Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 36 (2009) 844–851

Expert Systems with Applications

www.elsevier.com/locate/eswa

A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere

Der-Chiang Li *, Yao-Hwei Fang

Department of Industrial and Information Management, National Cheng Kung University, 1, University Road, Tain 701, Taiwan

Abstract

In manufacturing systems, only a small training dataset can be obtained in the early stages. A small training dataset usually leads to low learning accuracy with regard to classification of machine learning, and the knowledge derived is often fragile, and this is called the small sample problem. This research mainly aims at overcoming this problem using a special nonlinear classification technique to generate virtual samples to enlarge the training dataset for learning improvement. This research proposes a new sample generation method, named non-linear virtual sample generation (NVSG), which combines a unique group discovery technique and a virtual sample generation method using parametric equations of hypersphere. By applying a back-propagation neural network (BPN) as the learning tool, the computational experiments obtained from the simulated dataset and the real dataset quoted from the Iris Plant Database show that the learning accuracy can be significantly improved using NVSG method for very small training datasets.

Keywords: Classification; Parametric equation; Small sample; Virtual sample

1. Introduction

When provided with plenty of data, data mining techniques are widely and successfully used to extract knowledge for management proposing. However, in the early stages of a production system, only limited data can be obtained, so that the knowledge derived is naturally not stable enough to effectively make predictions. This is called the small sample problem (Raudys, 2006), and thus data analysis within a small sample is always difficult.

Similar cases occur in the analysis of dynamic manufacturing environments such as flexible manufacturing systems (FMS) (Li, Chen, & Lin, 2003, 2006; Li & Lin, 2006), since a production system could vary with time and enlarge the data ranges in the later stages. A similar problem is found with DNA microarray data since the number of features (genes) greatly exceed the number of instances (tissue samples). This characteristic is called the small sample size (SSS) problem (Fukunaga, 1990), and the microarray-based classification problem was considered a SSS problem (Dougherty, 2001; Wong & Hsu, 2008).

In the literature there are many methods to solve the small sample problem, such as virtual sample generation (VSG). Niyogi, Girosi, and Poggio (1998) used prior knowledge obtained from a given small training data set to create virtual samples to improve recognition ability in the field of pattern recognition (Niyogi et al., 1998). The method was, given a 3D view of an object, to create new images from any other angles through mathematical transformations. The new images generated were called virtual samples. Via applying these virtual samples, a learning machine could improve its accuracy.

Li et al. proposed another VSG method, called the functional virtual population (FVP) approach to learn scheduling knowledge in dynamic manufacturing environments (Li et al., 2003). They used the small data set collected to form a functional virtual population and drew more samples out

[*] Corresponding author. Tel.: +886 6 2757575x50501; fax: +886 6 2374252.

E-mail addresses: lidc@mail.ncku.edu.tw (D.-C. Li), yaohwei_fang@hotmail.com (Y.-H. Fang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.10.029

of this virtual pool. From their experimental studies, the information of virtual samples did raise the learning performance of neural networks. The FVP algorithm systematically expanded the domains of the system attributes and generated a number of virtual samples. Using these virtual samples, a new level of scheduling knowledge was constructed.

Although FVP could solve the small sample problem, faced with the nonlinear classification of the small sample problem FVP could not improve the learning ability. Later, Li and Lin proposed the intervalization kernel density estimation (IKDE) method (Li & Lin, 2006), which included the intervalization process, kernel density estimation and virtual sample generation to improve the small training data learning. The intervalization process modified the kernel density estimation to solve the nonlinear classification small sample problem. Using a BPN as the tool, virtual sample generation produced extra information for expediting the learning stability. With this useful extra information, their method had the ability to improve the scheduling knowledge for a system in the early stages.

In order to upgrade the density estimation based method of IKDE, this research developed a new method, named non-linear virtual sample generation (NVSG), by hybridizing a group discovery technique and a special kind of virtual sample generation using parametric equations of hypersphere to solve nonlinear small sample classification problems.

The detailed description of the method is addressed in Section 3; the experiments using a simulated dataset and real dataset from the Iris Plants Database are shown in Section 4; the conclusions are presented in Section 5.

2. Other related studies

In another literature, an SSS problem using the traditional linear discriminant analysis (LDA) classification tool was explored (Fukunaga, 1990). The traditional LDA is one of the most popular classification tools using the projection technique. It finds the set of the most discriminant projection vectors which can map high-dimensional samples onto a low-dimensional space. Using the set of projection vectors determined by LDA as the projection axes, all projected samples will form the maximum between-class scatter and the minimum within-class scatter simultaneously in the projective feature space, but with a major drawback is that the within-class scatter matrix becomes singular and can hardly find a solution when encountering an SSS problem. Chen et al. proposed a new LDA for a face recognition system (Chen, Mark Liao, Ko, Lin, & Yu, 2000). The new LDA process (null space based LDA) applies a theory from linear algebra to the calculation of the within-class scatter matrix in the null space that maximizes the between-class scatter matrix, and thus significantly improves the performance of a face recognition system.

In the small sample problem, Huang and Moraga added the concept of fuzzy theory and presented a diffusion-neural-network (DNN) that combined a conventional neural network and the information diffusion technique to improve learning (Huang & Moraga, 2004). The information diffusion approach uses fuzzy theories to derive new samples to solve the problem of data incompleteness. Though the DNN shows better learning than that of BPN, the research does not offer any determination of diffusion function and coefficients. Li et al. employed the data fuzzification concept to systematically expand a small amount of training data to improve learning accuracy (Li, Wu, Tsai, & Chang, 2006). In their research, mega-fuzzification, data trend estimation, and application of adaptive-network-based fuzzy inference systems (ANFIS) were used to obtain scheduling knowledge in flexible manufacturing system (FMS) scheduling and non-sufficient data environments.

3. The proposed method

With regard to a multi-class classification problem with small sample, the proposed model develops a group discovery technique and a virtual sample generation method, named non-linear virtual sample generation (NVSG). Principally, the group discovery technique determines groups in each class, and the virtual sample generation enlarges the training data set in each determined group of every class.

The purpose of grouping data is to precisely specify the related configuration of classified data. In other words, it is necessary to be certain about the data structure before generating virtual samples. For example, the dataset is composed of two groups of positive data in a non-linearly classified structure (see Fig. 1), and the dataset is composed of one negative group and two positive groups of data in a linearly classified structure (Fig. 2). Faced with the former structure, it is not reasonable to combine two positive groups to generate virtual samples.

3.1. Group discovery technique

At the very beginning of the algorithm, we search all training samples and compute the minimum distance

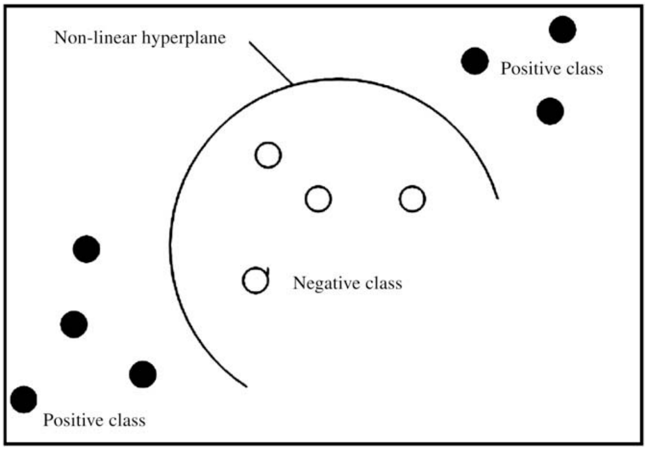

Fig. 1. Non-linear classification problem with two positive groups.