit and explicit approaches (Avnimelech & Intrator, 1999; Ebrahimpour, Sadeghnejad, Arani, & Mohammadi, 2013). Also, the explicit methods such as NCL and ME have complementary properties which can be combined (Masoudnia et al., 2012b; Masoudnia, Ebrahimpour, & Arani, 2012a).

In the follows, NCL and ME methods are described and the features of them are compared.

NCL (Liu & Yao, 1999a) is an ensemble of neural network experts with a correlation penalty function added to the error function of each expert. So each expert not only reduces the mean square error, but also decreases the correlation with the ensemble (Brown & Yao, 2001). NCL is defined briefly as follows:

Suppose $D = \{X, Y\}$ denote the training dataset where $X = \{x_{(n)}\}_{n=1}^{N}$ is the input set and $Y = \{y_{(n)}\}_{n=1}^{N}$ is the target set, N is the number of training data. Also, the artificial neural network (ANN) is applied as experts and gating model. The ensemble output for the n^{th} input is given by:

$$f_{ens}(x_n) = \frac{1}{M}\sum_{i=1}^{M} f_i(x_n) \tag{1}$$

where f_i is the output of the ith expert and M is the number of experts. The error function of the ith ANN is given with:

$$e_i = \sum_{n=1}^{N} (f_i(x_n) - y_n)^2 + \lambda p_i \tag{2}$$

where λ is a weighting parameter to control the effect of the correlation penalty function p_i. This parameter controls the trade-off between accuracy and diversity. $\lambda = 0$ is equivalent to train each network independently. The correlation penalty function p_i in NCL is defined as follows:

$$\begin{aligned} p_i &= \sum_{n=1}^{N} \left\{ (f_i(x_n) - f_{ens}(x_n)) \sum_{j\neq i} (f_j(x_n) - f_{ens}(x_n)) \right\} \\ &= -\sum_{n=1}^{N} (f_i(x_n) - f_{ens}(x_n))^2 \end{aligned} \tag{3}$$

NCL have been applied in regression problems (Fernandez-Navarro, Gutiérrez, & Hervás-Martínez, 2013) and classification problems (Oliveira, Morita, Sabourin, & Bortolozzi, 2005; Wang, Chen, & Yao, 2010) because of its good performance. One of the good ideas for NCL was proposed by McKay and Abbass (2001). They proposed an anti-correlation measure, namely root quartic negative correlation learning measure with the better performance than NCL on some problems (McKay & Abbass, 2001). In the next part ME is briefly described.

ME has been introduced by Jacobs Jacobs, Jordan, Nowlan, and Hinton (1991) as a modular architecture consisting of a set of experts and a gating network which learn to decompose the input space and to assign weights to the outputs of each expert according to their inputs. A hard and complex problem is partitioned into the simple sub problems based on the divide and conquer approach. The experts are trained to solve the sub-problems and the gating network combines the solutions of experts. Suppose we have M experts denoted with $j = 1, 2, \ldots, M$. The output of jth expert is $f_j(x)$. The gating network produces $g_j(x)$ for each expert with respect to the input vector x. $g_j(x)$ can be interpreted as the probability of selecting the output from expert j by the gating network. The following softmax function is used as the gating network which satisfies $g_j(x) \geq 0$, and $\sum_j g_j(x) = 1$.

$$g_j(x) = \frac{\exp(o_{g_j})}{\sum_j \exp(o_{g_j})} \tag{4}$$

where o_{g_j} is the jth output of the gating network. The output vector ME is combination of the gating networks and the expert output as follows (Jacobs, 2008):

$$f_{ens}(x) = \sum_{j=1}^{M} g_j(x) f_j(x) \tag{5}$$

The architecture of ME is depicted in Fig. 1.

Different error functions were presented for the mixture of experts (Jacobs et al., 1991). The following error function performs better in a great number of experiments (Jacobs et al., 1991):

$$E = -\log \sum_j g_j \exp\left(-\frac{1}{2}(y - f_j)^2\right) \tag{6}$$

where g_j is the output of gating network corresponding to jth expert, y is the target value and f_j is the output of jth expert.

Because of the good performance of ME, it is applied in various areas such as bioinformatics (Cao, Meugnier, & McLachlan, 2010; Qi, Klein-Seetharaman, & Bar-Joseph, 2007), robotic (Trentin & Cattoni, 1999), medical diagnosis (Yao, Walther, Beck, & Fei-Fei, 2009), activity recognition (Lee & Cho, 2014) and so on.

Several methods were proposed in order to distribute the experts in the input space and to improve task decomposition in