ARTICLE IN PRESS

Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Soft estimation by hierarchical classification and regression

Shih-Wen Ke[a], Wei-Chao Lin[b,*], Chih-Fong Tsai[c], Ya-Han Hu[d]

[a] Department of Information and Computer Engineering, Chung Yuan Christian University, Taiwan
[b] Department of Computer Science and Information Engineering, Asia University, Taiwan
[c] Department of Information Management, National Central University, Taiwan
[d] Department of Information Management, National Chung Cheng University, Taiwan

ARTICLE INFO

Communicated by: Zhaohong Deng

Keywords:
Prediction
Data mining
Classification
Regression
Hierarchical estimation

ABSTRACT

Classification and numeric estimation are the two most common types of data mining. The goal of classification is to predict the discrete type of output values whereas estimation is aimed at finding the continuous type of output values. Predictive data mining is generally achieved by using only one specific statistical or machine learning technique to construct a prediction model. Related studies have shown that prediction performance by this kind of single flat model can be improved by the utilization of some hierarchical structures. Hierarchical estimation approaches, usually a combination of multiple estimation models, have been proposed for solving some specific domain problems. However, in the literature, there is no generic hierarchical approach for estimation and no hybrid based solution that combines classification and estimation techniques hierarchically. Therefore, we introduce a generic hierarchical architecture, namely hierarchical classification and regression (HCR), suitable for various estimation problems. Simply speaking, the first level of HCR involves pre-processing a given training set by classifying it into k classes, leading to k subsets. Three approaches are used to perform this task in this study: hard classification (HC); fuzzy c-means (FCM); and genetic algorithms (GA). Then, each training data containing its associated class label is used to train a support vector machine (SVM) classifier for classification. Next, for the second level of HCR, k regression (or estimation) models are trained based on their corresponding subsets for final prediction. The experiments based on 8 different UCI datasets show that most hierarchical prediction models developed with the HCR architecture significantly outperform three well-known single flat prediction models, i.e., linear regression (LR), multilayer perceptron (MLP) neural networks, and support vector regression (SVR) in terms of mean absolute percentage error (MAPE) and root mean squared error (RMSE) rates. In addition, it is found that using the GA-based data pre-processing approach to classify the training set into 4 subsets is the best threshold (i.e., k=4) and the 4-class SVM+MLP outperforms three baseline hierarchical regression models.

1. Introduction

One major ultimate goal of data mining is prediction. The process can be categorized for classification or for estimation based on the data type involved in the prediction. Classification is one of the most common research problems in data mining. This problem is usually approached by (supervised) classification techniques. The aim of classification is to allocate an (unknown) instance represented by specific features into one correct class from a finite set of classes. To achieve classification, a learning (or training) task is necessary. This involves the computation of a classifier or model, which is achieved by approximating the mapping between input-output training examples, thereby enabling the correct labeling of the training set at a particular level of accuracy. After the model is generated or trained, it can be used to classify unknown instances assigning one of the class labels learned in the training set [14,29].

Similar to classification learning, numeric estimation also involves training a model that is learned from a set of training examples including the output attribute. However, the output attribute is continuous, i.e., numeric rather than discrete. Therefore, the goal of numeric estimation is to 'estimate' the output value.

Studies of classification and estimation problems generally apply a single (flat) prediction model. More specifically, many recent studies have shown that a hierarchical structure outperforms a flat structure for solving various classification problems (e.g., [15,38,35]) and estimation problems (e.g., [1,20,27,37,41,42]). Such hierarchical approaches have been proposed for specific problem domains. In particular, for estimation, they are based purely on combining several

* Corresponding author.
E-mail address: viclin@asia.edu.tw (W.-C. Lin).

http://dx.doi.org/10.1016/j.neucom.2016.12.037
Received 10 September 2015; Received in revised form 1 November 2016; Accepted 12 December 2016
0925-2312/

estimation models in a hierarchical manner. A search of the literature showed no hybrid based hierarchical solutions that combine classification and estimation techniques for numeric estimation (c.f., Section 2.2).

Therefore, to remedy this lack, this study introduces a generic hierarchical architecture, namely hierarchical classification and regression (HCR), designed to overcome the limitation of the flat estimation models. In addition, the HCR architecture can be used to solve a variety of estimation domain problems. The first level of HCR aims at classifying a new unknown case into a specific class (e.g., class *i*). This case is then input into an estimator trained to predict only the continuous values that have been categorized to class *i* (over a training set). To determine which output value belongs to which class in a given training set, three data pre-processing approaches are used in this paper. The first approach is the simplest one and is based on 'hard classification'. It divides the sum of the maximum and minimum output values by *k* predefined classes. The second and third methods can be regarded as 'soft classification' approaches, in which the former and the latter are based on using the fuzzy c-means clustering algorithm and genetic algorithm, respectively, to categorize the training data into one of the *k* predefined classes (c.f., Section 3.2).

The proposed HCR architecture is based on the divide-and-conquer principle of solving complex problems by first solving subproblems (i.e., simpler tasks) that can be solved with a classifier on the first level and by an estimator on the second level of HCR. This concept is similar to ensemble learning [21] that is inspired by the nature of information processing in the brain which is modular. That is, individual functions can be subdivided into functionally different subprocess or subtasks without mutual interference [17]. Based on this characteristic, our experimental results demonstrate the outperformance of HCR by combining specific classifier and regression models over single flat models and related baselines (c.f. Section 4.2).

The rest of this paper is organized as follows. Section 2 offers an overview of three well-known estimation techniques, linear regression, neural networks, and support vector regression. In addition, related works on hierarchical estimation are described. Section 3 introduces the proposed HCR architecture and three different data pre-processing approaches used to classify the training data into specific classes. Section 4 presents the experimental results and some conclusions are provided in Section 5.

2. Literature review

Estimation models or estimators are needed to infer the value of unknown parameters in statistical models. Restated, models are needed to estimate parameter values based on measurement data. The estimator uses the measured data as input for parameter estimation [25]. Estimation methods such as supervised prediction involve learning from a set of training examples that includes output attributes that are numeric (i.e., continuous values) rather than discrete values.

Three well-known methods of numeric output estimation, namely linear regression, neural networks, and support vector regression are described in the following subsection. For other related regression techniques applied in different domain problems, please refer to Deng et al. [12], Deng et al. [10], Jiang et al. [23], and Luo et al. [26].

2.1. Estimation techniques

2.1.1. Linear regression

Linear regression a statistical method for modeling the relationship between dependent variables (response variables) and independent variables (explanatory variables) [13]. In particular, a linear regression model assumes that this relationship is linear.

The general formula for multiple regression models is

$$Y = \beta_0 + \sum_{j=1}^{n} \beta_j X_j + \varepsilon \tag{1}$$

where Y is a dependent variable; β_0 is a constant; β_j is a regression coefficient (j=1, 2, ..., n); and ε is an error term. Simple regression analyses use only one independent variable (X_j) while multiple regression analyses use two or more variables.

2.1.2. Neural networks

Neural networks (or artificial neural networks) were inspired by the biological neural networks and the central nervous system in the human brain. In general, a neural network consists of information-processing neurons (or units), which are connected together [19].

One widely developed model is the multilayer perceptron (MLP) neural network. It is composed of multiple layers of nodes (or neurons), including an input layer, hidden layer(s), and an output layer. The nodes in the input layer represent the feature values of an instance whereas the nodes in the output layer are used to distinguish between different classes (or estimations).

In addition, a MLP is usually trained by the backpropagation learning algorithm. It focuses on weights tuning between the connections of the hidden layer nodes in order to minimize the classification error over a given training dataset.

2.1.3. Support vector regression

Support vector machines (SVMs) are one of the most popular supervised learning techniques, and have been widely used in many pattern recognition problems [39]. Its idea is to map a given training dataset containing *d*-dimensional features into a higher dimensional feature space where an optimal separating hyperplane is constructed to effectively distinguish between two classes.

The SVM is applicable in both classification and regression problems. Support vector regression (SVR) [36] is a version of SVM in which a linear regression model is constructed in the new higher dimensional feature space.

2.2. Hierarchical estimation

An examination of the literature shows that several hierarchical estimation models have been developed to solve different domain problems. For example, Hellier et al. [20] proposed an efficient hierarchical optimization framework for medical image registration, which is both multiresolution and multigrid. Note that the registration problem can be regarded as a motion estimation problem. In this framework, an anatomical segmentation of the cortex is introduced for the adaptive partitioning of the volume on which the process of multigrid minimization is based. This allows limiting the estimation to the areas of interest, to accelerate the algorithm, and to refine the estimation in specified areas.

Memin and Perez [27] introduced an energy-based framework for the estimation and segmentation of the apparent motion in image sequences. Two different methods are presented in order to reach the goal of mixing local non-parametric smoothing and a more global parametric representation. The first one is based on a constrained minimization technique used with an energy-based dense motion estimation model. The second one deals with an energy-based model for the joint estimation segmentation of the apparent motion.

Strijbosch and Moors [37] presented a hierarchical estimation approach for finding the total forecast where individual items are broken down to produce the desired individual demand forecasts. Then, the total demand for a number of items and the fraction of this total taken by individual items are estimated. Multiplying these two quantities can provide the hierarchical estimate for each individual demand. This proposed approach was assessed using two management related problems.

On the other hand, You and Ryu [41] proposed a hierarchical