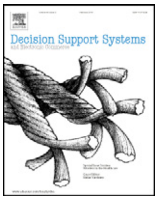
Decision Support Systems 66 (2014) 71–81

Contents lists available at ScienceDirect

Decision Support Systems

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



# Generating information for small data sets with a multi-modal distribution

Der-Chiang Li [*], Liang-Sian Lin

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

ARTICLE INFO

*Article history:*
Received 15 August 2013
Received in revised form 26 February 2014
Accepted 7 June 2014
Available online 16 June 2014

*Keywords:*
Multi-modal distribution
Small data set
Multi-modal virtual sample
Virtual sample size

ABSTRACT

Virtual sample generation approaches have been used with small data sets to enhance classification performance in a number of reports. The appropriate estimation of data distribution plays an important role in this process, with performance usually better for data sets that have a simple distribution rather than a complex one. Mixed-type data sets often have a multi-modal distribution instead of a simple, uni-modal one. This study thus proposes a new approach to detect multi-modality in data sets, to avoid the problem of inappropriately using a uni-modal distribution. We utilize the common *k*-means clustering method to detect possible clusters, and, based on the clustered sample sets, a Weibull variate is developed for each of these to produce multi-modal virtual data. In this approach, the degree of error variation in the Weibull skewness between the original and virtual data is measured and used as the criterion for determining the sizes of virtual samples. Six data sets with different training data sizes are employed to check the performance of the proposed method, and comparisons are made based on the classification accuracies. The results using non-parametric testing show that the proposed method has better classification performance to that of the recently presented Mega-Trend-Diffusion method.

## 1. Introduction

Companies can gain a competitive advantage by speedily providing new products, but when these are in the pilot run stage there is generally only a small amount of data that can be used to improve their performance, due to financial and time limitations. It is thus important to develop analysis methods for use with small data sets, in order to achieve better classification performance [19,20,23,25,28]. Many approaches have been proposed to deal with this issue, with, for example Das and Nenadic [8] and Xu et al. [33] creating algorithms for certain data sets. While this approach is very effective for specific classifiers, the classification is less accurate when the items in the data sets have various characteristics [12]. Other researchers have utilized virtual sample generation (VSG) methods to enlarge data sets, such as those in Yang et al. [34], Li et al. [19] and Li and Lin [21], which all used VSG methods based on estimated density functions. Poggio and Vetter [29] first proposed the concept of using virtual samples to increase the recognition rate in 2D models, and this approach has since been applied in many fields. Cho et al. [7] presented a scheme to select close samples from the population of a network. Li et al. [22] proposed a uniform data generation method to produce a functional virtual population to learn more from small data sets, and provided a criterion for expanding the domain of each feature in a data set to produce the virtual data. Li et al. [24] applied the Mega-Trend-Diffusion (MTD) method to improve small data set learning using early flexible manufacturing system scheduling knowledge. They used a linear, triangular membership function to generate the virtual samples. Presenting a nonlinear model, Yang et al. [34] used the Gaussian distribution to generate virtual samples and achieved data smoothness. The results of these earlier papers demonstrate that various VSG approaches can be used to help solve the problem of small data sets. Nevertheless, these studies all aimed to improve small data set classification based on the assumption of a uni-modal distribution, such as the Gaussian or uniform distributions, which may sometimes lead to an analytical bias in the results. In practice, a complex data set may have a distribution based on a multi-modal density function, in which the mode size of the data is more than one.

Many methods have been developed to find multi-modality in a distribution. For example, Hartigan and Hartigan [15] presented a dip test to measure multi-modality with uniform samples. In this approach, the dip statistic is equal to the maximum difference between the empirical cumulative distribution function (CDF) and the theoretical CDF within all sample points, where the null hypothesis is that the sample has a uniform distribution. Müller and Sawitzki [27] proposed an excess mass test to search for the maximum mode size when the distribution is assumed to be uniform. Without using the uniform assumption, Cheng and Hall [6] proposed a calibrated excess mass test based on other unimodality models, including those with beta, normal, and t distributions.

Many studies have employed these modality test methods to improve performance in classification problems. For example, Polonik and Wang [30] adopted the excess mass test to estimate the modality of each cluster

[*] Corresponding author. Tel.: +886 2757575x53134; fax: +886 2374252.
*E-mail addresses:* lidc@mail.ncku.edu.tw (D.-C. Li), r38991052@mail.ncku.edu.tw (L.-S. Lin).

http://dx.doi.org/10.1016/j.dss.2014.06.004
0167-9236/© 2014 Elsevier B.V. All rights reserved.

before implementing classification. Using mixed data, Chan and Hall [4] proposed a non-parametric approach, before performing clustering algorithms, for selecting the main features according to the testing modality of the density function. In the literature, the mode-testing technique is often used as a criterion to assess whether it is necessary to implement data preprocessing by using cluster analysis. Unfortunately, in these earlier papers the mode-testing methods including the dip test, excess mass test, and calibrated excess mass test, operate on the assumptions that there are sufficient samples and that the data has a uni-modal distribution. However, these often do not apply to small data sets, as the distributions of these are often inflexible. Therefore, in the current work we use a two-parameter Weibull distribution to fit the data and thus achieve greater flexibility, and propose a new modality test by constructing a hypothesis testing procedure to evaluate the fitness of the resulting distribution. We use the Cramer–von Mises statistic in the proposed testing method, which is a common goodness-of-fit technique for small data sets [11,18]. With a given significance level, $\alpha$, we use the statistic to compute the difference between the empirical CDF $\widetilde{F}(x)$ and the theoretical CDF $F(x)$, where the $\widetilde{F}(x)$ is the population distribution function of a small data set from a single Weibull distribution. A uni-modal Weibull distribution is chosen as the $F(x)$, because when the sample size is small data may come from an arbitrary probability distribution. This paper uses the Weibull distribution recommended by Little [26] to form various shapes of a density function and represent uni-modal distributions, including skewed and mound-shaped curves. The Weibull density function can depict the shape of a small data distribution with various shape parameters, as seen in Abernethy [1], Zhang et al. [35], Wahed et al. [32], and Li and Lin [21]. These works all suggest using the flexible shape of the Weibull distribution to handle small data set problems.

In the null hypothesis, we first assume that all data can be fitted by a single Weibull distribution. If the null hypothesis is rejected, we determine that the data is beyond a single Weibull distribution, and thus use two or three Weibull distributions in order to make the fitting process more flexible. Based on the proposed approach, when the density function of a data set indicates multi-modality, we employ the $k$-means clustering approach to set the modality size. Since deciding the exact size is not the main aim of this study, we examine only two- and three-modality cases, as there are few small data sets for which the modality size is more than three, as noted by Good and Gaskins [14] and Silverman [31]. In addition, using the simulated data from a four-modality distribution, Davies and Kovac [9] showed that the peaks are not significant with small samples. In this study we assume that all samples follow a two-parameter Weibull distribution, and the parameters can be evaluated by using the Maximal $p$-Value (MPV) method recommended in Li and Lin [21].

Based on the estimated Weibull distribution with consideration of multi-modality, we generate virtual samples to improve classification performance with small data sets. The determination of the appropriate virtual sample size is another important task in this paper, since adding too many virtual samples to training data sets does not always improve the classification accuracy, while it can significantly decrease the computational efficiency. For this reason, this paper uses the error variation of the Weibull skewness between the original and virtual data in order to measure the structure of virtual data sets, and thus find the suitable virtual sample size (the low variation indicates that the distribution of virtual samples is suitable to depict the original data).

Finally, four real and two simulated data sets are employed in this work to illustrate the effectiveness of the proposed method by comparing its classification accuracy with that of the MTD method. In addition, the technique recommended in Demšar [10] is applied to test the statistical significance and classification performance of the following four classifiers, namely Fisher's linear discriminant analysis (LDA), K nearest-neighbor (KNN), and two types of support vector machines (SVMs) [16].

The remainder of this study is organized as follows: Section 2 reviews the MTD method for small data set problems, and also describes the use of the MPV approach for evaluating the two-parameter Weibull distribution. Section 3 presents a VSG scheme with the proposed testing steps for multi-modality, and explains how to determine the virtual sample size. Section 4 uses six data sets, and compares the results obtained from the proposed method, the MTD, and REAL (a method using only the existing real data). Finally, we present the conclusions of this work in Section 5.

## 2. Related studies

Modeling with more training data can usually achieve better classification accuracy, and so many researchers have suggested using virtual sample generation (VSG) approaches when dealing with small data sets. The current study aims to generate multi-modal virtual samples using an estimation method (i.e., MPV) for the two-parameter Weibull distribution. The related studies are reviewed in detail in the following subsections.

### 2.1. The MTD method

As mentioned above, Li et al. [24] proposed the MTD for small data learning in an early manufacturing system, using this approach to data trend estimation to create virtual data. As shown in Fig. 1, the MTD method constructs a triangular membership function to calculate the range of virtual data, which is the interval from $\alpha$ to $b$, described mathematically as:

$$a = u_{set} - Skew_L \times \sqrt{-2 \times s_x^2/N_L \times \ln\left(10^{-20}\right)}, 1<N_L<\infty, \tag{1}$$

$$b = u_{set} + Skew_U \times \sqrt{-2 \times s_x^2/N_U \times \ln\left(10^{-20}\right)}, 1<N_U<\infty. \tag{2}$$

Note that the $Skew_L = N_L/(N_L + N_U)$ is the left of the skewness degree of $\sqrt{-2 \times s_x^2/N_L \times \ln\left(10^{-20}\right)}$, and $Skew_U = N_U/(N_L + N_U)$ is the right of the skewness degree of $\sqrt{-2 \times s_x^2/N_U \times \ln\left(10^{-20}\right)}$. Where $N_U$ and $N_L$ denote the number of data less and more than $u_{set} = (min + max)/2$, respectively, and the $min$ and $max$ are the minimum and maximum values in real data sets. The lower bound $\alpha$ and the upper bound $b$ can be calculated from Eqs. (1) and (2). After obtaining the values of $\alpha$ and $b$, generate virtual samples that are distributed following the triangular density function in the interval [$\alpha$, $b$], and add them to the original data set. Li et al. [24] set the number of virtual samples that are generated at 100.

### 2.2. The MPV method for parameter estimation

Given a random variable $X$ that is denoted by a two-parameter Weibull distribution, the probability density function (PDF) of the Weibull distribution is expressed as:

$$f(x, \lambda, \beta) = \frac{\beta}{\lambda}\left(\frac{x}{\lambda}\right)^{\beta-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{\beta}\right\}, x \geq 0, \lambda > 0, \beta > 0 \tag{3}$$

where $\lambda$ is the scale parameter and $\beta$ is the shape parameter.

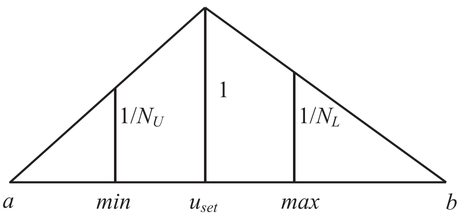


**Fig. 1.** Data trend estimation.