PAPER • OPEN ACCESS

Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review

To cite this article: Masitah Abdul Lateh *et al* 2017 *J. Phys.: Conf. Ser.* **892** 012016

View the article online for updates and enhancements.

You may also like

- Comparative study of selected indoor concentration from selective laser sintering process using virgin and recycled polyamide nylon (PA12)
A A M Damanhuri, A S A Subki, A Hariri et al.

- Embracing tourist behaviour in managing Melaka WHS
N D Md Khairi, H N Ismail and S M R S Jaafar

- THE PRODUCTIVITY IMPROVEMENT OF HOME –MADE PRODUCTION LINE BY DEVELOPING OF SEMI-AUTO BATTER DISPENSING SYSTEM: KUIH BAHULU
A. M. Kamarul, A. R. Soufhwee, Haeryip Sihombing et al.

IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 189.250.118.26 on 12/07/2022 at 05:06

Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review

Masitah Abdul Lateh[1], Azah Kamilah Muda[1], Zeratul Izzah Mohd Yusof[1], Noor Azilah Muda[1] and Mohd Sanusi Azmi[1]

[1]Computational Intelligence and Technologies Lab (CIT), Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, 76100 Melaka, Malaysia

azah@utem.edu.my

Abstract. The emerging era of big data for past few years has led to large and complex data which needed faster and better decision making. However, the small dataset problems still arise in a certain area which causes analysis and decision are hard to make. In order to build a prediction model, a large sample is required as a training sample of the model. Small dataset is insufficient to produce an accurate prediction model. This paper will review an artificial data generation approach as one of the solution to solve the small dataset problem.

1. Introduction

In recent years, small dataset problem is widely discovered and getting more attention from researchers. There are many organizations that work with small datasets including manufacturing and medical areas. Manufacturing area usually dealing with small dataset problems in the early stage of manufacturing such as high cost of products while in the medical area the special medical record as an example, spinocerebellar ataxia, which is a kind of rare hereditary disease with very few records around the world.

There are many research works working on prediction model. To obtain a precise prediction model, a large sample set of data is required for the learning process. Otherwise, the model built is considered unreliable and the information gained is fragile [1]. For example, in the case of TFT-LCDs, their forecasting problem is related to a limited number of sample in the production line and a good prediction model could reduce the manufacturing cost. Thus, a solution is needed to handle a small dataset to build a precise prediction model and reduce a manufacturing costs. In Fig.1, a small dataset is one of the factors affecting inaccurate prediction model. So, if a small sample is used as training sample of a model, it might significantly affect the prediction uncertainty because of lack of information [2]. Consequently, the knowledge gained from small sample of data is considered unreliable and imprecise for learning system [3]. In addition, in the context of computational learning theory, small sample size gives a major effect to learning performances and is found as one of encountered in machine learning and data mining. This is because insufficient data size of training dataset is liable to poor performances of learning[4].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

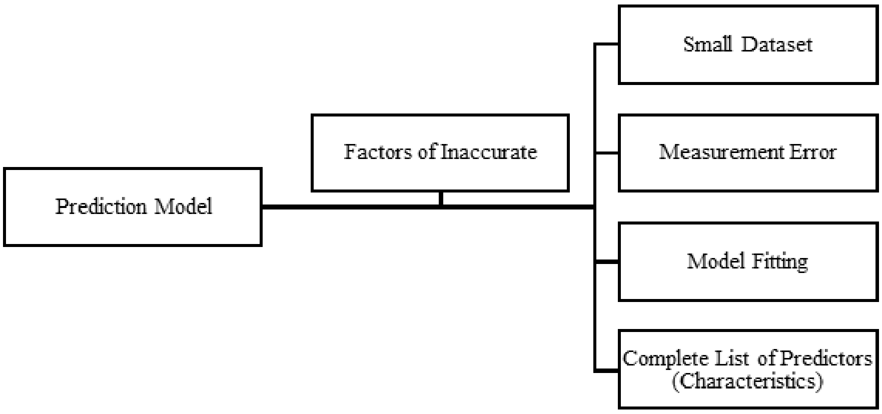

Figure 1. Factors Effecting Inaccuracy of Prediction Model [5]

An effective way is needed to improve the learning accuracy and at the same time, a learning stability is gained. From the literature, one of the effective approach to deal with this problem is adding artificial data to the system. From previous research works, the accuracy of learning algorithm is improved when the amount of training data examples increases. Therefore, this paper is presented to review artificial data generation approach which contain several techniques that robustly handle a small dataset problem for prediction model.

The remainder of this paper is organized as follows. In section 2, related work in small dataset problem is reviewed. In section 3, the artificial data generation approach is explained. Next section will review the general concept of artificial data generation approach to the prediction model. Then this paper is concluded in section 5.

2. Characteristics of Small Data

Data can be defined as a collection of raw facts that has not yet been processed. Then from the processed data, a meaningful form is derived which defined as information. However, the information in small data size is scarce and have some learning limit. The main reason why small dataset cannot provide enough information is that gaps are existed between samples. The gaps between sample and observations are called the information gaps, which cause most of the learning tools are difficult to predict [5]. The Fig.2 shows the example of the information gap that are existed between samples.

According to Vapnik [6], a sample size is considered as small if the ratio of the number of training sample to the Vapnik- Chervonenkis (VC) dimensions of a learning machine is less than 20 [8-10]. VC dimensions is a measure of the capacity for a classificatory, defined as the cardinality of the largest set of points that the algorithm can shatter[8]. However, these theories focus on general machine learning with a large number of training samples, which cannot be applied to practical cases with the small data set learning model [7].

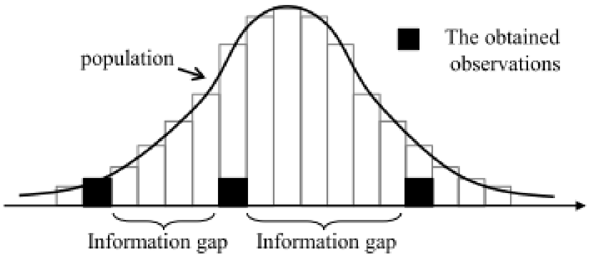

Figure 2. The distribution of a small dataset relative to its population [6]

3. Artificial Data Generation Approach

3.1. Fuzzy Theories

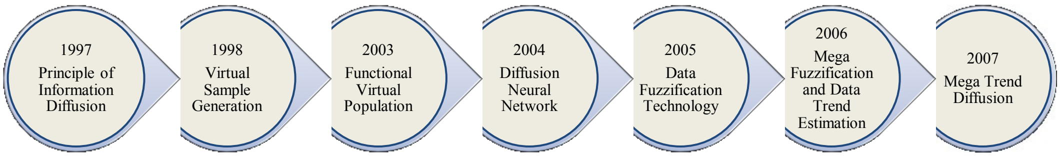
1997
Principle of Information Diffusion
1998
Virtual Sample Generation
2003
Functional Virtual Population
2004
Diffusion Neural Network
2005
Data Fuzzification Technology
2006
Mega Fuzzification and Data Trend Estimation
2007
Mega Trend Diffusion

Figure 3. Chronological of Techniques employ Fuzzy Theory

Enhancing learning process in machine learning is important to make sure the performance is improved. One of the approach to enhance the learning process is by adding some synthetic data to the system[9]. Based on the literature, artificial data generation approaches employ some fuzzy theories and neural network techniques which are used to estimate or approximate functions to generate more sample for training sets.

The concept of synthetic data was originally proposed by Niyogi [10] in pattern recognition work which is to improve a 3-D view recognition of an object by increasing the amount of sample data using the prior knowledge gained from small training set. When the amount of the synthetic data is added, it is found that, a 3-D view recognition of an object is improved. However, the estimation of the data effect become the priority in generating a synthetic data. Hence, in the method of a diffusion neural network (DNN) introduced by Huang and Moraga [11], they applied the principle of information diffusion by Huang[12]. Huang proposed a solution for small sample size problem by using fuzzy theories. This is because the incompleteness of data can precisely be represented using fuzzy membership function to represent the similarities between samples. Principle of information diffusion has presented a concept that new artificial examples were used to fill the information gaps which was caused by data incompleteness. The unique diffusion formula is as follows:

$$x_i = u_i \pm \sqrt{-2 \times h \times \ln\big(\varphi(x_i)\big)} \qquad (1)$$

DNN is formed by the principle of information diffusion and traditional neural network. In neural network, the information is processed by assuming two conditions which are 1) the patterns are compatible and 2) the learning pattern for training neural network is sufficient. Otherwise, it is difficult to identify a non-linear system or there will exist a non-negligible error between the real function and estimated function from a trained network [11]. So, the information diffusion is implemented in this technique to derive more pattern to partly fill the gaps between samples which later can produce a better result in identifying non-linear function.

Flexible Manufacturing System (FMS) is a scheduling system which commonly used in production planning and controlling processes. The purposes are to reduce the cost and balance the machine loads. However, it is difficult to decide because the data is not enough for learning process. Li [13] invented the first method named Functional Synthetic Population (FVP) to increase the accuracy of machine learning for FMS scheduling. By following the synthetic data concept, FVP will generate synthetic samples to improve the FMS scheduling knowledge in small data sets learning. The main contribution of this method is the algorithm was developed to expand the domain of the system attributes and added some synthetic samples since for constructing early new scheduling knowledge by adopting neural network algorithm. Since neural network needs enough training set for the learning process, the scheduling knowledge is improved by the added sample. Unfortunately, FVP require many steps to complete [7].

Later, Li employed a data fuzzification technology [13] concept to expand the small dataset obtained to further improve learning accuracy. In this method, the approaches used include data-fuzzifying, domain range expansion, and adaptive-network-based fuzzy inference systems (ANFIS) as neuro-fuzzy learning. After input data were fuzzified, the domain range expansion technique was administered to enhance the domain range allowance in a dynamic and non-enough data environment. The results indicate that the learning accuracy under this strategy is significantly better than that of a traditional crisp data neural networks. However, their proposed method has limitations of the domain range which need to be pre-determined and it must consider the data behavior.

Subsequently, the extended strategy to improve the accuracy of FMS scheduling in small dataset problem is by introducing a new method named Mega Fuzzification and Data Trend Estimation. Because of the limitation existed in previous technique, a new technique is developed by Li [1] intends a combining data fuzzification, data trend estimation, and ANFIS. To build a precise and useful knowledge, considerable quantities of data sets are needed before proceed to the learning process. So, mega-fuzzification or also called as data fuzzification method is presented to determine the possible coverage of a data set as shown in Fig.4.

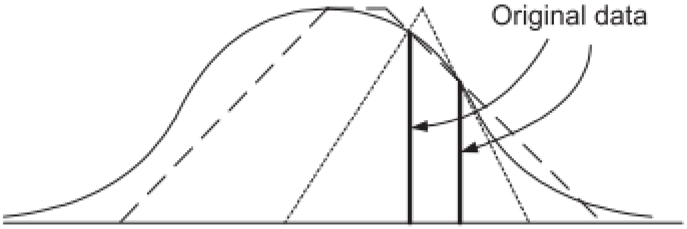

Figure 4. Determining possible coverage of data set using fuzzifying technique [1]

Asymmetric domain range expansion is utilized to estimate the data trend as shown in Fig.5, while expanding the data domain ranges. According to the author, there are some of the matter that should be further studied such as determining membership function type and appropriate height for min and max in the triangular [1]. Besides, the height of min and max in the triangular did has a direct relationship with the attribute domain range estimation. Otherwise, it would lead to either over-estimating or underestimating ranges which cause poor learning process. Thus, to determine the domain ranges of the attribute, providing a suitable method is important.

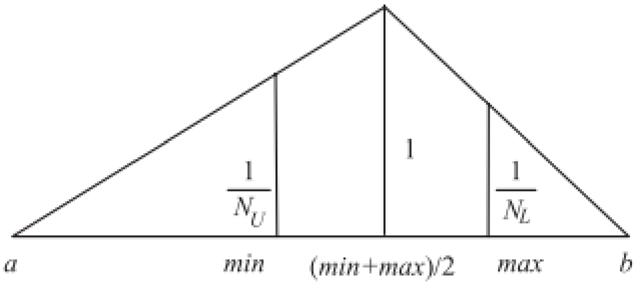

Figure 5. Data Trend Estimation [1]

Later, the Mega Trend Diffusion (MTD) is proposed by Li [3] which employs data trend estimation concept. The mega diffusion produced mega trend diffusion which aimed to avoid over-estimating. The main idea of MTD function is to generate synthetic samples. In statistics view, a normal distribution is necessary before the data analysis is done. However, when the dataset is small, it is often difficult to show the data following the normal distribution [14]. Alternatively, Li [3] used the membership function in fuzzy set theory to calculate the possibility values of synthetic samples than the probability in statistics to avoid the normal distribution assumption. Fig.6 shows the concept of the fuzzy theorem applied to the MTD function.

The conventional method (principle of information diffusion) diffuse each sample individually whereas the MTD method diffuses a set of data using a common diffusion function. MTD method employs information diffusion based on fuzzy possibility distribution to determine possible data set coverage on an attribute domain. Fig. 1 shows the concept of applying the fuzzy theorem to this context. The triangular shape is the membership function, and values a and b are the boundary of the MTD function. The heights of samples m and n are the possibility values of the membership function, and the possibility value is in the interval 0–1. Given a sample set $X = \{x_1 , x_{2,} , \ldots , x_n\}$, the boundaries a (lower boundary) and b(upper boundary) are defined as below:

$$a = u_{set} \; - Skew_U \times \sqrt{-2 \times \frac{\hat{s}_x^2}{N_U} \times \ln(10^{-20})} \tag{2}$$

$$b = u_{set} + Skew_L \times \sqrt{-2 \times \frac{\hat{s}_x^2}{N_U} \times \ln(10^{-20})} \tag{3}$$

where u_{set}= (min + max) / 2, with min and max being minimum and maximum values of observation, respectively, $\hat{s}_x^2$is the variance of observations, and ln (10^{-20}) is the diffusion coefficient. N_L and N_U denoted as number of observations which are smaller and greater than u_{set} respectively, and $Skew_L$= N_L / (N_L + N_U) and $Skew_U$= N_L / (N_L + N_U) for the skewness of the possible distribution of the data.

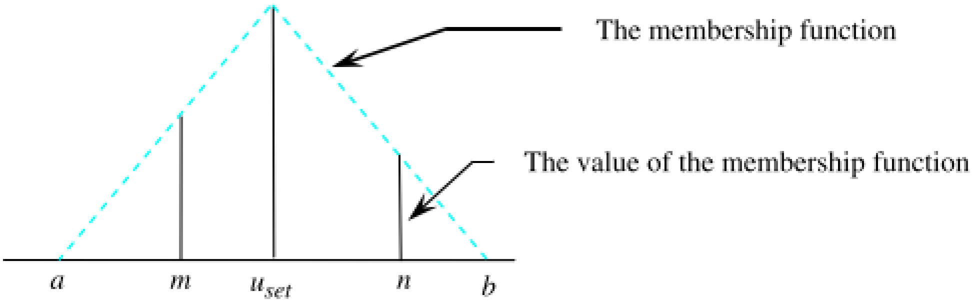

Figure 6. MTD function [3]

Table 1. Computational Result

Technique	**Number of Training Set**	**Learning Accuracy (%)**
Mega Fuzzification and Data Trend Estimation [1]	5	69.3
	10	76.7
	20	80.7
	30	88.3
	40	92.2
	50	93.8
	100	94.7
Mega Trend Diffusion [3]	5	78.23
	10	88.53
	20	89.08
	30	91.74
	40	93.41
	50	94.33
	100	95.33

Based on the computational result on Table 1, the learning accuracy of MTD method showed that it is a superior technique and significant to be applied in the real world. The learning accuracy for 30 samples in Mega Fuzzification and Data Trend Estimation are equal to learning accuracy of 10 samples in Mega Trend Diffusion. As a result, the MTD method is widely used in many application including [17-19]. The main idea of MTD is to generate a synthetic sample. However, MTD did not consider the relation among attributes but only for independent attributes. From the finding, one of the improvement is by considering the relationship between attributes. The following section introduce some of the improved technique of MTD function.

3.1.1. Heuristic Algorithm: Genetic Algorithm.

Genetic Algorithm (GA) is one of the optimization algorithm. Even DNN and MTD method discussed previously could solve the small data learning problem by synthetic sample approach, they only considered the data for independent attribute only. Thus, a genetic algorithm was applied which aim to consider the relation of integrated effects and constraints of data attributes [16]. Three main steps were involved and GA was applied in the second step to generate the number of most feasible synthetic samples. The proposed method showed better performance in prediction than without synthetic sample.

3.1.2. Data Transformation.

Data transformation is introduced to bring the data close to the normal distribution to generate a synthetic sample since the small data size is difficult to follow a normal distribution [17]. Although there a many transformation developed, Li introduced Johnson Transformation which is suitable for large data sets (n>30). To make it suitable for small data size, Li introduced Small Johnson Data Transformation (SJDT) function [14] dedicated for small datasets. Johnson system comprises of a family of three curves: bounded system (SB), the log-normal system (SL) and unbounded system (SU) and is represented as follow:

$$Z = \gamma + \eta \times k_i(x, \lambda, \varepsilon) \tag{4}$$

where Z is a random variable that obeys to the standard normal distribution and $k_i(x, \lambda, \varepsilon)$ is the function chosen to cover a wide range of possible occurrence [14]. The function of SB, SL, and SU are defined as follow:

$$\text{Bounded System (SB): } y = \gamma + \eta * \text{Ln}\left(\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right) \tag{5}$$

$$\text{Log-normal system (SB): } y = \gamma + \eta * \text{Ln}\left(\frac{x-\varepsilon}{\lambda}\right) \tag{6}$$

$$\text{Unbounded system (SU): } y = \gamma + \eta * \sin^{-1}\left(\frac{x-\varepsilon}{\lambda}\right) \tag{7}$$

where y is a transformed data, x is a raw data and $(\varepsilon, \gamma, \eta, \lambda)$ are the Johnson parameters. Following to Slifker and Shapiro [18] the parameter of three curves above is estimated based on sample quantiles x_{-3z} ,x_{-1z},x_{1z} and x_{3z} . However, Johnson transformation is suitable for the large dataset (n>30). So, to make Johnson Transformation is fit for a small dataset, the author redefines the four sample quantiles using the Eq.2 and Eq.3 in MTD function. In the proposed SJDT function [14], it is found that the 97% of data is located within interval [x_{-3z} , x_{3z}] when z=1 which is similar to MTD function about 99.73% data. Therefore, the boundaries of a and b in MTD function is used to replace the data interval x_{-3z} and x_{3z}. The four quantiles are now defined as:

$$x_{-3z} = a = u_{set} \text{ - } Skew_U \times \sqrt{-2 \times \frac{\hat{s}_x^2}{N_U} \times \ln(10^{-20})} \tag{8}$$

$$x_{-1z} = \text{min} \tag{9}$$

$$x_{1z} = \text{max} \tag{10}$$

$$x_{3z} = b = u_{set} + Skew_L \times \sqrt{-2 \times \frac{\hat{s}_x^2}{N_U} \times \ln(10^{-20})} \tag{11}$$

To differentiate the type of transformation action to perform, the selection criteria below are follow:

$$\frac{mn}{p^2} > 1 \text{ for all SU distribution}$$
$$\frac{mn}{p^2} < 1 \text{ for all SB distribution}$$
$$\frac{mn}{p^2} = 1 \text{ for all SL distribution}$$

where $m = x_{3z} - x_{1z}, n = x_{-1z} - x_{-3z}, p = x_{1z} - x_{-1z}$.

Later, the corresponding parameter $(\varepsilon, \gamma, \eta, \lambda)$ are estimated according to three distribution below:

The parameter estimated for SU Distribution are:

$$\eta = \frac{2 \times z}{cosh^{-1}\left[\frac{1}{2}\left(\frac{m}{p}+\frac{n}{p}\right)\right]}, \eta > 0$$

$$\gamma = \eta \times sinh^{-1}\left[\frac{\frac{n}{p}-\frac{m}{p}}{2\left(\frac{m}{p} \times \frac{n}{p} - 1\right)^{1/2}}\right]$$

$$\lambda = \frac{2p\left(\frac{m}{p}\times\frac{n}{p}-1\right)^{1/2}}{\left(\frac{m}{p}+\frac{n}{p}-2\right)\times\left(\frac{m}{p}+\frac{n}{p}+2\right)^{1/2}}, \lambda > 0$$

$$\varepsilon = \frac{x_{1z} + x -_{1z}}{2} + \frac{p \times \left(\frac{n}{p} - \frac{m}{p}\right)}{2\left(\frac{m}{p} + \frac{n}{p} - 2\right)}$$

The parameter for SB distribution are:

$$\eta = \frac{z}{cosh^{-1}\left\{\frac{1}{2}\left[\left(1+\frac{p}{m}\right)\left(1+\frac{p}{n}\right)\right]^{1/2}\right\}}, \eta > 0$$

$$\gamma = \eta \times sinh^{-1}\left\{\frac{\left(\frac{p}{n}-\frac{p}{m}\right)\left[\left(1+\frac{p}{m}\right)\left(1+\frac{p}{n}\right)-4\right]^{1/2}}{2\left(\frac{p}{m}\frac{p}{n}-1\right)}\right\}$$

$$\lambda = \frac{p \times \left\{\left[\left(1+\frac{p}{m}\right)\left(1+\frac{p}{n}\right)-2\right]^2 - 4\right\}^{\frac{1}{2}}}{\frac{p}{m}\frac{p}{n} - 1}, \lambda > 0$$

$$\varepsilon = \frac{x_{1z} + x_{-1z}}{2} - \frac{p}{2}\left(\frac{\frac{m}{p}+1}{\frac{m}{p}-1}\right)$$

and the parameters estimated for SL distribution are:

$$\eta = \frac{2z}{\ln\left(\frac{m}{p}\right)}$$

$$\gamma = \eta \times \ln\left[\frac{\frac{m}{p} - 1}{p\left(\frac{m}{p}\right)^{1/2}}\right]$$

$$\varepsilon = \frac{x_{1z} + x_{-1z}}{2} - \frac{p}{2}\left(\frac{\frac{m}{p} + 1}{\frac{m}{p} - 1}\right)$$

3.2. Resampling Mechanism

Instead of using fuzzy theories to increase the amount of observation, the statistical method is also employed as an alternative approach. Bootstrap is the most well-known synthetic sample generation method developed based on statistical resampling mechanism. Bootstrap randomly resample the original observations with replacement [7] which is differ from previous techniques which applies fuzzy theories. Bootstrap is introduced by Efron [21] in early 1979. However, it was not quite favorable because, according to Tsai and Li [5], bootstrap aimed to increase the amount of obtained observation rather than filling the information gaps of new observations. Although it is unfavorable technique, the learning accuracy increase in the prediction of bladder cancer cell for example [7].

3.3. Data Clustering

From the point of fuzzy membership generation, the fuzzy c-means (FCM) is similarly to MTD method. The difference is just in the calculation procedure. FCM is a data clustering technique proposed by Bezdek which applied fuzzy approach in clustering procedure. Generally, the FCM algorithm allows one piece of data to belong to different cluster with different membership degree of each cluster. Sezer used FCM as a method to produce synthetic training set for small dataset problem [22]. The author [22]found that most of the previous studies used the methods which are hard to understand. So, a better and easy method is more preferred to represent the information within the data. Hence, FCM is chosen as the alternative method to produce a synthetic data using centroid calculation scheme of FCM to reflect all the closeness relationships within the data. The instance is considered to have some closeness with another instance in another cluster if it is in the boundary of the cluster. Thus, centroid calculation scheme will identify whether the data is a good candidate to be synthetic data based on the centroid of the classes in different random samplings [22]. FCM need a specification of clusters number at the beginning. Each of cluster centroid is then randomly initialized. The iterative process is followed for computing the centroid of each cluster by using centroid calculation in Eq.12 until convergence criteria are met.

$$ce_i = \frac{\sum_{j=1}^{N}\left(u_{c_i}(x_j)\right)^m . x_j}{\sum_{j=1}^{N}\left(\mu_{c_i}(x_j)\right)^m} \tag{12}$$

where $u_{c_i}(x_j)$ is the membership degree of the *j*th instance "*x*" to cluster "*c*", *N* is the number of instances and ce_i is the centroid of *i*th cluster. After that, the membership of each instance to the cluster is calculated by using Eq.13.

$$u_{c_i}(x_j)=\frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_j-ce_i\right\|}{\left\|x_j-ce_k\right\|}\right)^{2/(m-1)}} \tag{13}$$

where "c" is the total of cluster number, ce_k is the centroid of the *k*th cluster while "*m*" is the degree of fuzziness.

4. Artificial Data Generation Concept to Prediction Model

From the findings, increasing the number of synthetic samples can increase the learning accuracy of a model. This is because the gaps between samples are filled with artificial data generation approach as discussed previously. We can conclude the overall concept of artificial data generation approach in small dataset is:

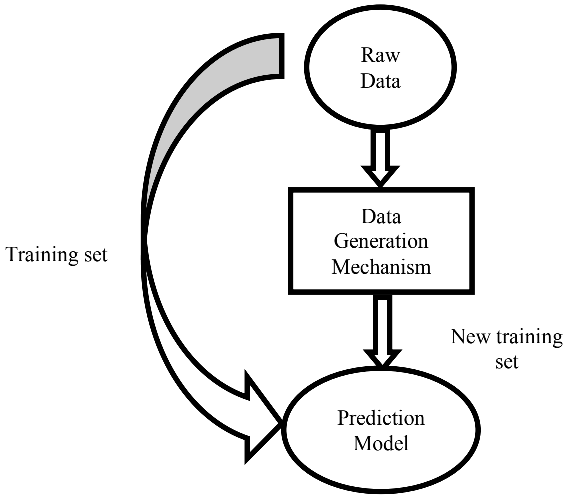

Figure 7. Data Generation Concept in Prediction Model

Referring to Fig.7, the concept generally can be explained from the data obtained, the data generation mechanism increases the number of training set to be used in the learning process to produce a model. The new form of training set then is combined with the original data to be used as a training sample of a model.

5. Conclusion

This paper presented a review on small dataset problem occur in certain areas where data is hard to obtain. Small dataset problem which cause poor learning performance and significant uncertainty to prediction model, is the reason why an efficient way is needed to build a precise prediction model. From literature, artificial data generation approach is one of the effective methods to cope with small dataset problem. Different techniques have been introduced to overcome a small dataset problem. Based on the techniques discuss earlier, MTD technique is widely used as a synthetic sample generation method to increase the amount of training sample. Besides, the improvement of MTD method has also been explored and introduced.

Acknowledgement

This work is funded by the Ministry of Higher Education Malaysia and Universiti Teknikal Malaysia Melaka (UTeM) through the PJP High Impact Research Grant [PJP/2016/FTMK/HI3/SO1473]

References

[1] D. C. Li, C. Sen Wu, T. I. Tsai, and F. M. Chang, "Using mega-fuzzification and data trend estimation in small data set learning for early FMS scheduling knowledge," *Comput. Oper. Res.*, vol. 33, no. 6, pp. 1857–1869, 2006.

[2] D. C. Li, C. C. Chang, C. W. Liu, and W. C. Chen, "A new approach for manufacturing forecast problems with insufficient data: The case of TFT-LCDs," *J. Intell. Manuf.*, vol. 24, no. 2, pp. 225–233, Jul. 2013.

[3] D. C. Li, C. Sen Wu, T. I. Tsai, and Y. S. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," *Comput. Oper. Res.*, vol. 34, no. 4, pp. 966–982, 2007.

[4] A. M. Al-bakary and S. H. Ali, "Genetic Programming Data Construction Method to Handle Data Scarcity Problem," no. February, pp. 1–10, 2010.

[5] C.-H. Tsai and D.-C. Li, "Improving Knowledge Acquisition Capability of M5' Model Tree on Small Datasets," *2015 3rd Int. Conf. Appl. Comput. Inf. Technol. Int. Conf. Comput. Sci. Intell.*, pp. 379–386, 2015.

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, vol. 8, no. 6. 2000.

[7] G. Chao, T. Tsai, T.-J. Lu, H. Hsu, B. Bao, W. Wu, M. Lin, and T. Lu, "A new approach to prediction of radiotherapy of bladder cancer cells in small dataset analysis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 7963–7969, 2011.

[8] R. Andonie, "Extreme data mining: Inference from small datasets," *Int. J. Comput. Commun. Control*, vol. 5, no. 3, pp. 280–291, 2010.

[9] T. I. Tsai and D. C. Li, "Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1293–1300, 2008.

[10] P. Niyogi, "niyogi-poggio-ProcIEEE-1998.pdf," *Proc. IEEE 86(11)*, 1988.

[11] C. Huang and C. Moraga, "A diffusion-neural-network for learning from small samples," *Int. J. Approx. Reason.*, vol. 35, no. 2, pp. 137–161, 2004.

[12] C. Huang, "Principle of information diffusion," *Fuzzy Sets Syst.*, vol. 91, no. 1, pp. 69–90, 1997.

[13] D. C. Li, C. Wu, and F. M. Chang, "Using data-fuzzification technology in small data set learning to improve FMS scheduling accuracy," *Int. J. Adv. Manuf. Technol.*, vol. 27, no. 3–4, pp. 321–328, 2005.

[14] D.-C. Li, I.-H. Wen, and W.-C. Chen, "A novel data transformation model for small data-set learning," *Int. J. Prod. Res.*, vol. 7543, no. June, pp. 1–11, 2016.

[15] D.-C. Li, W.-T. Huang, C.-C. Chen, and C.-J. Chang, "Employing virtual samples to build early high-dimensional manufacturing models," *Int. J. Prod. Res.*, vol. 51, no. 11, pp. 3206–3224, 2013.

[16] D. C. Li and I. H. Wen, "A genetic algorithm-based virtual sample generation technique to improve small data set learning," *Neurocomputing*, vol. 143, pp. 222–230, 2014.

[17] D.-C. Li, C.-C. Chang, and C.-W. Liu, "Using structure-based data transformation method to improve prediction accuracies for small data sets," *Decis. Support Syst.*, vol. 52, no. 3, pp. 748–756, 2012.

[18] J. F. Slifker and S. S. Shapiro, "The johnson system: selection and parameter estimation," *Technometrics*, vol. 22, no. 2, pp. 239–246, 1980.

[19] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1. pp. 1–26, 1979.

[20] E. a. Sezer, H. a. Nefeslioglu, and C. Gokceoglu, "An assessment on producing synthetic samples by fuzzy C-means for limited number of data in prediction models," *Appl. Soft Comput. J.*, vol. 24, pp. 126–134, 2014.