Pattern Recognition 47 (2014) 2558–2567

Contents lists available at ScienceDirect

Pattern Recognition

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

A multiple criteria active learning method for support vector regression

Begüm Demir *, Lorenzo Bruzzone

Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 14, I-38123 Trento, Italy

ARTICLE INFO

Article history:
Received 5 August 2013
Received in revised form
31 January 2014
Accepted 4 February 2014
Available online 13 February 2014

Keywords:
Regression
Parameters estimation
Active learning
Support vector regression

ABSTRACT

This paper presents a novel active learning method developed in the framework of ε-insensitive support vector regression (SVR) for the solution of regression problems with small size initial training data. The proposed active learning method selects iteratively the most informative as well as representative unlabeled samples to be included in the training set by jointly evaluating three criteria: (i) relevancy, (ii) diversity, and (iii) density of samples. All three criteria are implemented according to the SVR properties and are applied in two clustering-based consecutive steps. In the first step, a novel measure to select the most relevant samples that have high probability to be located either outside or on the boundary of the ε-tube of SVR is defined. To this end, initially a clustering method is applied to all unlabeled samples together with the training samples that are inside the ε-tube (those that are not support vectors, i.e., non-SVs); then the clusters with non-SVs are eliminated. The unlabeled samples in the remaining clusters are considered as the most relevant patterns. In the second step, a novel measure to select diverse samples among the relevant patterns from the high density regions in the feature space is defined to better model the SVR learning function. To this end, initially clusters with the highest density of samples are chosen to identify the highest density regions in the feature space. Then, the sample from each selected cluster that is associated with the portion of feature space having the highest density (i.e., the most representative of the underlying distribution of samples contained in the related cluster) is selected to be included in the training set. In this way diverse samples taken from high density regions are efficiently identified. Experimental results obtained on four different data sets show the robustness of the proposed technique particularly when a small-size initial training set are available.

1. Introduction

The automatic and accurate estimation of parameters from specific data, such as biomedical parameters (e.g., glucose concentration in diabetic patients from biomedical data) and geo/biophysical parameters (e.g., forest parameters from remote sensing data) represents an important research field in machine learning and pattern recognition. A possible and effective way to deal with the estimation process is to exploit regression methods, which estimate a functional relationship between a set of variables and corresponding target values (i.e., the measurement of the parameters of interest). Different machine learning methods have been used for addressing regression problems, e.g., neural networks [1–3], Gaussian Process (GP) regression [4], nearest neighborhood regression [5], and Support Vector Regression (SVR) [6–8]. These methods rely on the availability of samples for which reference measures are available (i.e., training samples) to be used in the learning phase of the regression algorithm. The amount and the quality of these samples are important to obtain accurate estimations. An insufficient number of training samples or the availability of biased samples may result in a decrease of the estimation performance in terms of accuracy and generalization ability. Nevertheless, gathering labeled samples is often expensive and complex as producing reference measures may require significant time.

An effective approach to reduce the labeling effort in regression systems is active learning (AL) [9,10]. AL attempts to iteratively select the most informative samples for collecting reference measures and optimize the training set T with a minimum number of high quality samples. At each iteration the most informative samples for the considered regression algorithm are chosen among a pool U of unlabeled samples. Then, the true target values are assigned to the selected samples by a human expert according to proper measurements, and these samples are added to the training set T. Finally, the regression algorithm is retrained with the additional labeled samples. The process is iterated until convergence that is reached when the desired estimation performance (i.e., a pre-defined lower bound for the estimation error) is met.

* Corresponding author.
E-mail addresses: demir@disi.unitn.it (B. Demir), lorenzo.bruzzone@ing.unitn.it (L. Bruzzone).

http://dx.doi.org/10.1016/j.patcog.2014.02.001
0031-3203

At convergence, the training set T consists of a minimum number of most informative samples for the considered regression algorithm.

AL can be considered as an interactive expert-guided regression approach and can be only applicable when parameters being estimated consist of stable variables. In other words, AL is impractical in the case of considering a regression problem with unstable parameters. This is due to the fact that unstable variables may cause wrong decisions on selected informative samples and thus provides a risk to query samples that are not informative anymore. Accordingly, this may result in an increased labeling cost without any improvement on the estimation accuracy.

Most of the previous studies in AL have been developed in the context of classification problems [11–13], whereas AL has been marginally considered in regression problems. A query by committee approach is presented in [14], which generates a committee of regressors and then selects the samples with the maximal disagreement among the regressors (those that have the highest variance on the different predictions) to be included in the training set. In [15], statistical AL methods are discussed in the context of multilayer perceptron models. An AL algorithm for kernel-based linear regression and classification is presented in [16]. This algorithm employs a minimum-entropy criterion derived using a Bayesian interpretation of ridge regression. In [17], an AL technique has been presented in the context of feedforward neural networks, mixtures of Gaussians and locally weighted regression. This method selects the unlabeled sample that reduces the generalization error by minimizing output variance when it is included into the training set.

The central part of any AL algorithm is the sample selection strategy, which should be capable of selecting informative samples while avoiding unnecessary and redundant labeling. The above-mentioned AL methods evaluate the informativeness of the unlabeled samples by either (i) their relevancy or (ii) their relevancy together with their diversity to each other. The relevancy of samples is modeled by either (1) their similarity to the current labeled training samples to limit the selection of similar samples, or (2) the maximal disagreement among different regressors. The aim behind the maximal disagreement is to select unlabeled samples that have the lowest confidence on their estimated target values among all unlabeled samples. This is because of that their inclusion in the training set is expected to be important for better modeling the regression learning function. The diversity criterion aims at reducing the redundancy among the selected relevant samples and its utilization is necessary in addition to the relevancy criterion to select a batch of samples. This is due to the fact that the use of only relevancy criterion is effective to select one sample at each iteration of AL, whereas it may result in poor estimation performances in the case of choosing a batch of samples due to possible redundancy among the selected samples. An important drawback of the AL methods presented in the literature in the context of regression problems is that they do not assess the representativeness of samples in terms of their density, i.e., in terms of their prior. However, unlabeled samples that fall into the high density regions of the feature space are highly important for regression problems particularly when a small number of initially labeled samples is available. This is due to the fact they are statistically very representative of the underlying sample distribution. Accordingly, the estimation results on them affect much more the overall accuracy of the estimation process than those obtained on samples within low density regions.

To overcome these problems, in this paper we propose a novel multiple criteria AL (MCAL) method that selects the most informative as well as representative unlabeled samples in the context of regression problems with a small size initial training set. The proposed MCAL technique is defined in the context of the ε-insensitive SVR. The main motivations for which we consider the SVR are its good properties, such as (i) good generalization capability; (ii) ability to handle high dimensional input spaces also when few training samples are available; and (iii) relatively limited computational load in the training phase [18–20]. It is worth noting that the most of the previous AL studies in the context of Support Vector Machine (SVM) are defined for classification problems. However, these techniques cannot be directly used for SVR problems. This is due to (i) the decision rule of SVM classification is in general different from that of SVR and (ii) classification and regression problems are intrinsically different. Thus, in this paper we focus our attention on extending the use of AL to the framework of SVR.

The proposed MCAL method is based on the evaluation of three criteria for the selection of samples to be labeled, namely relevancy, diversity and density. In order to assess the above-mentioned three criteria, the proposed MCAL method exploits a two step procedure. Both steps rely on a novel clustering-based procedure. The first step is devoted to select the most relevant samples according to the SVR properties. To this end, initially a clustering method is applied to the unlabeled samples together with the training samples that are not support vectors (that are located inside the ε-tube, i.e., non-SVs). Since the unlabeled samples contained in the same clusters with any non-SVs have high probability to lie inside the ε-tube, they can be assumed as irrelevant (since they have the highest confidence on their target value). Accordingly in the first step the clusters that have non-SVs inside are neglected, whereas all the unlabeled samples of the remaining clusters are taken as the most relevant samples. The second step is devoted to choose the most diverse samples among the relevant ones from high density regions in the feature space. To this end, initially the densities of clusters identified at the first step are estimated and the highest density clusters are selected. Then, from each chosen cluster only one sample, which is associated with the portion of feature space with the highest density is selected to be included in the training set. Due to this choice, on one hand the sample that is the most representative of the underlying distribution of samples contained in the related cluster is chosen. On the other hand, diverse samples are identified due to the selection of only one sample per cluster. Besides the specific procedure presented, the main novelty from a conceptual viewpoint in the proposed MCAL method are: (i) the idea to use clustering and non-SV for identifying relevant and diverse unlabeled samples; (ii) the exploitation of the prior term of the distributions based on the density of unlabeled samples in the feature space for driving the selection of samples to include in the training set.

In the experiments, the performance of the proposed AL method is demonstrated with respect to estimation of (i) housing values from the variables that model the properties of houses, (ii) tree parameters (i.e., stem volume and stem diameter) from a set of features extracted by Light Detection And Ranging (LiDAR) data acquired on a forest, and (iii) age of the abalones from the variables that model the physical characteristics of abalones.

The remaining part of this paper is organized as follows. Section 2 formulates the considered problem and Section 3 introduces the proposed AL method. Section 4 describes the considered data sets and the design of experiments, whereas Section 5 illustrates the experimental results. Finally, Section 6 draws the conclusion of this work.

2. Problem definition

Let us assume that a training set T made up of N pairs $(\mathbf{x}_i, y_i)_{i=1}^{N}$ is initially available, where $\mathbf{x}_i \in \Re^d$ are the training samples, $y_i \in \Re$ are the corresponding target values, and N is the number of training samples. In this paper, we propose an AL method that