Pattern Recognition 47 (2014) 3451–3465

Contents lists available at ScienceDirect

Pattern Recognition

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

A compressed sensing approach for efficient ensemble learning

Lin Li [a,*], Rustam Stolkin [b], Licheng Jiao [a,*], Fang Liu [a], Shuang Wang [a]

[a] *Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an 710071, China*

[b] *School of Mechanical Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

A R T I C L E I N F O

Article history:
Received 4 June 2013
Received in revised form
5 March 2014
Accepted 16 April 2014
Available online 28 April 2014

Keywords:
Ensemble learning
Classification
Classifier ensemble
Sparse reconstruction
Compressed sensing
Roulette-wheel selection
Kappa-error

A B S T R A C T

This paper presents a method for improved ensemble learning, by treating the optimization of an ensemble of classifiers as a compressed sensing problem. Ensemble learning methods improve the performance of a learned predictor by integrating a weighted combination of multiple predictive models. Ideally, the number of models needed in the ensemble should be minimized, while optimizing the weights associated with each included model. We solve this problem by treating it as an example of the compressed sensing problem, in which a sparse solution must be reconstructed from an under-determined linear system. Compressed sensing techniques are then employed to find an ensemble which is both small and effective. An additional contribution of this paper, is to present a new performance evaluation method (a new pairwise diversity measurement) called the roulette-wheel kappa-error. This method takes into account the different weightings of the classifiers, and also reduces the total number of pairs of classifiers needed in the kappa-error diagram, by selecting pairs through a roulette-wheel selection method according to the weightings of the classifiers. This approach can greatly improve the clarity and informativeness of the kappa-error diagram, especially when the number of classifiers in the ensemble is large. We use 25 different public data sets to evaluate and compare the performance of compressed sensing ensembles using four different sparse reconstruction algorithms, combined with two different classifier learning algorithms and two different training data manipulation techniques. We also give the comparison experiments of our method against another five state-of-the-art pruning methods. These experiments show that our method produces comparable or better accuracy, while being significantly faster than the compared methods.

1. Introduction

The idea of ensemble learning methods is to build an improved predictive model by integrating the decisions of multiple classifiers. An overall decision is computed by taking a weighted vote of the decisions of each individual classifier within the ensemble. Given the same amount of information, an ensemble decision can often be better than the decision from any single classifier. Firstly, several different weak individual classifiers are generated, then they are weighted and a better classifier is obtained by combining their predictions.

Ensemble methods have already generated significant attention, and a number of ensemble algorithms has been proposed during the past two decades [1–5]. However, a difficulty with existing ensemble learning algorithms is that they often generate a large number of classifiers within each ensemble. Much more memory [6,7] is needed to store these large ensembles and much more computation time is also needed to predict the classification of each new unlabelled data point. These two extra costs may appear negligible when ensemble learning is applied to small data sets, but they may become serious with large scale data sets. It is not uncommon to see a large scale implementation of ensemble learning generating thousands of individual classifiers, e.g. [8].

However, both theoretical and empirical evidence [1,7,9–11] suggest that larger ensembles are not necessarily better, and indeed a smaller ensemble size can often achieve better performance than a larger ensemble. Therefore, it is useful to obtain an ensemble which both minimizes the number of individual classifiers, and also maximizes the classification accuracy of the overall ensemble. Many pruning methods have been proposed to prune unnecessary classifiers and to obtain the optimal sub-set of classifiers for ensemble. The pruning ensemble methods can be classified into two groups based on their optimization approaches [11,12]. The first group of methods includes greedy search methods, such as [13–16]. Although these algorithms have generated significant attention (because they can obtain comparable results

* Corresponding authors. Tel.: +86 29 8820 9786; fax: +86 29 8820 1023.
E-mail addresses: xdlinli86@163.com (L. Li), lchjiao@mail.xidian.edu.cn (L. Jiao).

http://dx.doi.org/10.1016/j.patcog.2014.04.015
0031-3203/

at much smaller computational costs), they cannot guarantee to obtain a globally optimal ensemble. The second group consists of global search methods, such as [7,6,9]. These algorithms can obtain highly accurate results, but they are computationally expensive [12,11]. Moreover, after pruning unnecessary classifiers, these methods typically use a simple majority voting rule for the remaining classifiers, in which all classifiers carry equal weight. In contrast, it would be useful if this small subset of classifiers could be weighted in some way according to their relative importance.

Motivated by the performance trade-offs inherent in the techniques described above, we propose a novel ensemble learning framework which explores the global optimal sub-set of classifiers for ensemble at low computation cost by posing the ensemble learning problem in terms of another problem known as compressed sensing [17–19]. Compressed sensing methods seek sparse solutions to under-determined linear systems, sparse solutions being those which contain large numbers of zero entities. We consider such a sparse solution as being the set of weightings of the individual classifiers of an ensemble classifier. Thus solving this compressed sensing problem, firstly generates a sparse weighting vector which produces accurate classification results while containing many zeros (i.e. an ensemble with a minimum number of classifiers to achieve good performance) and, secondly, provides appropriate weights for the remaining small number of classifiers according to their relative importance.

To help evaluate our method, an additional contribution of this paper is a new kind of method for measuring the pairwise diversity of the two classifiers, which we call a roulette-wheel kappa-error. The roulette-wheel not only considers both weightings of the two classifiers of each pair but also improves the clarity and meaningfulness of such pairwise diagrams by selecting pairs according to their significance, based on roulette-wheel selection. In contrast, conventional methods, e.g. [20–22], show all possible pairs of large ensembles on a single diagram with apparently equal significance, even when many of these pairs have insignificantly small weightings.

The reminder of this paper is organized as follows. Section 2 gives a brief introduction to compressed sensing (CS), and then presents the formulation and solution for the compressed sensing ensemble selection method. Section 3 describes the roulette-wheel kappa-error for evaluating such methods. Section 4 presents the results of empirically testing the method on a large number of public datasets. Section 5 provides concluding remarks and suggests directions for future work.

2. Compressed sensing ensemble

Consider a supervised ensemble learning problem, where $C_i(\boldsymbol{x})$ is the ith classifier in the ensemble $\{C_1(\boldsymbol{x}), C_2(\boldsymbol{x}), \ldots, C_N(\boldsymbol{x})\}$. $\boldsymbol{x}$ is the input vector from a data set $\boldsymbol{X}$ and y is the output corresponding to a given input $\boldsymbol{x}$. $C_i(\boldsymbol{x})$ is usually obtained by using a training dataset. A training dataset $\boldsymbol{X}_{tr}$ is usually denoted as $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_M, y_M)\}$ and $\boldsymbol{x}_i \in R^d$, whose elements are real-valued or discrete such as weight, color, age, height and so on. The y values are drawn from a discrete set of class labels $\{1, 2, \ldots, c\}$. Note that here we only handle classification with two-class labels and we set the labels as $\{1, -1\}$. A learning algorithm uses a training dataset to generate a classifier, which is an estimate of the unknown function $y = f(\boldsymbol{x})$. The classifier $C_i(\boldsymbol{x})$ is a hypothesis $f_i(\boldsymbol{x})$ about the true function f. It can predict the class, y, for a new input vector, $\boldsymbol{x}$, or a set of classes for a testing dataset, $\boldsymbol{X}_{ts}$. Let f_{ij} be the prediction of the ith learner $C_i(\boldsymbol{x})$ for the jth sampling of the training dataset $\boldsymbol{x}_j$, that is $f_{ij} = C_i(\boldsymbol{x}_j)$.

An ensemble of the individual classifiers consists of a method for combining the decisions of each classifier in some way such as assigning weightings w_i to each of the classifiers $C_i(\boldsymbol{x})$. The ensemble result for new input data $\boldsymbol{x}$ can be denoted as $y = \boldsymbol{w}^T \boldsymbol{F}(\boldsymbol{x})$, where $\boldsymbol{w}$ is the vector of the weightings w_i and $\boldsymbol{F}(\boldsymbol{x})$ is the vector of the hypotheses $f_i(\boldsymbol{x})$ generated by each learning algorithm. One of the key problems in ensemble learning is that of how to find an appropriate choice of weighting for each learner.

The main contribution of this paper is to show how a compressed sensing algorithm can be applied to ensemble learning, in order to find a sparse but effective set of weightings for an ensemble of the classifiers. Compressed sensing [17–19] is a signal processing method that takes advantage of the signal's sparseness in some domain and reconstructs the entire signal by solving the under-determined linear system. Compressed sensing has attracted considerable attention in applied mathematics, electrical engineering, statistics and computer science. Many applications have been found in computer vision, coding theory, signal processing, image processing and algorithms for efficient data processing.

More specifically, compressed sensing tries to find sparse solutions to under-determined linear systems. Consider a matrix $\boldsymbol{F} \in R^{M \times N}$ with $M < N$, and an under-determined linear system of equations defined by the following equation:

$$\boldsymbol{F}\boldsymbol{w} = \boldsymbol{y}. \tag{1}$$

A sparse solution $\boldsymbol{w}$ can be solved from the above under-determined system defined by Eq. (1). When noise is also considered, compressed sensing becomes the problem of finding a vector $\boldsymbol{w}$ according to

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 \quad \text{s.t.} \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{w}\|_2^2 \le \sigma, \tag{2}$$

where $\boldsymbol{y} \in R^M$ is the observed vector and $\boldsymbol{F}$ is the sensing matrix. $\boldsymbol{y}$ can therefore be regarded as being approximated by a linear combination of the column vectors in $\boldsymbol{F}$, that is

$$\boldsymbol{y} \approx (\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_N}) \times (w_1, w_2, \ldots, w_N)^T = \sum_{i=1}^{N} w_i \boldsymbol{f_i} \tag{3}$$

where the weighting factors w_i are stored in $\boldsymbol{w}$. $\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{w}\|_2^2$ is known as the measurement error.

The following sections explain in detail how to pose the ensemble learning problem in terms of a compressed sensing problem, and how to then solve such ensemble problems by making use of sparse reconstruction algorithms, adapted from the compressed sensing literature.

2.1. Problem formulation and solution technique

In ensemble learning, we try to obtain a subset of models that contains the minimum number of classifiers, but which also produces the maximum accuracy of all possible subsets of the ensemble. In comparison, two objectives are typically emphasized in compressed sensing problems: one is to minimize the sparsity of the solution and the other is to minimize the measurement error. This comparison, between the ensemble learning and compressed sensing problems, is summarized in Table 1. It is evident that minimizing the number of classifiers in the ensemble can be considered as minimizing the zero-norm of $\boldsymbol{w}$, if $\boldsymbol{w}$ denotes the weightings for each individual classifier. Note that $\|\boldsymbol{w}\|_0$ counts the non-zero elements in the vector $\boldsymbol{w}$. The second task for efficient ensemble learning is to maximize the accuracy of the

Table 1
Comparison between the ensemble classifier and compressed sensing problems.

Two tasks for efficient ensemble learning	Two objectives of compressed sensing
Minimize the number of classifiers	Minimize the zero-norm of $\boldsymbol{w}$
Maximize the accuracy of these classifiers	Minimize the measurement error