Applied Soft Computing 68 (2018) 626–635

Contents lists available at ScienceDirect

Applied Soft Computing

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

Multi-Manifold based Rotation Forest for classification

M. Khamar, M. Eftekhari*

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

Article history:
Received 13 July 2017
Received in revised form 1 April 2018
Accepted 18 April 2018

Keywords:
Rotation Forest
Ensemble learning
Manifold learning

ABSTRACT

Rotation Forest (RF) is a powerful ensemble classifier which has attracted substantial attention due to its performance. The RF algorithm uses Principal Component Analysis (PCA) for constructing the rotation matrix and extracting new features. In this paper, with the aim of extracting new features, three well-known manifold learning techniques are utilized to extract new features and incorporate into PCA for feature extraction. This new RF algorithm is hereby called Multi-Manifold RF (MMRF), and several experiments are conducted in the present study in order to evaluate its performance. The obtained results reported for nineteen datasets show the high efficiency of MMRF compared to fourteen state-of-the-art ensemble methods in terms of classification accuracy and computational effort. Furthermore, two statistical non-parametric tests (Friedman and Wilcoxon) are carried out to compare the average classification accuracies of MMRF with those of the other methods The experimental results demonstrate that MMRF outperforms twelve of these methods, while there is no significant difference between MMRF and the other two powerful ensemble-based methods, namely the SES-NSGAII and the IDES-P.

1. Introduction

Ensemble methods are learning algorithms which leverage a set of classifiers [1] rather than using a single one. These classifiers called base models are trained to jointly solve one specific classification task [2]. In ensemble methods, the final classification label for a sample is usually determined by applying a voting scheme in order to aggregate the individual votes of the base models [3]. In practice, ensemble learning is usually more accurate than individual base models. An ensemble-based model is comprised of three parts: sample selection strategy, training the base classifiers to compose the Base Classifier Pool (BCP), and combining the classifiers of the BCP [2,4]. Some of the ensemble-based algorithms are: bagging, boosting and RF [5].

Generally, Ensemble Learning Systems (ELSs) are divided into two categories: Static Classifier Ensemble (SCE) and Dynamic Classifier Ensemble (DCE). Within the SCE approach, a fixed ensemble scheme, learned during the training phase, is utilized for all the test samples. There are three types of SCE methods: Classifier Fusion (CF), Static Classifier Selection (SCS) [6,7], and Static Ensemble Selection (SES) [4,8]. On the other hand, the basic idea in DCE is to estimate the accuracy of each classifier in a local region of the feature space around a given test sample and to select one (or more) classifier(s) with the highest value of the local accuracy in order to separately classify a test sample [9]. The DCE approaches are divided into two strategies: Dynamic Classifier Selection (DCS) [9–12] and Dynamic Ensemble Selection (DES) [9].

One objective behind embedding the data into a low dimension space is to preserve the local structure of the new projected data. Therefore, the projection may reduce the effects of noise and outliers [13]. In order to project data into a lower dimension space, various mapping methods have been devised, such as PCA, LDA, LLE, and LEM (described in Section 2) [13]. The RF algorithm is a method for generating classifier ensembles on the basis of feature extraction. To construct training data for the base classifiers (base classifier is DT), the feature set is randomly split into K subsets, and the PCA is applied to each subset. The training data are projected along with the eigenvectors of the covariance matrix, and then, the models are constructed based on the new samples [13,15].

The main contribution of this paper is to extend the RF algorithm such that three well-known data transformation techniques, namely Local Linear Embedding (LLE), Laplacian Eigen Maps (LEM) and Linear Discriminant Analysis (LDA) are utilized for feature extraction and incorporating into PCA. Accordingly, some idea from manifold learning is used within RF for projecting the feature subsets into the new spaces.

The rest of the paper is organized as follows. A brief review of ensemble learning systems and feature extraction methods are conducted in Sections 2 and 3, respectively. In Section 4, a framework is described for the proposed method, which is based on

* Corresponding author.
E-mail address: m.eftekhari@uk.ac.ir (M. Eftekhari).

https://doi.org/10.1016/j.asoc.2018.04.026
1568-4946/

different manifold learning approaches. A variety of experiments on nineteen datasets are presented and discussed in Section 5. Finally, Section 6 draws the conclusion.

2. Related works

According to review of contributions in the field of ELSs, there are five methods for combining and generating the classifiers: CF [4,8], SCS [5,8,24], SES [7,8], DCS [10,11,25] and DES [8,9,24]. CF, SCS, and SES are static classifiers ensembles, while DCS and DES are dynamic. These strategies are briefly explained below.

2.1. Classifier Fusion (CF)

CF utilizes all classifiers in the ensemble which are trained in the training phase for decision making as Majority Voting (MV), Bagging, Boosting, Adaboost, and RF. One can refer to [4,8] for more details about these methods. Since the objective in the present study is to extend the original RF algorithm, some recently-introduced extensions of RF are reviewed in the following.

Lu et al. [16] proposed a cost-sensitive RF(C-RF) algorithm for gene expression data classification, with emphasis on misclassification cost, test cost, and rejection cost. Costs were embedded into the RF algorithm. This algorithm is explained in more details in [15]. Using some heterogeneous classifiers, including DT, a heterogeneous RF algorithm have been proposed in the literature [17,18]. The Anticipative Hybrid Extreme RF (AHERF) is a method introduced in [16], being constructed from a pool of DT, Extreme Learning Machines (ELM), Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Adaboost, Random Forests and Gaussian Naive Bayes (GNB) classifiers. An improved heterogeneous RF was proposed by Mousavi et al. for MicroRNA target prediction [18]. Self-training method is among the most common schemes of semi-supervised techniques. In this regard, self-trained RF has been proposed by Fazakis et al. for semi-supervised learning [19] as a combination of self-training scheme and RF. The steps of this method are as follows:

- The RF classifier is trained as the base classifier for selecting a subset of samples called X_{MCP} with the highest confidence prediction.
- X_{MCP} is eliminated from the primary labeled samples (U) and is added to the primary unlabeled samples (L).
- In each iteration, a few samples per class are selected from U and are added to L.
- The RF method is trained again on L samples.
- At the end of each iteration, the RF trained on L is used for evaluating the test samples.

In order to solve noise and outlier problems, a feature-weighted RF algorithm (FWRF) was proposed by Wang et al. [20] which investigates the interactions among the proteins. Here, features with low weight values are removed and the elimination of extra information facilitates the application of useful features and prediction of the interactions among proteins. Xia et al. [21] presented a spectral and spatial Rotation Forest (SSRF) which contains the following steps:

- Pixels are smoothed by the multi-scale (MS) spatial weight mean filtering.
- The spectral-spatial data transformation is employed in the RF.
- Classification results are obtained through majority voting.

Su et al. [22] have improved RF based on Hellinger Distance (HD) for classification of highly imbalanced data. More recently, fuzzy-based ideas have been applied for improving RF classification performance across imbalanced datasets [23,24].

2.2. Static Classifier Selection (SCS)

SCS is a method which, firstly, selects the best classifier for each region of competence in the feature space during the training phase. After that, each test sample is classified with a classifier related to its own region. In the same context, Kuncheva [6] has proposed a method for classifier combination based on a selection technique. Also, single Best approach (SB) can select the best classifier in the ensemble with the lowest training error for all test samples [25].

2.3. Static Ensemble Selection (SES)

SES selects an optimal set of classifiers for all test samples. Here, Yang [8] proposes two methods: classifiers that are selected based on accuracy (SA), and those selected based on accuracy and diversity (SAD). The first method (SA) selects 75% of classifiers with the lowest error in the validation phase for all test samples, while the latter (i.e. SAD) selects 90% of the classifiers with the lowest error in the validation phase. Eventually, 75% of the most accurate and diverse classifiers are selected.

2.4. Dynamic Classifier Selection (DCS)

DCS method selects one classifier with the highest accuracy for each test sample based on the validation set [26]. Here, two methods are introduced as follows: DCS-Local Accuracy (DCS-LA) [10] and DCS-Multiple Classifier Behavior (DCS-MCB) [11]. The value of "k" adapts dynamically in these two methods.

- *DCS-LA*

The accuracy of each classifier is estimated in a local region defined as k-nearest neighbors of the test instance taken from the validation set. The classifier with the highest value of this local accuracy is selected for classifying the test sample [4,10].

- *DCS- MCB*

DCS-MCB is a method proposed by Giacinto and Roli [11] based on the concepts of classifiers LA (CLA) and MCB. This method firstly estimates the accuracy of each classifier in a local region of the feature space surrounding a test instance, and then selects the classifier with the highest value of this LA to classify the test instance. This algorithm includes the following parts:

- First part: the k-nearest neighbors in the training data are computed for each test instance by MCBs (the MCB is defined as a vector of class labels assigned to the test instance x).
- Second part: this strategy allocates the class labels to the test instance x and its k-nearest neighbors between all classifiers in the BCP. Next, the similarities between MCBs are computed using Hamming distance. The k-nearest neighbors of a test instance whose similarities are higher than a similarity threshold are selected [4,11].

2.5. Dynamic Ensemble Selection (DES)

The DES approaches utilize an optimal classifier ensemble and have better performance for classifying each test instance, compared to single classifiers. Several DES approaches have recently been proposed [4], including DES-Knora-Eliminate (DES-KE) [4,9], DES-Performance (DES-P), DES-Kullback–Leibler (DES-KL) and