Neurocomputing 286 (2018) 75–87

Contents lists available at ScienceDirect

Neurocomputing

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

An attribute extending method to improve learning performance for small datasets

Liang-Sian Lin[a], Der-Chiang Li[b,*], Hung-Yu Chen[b], Yu-Chun Chiang[b]

[a] *Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu 31040, Taiwan, ROC*
[b] *Department of Industrial and Information Management, National Cheng Kung University, University Road, Tainan 70101, Taiwan, ROC*

ARTICLE INFO

Article history:
Received 24 October 2016
Revised 29 December 2017
Accepted 30 January 2018
Available online 2 February 2018

Communicated by Dr. Nianyin Zeng

Keyword:
Small dataset
High dimension
Sample extending attribute
Predictive model

ABSTRACT

A small dataset often makes it difficult to build a reliable learning model, and thus some researchers have proposed virtual sample generation (VSG) methods to add artificial samples into small datasets to extend the data size. However, for some datasets the assumption of the distribution of data in the VSG methods may be vague, and when data only has a few attributes, such approaches may not work effectively. Other researchers thus proposed attribute extension methods to generate attributes to convert data into a higher dimensional space. Unfortunately, the resulting dataset may become a sparse dataset with many null or zero values in extended attributes, and then a large quantity of such attributes will reduce the representativeness of instances for the learning model. Therefore, based on fuzzy theories, this paper proposes a novel sample attribute extending (SEA) method to extend a suitable quantity of attributes to improve small dataset learning. In order to verify the validity of the SEA method, using SVR and BPNN, this paper adopts two real cases and two public datasets to conduct the learning of the predictive model, and uses the paired *t*-test to statistically examine the significance of improvement. The experimental results show that the proposed SEA method can effectively improve the learning accuracy of small datasets.

1. Introduction

In the early phases of manufacturing, the sample size used to build a learning model is usually small due to today's shorter lead times and cost considerations. As such, the small sample size often results in a poor learning model, and may lead managers to make suboptimal decisions.

The reasons for the poor learning ability with a small dataset is that the data has a loose structure, with many gaps, and so a lot of potentially useful learning information is missing. In order to find more hidden information, some scholars have proposed virtual sample generation (VSG) methods which are data preprocessing techniques to systematically produce virtual samples to fill in the data gaps, as seen in Hung and Chan [1], Li and Wen [2], Xu et al. [3], Gao et al. [4], Tang et al. [5], Sezer et al. [6], Berrones et al. [7], and Krawczyk et al. [8]. A key assumption in VSG methods is that the data follows a random distribution. The assumed distribution is then used to evaluate the range of data values in order to generate virtual samples within this. For example, Huang and Moraga [9] proposed a diffusion-neural-network to generate virtual samples, in which a dataset is assumed to follow a fuzzy normal distribution, and then added the virtual samples into the original dataset to carry out learning. Li et al. [10] used the mega-trend-diffusion (MTD) technique to construct a linear triangular membership function (MF) to estimate the data range and generate a uniformly distributed set of virtual samples within this. For nonlinear data, Yang et al. [11] shaped the data into a Gaussian distribution and used it to generate normally distributed virtual samples. With regard to data that has an abnormal distribution, Li and Lin [12] proposed the maximal p-value method to generate abnormally distributed virtual samples, and the data was assumed to follow a Weibull distribution. These studies showed that the VSG method can offer significant improvements for small sample learning. However, because the data distribution is usually arbitrary, this may produce misleading results depending on the actual characteristics of the real data. Moreover, when the data only has a small number of attributes, VSG methods may not be able to effectively improve the learning performance.

In order to overcome this situation, some scholars recommend using attribute extension methods to increase the dimensionality of the data, and so enhance the learning ability for small datasets. For example, Li and Liu [13] proposed the extending attribute information (EAI) method to carry out attribute extension according to overlapping areas of the MTD function between classes. Unfortunately, their method constructed a large quantity of attributes,

[*] Corresponding author.
E-mail address: lidc@mail.ncku.edu.tw (D.-C. Li).

https://doi.org/10.1016/j.neucom.2018.01.071
0925-2312/© 2018 Elsevier B.V. All rights reserved.

thus making the original dataset fall into a high dimensional space, wherein the data may encounter the Hughes phenomenon [14] that decreases the resulting learning abilities. Furthermore, datasets changed by the EAI method have many zero values and also become sparse datasets, meaning that the distance used to distinguish the different data points is ineffective and meaningless. Ramezani et al. [15] proposed the remove redundancy method to change a sparse dataset into a sub-dataset to reduce data dimensionality. However, with this method there are some key attribute values that are removed, and so this approach reduces the learning effects for some datasets. In order to keep these values, some researchers thus developed methods based on the principle component analysis (PCA), which is a popular technique to transform a high dimensional dataset into a low dimensional one without removing any attribute values, as seen in Fan et al. [16], Li et al. [17], and Hu et al. [18]. When there is sufficient data, these algorithms can obtain great learning performance with regard to both classification and prediction. Unfortunately, for small datasets, the use of PCA may encounter the problem of over-fitting.

As mentioned above, when the number of extended attributes in data is large, such high dimensional data may decrease its representativeness, and the data is not able to further improve the learning accuracy of models. Although the reduction of data dimensionality can deal with this problem, it may cause the problems of over-fitting when it comes to learning with small data. For these reasons, this paper proposes a new method to generate integrated attributes according to the relationships among attributes in data. The method aims to address the issue of predictive learning using small data with numerical attributes, with the related mathematical formulas only suitable for calculations of numerical attributes.

This paper proposes the sample extending attribute (SEA) method to extend a suitable quantity of attributes to improve the learning performance of small datasets and prevent the data from becoming sparse, in which the extended attributes integrate all the membership function values based on the antecedents of fuzzy rules. With regard to the overlapping of the membership functions, we use the fuzzy c-means (FCM) approach to subordinate data into clusters, where the clustering method gives different levels of weights to each data in each cluster. In addition, for the data range estimation, a non-parametric method is applied to shape the data distribution in each cluster, using the method proposed by Li et al. [19], where the box-and-whisker plots method is applied to estimate the corresponding value range of the data to avoid overestimation. Moreover, this paper applies the α-cut technique to integrate all the data weights to carry out the data partitioning, and the operating principle is that when the data weights are greater than the set α values, then this data is classified into the corresponding data clusters. According to the setting of different α values, each cluster's dataset would have a different distribution density (low or high), and then we deploy different clustering conditions to examine their influence on the learning outcomes using small sample sizes. In addition, this paper uses a fuzzy rules operator to integrate the MF values, and the resulting values are then set as new extended attributes.

Four datasets are adopted to verify the validity of the proposed SEA method. This paper then uses paired t-tests to examine whether the perfomance of the SEA method has significant differences among the RAW (only using original dataset) and EAI methods with regard to the root mean square error (RMSE). Support vector regression (SVR) and a back propagation neural network (BPNN) were used to build the predictive learning model.

The rest of this paper is organized as follows. Section 2 reviews the EAI method, MTD technique, and SVR model. Section 3 introduces the proposed SEA method, including the fuzzy rule antecedents, attribute extension and data partitioning approaches. Section 4 introduces two manufacturing cases and two public datasets and explains this paper's experimental procedures and results. Section 5 then presents the conclusion of this work.

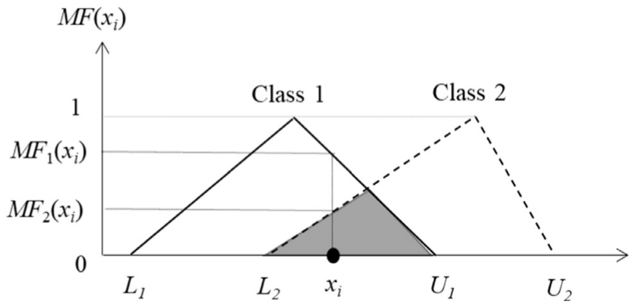

Fig. 1. The overlapping area in gray with two classes.

2. Related studies

In most past studies, the number of samples are increased to improve learning when there is a small sample size. Therefore, by focusing on a small sample size to carry out the attribute extension, this section will introduce the EAI method for attribute extension and the MTD technique for small data range estimation.

2.1. The EAI method

Li and Liu [13] proposed the EAI method to build attributes by using the fuzzy-based transformation function and attribute construction for the lower and higher overlapping areas of MTD function, respectively. When the overlapping area of the MTD function between classes is smaller (larger), the extended attribute is more effective (ineffective) with regard to improving the classification accuracies, as seen in Fig. 1.

In order to deal with the overlapping areas, and based on the MTD function, the data x with k-class is transformed by the fuzzy-based transformation function from one dimension to $k+1$ dimensions. In addition, the attribute construction uses the following non-linear operations, $\{X_i \times X_j,\ X_i/X_j,\ X_j/X_i\}$, for M pair of different attributes X_i and X_j, where the number of extended attributes is $C_2^M + P_2^M$, in which C is a combination operation and P is a permutation operation. After the original data is changed by the EAI method, the number of extended attributes will be $k + C_2^M + P_2^M$.

2.2. The MTD technique

Li et al. [10] proposed the MTD technique to generate virtual samples within an estimated data range to increase the sample size. In the MTD method, the estimation of the data range is [L, U], where L and U are calculated as follows:

$$L = \begin{cases} u_{set} - Skew_L \times \sqrt{-2 \times \frac{{s_x}^2}{N_L} \times \ln(\varphi(L))}, & N_L \neq 0 \\ \frac{\min}{5}, & N_L = 0 \end{cases} \tag{1}$$

$$U = \begin{cases} u_{set} + Skew_U \times \sqrt{-2 \times \frac{{s_x}^2}{N_U} \times \ln(\varphi(U))}, & N_U \neq 0 \\ max \times 5, & N_U = 0 \end{cases} \tag{2}$$

where u_{set} is $(\max + \min)/2$, max and min are the maximum and minimum values in a dataset, N_L and N_U are the number of values greater or smaller than u_{set}, s_x^2 is the sample variance, n is the number of samples, and $\varphi(L)$ and $\varphi(U)$ are an extreme minimum value: 10^{-20}. For *Skew*, Li et al. [10] thought that the real data may be skewed to the right, skewed to the left or not skewed, and thus