A Mega-Trend-Diffusion and Monte Carlo based virtual sample generation method for small sample size problem

Xiaoru Yu, Yanlin He, Yuan Xu and Qunxiong Zhu*
College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China
*Corresponding author’s e-mail: zhuqx@mail.buct.edu.cn

Abstract. Data-driven modeling has attracted wide attention in academia because of its effectiveness. However, Due to the lack of data, some traditional modeling methods, such as extreme learning machine (ELM), can’t achieve high learning accuracy. A novel approach based on Mega-Trend-Diffusion (MTD) and Monte Carlo is presented in this paper to deal with the problem, named Monte Carlo Mega-Trend-Diffusion (MCMTD). The proposed approach utilizes MTD to estimate the acceptable range of the attributions and Latin hypercube sampling method to sample. ELM is employed to establish the prediction model. In this paper, two real data sets, the multi-layer ceramic capacitors (MLCC) and the purified terephthalic acid (PTA), are used to verify the effectiveness and reasonability of MCMTD. The experimental results show that MCMTD can significantly enhance the accuracy and ability of the forecasting model.

1. Introduction

Data-driven modeling has attracted wide attention in academia on account of their effectiveness. Sufficient data is the main condition to ensure accurate prediction model. In the era of big data, the difficulty of data collection can’t be ignored. In real life, many data have low probability and high cost. This leads to a matter related to sample size, called "big data, small sample set". The problem of small sample size appears in many practical applications, for example, using data-driven model to diagnose bladder cancer genes, when adding virtual samples, the accuracy of diagnosis is steadily improved[1]. Effectiveness and efficiency are important factors for manufacturing systems to maintain a strong competitive advantage. However, due to the lack of samples, establishing an appropriate forecasting model is a difficult task in the early stages of manufacturing systems[2-5].

Many methods used to solve small sample size problems have been proposed, such as grey model[15], feature selection[16], virtual sample generation (VSG) technology. It’s an effective approach to expand the original sample set and enhance the prediction ability of small samples by VSG. In 1992, Poggio and Vetter put forward the idea of virtual sample[6]. With the development of VSG, Li et al.[7] presented a approach called Functional Virtual Population for solving the problem of dynamic scheduling and insufficient samples in flexible manufacturing system. Adequate training samples are an important condition to ensure the high accuracy of the neural network model. Huang et al.[8] developed a diffusion neural network (DNN) for the purpose of improving the performance of neural network on small sample set. The DNN generates derived patterns from the original patterns based on information diffusion. The Mega-Trend-Diffusion (MTD)[9] technology uses the triangular membership function to express the importance of samples, so as to mine the trend information of the sample. Zhu et al.[10] proposed a multi-distribution mega trend diffusion technology for averting the...
shortcoming of single distribution. The technique generates virtual samples in the observation area and extended area according to triangular and uniform distribution, separately.

For overcoming the problem of small sample size and increasing the ability and accuracy of prediction model, a VSG method based on the MTD and Monte Carlo is proposed in this paper, named Monte Carlo Mega-Trend-Diffusion (MCMTD). A prediction model is established by standard Extreme Learning Machine (ELM).

2. Preliminaries
In this section, MTD and Monte Carlo will be reviewed briefly.

2.1. Mega-Trend-Diffusion
The MTD is a technique for estimating the acceptable range of sample attributes, diffusing the information of the data and then generating the virtual samples. It asymmetrically expands the attribute domain of samples, and takes triangular membership function (MF) value as the possibility of the sample. The MF value also presents the importance of the single datum. As an effective VSG technology, The MTD has been applied in many fields, such as medical domain. The specific steps are as below:

Step 1: Consider a data set \( X = \{x_1, x_2, x_3, \ldots, x_n\} \), \( \text{max} \) and \( \text{min} \) is set to be the maximum and minimum of \( X \), respectively. \( N_U \) and \( N_L \) indicate the amount of samples greater than and smaller than \( CL \), respectively. \( CL \) represents the centre of an attribute, and it can be calculated as:

\[
CL = \frac{(\text{min} + \text{max})}{2} \quad (1)
\]

Step 2: Calculate the upper boundary \( UB \) and lower boundary \( LB \) using equation (2) and (3).

\[
UB = \left\{ \begin{array}{ll}
CL + \text{Skew}_U \times \frac{-2 \times \hat{S}^2}{N_U \times \ln(10^{-30})}, & UB \geq \text{max} \\
\text{max}, & UB < \text{max}
\end{array} \right. \quad (2)
\]

\[
LB = \left\{ \begin{array}{ll}
CL - \text{Skew}_L \times \frac{-2 \times \hat{S}^2}{N_L \times \ln(10^{-30})}, & LB \leq \text{min} \\
\text{min}, & LB > \text{min}
\end{array} \right. \quad (3)
\]

\[
\text{Skew}_L = \frac{N_L}{N_U + N_L} \quad (4)
\]

\[
\text{Skew}_U = \frac{N_U}{N_U + N_L} \quad (5)
\]

\[
\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (6)
\]

In equations (4-6), \( n \) denotes the number of samples, \( \hat{S}^2 \) is variance of samples, \( \text{Skew}_L \) and \( \text{Skew}_U \) is left and right skewness magnitudes based on \( CL \), respectively.

Step 3: Randomly pick up \( n_s \) samples within the acceptable range according to uniform distribution.

Step 4: Obtain the MF values of observation \( x_i \) using equation (7).

\[
MF = \left\{ \begin{array}{ll}
\frac{(x - LB)}{(CL - LB)}, & x \leq CL \\
\frac{(UB - x)}{(UB - CL)}, & x > CL
\end{array} \right. \quad (7)
\]

The samples generated by the MTD are randomly sampled by uniform distribution, which can not represent the complete information of sample distribution.

2.2. Monte Carlo
Monte Carlo[14] sampling techniques provide more reliable samples using simulation and experimental data. The core of Monte Carlo is to sample a distribution. When the iteration is less, it is easy to cause the problem of data over-aggregation, that is, high probability samples are easier to be extracted. If the low probability sample has high importance, it will have a great impact on the results because it is not easy to be extracted. Latin hypercube sampling (LHS) solves the problem. LHS technique is an efficient sampling method. It can accurately reconstruct the input distribution by sampling with lesser iterations. The LHS method divides the cumulative distribution function (CDF) into equal intervals on the [0, 1], that is stratification, and then makes random sampling on each interval. The sampling value is the probability of random sample, which is transformed into the value...
of the target distribution through the inverse function of the CDF. The stratification technology ensures that each layer will be extracted, and will not be repeated extraction, thus solving the problem of data over-aggregation.

The specific steps of LHS are as follows:

Step 1: Divide \([0, 1]\) into \(n\) equal intervals, and a random number is generated according to the uniform distribution on each interval.

Step 2: Disturb the \(n\) random numbers generated in Step 1.

Step 3: The inverse function of CDF is used to generate the value of the target distribution.

3. The proposed method

This section introduces the presented approach MCMTD in detail.

3.1. Establishing forecasting model

In the first step a certain amount of samples is randomly extracted from the original dataset as training set \(\text{Strain}\) and the rest samples as the testing set \(\text{Stest}\). ELM is chosen to build the prediction model due to the advantages of high learning efficiency. The trial-and-error method is applied to set the number of nodes in the hidden layer \(\text{Nh}\).

3.2. Determining the attribute ranges

The second step is to determine the acceptable domain of the attribute. As mentioned in 2.1, the MTD extends the single point diffusion to the mega diffusion. It can take into account the overall distribution trend of samples and use it to produce virtual samples, which can better display the potential information of samples. For a given dataset, \(\text{LB}\) and \(\text{UB}\) are calculated as equations (1-6).

3.3. Generating virtual samples

In 3.2, \(\text{UB}\) and \(\text{LB}\) of the samples are obtained by the MTD. In \([\text{LB}, \text{UB}]\), the LHS method is used to sample according to the triangular distribution. The sampled values are used as the input \(\text{virP}\) of virtual samples. Through the ELM model trained in 3.1, the output \(\text{virT}\) of \(\text{virP}\) is predicted and used to compose the virtual sample set \(\text{Svir}\).

Finally, the original training set \(\text{Strain}\) and the virtual sample set \(\text{Svir}\) are combined into a new training set \(\text{Snew}\). The ELM model was retrained with \(\text{Snew}\) and validated with the same testing set \(\text{Stest}\).

4. Case study

Two real datasets will be used to test the reasonability and effectiveness of MCMTD in this section. The first is the multi-layer ceramic capacitors (MLCC) data set, and the second is the purified terephthalic acid (PTA) data set.

Each computational result is the mean of 10 independent experiments. The formulas for \(\text{AveEIR}\) and \(\text{AveMAPE}\) are as below.

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% 
\]

\[
\text{AveMAPE} = \frac{1}{m} \sum_{i=1}^{m} \text{MAPE}_i
\]

\[
\text{EIR} = \frac{\text{MAPE}_{\text{before}} - \text{MAPE}_{\text{after}}}{\text{MAPE}_{\text{before}}} \times 100\% 
\]

\[
\text{AveEIR} = \frac{1}{m} \sum_{i=1}^{m} \text{MAPE}_i
\]
4.1. MLCC
MLCC[12][13] has 44 groups of samples. There are 12 inputs affect the real K-value (RK). For the aim of investigating the influence of training set size on the accuracy of prediction model, the size is set to 5, 10, 15, 20 and 25 in turn, and the rest samples are treated as the testing set. \( N_h \) is set to 8.

Table 1. The improvements of MCMTD with and without 100 virtual samples.

| Size  | Before | After | Improvement (%) |
|-------|--------|-------|-----------------|
| 5     | 7.7418 | 5.2007| 32.8224         |
| 10    | 6.3793 | 4.5828| 28.1606         |
| 15    | 6.4189 | 4.0030| 37.6375         |
| 20    | 6.6632 | 3.8253| 42.5908         |
| 25    | 5.3452 | 3.6855| 31.0504         |
| 30    | 5.1301 | 3.4216| 33.3045         |

Figure 1. The average accuracy of MCMTD before and after adding 100 virtual samples.

Figure 2. The improvements of MCMTD before and after adding 100 virtual samples.

Figure 3. Comparison between different methods with different size of training set after adding 100 virtual samples.

The experimental results are summarized in table 1, shown in figure 1 and figure 2. Obviously, the performance of the prediction model has been significantly improved after adding virtual samples. For proving that the presented method has better performance, it is compared with Chen et al.[11] and the MTD. The results present that the proposed MCMTD is superior to the other two methods in different training set sizes, as presented in table 2 and figure 3. In a word, the prediction model has better performance with virtual samples than without.

4.2. PTA
The PTA[14] data set consists of 260 groups of samples, including 17 input variables and one output. 50 groups of samples are extracted randomly, of which 30 groups are used to be training set, and the remaining ones are used as testing set. \( N_h \) is set to 6 in hidden layer. The number of virtual samples is set to 100, 150, 200, 250, 300, 350, sequentially, to explore the affect of the number of virtual samples on the prediction model performance. Table 3 summarizes the results before and after adding diverse amount of virtual samples. Apparently, it is found that the performance of the forecasting model has
been significantly improved after adding virtual samples. Figure 4 and figure 5 display that the more the amount of virtual samples, the better the performance of the prediction model. Table 4 and figure 6 demonstrate that the presented method is superior to the other two methods in various numbers of virtual samples. In brief, the proposed MCMTD is a useful solution to small sample set problem.

Table 3. The improvements of MCMTD before and after increasing virtual samples of various sizes.

| Size  | Before   | After   | Improvement (%) |
|-------|----------|---------|-----------------|
| 100   | 1.2499   | 1.1218  | 10.2505         |
| 150   | 1.2499   | 1.0788  | 13.6877         |
| 200   | 1.2499   | 1.0165  | 18.6778         |
| 250   | 1.2499   | 0.9447  | 24.4195         |
| 300   | 1.2499   | 0.8620  | 31.0358         |
| 350   | 1.2499   | 0.8365  | 33.0725         |

Figure 4. The accuracy of MCMTD with and without virtual samples.

Figure 5. The improvements of MCMTD before and after adding virtual samples with different size.

Figure 6. Comparison between different methods with different number of virtual samples.

Table 4. Comparison of different methods in various numbers of virtual samples.

| Method | Number of virtual samples. |
|--------|----------------------------|
|        | 100 | 150 | 200 | 250 | 300 | 350 |
| MCMTD(%) | 1.1218 | 1.0788 | 1.0165 | 0.9447 | 0.8620 | 0.8365 |
| PSOVSG(%) | 1.3090 | 1.2190 | 1.1220 | 1.1100 | 1.0880 | 1.0220 |
| MTD(%)   | 1.4820 | 1.4680 | 1.4460 | 1.4560 | 1.4540 | 1.4330 |

5. Conclusions
The lack of training samples makes the data-driven model unable to achieve better performance and accuracy. To solve this problem, this paper put forward an approach of generating virtual samples based on MTD and Monte Carlo. Two real datasets are used in the experimental part for confirming the effectiveness and reasonability of the presented method. The computational results indicate that MCMTD can enlarge the raw training set using virtual samples so as to reduce the information gaps between samples. To sum up, MCMTD can be regard as a stable and effective forecasting technique for small sample set problems.

The amount of virtual samples plays a significant role in the performance of the prediction model. How to set the best amount of virtual samples will be focused in the future research work.

Acknowledgments
This research is supported by the National Natural Science Foundation of China under Grant Nos. 61703027 and 61533003 and Fundamental Research Funds for the Central Universities under Grant Nos. JD1914 and XK1802-4.
References

[1] Li, D.C., Hsu, H.C., Tsai, T.I., Lu, T.J., Hu, S.C. (2007) A new method to help diagnose cancers for small sample size. Expert Systems with Applications, 33(2): 420-424.

[2] Chang, C.J., Li, D.C., Dai, W.L., Chen, C.C. (2014) A latent information function to extend domain attributes to improve the accuracy of small-data-set forecasting. Neurocomputing, 129: 343-349.

[3] Li, D.C., Wu, C., Chang, F.M. (2005) Using data-fuzzification technology in small data set learning to improve fms scheduling accuracy. The International Journal of Advanced Manufacturing Technology, 27(3-4): 321-328.

[4] Li, D.C., Wu, C.S., Tsai, T.I., Chang, F.M. (2006) Using mega-fuzzification and data trend estimation in small data set learning for early fms scheduling knowledge. Computers and Operations Research, 33(6): 1857-1869.

[5] Li, D.C., Chang, C.C., Liu, C.W. (2012) Using structure-based data transformation method to improve prediction accuracies for small data sets. Decision Support Systems, 52(3): 748-756.

[6] Poggio, T., Vetter, T. (1992) Recognition and structure from one 2d model view: observations on prototypes, object classes and symmetries. Laboratory Massachusetts Institute of Technology, 1347.

[7] Li, D.C., Chen, L.S., Lin, Y.S. (2003) Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments. International Journal of Production Research, 41(17): 4011-4024.

[8] Chongfu, H.A., Claudio, M.B. (2003) A diffusion-neural-network for learning from small samples. International Journal of Approximate Reasoning, 35(2): 137-161.

[9] Li, D.C., Wu, C.S., Tsai, T.I., Lina, Y.S. (2007) Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. Computers & Operations Research, 34(4): 966-982.

[10] Zhu, B., Chen, Z., Yu, L. (2016) A novel mega-trend-diffusion for small sample. CIESC Journal.

[11] Chen, Z.S., Zhu, B., He, Y.L., Yu, L.A. (2017) A pso based virtual sample generation method for small sample sets: applications to regression datasets. Engineering Applications of Artificial Intelligence, 59: 236-243.

[12] Li, D.C., Chen, C.C., Chang, C.J., Lin, W.K. (2012) A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing systems. Expert Systems with Applications, 39(1): 1575-1581.

[13] Li, D.C., Wen, I.H. (2014) A genetic algorithm-based virtual sample generation technique to improve small data set learning. Neurocomputing, 143: 222-230.

[14] Gong, H.F., Chen, Z.S., Zhu, Q.X., He, Y.L. (2017). A monte carlo and pso based virtual sample generation method for enhancing the energy prediction and energy optimization on small data problem: an empirical study of petrochemical industries. Applied Energy, S0306261917303811.

[15] Chang, C.J., Li, D.C., Huang, Y.H., Chen, C.C. (2015) A novel gray forecasting model based on the box plot for small manufacturing data sets. Applied Mathematics and Computation, 265: 400-408.

[16] Dernoncourt, D., Hanczar, B., Zucker, J.D. (2014) Analysis of feature selection stability on high dimension and small sample data. Elsevier Science Publishers B. V.