Contents lists available at ScienceDirect

Information Fusion

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

Hierarchical distance learning by stacking nearest neighbor classifiers

Mete Ozay [a,*], Fatos Tunay Yarman-Vural [b]

[a] Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi, Japan
[b] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 27 December 2014
Revised 9 August 2015
Accepted 20 September 2015
Available online 26 September 2015

Keywords:
Decision fusion
Classification
Nearest neighbor rule
Ensemble learning
Hierarchical distance learning

ABSTRACT

We propose a two-layer decision fusion technique, called Fuzzy Stacked Generalization (FSG) which establishes a hierarchical distance learning architecture. At the base-layer of an FSG, fuzzy k-NN classifiers receive different feature sets each of which is extracted from the same dataset to gain multiple views of the dataset. At the meta-layer, first, a *fusion space* is constructed by aggregating decision spaces of all the base-layer classifiers. Then, a fuzzy k-NN classifier is trained in the fusion space by minimizing the difference between the large sample and N-sample classification error. In order to measure the degree of collaboration among the base-layer classifiers and the diversity of the feature spaces, a new measure called, *shareability*, is introduced. *Shearability* is defined as the number of samples that are correctly classified by at least one of the base-layer classifiers in FSG. In the experiments, we observe that FSG performs better than the popular distance learning and ensemble learning algorithms when the *shareability* measure is large enough such that *most* of the samples are correctly classified by at least one of the base-layer classifiers. The relationship between the proposed and state-of-the-art diversity measures is experimentally analyzed. The tests performed on a variety of artificial and real-world benchmark datasets show that the classification performance of FSG increases compared to that of state-of-the art ensemble learning and distance learning methods as the number of classes increases.

1. Introduction

Stacked Generalization (SG) is a widely used ensemble learning technique which is proposed by Wolpert [1] and used by many others [2,3]. Although gathering classifiers under an SG algorithm significantly boosts classification performance in some application domains, it is observed that the performance of the overall system may get worse than that of the individual classifiers in some other cases. Wolpert [1], Ting and Witten [2] defined the problem of designing feature spaces and classifiers, and modeling the relation between the classification performance and various parameters of the SG as a *black art*.

In this paper, the *black art* problem is reformulated as a hierarchical distance learning problem. We propose a Fuzzy Stacked Generalization (FSG) method, which learns a distance function at the meta-layer feature space by minimizing the difference between the large sample and N-sample classification error of the Nearest Neighbor classifier. At the base-layer of an FSG, a set of fuzzy k-Nearest Neighbor (k-NN) classifiers is trained in different feature spaces in order to extract complementary information and gain *expertise* on various properties of samples using multiple distinct features extracted from the samples (i.e., their multiple views) [4]. A fuzzy membership value of a sample for a class obtained at the output of a base-layer classifier is considered as the decision value of the classifier on the class membership of a sample. Decision values enable us to obtain information about the uncertainty of the classifier decisions and the belongingness of the samples to classes [5]. Fusion of the decision values is accomplished by aggregating them under the same space, called fusion space. Finally, a meta-layer fuzzy (k-NN) classifier is employed to learn the distance function for the fusion space. This task is achieved by formulating the classification error of the proposed FSG in two parts, namely (i) N-sample error which is the error of a classifier employed on a training dataset of N samples, and (ii) large-sample error which is the error of a classifier employed on a training dataset of the large number of samples such that $N \to \infty$. A distance learning approach proposed by Short and Fukunaga [6] is extended into the hierarchical FSG architecture for decision fusion in order to minimize the difference between N-sample and large sample error.

In the experiments, it is observed that the performance of the FSG depends on the degree of collaboration among the base-layer classifiers. In addition, FSG may boost the overall performance using a set of weak base-layer classifiers instead of using a set of strong classifiers, if decisions of the weak classifiers are shared among the classifiers in order to recognize all the samples by a meta-layer classifier in

[*] Corresponding author. Tel.: +90 5353356295.
E-mail address: meteozay@gmail.com, mozay@vision.is.tohoku.ac.jp (M. Ozay).

http://dx.doi.org/10.1016/j.inffus.2015.09.004
1566-2535/© 2015 Elsevier B.V. All rights reserved.

the dataset. This property is assessed by introducing a new criterion, called *shareability* measure.

Related work and motivation

Among a wide range of Stacked Generalization methods, we suffice to review the ones which motivate us for the development of the suggested FSG architecture, where decisions of classifiers are fused by the vector concatenation or linear combination operation in an ensemble.

Analysis of the black art problem, and shareability of features for collaboration of classifiers:. Due to numerous nonlinear relations among the parameters and incompatibilities among the classifiers of SG, tracing the feature mappings from the input spaces of base-layer classifiers to a meta-layer output space becomes an intractable problem. In a heterogeneous SG, classifiers generate different types of information about the decisions, such as crisp, fuzzy or probabilistic class labels [7–9]. A detailed comparison of classification performances of SG methods is reported by Sigletos et al. [10] where the performances are shown to be inconsistent with the literature [11]. The contradictory results can be attributed to many non-linear relations among the parameters and classification performances of the SG methods, such as the dimension of feature spaces, and the number and diversity of base-layer classifiers. Stacking is used to identify and incorporate the dependencies between class labels and features in [12,13]. Sen and Erdogan analyze [3] various weighted and sparse linear combination methods which aggregate decisions of heterogeneous base-layer classifiers, such as decision trees and k-NN. Several SG algorithms are reported in the literature for data fusion, regression, classification and Natural Language Processing [14–22]. An experimental study of the analysis of diversity of linear classifiers in bagging and boosting is provided by Kuncheva et al. [23]. In this study, most of the intractable problems are avoided by designing a homogeneous Stacked Generalization method, called Fuzzy Stacked Generalization (FSG). Inspiring from these studies, we introduce a new criterion, called shareability, to measure the collaboration among the base-layer classifiers, and investigate the dependencies between the feature and decision spaces of base-layer and meta-layer classifiers.

Hierarchical distance learning, and minimization of the difference between the large sample error and N*-sample error:.* In the literature, distance learning methods have been employed for selection and weighted combination of samples and features by computing the weights associated with the samples, feature vectors and distributions [24–27]. The weights are used in weighted distance functions to transform the feature space of a classifier into a more discriminative space [28,29], and decrease the N-sample classification error of the classifiers [30]. A detailed literature review of prototype selection and distance learning methods for nearest neighbor classification was given in [26]. Unlike the aforementioned methods, we propose a hierarchical distance learning approach in order to minimize the difference between the large sample error and N-sample error. Therefore, our theoretical and experimental results are complementary with the results of Wang et al. [20] who analyzed the fusion of fuzzy decisions for the minimization of the Bayes risk.

2. *N*-sample and large sample errors for a single *k*-NN classifier

We start by defining the large sample and N-sample classification error of a k-NN classifier. Then, we minimize the expected square of the difference between the large sample and N-sample classification error to learn the distance function employed in a k-NN classifier.

Suppose that a training dataset $S = \{(s_i, y_i)\}_{i=1}^{N}$ of N samples is given, where $y_i \in \{\omega_c\}_{c=1}^{C}$ is the label of a sample s_i which is represented in a feature space F by a feature vector $\bar{x}_i \in \mathbb{R}^D$. Given a new test sample s'_i with $\bar{x}'_i \in F$, let $\eta_k(\bar{x}'_i) = \{\bar{x}_{l(n)}\}_{n=1}^{k}$ be a set of k-nearest neighbors of $\bar{x}'_i$ such that

$$d(\bar{x}'_i, \bar{x}_{l(1)}) \le d(\bar{x}'_i, \bar{x}_{l(2)}) \le \cdots \le d(\bar{x}'_i, \bar{x}_{l(k)}). \tag{1}$$

The k-nearest neighbor rule simply estimates the label of $\bar{x}'_i$ as

$$\hat{y}'_i = \arg\max_{\omega_c} \mathcal{N}(\eta_k(\bar{x}'_i), \omega_c), \tag{2}$$

where $\mathcal{N}(\eta_k(\bar{x}'_i), \omega_c)$ is the number of samples which belong to ω_c in $\eta_k(\bar{x}'_i)$.

The probability of error $\epsilon(\bar{x}_i, \bar{x}'_i) \triangleq P_N(error|\bar{x}_i, \bar{x}'_i)$ of the nearest neighbor rule ($k = 1$) is computed as

$$\epsilon(\bar{x}_i, \bar{x}'_i) = 1 - \sum_{c=1}^{C} P(\omega_c|\bar{x}_i) P(\omega_c|\bar{x}'_i), \tag{3}$$

where $P(\omega_c|\bar{x}_i)$ and $P(\omega_c|\bar{x}'_i)$ represent posterior probabilities for ω_c [31]. In the asymptotic of number of training samples, if $P(\omega_c|\bar{x}_i)$ is continuous at $\bar{x}'_i$, then large sample error $\epsilon(\bar{x}'_i) = \lim_{N\to\infty} P_N(error|\bar{x}'_i)$ is computed as

$$\epsilon(\bar{x}'_i) = 1 - \sum_{c=1}^{C} P^2(\omega_c|\bar{x}'_i). \tag{4}$$

Therefore, the difference between the N-sample error (3) and the large sample error (4) is computed as

$$\epsilon(\bar{x}_i, \bar{x}'_i) - \epsilon(\bar{x}'_i) = \sum_{c=1}^{C} (P(\omega_c|\bar{x}'_i))(P(\omega_c|\bar{x}'_i) - P(\omega_c|\bar{x}_i)). \tag{5}$$

Cover and Hart [32] remarked the elegant relationship between the Bayes probability of error of classification, i.e. the Bayes risk (e^*), and N-sample and large sample classification error of k-NN as follows:

$$e^* \le \epsilon(\bar{x}'_i) \le \epsilon(\bar{x}_i, \bar{x}'_i) \le 2e^*. \tag{6}$$

Note that, if k grows with N, such that $\lim_{N\to\infty} k \to \infty$ and $\lim_{N\to\infty} \frac{k}{N} \to 0$, then the classification error of k-NN converges to the Bayes probability of error (i.e. the Bayes risk) [32,33]. Therefore, a distance function which minimizes

$$E_N\Big\{\big(\epsilon(\bar{x}_i, \bar{x}'_i) - \epsilon(\bar{x}'_i)\big)^2\Big\}, \tag{7}$$

where the expectation is computed over the number of training samples N, enables us to get closer to the Bayes error (e^*). Short and Fukunaga [6] show that (7) can be minimized by either increasing N or designing a distance function $d(\bar{x}'_{i,j}, \cdot)$ which will be employed for the computation of the neighborhood system of the classifier. In a classification problem, an appropriate distance function which minimizes (7) is

$$d(\bar{x}'_i, \bar{x}_i) = \|\bar{P}(\omega|\bar{x}_i) - \bar{P}(\omega|\bar{x}'_i)\|_2^2, \tag{8}$$

where $\|\cdot\|_2^2$ is the squared ℓ_2 norm, and

$$\bar{P}(\omega|\bar{x}_i) = \big[P(\omega_1|\bar{x}_i), \ldots, P(\omega_c|\bar{x}_i), \ldots, P(\omega_C|\bar{x}_i)\big]. \tag{9}$$

The main goal of this paper is to design an ensemble learning architecture which minimizes the variance of the difference between the N-sample and large sample error of the base-layer classifiers. For this purpose, first (7) is extended for an ensemble of classifiers in the next section. Then, a hierarchical ensemble learning architecture, called Fuzzy Stacked Generalization, is proposed to minimize the variance of the error difference by employing a hierarchical distance learning approach in Section 4.