Engineering Applications of Artificial Intelligence 37 (2015) 34–42

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Proposing a classifier ensemble framework based on classifier selection and decision tree

Hamid Parvin [a], Miresmaeil MirnabiBaboli [b], Hamid Alinejad-Rokny [c,d,e,*]

[a] School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[b] Institute of Informatics and Automation Problem, NAS RA, Armenia
[c] Faculty of Medicine, The University of New South Wales, Sydney, Australia
[d] School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia
[e] Faculty of Computer Engineering, University of Mazandaran, Mazandaran, Iran

ARTICLE INFO

Article history:
Received 4 January 2013
Received in revised form
7 July 2014
Accepted 8 August 2014
Available online 19 September 2014

Keywords:
Decision tree
Classifier ensembles
Clustering
Bagging
Learning
Ada Boosting

ABSTRACT

One of the most important tasks in pattern, machine learning, and data mining is classification problem. Introducing a general classifier is a challenge for pattern recognition communities, which enables one to learn each problem's dataset. Many classifiers have been proposed to learn any problem thus far. However, many of them have their own positive and negative aspects. So they are good only for specific problems. But there is no strong solution to recognize which classifier is better or good for a specific problem. Fortunately, ensemble learning provides a good way to have a near-optimal classifying system for any problem. One of the most challenging problems in classifier ensemble is introducing a suitable ensemble of base classifiers. Every ensemble needs diversity. It means that if a group of classifiers is to be a successful ensemble, they must be diverse enough to cover their errors. Therefore, during ensemble creation, a mechanism is needed to ensure that the ensemble classifiers are diverse. Sometimes this mechanism can select/remove a subset of base classifiers with respect to maintaining the diversity of the ensemble. This paper proposes a novel method, named the Classifier Selection Based on Clustering (CSBS), for ensemble creation. To insure diversity in ensemble classifiers, this method uses the clustering of classifiers technique. Bagging is used to produce base classifiers. During ensemble creation, every type of base classifier is the same as a decision tree classier or a multilayer perceptron classifier. After producing a number of base classifiers, CSBC partitions them by using a clustering algorithm. Then CSBC produces a final ensemble by selecting one classifier from each cluster. Weighted majority vote method is used as an aggregator function. In this paper we investigate the influence of cluster number on the performance of the CSBC method; we also probe how we can select a good approximate value for cluster number in any dataset. We base our study on a large number of real datasets of UCI repository to reach a definite result.

1. Introduction

Classification is the most important task in pattern recognition institute. Therefore, since the beginning of the pattern recognition science, one of the most challenging problems in this field was introducing a general classifier that can learn properly every dataset of any given problem. Many classifiers have been proposed to learn problems thus far. However, all of them have their own positive and negative aspects. So they are good only for specific problems. But there is no strong solution to recognize which classifier is a better or good classifier for a specific problem. Since finding the best classifier is an impractical problem, we must use another approach. Thus might be using many inaccurate classifiers, where each of them is assigned to a subspace of dataset as an ensemble (i.e. use their gathering vote as the decision of ensemble). Ensemble learning is a strong approach to produce a near-to-optimal classifier for any problem. This method reinforces the ensemble in error-prone subspaces, and hence can lead to better performance of classification. In general the following sentence can be true to say that the result of combination of diverse classifiers, is better classification. Diversity is an important factor for each ensemble to be successful. The existence of diversity in an ensemble ensures that those classifiers are independent of each other. It means that misclassifications do not occur simultaneously. Kuncheva showed that increasing the number of diverse classifiers

* Corresponding author.
E-mail addresses: H.Alinejad@ieee.org, Hamid.AlinejadRokny@UoN.edu.au (H. Alinejad-Rokny).

http://dx.doi.org/10.1016/j.engappai.2014.08.005
0952-1976/

can lead to better performance (even perfect accuracy) (Kuncheva, 2005; Minaei-Bidgoli et al., 2004; Alizadeh et al., 2011). Also, ensemble philosophy is applicable to Bayesian Networks (Peña, 2011). The main challenge in the creation of classifier ensemble is to provide a general approach to ensure diversity, which is an important factor for an ensemble. It means that if an ensemble of classifiers has to be a successful ensemble, they should be diverse enough to cover their errors. Creating some suitable diverse classifiers that can participate in an ensemble is a challenging problem. There are a number of ways to obtain a desired diversity in an ensemble. Kuncheva proposed some approaches based on the metrics that indicates the amount of similarities or differences among classifiers' outputs (Kuncheva and Whitaker, 2003; Parvin et al., 2013a, 2013b, 2013c, 2011c, 2011d; Rezaei et al., 2011).

Clustering is a process of assigning a group of objects into clusters. So objects in the same cluster are more similar to each other than the objects in other clusters. This is used a lot in some applications of data mining, especially for information retrieval, text categorization and text ranking (Yang, 2006; Dasgupta and Ng, 2010; Amigó et al., 2011; Kurland and Krikon, 2011).

Giacinto and Roli, by producing a large number of artificial neural network classifiers by different initializations of their parameters and then selecting a subset of them based on their distances in output space, proposed an approach to hosting the classifiers with a high degree of diversity.

Hamid et al. were inspired from the clustering and selection method and proposed a new clustering and selection method that enables one to reduce the drawbacks of the simple ensemble methods in creating diversity (Parvin et al., 2001, 2013c, 2013e). They consider how the base classifiers are created. They also investigated the usage of Boosting and Bagging method as a source of diversity generation on Giacinto and Roli's method. At first, they trained a large number of classifiers using the Boosting and Bagging method, and then partitioned them based on the output over the training set. Finally they chose a classifier randomly and then inserted it into the ensemble. The weighted majority voting mechanism was used as the consensus function of the ensemble.

In this paper a novel method, named the Classifier Selection Based on Clustering (CSBC), has been proposed. This method can provide the necessary diversity between ensemble classifiers by using the clustering technique. And it uses the Bagging method as a generator of base classifiers. Base classifiers are still fixed in the decision tree classifier or the multilayer perceptron classifier during the creation of an ensemble. Then the clustering algorithm partitions the classifiers. Weighted majority vote mechanism was used as the consensus function of ensemble. We investigated how the number of clusters can affect the performance of the CSBC method. We based our study on a large number of real datasets of the UCI repository to reach a definite result. In this paper we investigate the better selection of the classifier from each cluster, how to select a good parameter according to the dataset and the effect of the training set ratio of each base classifier on CSBC's performance.

Machine learning is an important paradigm in artificial intelligence, and Artificial Neural Network (ANN) is a common approach to learning. Unlike the traditional approach, ANN has some properties such as self-adaptivity, ability to generalize, and so on. But the source of these properties is not explained well. They are almost explained by the comparison between ANNs and real neural networks. From bionic point of view, these formal and biological neurons do not have anything in common.

An ANN is a model that enables one to obtain any input and produce the desired set of outputs. An ANN includes two base elements: neurons and connections. An ANN is a set of neural network with connections between them. From another perspective an ANN includes two distinct views: topology and learning. Topology is related to the existence or nonexistence of a connection. Learning in an ANN indicates the power of topology connections. Multi-Layer Perceptron (MLP) is one of the most representative of ANNs. There are different methods to set the power of connections in MLP. One method is setting the weights using a priori knowledge. Another method is to train the MLP, and then using the teaching patterns and finally changing the weights based on some learning rules. In this paper we use MLP as the base classifier.

However, there is no common theory for ANN learning; some training algorithms are proposed for any given ANN architecture. However, these algorithms are all over classical and external to ANNs. Neural networks, which include coded learning algorithm within astrocytic nets, are rather interesting, because different learning rules can be made internal for them. But the source and structure of these rules remain unclear. Therefore, it seems that there is a paradox between the self-learning capabilities of ANNs and their distinctions from the traditional algorithms. Outwardly, ANNs cannot solve the machine learning problems. Particularly, over-learning is one difficult subject and there is no good explanation in the ANN theory for it. Indeed limiting the training time prevents over-learning. Despite all that, the popularity of ANN is not by chance. However, their benefits must be considered, in order to improve them in the future. Decision tree (DT) is one of the versatile classifiers in the machine learning field. DT is an unstable classifier that can introduce different outputs in successive trainings on the same condition. DT uses a tree-like graph or model for decision. The type of presentation helps experts understand the classifiers (Yang, 2006; Parvin et al., 2011a, 2011b). The natural instability of this method can be a source of diversity for classifier ensemble. An ensemble of a number of DTs is similar to a Random Forest (RF) algorithm, which is one of the powerful ensemble algorithms. This algorithm was developed by Breiman (1996). In this paper, DT is used as the base classifier.

The rest of this paper is organized as follows. Section 2 is related works. In Section 3, the proposed method is explained. Section 4 demonstrates the results of our proposed method against traditional methods. Finally, the conclusion is presented in Section 5.

2. Related work

In general, there are two challenging approaches to mix a number of classifiers that use different training sets: Bagging and Boosting. Both of them are two sources for diversity generation and they are the best ensemble methods.

We suppose that a training set is represented by TS. And the ith data item in TS is represented as O_i and m is the number of data items in TS. Training phase of CSBC is shown in Fig. 1, which uses the Bagging method as the base classifier generation.

Breiman is one of the first researchers to have used Bagging for Bootstrap AGGregatING.

The idea of Bagging is simple and attractive: the classifiers of the ensemble are made by bootstrap copies of the training set. Using different training sets can ensure the necessary diversity of ensembles. It is noticeable that Bagging cannot ensure the necessary diversity.

Another kind of Bagging named “Random Forest” has been proposed by Breiman. RF is a method for ensemble creation that uses a decision tree as the base classifier generator. In “Random Forest”, an ensemble of decision trees must be built by creating independent identically distributed random vectors and each vector is used to grow a decision tree. Random Forest also cannot ensure the necessary diversity of ensembles. In this paper