Knowledge-Based Systems 89 (2015) 278–287

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Classifier ensemble creation via false labelling

Bálint Antal

Faculty of Informatics, University of Debrecen, 4010 Debrecen, POB. 12, Hungary

ARTICLE INFO

Article history:
Received 15 October 2014
Received in revised form 25 June 2015
Accepted 10 July 2015
Available online 17 July 2015

Keywords:
Ensemble learning
Diversity
Hidden Markov Random Fields
Simulated annealing
Bioinformatics

ABSTRACT

In this paper, a novel approach to classifier ensemble creation is presented. While other ensemble creation techniques are based on careful selection of existing classifiers or preprocessing of the data, the presented approach automatically creates an optimal labelling for a number of classifiers, which are then assigned to the original data instances and fed to classifiers. The approach has been evaluated on high-dimensional biomedical datasets. The results show that the approach outperformed individual approaches in all cases.

1. Introduction

Classification is a fundamental task in machine learning. In numerous application fields very complex data needs to be classified which is often a difficult task for a single machine learning classifier [1,2]. There are tremendous amount of research on improving the classification performance in such cases. One highly investigated field for this problem is ensemble learning [3], where multiple prediction are fused the produce a more efficient classification approach. One fundamental requirement for the creation of classifier ensembles is diversity among them [4], that is, the classifiers included in the ensemble need to complement each other to provide more generalization capabilities than a single learner. Bagging [5] uses randomly selected training subsets with possible overlap (bootstrapping [6]) to ensure diversity among the member of the ensemble. Other diversity creation techniques may involve disjoint random sampling (random subspace methods [7], for example, some variants of Random Forest algorithms [8]), while Adaboost [9] based techniques aims to increase the accuracy of a weak learner iteratively (boosting [10]) using targeted sampling: each iteration considers the misclassified instances of the training data to be more important, and drives the iteration process to include them in the current training set. Another approach to create diverse ensembles is ensemble selection [11], where diversity of classifiers trained on the same dataset is measured and an optimal subset is selected.

A more comprehensive review on the above described techniques can be found in [12]. The relationship of classifier diversity and ensemble accuracy is highly investigated in the ensemble learning community. Although the definite connection between diversity measures and ensemble accuracy is an open question [13], a decomposition of majority voting error into good and bad diversity is proposed in [14].

In this paper, a novel approach for ensemble creation based on this theoretical result is presented. The proposed approach takes the predictions of a single classifier on a training set. Then, an optimal labelling complimenting the predictions of the classifiers is created. Thus, an optimal but false labelling set is created for a number of classifiers. The data with each false labelling is trained to a classifier thus forming an ensemble. We define a Markov Random Field problem to create an optimal ensemble with this method. The approach has been tested on high-dimensional biomedical datasets where a large improvement over a single learner is achieved. Other aspects of the algorithm including its performance comparison with different number of ensemble members are also discussed. The outline of the proposed algorithm can be seen in Fig. 1.

The rest of the paper is organized as follows: Section 2 contains the mathematical background behind the proposed method, while Section 3 defines an optimization problem to solve it and proposes an implementation. Section 4 contains our experimental details, while the results are presented and discussed in Section 5. Finally, conclusions are drawn in Section 6.

E-mail address: antal.balint@inf.unideb.hu

http://dx.doi.org/10.1016/j.knosys.2015.07.009
0950-7051/

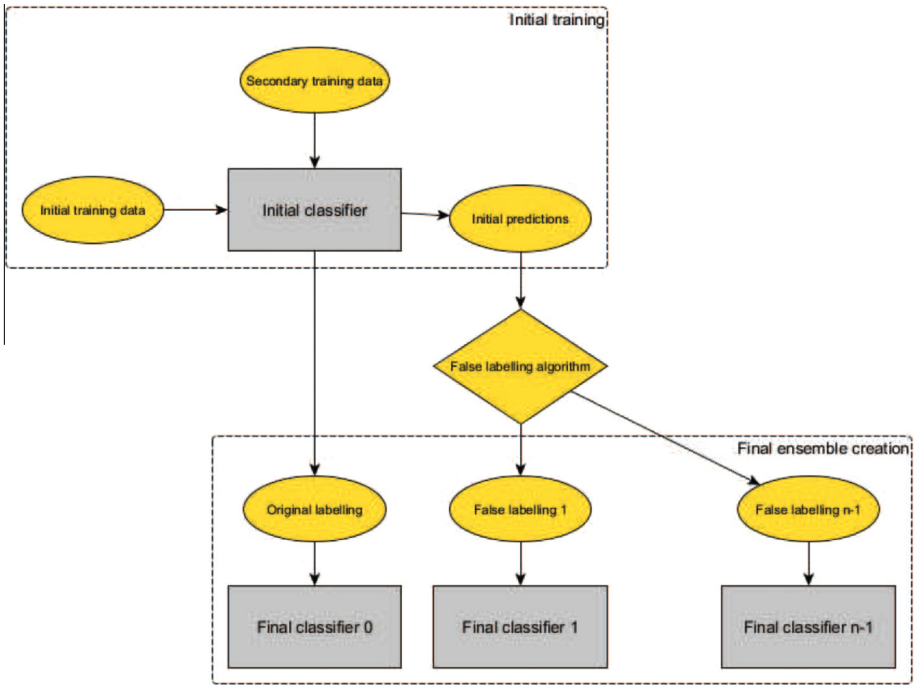

Fig. 1. Flowchart of ensemble creation via false labelling.

2. Ensemble creation via false labelling

The presented false labelling based ensemble creation are presented is restricted to binary classification problems. In this section, the mathematical background behind the algorithm is presented. Moreover, an optimization problem is defined to provide an efficient solution for the false labelling problem. For the basic machine learning and ensemble definitions, we relied on the classic literature [3,14].

Let $\Omega = \{-1, +1\}$ be a set of class labels. Then, a function

$$D : \mathbb{R}^n \to \Omega \tag{1}$$

is called a classifier, while a vector $\vec{\chi} = (\chi_1, \chi_2, \ldots, \chi_n) \in \mathbb{R}^n$ is called a feature vector. A dataset $T \in \{R^n \times \Omega\}^l$ can be defined as follows:

$$T = \{\langle\vec{\chi}_0, \omega_0\rangle, \langle\vec{\chi}_1, \omega_1\rangle, \ldots, \langle\vec{\chi}_k, \omega_k\rangle\}, \tag{2}$$

where $\vec{\chi}_i \in \mathbb{R}^n$, $\omega_k \in \Omega$, $i = 1, \ldots, k$ are feature vectors and labels, respectively.

Let $D_1, D_2, \ldots, D_L$ be classifiers and $d_t(\vec{\chi}) \in \Omega$, $t = 1, \ldots, L$ their output on the feature vector $\vec{\chi}$. Then, the output of the majority voting ensemble classifier $\mathcal{D}_{maj} : \mathbb{R}^n \to \Omega$ can be defined as follows:

$$d_{maj}(\vec{\chi}) = sign\left(\frac{1}{L}\sum_{t=1}^{L} d_t(\vec{\chi})\right). \tag{3}$$

The creation of an ensemble $\mathcal{D}_{maj}$ of L classifiers (Eq. (1)) starts by training a base classifier on the half of the training dataset (Eq. (2)) T (T_0). We take the output C_{orig} of the classifier D_{orig} on the other half of the training set (T_1) and create $L-1$ optimal labellings for a the remaining base classifiers $D_i, i = 2, \ldots, L$. Then, we train these classifiers on T_1 with their respective false labellings $\mathcal{C}^i_{false}$.

The outline of the ensemble creation method is summarized in Algorithm 1, while the mathematical formulation is presented in the rest of the section.

Algorithm 1. Outline of ensemble creation via false labelling

Require: a dataset $T \neq \emptyset$, a label set $\mathcal{C} \neq \emptyset$, a classifier D_{orig}, the number of ensemble members $L > 2$ (L is odd).
Ensure: an ensemble of trained classifiers $\mathcal{D}_{maj}$.
1: Split T into T_0 and T_1 randomly.
2: Train D_{orig} on T_0.
3: $C_{orig} \leftarrow D_{orig}(T_1)$
4: $C_{cl} \leftarrow F(C_{orig}) = \{\mathcal{C}^2_{false}, \mathcal{C}^3_{false}, \ldots, \mathcal{C}^L_{false}\}$
5: **for** $i \leftarrow 2, \ldots, L$ **do**
6: Train a classifier D_i on $LC\left(T_1, \mathcal{C}^i_{false}\right), \mathcal{C}^i_{false} \in C_{cl}$.
7: **end for**
8: **Return** $\{D_{orig}, D_2, \ldots, D_L\}$

2.1. Ensemble creation

The proposed ensemble creation depends on the output of one classifier D_{orig} for a given training dataset T.

First, we split T into two equal parts $T^{(0)}$ and $T^{(1)}$ randomly. We train D_{orig} on $T^{(1)}$ and classify all $\vec{\chi}^1_j \in T^{(0)}, j = 1, \ldots, k/2$ element of $T^{(1)}$:

$$\mathcal{C}^1_{orig} = \left\{\omega_j | \omega_j = D_{orig}\left(\vec{\chi}^1_j\right),\ \vec{\chi}^1_j \in T^1,\ j = 1, \ldots, k/2\right\}. \tag{4}$$

Then, we create a majority voting classifier ensemble of L members:

$$\mathcal{D}_{maj} = \{D_1 = D_{orig}, D_2, \ldots, D_L\}. \tag{5}$$

To train $D_2, \ldots, D_L$, we will define a false labelling function $F : \Omega^{k/2} \to \Omega^{k/2\cdot(L-1)}$. That is

$$F\left(\mathcal{C}^1_{orig}\right) = \{\mathcal{C}^2_{false}, \mathcal{C}^3_{false}, \ldots, \mathcal{C}^L_{false}\}, \tag{6}$$