Decision Support Systems 52 (2012) 748–756

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

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

Using structure-based data transformation method to improve prediction accuracies for small data sets

Der-Chiang Li [*], Chih-Chieh Chang, Chiao-Wen Liu

Department of Industrial and Information Management, National Cheng Kung University, Taiwan

ARTICLE INFO

Article history:
Received 12 November 2010
Received in revised form 7 October 2011
Accepted 18 November 2011
Available online 1 December 2011

Keywords:
Manufacturing
Small data set
Cluster
DBSCAN
SVR

ABSTRACT

Small data set problems have been widely considered in many fields, where increasing the prediction ability is the most important goal. This study considers the data structure to identify new data points in a more precise manner, and is thus able to achieve improved prediction capability. The proposed method, named structure-based data transformation, consists of two steps. The first step is using the density-based spatial clustering of applications with noise (DBSCAN) algorithm to separate data sets into clusters, which generates the number of clusters dynamically. The second step is to build up the data transformation function, in which the new attributes are computed using fuzzy membership functions obtained by the corresponding membership grades in each cluster. Three real cases are selected to compare the proposed forecasting model with the linear regression (LR), backpropagation neural network (BPNN), and support vector machine for regression (SVR) methods. The result show that the structure-based data transformation method has better performance than when using the raw data with regard to the error improving rate, mean square error (MSE), and standard deviation (STD).

1. Introduction

Small data set problems have been widely considered in many fields, such as manufacturing systems for high cost products or special medical cases. Taking polarizers as an example, they are one of the key parts in Thin Film Transistor-Liquid Crystal Displays (TFT-LCD), and the production cost is very expensive. Thus, how to use the small data sets obtained in the pilot runs to build a management model to reduce overall manufacturing costs is a very important task. Similarly, spinocerebellar ataxia is a rare hereditary disease, with only few records from all over the world, and thus is also defined as a small data set problem. Theoretically, it is hard to make precise predictions using a small data set, and so the main goal of this study is how to decrease the error rate between predicted and real data.

Several approaches have been proposed to deal with small data set problems, one of which is virtual sample generation. The original idea was proposed by Niyogi et al. [27], who used prior knowledge obtained from a given small training set to create virtual samples to improve recognition performance. This work generated new views of a given 3-D object from any other direction through a mathematical transformation, with the samples thus generated called virtual ones. Later, Li et al. [16] used a functional virtual population to solve the scheduling problem in early flexible manufacturing systems, where they utilized a virtual sample generation technique to increase the amount of training data to improve the classification accuracy of a BPNN. Huang and Moraga [11] proposed a diffusion neural network (DNN) which, compared with other such networks, had more nodes in the inputs and layers, and was trained by derived patterns instead of the original ones. The DNN method's error rate, 48%, was found to be better than that of the conventional BPNN. Recently, Yang et al. [35] presented a Gaussian distribution virtual sample generation method, which showed good performance in both in small sample problems and imbalanced sample problems. Khan et al. [26] presented a novel model to deal with the face recognition classification problem, which was also small data set problem.

Using kernel methods, such as a support vector machine, is another effective approach to solve small data problems. A kernel $\kappa: X \times X \to \mathbb{R}$ is defined as the Mercer kernel viewed as a measure of the similarity between two paired data, where X is the input data set and $\mathbb{R}$ is a real number [5]. A typical kernel representation uses a feature map $\phi: \Omega \to F$ to transform data from the input space Ω into a feature space F, and then constructs linear algorithms in the feature space to solve nonlinear problems in the input space, as shown in Fig. 1. It is usually supposed that the transformed data in the feature space has a better linear estimation than that of the original data in the input space. The map ϕ is usually represented implicitly by a kernel function, such as the inner product denoted as $\kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}), \phi(\mathbf{x}')$ [4]. Hence, the main purpose of a kernel method is extending the data set into a higher dimensional space by using a functional kernel.

Although using a kernel function to extend data into a higher dimension feature space is effective for data analysis, it is difficult to find the best kernel function for a problem. Therefore, this research

[*] Corresponding author. Tel.: +886 6 2757575x53134.
E-mail address: lidc@mail.ncku.edu.tw (D.-C. Li).

0167-9236/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2011.11.021

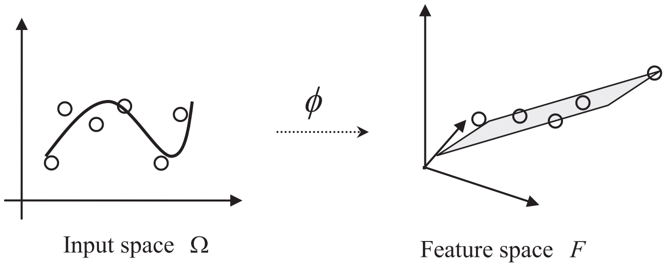

Fig. 1. The mapping ϕ from Ω to F.

follows the concept of extending the original data set into a higher dimensional feature space and develops a data transformation procedure as the kernel function. In fact, the proposed method, named structure-based data transformation, uses the data clustering technique to place similar data into groups based on data density to further obtain the geographic structure of the data set, where DBSCAN is employed as the data clustering tool to generate clusters dynamically and also to detect any noise. Based on these clusters, the second step of the proposed approach is to build up the attribute creation procedure to generate new attributes using the fuzzy membership function. Through the fuzzy membership function, one attribute could generate several cluster-possibility attributes using the clusters constructed in the first step. Finally, after mapping the data into a higher dimensional space, the data with the newly generated attributes and values will be input into the classifier as the training data set.

In the proposed method, the main difference from the kernel method is that we consider the data structure using DBSCAN clustering, and based on this the new attributes are constructed to increase the attribute information. It is thus a data structure oriented approach to extend the data dimensions. Also, different from the virtual sample generation methods, the proposed method generates more effective data attribute (for prediction accuracy) into the small data set. Although virtual sample generation can increase the data quantity, there is still a risk of data bias when oversupplying virtual samples. The proposed method can artificially increase the prediction accuracy by creating attributes.

Three real case studies are examined in this study to evaluate the proposed approach. The first case uses multi-layer ceramic capacitors (MLCC), which are a product made of ceramic powder. The second case is a concrete slump test, which is commonly used in the literature and drawn from the UCI database. The last case study is used to predict the quality of TFT-LCD. The error improving rate, MSE, STD, and *t*-test statistic are used to verify the effectiveness of the forecasting model, with the results compared to the LR, BPNN, and SVR methods. The results show that the proposed method is superior to other approaches with regard to its prediction capability.

The rest of this paper is organized as follows: Section 2 describes the modeling concept of the data transformation method. The structure-based transformation function is shown in Section 3. In Section 4, three real cases are demonstrated to explain the proposed method. Finally, the conclusions and suggestions for future studies are presented in Section 5.

2. Related studies

2.1. The data transformation methods

Mapping the original data set by a nonlinear function to deal with nonlinear prediction problems is widely used in the past research, as it is a useful preprocessing technique to map data to a higher dimensional space [23]. For example, the residuals are structureless design of experiment if the analysis model is correct and the assumptions are satisfied [25]. However, the residuals in some of the cases are not structureless, and an outward-opening funnel or megaphone will occur when these get larger as the number of observations increases. To deal with this problem, the population of observations is generally transformed, using method such as Box–Cox transformation, logarithmic transformation, and arcsin transformation.

Artificial neural networks (ANN) are allowed to transform data to solve nonlinear problems. They use activation functions to transform linear combinations of the weights and input data sets [32]. A representative model is multilayer perceptrons, which uses multiple layers with several activation functions to transform the data into a feature space that can produce a forecast model with less prediction error. Fig. 2 shows a simple artificial neuron model, where $\mathbf{x}=(x_1,x_2,\cdots,x_M)$ is the input, y is the output, b is the bias, $f(\cdot)$ is an activation function, and $\mathbf{w}=(w_1,w_2,\cdots,w_M)$ is the weights. The mathematical representation of the neuron is $y=f(\sum_{i=1}^{M}w_ix_i+b)$. It is easy to see that when the activation function $f(\cdot)$ is nonlinear, the input data will be transformed to another feature space.

Another transformation is the kernel method, and support vector machines (SVM) are commonly used to achieve this. In an SVM, in order to deal with a nonlinear problem, the kernel usually transforms the data set into a higher dimension feature space, and this is then used to find an appropriate hyperplane for the problem [32]. For instance, given two samples $\mathbf{x}=(x_1,x_2)$ and $\mathbf{z}=(z_1,z_2)$, which are the two-dimensional input space data, when a polynomial kernel is used to expand the data the transformed feature space is obtained as a three-dimensional space, as shown below:

$$\begin{aligned}\kappa(\mathbf{x},\mathbf{z}) &= \left(\mathbf{x}^T\mathbf{z}\right)^2 = (x_1z_1+x_2z_2)^2 = x_1^2z_1^2+2x_1z_1x_2z_2+x_2^2z_2^2\\ &= x_1^2z_1^2+2x_1z_1x_2z_2+x_2^2z_2^2 = \left(x_1^2,\sqrt{2}x_1x_2,x_2^2\right)\left(z_1^2,\sqrt{2}z_1z_2,z_2^2\right)^T.\end{aligned} \tag{1}$$

2.2. Support vector machines for regression

SVM have been widely used in many fields, such as to find tighter error bounds on performance of classification [9], and to predict human wine taste preferences [7]. Unlike traditional methods, SVM minimize the upper bound of the generalization error by maximizing the margin between the separating hyperplane and the data [2]. In its original form, SVM learning leads to a quadratic programming problem with a convex constrained optimization property, and thus there is a unique solution to it. Given a training set of N samples, $(\mathbf{x}_1,t_1),(\mathbf{x}_2,t_2),\cdots,(\mathbf{x}_N,t_N)$, where $\mathbf{x}_i\in\mathbb{R}^p$ is the input vector corresponding to the *i*th sample labeled by $t_i\in\{-1,+1\}$ depending on its class. The SVM problem can be formulated as a quadratic programming optimization problem that will find the weight parameter $\mathbf{w}$ and the bias parameter b that maximize the margin while ensuring that the training samples are

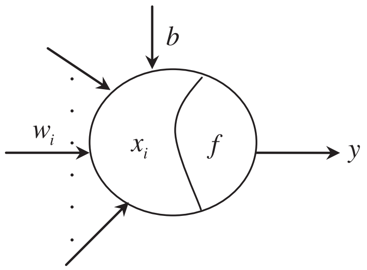

Fig. 2. The artificial neuron model [14].