Neurocomputing 143 (2014) 222–230

Contents lists available at ScienceDirect

Neurocomputing

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

A genetic algorithm-based virtual sample generation technique to improve small data set learning

Der-Chiang Li *, I-Hsiang Wen

Department of Industrial and Information Management, National Cheng Kung University, 1 University Road, Tainan 70101, Taiwan

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form
8 November 2013
Accepted 7 June 2014
Communicated by: A. Abraham
Available online 17 June 2014

Keywords:
Small data set
Genetic algorithm-based virtual sample generation (GABVSG)
Feasibility-based programming (FBP) model

ABSTRACT

While back-propagation neural networks (BPNN) are effective learning tools for building non-linear models, they are often unstable when using small-data-sets. Therefore, in order to solve this problem, we construct artificial samples, called virtual samples, to improve the learning robustness. This research develops a novel method of virtual sample generation (VSG), named genetic algorithm-based virtual sample generation (GABVSG), which considers the integrated effects and constraints of data attributes. We first determine the acceptable range by using mega-trend diffusion (MTD) functions, and construct the feasibility-based programming (FBP) model with BPNN. A genetic algorithm (GA) is then applied to accelerate the generation of the most-feasible virtual samples. Finally, we use two real cases to verify the performance of the proposed method by comparing the results with two well-known forecasting models, BPNN, support vector machine for regression (SVR) and one newly published approach MTD method [1]. The experimental results indicate that the performance of the GABVSG method is superior to that of using original training data without artificial samples. Consequently, the proposed method can improve learning performance significantly when working with small samples.

1. Introduction

Small data set problems are considered one of the important issues in early prediction. Such problems arise in some medical records, such as spinocerebellar ataxia, which is a rare hereditary disease with very few records around the world. Small data set problems also appear in the manufacturing industry, with high cost products or some pilot run testing samples. For example, polarizers are one of the key parts of Thin Film Transistor-Liquid Crystal Displays (TFT-LCD), which have a very short testing period of pilot runs and a very high manufacturing cost. Therefore, how to improve the forecasting accuracy when working with early obtained small data sets is a key issue in manufacturing management.

Several approaches have been proposed by researchers to deal with small data set problems. The virtual sample generation approach was originally proposed by Niyogi et al. [2], they used the prior knowledge obtained from a given small training set to create virtual samples to improve recognition performance. Their work generated new views of a given 3D-object from any other direction through a mathematical transformation, and with the newly generated samples being called virtual samples. Using virtual attributes is another effective way to deal with small data set problems, which extends the data into a high dimensional space to achieve better classification performance. Li and Liu [3] proposed a new method of extending attributes using the mega-trend diffusion (MTD) function, and with the result showing good performance when dealing with small data sets in classification problems. Later, Li et al. [4] improved the predictive ability of this approach for the case of TFT-LCDs through data clustering, and thus discovering the data structure and extending it into higher dimensions. In addition, Li et al. [5] proposed a modified grey-based system to deal with small data set problems while Yang and Kecman [6] presented an adaptive local hyperplane algorithm, and with the results showing that this method outperformed other approaches on two cancer data sets. Liao [7] proposed a feature selection method as a way to deal with bladder cancer data sets with small sample sizes.

A BPNN is an effective non-linear forecasting model, but is not robust when used to construct small-data-set models. Face to this problem, we try to use a virtual sample (VSG) method to stabilize the constructed BPNN small-data-set models. The purpose of the VSG is to fill the information gaps that exist in sparse data in order to improve the forecasting performance. This study uniquely develops a new VSG method, which is named genetic algorithm-based virtual sample generation (GABVSG). The main difference

* Corresponding author. Tel.: +886-6-2757575x53134; fax: +886-6-2341171.
E-mail address: lidc@mail.ncku.edu.tw (D.-C. Li).

http://dx.doi.org/10.1016/j.neucom.2014.06.004
0925-2312/

between the proposed method and the previous approaches of VSG is that the new method explores the integrated effects of attributes, while earlier methods usually deal with the attributes independently. This approach is constructed in three steps. The first is randomly selecting samples with sizes of 5, 10, 15, 20, and 25 as training data sets to construct the feasibility-based programming (FBP) model based on a BPNN, and to determine the acceptable range in each attribute by using MTD functions. The second step is then using a genetic algorithm (GA) to find a number of the most-feasible virtual samples. In the last step, we calculate the average errors by computing the difference between the predicted values, $\hat{Y}$, and the real ones, Y. This study selects a BPNN and a support vector machine for regression (SVR) as the prediction models and compares the results with the newly published approach, MTD method [1]. The comparison results show that the performance of the GABVSG method is superior to that of using original training data without any virtual samples and previous approach.

Two real cases are examined in this study to evaluate the effectiveness of the proposed approach. First, we use a multi-layer ceramic capacitor (MLCC) case [8], which is a product made of ceramic powder, with the data obtained from. Second, we use bladder cancer case [9], with the data obtained from a medical research center. The average errors and statistical t-test are used to verify the significance between the predicted and the real values. The results show that the proposed method is superior to other approaches that do not use virtual samples and is also better than MTD method.

The rest of this paper is organized as follows: Section 2 reviews some of the literature on small data sets approaches, virtual sample generation, genetic algorithms, and forecasting models. The genetic algorithm-based virtual sample generation technique is then presented in Section 3. In Section 4, two real cases are used to demonstrate for the proposed method. Finally, the conclusions of this work and suggestions for future studies are presented in Section 5.

2. Literature reviews

In this section, we will first review the literature on virtual sample generation. This is followed by a review of some works on genetic algorithms, and finally some of the studies that have considered the BPNN forecasting model.

2.1. Virtual sample generation

Virtual sample generation is a preprocessing method to enhance the predictive performance for small data set problems. The original idea of virtual data generation was proposed by Niyogi et al. [2], who used the prior knowledge obtained from a given small training set to create virtual samples to improve the recognition performance. Similarly, Li et al. [10] proposed a functional virtual population (FVP) to generate more virtual samples and solve scheduling problems in early flexible manufacturing systems. Specifically, they utilized VSG to increase the amount of training data to improve the classification accuracy of a BPNN. Huang and Moraga [11] then proposed a diffusion neural network (DNN) which was trained by derived patterns with more nodes in the inputs and layers, instead of the original ones. The DNN method's error rate of 48% was shown to be better than that of the conventional BPNN. Li and Lin [12] used an intervalization process to improve the kernel density estimation, and virtual sample generation to produce extra information to expedite the learning. Their results showed that this approach had good performance when using a BPNN as the learning tool with virtual samples. Later, Li et al. [1] proposed the mega-trend-diffusion (MTD) function to deal with small data set problem for scheduling strategies in early flexible manufacturing systems (FMS) by generating virtual samples. From a statistical viewpoint, the assumption of a normal distribution is a necessary condition before data analysis. However, it is often difficult to show that a data set follows a normal distribution when the data size is small. Consequently, Li et al. [13] used the membership function in fuzzy set theory to calculate the possibility values of virtual samples instead of the probability ones in statistics, to avoid the normal distribution assumption. Virtual samples have also been successfully applied for use with medical data, for which Li et al. [14] proposed the method Group Virtual Sample Generation (GVSG) method, which is a non-linear virtual sample generation algorithm that can significantly improve learning accuracy in the early stage of DNA microarray data. Lin and Li [15] proposed an extended algorithm named the Generalized-Trend-Diffusion (GTD) method, based on fuzzy theories, which includes a unique backward tracking process for exploring predictive information through the strategy of shadow data generation. In this approach, the extra information that is produced can both accelerate the learning task and dynamically correct the knowledge that is extracted from the shadow data. Recently, Yang et al. [16] presented a Gaussian distribution virtual sample generation method, which showed good performance in both small sample problems and imbalanced sample ones. Chao et al. [9] presented a new approach to improve the accuracy of predicting the outcomes of radiotherapy using small data sets. Based on these previous studies, many virtual sample generation approaches have shown good ability in improving forecasting and classification performance. However, none of them consider the relations that exists in the data structure. Accordingly, in this work we present a new virtual sample approach that also considers the relationship among each the attributes.

2.2. Forecasting models

In recent years, many studies have presented new forecasting. Li and Liu [17] proposed a neural network weight determination model by applying the concept of cognitive theory to develop a weighted learning algorithm. They showed that the resulting mean absolute percentage errors were less than those obtained with the BPNN, linear regression, x-bar, and Li and Yeh [18] model. Carrizosa et al. [19] proposed an SVM-based method that automatically detects the important predictor variables, and it showed good classification ability. Li et al. [20] proposed a yield forecasting model based on past manufacturing experience, and compared it with methods using regression, BPNN, RBFNN, and SVR. Later, Li et al. [21] demonstrated that SVR showed good ability to improve the non-linear quality in manufacturing TFT-LCDs.

This study will compare the proposed approach with models based on SVR and BPNN [22]. SVR, proposed in Vapnik [23], is a promising pattern recognition technique. Unlike traditional methods, which minimize the training error, SVR aims at minimizing an upper bound of the generalization error by maximizing the margin between the separating hyperplane and the data, as in Amari and Wu [24]. In its original form, SVR learning leads to a quadratic programming problem which is a convex constrained optimization one and thus has a unique solution. There are many studies that discuss the performance of SVR for different kernels [25,26]. Ali and Smith-Miles [25] found that the radial-basis function (RBF) kernels have superior performance, and thus we use these in this paper. BPNN have been widely discussed since the first being proposed in 1986 [27]. A back-propagation algorithm is also called a generalized delta rule, and from the perspective of