The Characteristics of Learning in Limited Data and the Comparative Assessment of Learning Methods

FENGMING M. CHANG
Department of Information Science and Applications
Asia University
Wufeng, Taichung 41354
Taiwan
paperss@gmail.com

Abstract: - Many studies about learning in limited data were made in recent years. Without double, small data set learning is a challenging problem. Information in data of small size is scarce and has some learning limit. While discussing the learning accuracy in limited data, different classification method causes different results for different data because each classification method has its property. A method is the best solution for one data but is not the best for another. Therefore, this study analyzes the characteristics of small data set learning by the comparison of classification methods. The Mega-fuzzification method for small data set learning is applied mainly. The comparison of different classification methods for small data set learning with several kinds of data is also presented.

Key-Words: - Small data set, Mega-fuzzification, Machine learning, Classification method, Paucity of data

1 Introduction

Some satisfactory learning results can be found in the conditions of large sample size but learning accuracy is limited using small size sample. Hence learning in limited data becomes a challenging problem. However, there have been many applications of neuro-fuzzy methods in the condition of the paucity of data [1-7]. In the early time of a new system development, data on hand are not enough. Therefore, data characteristics such as data distribution, mean, and variance are unknown. In such a limited data environment, a decision is hard to make under the limit data condition. On the other hand, learning methods also affect learning results. Each classification method has its property. A method is the best solution for one data but is not the best for another because each set of data does not satisfy each method's assumptions [8].

This study discusses and presents the characteristics of small data set learning by the comparison of different kinds of classification methods. Several data, such as data of the chaotic Mackey-Glass differential delay equation, credit card, Monk3, Hayes-Roth, and Nbus data are used to realize the characteristics of small data set learning. It is shown that some classification methods, such as neuro-fuzzy and mega-fuzzification methods are hard to perform for large number of attributes data. Therefore other methods than fuzzy methods are used for comparison.

2 Relative Works

There were some developments for learning in limited data, they are reviewed in this section.

The computational learning theory develops mathematical models to describe the learning data size and number of training in machine learning [9]. It can also be applied to small data set learning. However, it still leaves some practical problems. Although it offers a probably approximately correct (PAC) model to estimate the relation about predict accuracy and sample size, it is hard to calculate the sample space in the model. However, it indeed builds a theoretical model to describe the machine learning problem.

Using artificial data to increase the sample size to increase the predict accuracy in machine learning is a good idea. The virtual data concept is used in many small data set learning methods. It was first proposed by Niyogi et al. [10] in the study of human face recognition. In their study, virtual face recognition data from any other direction can be generated using given view data through mathematical functions. With these virtual samples, a learning machine can verify an instance more precisely.

Functional Virtual Population (FVP) was proposed to increase production scheduling predict accuracy in dynamic manufacturing environment by Li et al. [1]. In their study, the domains of the

system attributes are expanded by the FVP algorithm to generate a number of virtual samples. Using these virtual samples, a new scheduling knowledge is constructed and the prediction accuracy is increasing.

Huang and Moraga [11] combined the principle of information diffusion [12] with a traditional neural network, named diffusion-neural-network (DNN), for function learning. According to the results of their numerical experiments, the DNN improved the accuracy of the back-propagation neural networks (BPNN).

Mega-fuzzification method was proposed for the purpose of solving the prediction of the best strategy problem in the Flexible Manufacturing System (FMS) when data are small [2, 4, 6]. In the studies of [2, 4, 6] for the Mega-fuzzification, several concepts were offered, such as data fuzzified, continuous data band, domain external expansion, and data bias adaptation.

The concept of the continuous data band was first proposed by Li, et al. [2]. Such a data continuing technology aims to fill the gaps between individual data and make incomplete data into the more complete status as presented by Huang and Moraga [11-12].

Furthermore, in the studies of Li, et al. and Chang [2, 4, 6], the domain range external expansion concept was also proposed into the procedure of the continuous data band method. In addition to filling the data gaps within the data set, the purpose of domain external expansion is also to add more data outside the current data limits, because possible data are expected not only inside but also outside the current data range.

3 Learning Methods Used

Some machine learning classification methods are used in this study for comparisons.

3.1 Bayesian networks

Bayesian networks (BN) are graphical models that combine both graph and probability theory [13]. BN is defined as a directed acyclic graph or a probabilistic graphical model that presents a set of variables and their causal influences. The causal dependencies between the variables are expressed by conditional probability distributions. Numeric values are usually assumed that they are normal or Gaussian distribution in BN.

3.2 Decision tree ID3 and C4.5

ID3 is a decision tree method [14]. A decision tree performs categorical split following the value number of input attributes. A decision tree is a predictive model mapping from observations to predict the target values. C4.5 is an improved version of ID3. It was written by C. It can deal with both symbolic and numeric values of input attributes and then outputs a classification tree.

3.3 Support vector machine

SVM is developed to classify data points into linear separable data sets. It is based on statistical theory and was first developed for binary classification problems. SVM tries to find out the optimal hyperplane by minimizing an upper bound of the errors and maximizing the distance, margin, between the separated hyperplane and data [15]. A maximal separating hyperplane is built by SVM to map input vectors to a higher dimensional space. Two parallel hyperplanes are built and the data are separated on each side of the hyperplane. Given training dataset (x_i, y_i) , $i = 1,\ldots,k$, where $x_i \in R^n$ and $y \in \{1,-1\}^k$, SVM tries to find out the optimal solution problem using the following form mainly:

$$\min_{w,b,\varepsilon} \ \frac{1}{2} w^T w + C\sum_{i=1}^{k} \varepsilon_i$$

subject to

$$y_i(w^T \varphi(x_i) + b) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0$$

3.4 Artificial neural network

An ANN is a computational model that consists of nodes that are interconnected. Each node uses a mathematical model or computational model to compute its output from its inputs. In more practical terms, an ANN is a non-linear tool. It can model complex relationships between inputs and output data. In most of times, an ANN is an adaptive system that can adjust the parameters to improve its performance for a given task. There are three types of neural network learning algorithms: supervised, reinforcement, and unsupervised leaning. Back-propagation neural network (BPN) is the best known supervised learning algorithm.

3.5 Neuro-fuzzy

A neuro-fuzzy performs neural learning using fuzzy typed numbers. Given a set of input and output data, the ANFIS can constructs a fuzzy inference system with membership functions values adapted using a backpropagation algorithm or in

combination with a least square method. The adaptation function of the ANFIS, a FNN tool, provides the machine learning system with FNN characters.

The basic model of the FNN is the Sugeno fuzzy model [16]. In the model, assuming x and y are two input fuzzy sets and z is the output fuzzy set and the fuzzy if-then rules is formatted as:

If $x = P$ and $y = Q$ then $z = f(x, y)$

Fig. 1 shows FNN structure with a five-layer artificial neural network.

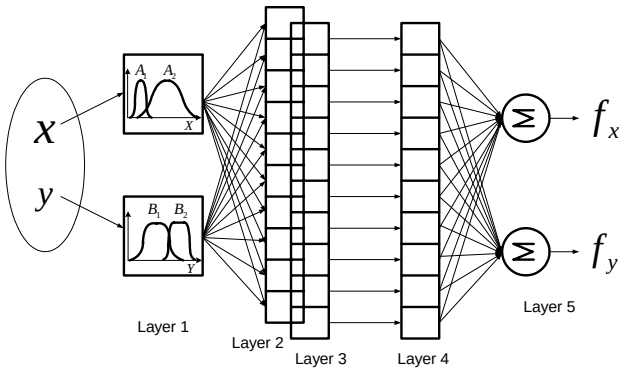

Fig. 1. The FNN structure.

4 Data used in this study

In this study, 5 data sets are used for comparison: chaotic data by Mackey-Glass differential delay equation [17], credit, Monk3, Hayes-Roth, data download from University of California Irvine [18], and Nbus data from Rose2 software [19-21]. Mackey-Glass differential delay equation was used in the introduction of ANFIS by Jang [16]. Credit data are credit approval data with total 690 instances that has 15 inputs and one output attributes. Monk3 data are donated by Sebastian Thrun of Carnegie Mellon University at Pittsburgh with 432 instances, 6 inputs and 1 output attributes. Hayes-Roth data have 132 instances, 4 inputs and 1 output attributes. As well as Nbus data have 76 instances, 8 input and 1 output attributes. Some of these data can be performed in neuro-fuzzy, mega-fuzzification, and other learning methods, some of them have a large number of attributes and can not be performed in fuzzy based methods.

5 The Proposed Method

Following the previous research in literature review, the proposed method invents a concept of transforming crisp data into continuous in order to create virtual data and to fill the gaps between crisp data. Fuzzy theory and FNN methods are applied. Data range expansion method is proposed in this study also.

5.1 Continuous data

To fill the gaps between crisp data, crisp learning data are transformed into continuous data as Figure 1 illustrates. In Fig. 2, there are five original crisp data. When these data are transformed into continuous type, virtual data between the crisp data are thus generated.

Fig. 2. Crisp data are transformed into continuous

5.2 Data effect estimation

The data effects have to be estimated also. In this research, continuous data are presented in fuzzy membership function forms. The fuzzy membership function can be a general bell, triangle, or even anomalous type and is the data's effective weight in the FNN learning later. Most of the time, an asymmetric fuzzy membership function is initiated. For example, a triangular fuzzy membership can be:

$$\mu_{\tilde{A}}(x_i) = \begin{cases} 0 & , x_i < \min \\ \dfrac{x_i - \min}{mid - \min} & , \min \le x_i \le mid \\ \dfrac{\max - x_i}{\max - mid} & , mid \le x_i \le \max \\ 0 & , x_i > \max \end{cases}$$

where $\min$ is the minimum value and $\max$ is the maximum value of the crisp learning data, and $mid = \dfrac{\min + \max}{2}$.

5.3 Data range external expansion

When crisp are transformed into continuous, boundaries of the continuous data band are minimum and maximum value of the original crisp data. However, the real data range is possible outside this data band. In order to build up real knowledge, the data band is externally expanded to the possible range in this study as illustrated in Fig. 3.

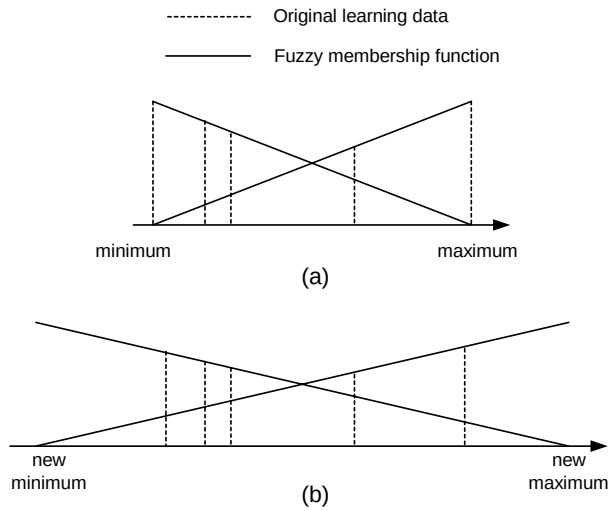

Fig. 3. Two triangular membership function values: (a) before, (b) after domain range external expansion

5.4 FNN learning

Because data have been transformed into a continuous data band, FNN are used in this study. During the learning process, the shape of fuzzy membership function is adapted to fit the best learning results.

6 Characteristics of Learning in Limited Data

In this section, chaotic data are used to describe some characteristics of learning in limited data. More characteristics are presented in the next section.

6.1 Chaotic data

A data set of time series values used in [16] is generated by the chaotic Mackey-Glass differential delay equation [17] is defined as:

$$\dot{y}(t) = \frac{0.2y(t-\tau)}{1+y^{10}(t-\tau)} - 0.1y(t)$$

For the purpose of explaining the learning procedure, 1000 data, $t = 118$ to 1117, was generated. The format of input and output datum elements is set as:

$$[\, y(t-18),\ y(t-12),\ y(t-6),\ y(t),\ y(t+6),\,]$$

where the first 4 items are the inputs, such as Input1= $y(t-18)$, and the last item is the output. The If-Then rule is:

If
(Input1=a) & (Input2=b) & (Input3=c) & (Input4=d)
Then
(Output=k)

6.2 Data size and type

In Table 1, using chaotic data, results are shown from using different size of random training data selected from the data, and different membership function types of input data including Generalized Bell-shaped (Gbellmf), Trapezoidal-shaped (Trapmf), and Triangular-shaped (Trimf) membership functions. The measure way of error is root mean squared error (RMSE). The define of RMSE is:

$$RMSE = \sqrt{\frac{\sum_i (p_i - a_i)}{n}}$$

Where p_i is the prediction value and a_i is the actual value, $1 \le i \le n$. For each type of the membership functions, two kinds of data domains were tested: without expansion and external expansion. External expansion is a procedure in the Mega-fuzzivication Without expansion, the data domains are set as the input ranges of selected training data, one domain for one input variable; in external expansion, data domains are externally expanded to [0.2192, 1.3137] for all the input variables. In addition, the results in Table 1 are plotted as curves in Fig. 4.

Table 1. RMSE values of different sizes and types of training data.

Training data size	Gbellmf		Trapmf		Trimf	
	Without expansion	External expansion	Without expansion	External expansion	Without expansion	External expansion
5	0.5028	0.1322	0.5000	0.3080	0.4791	0.2485
10	0.1181	0.0972	0.2265	0.2464	0.2240	0.0839
20	0.0499	0.0438	0.0647	0.0544	0.0790	0.0780
30	0.0218	0.0216	0.0485	0.0519	0.0487	0.0291
40	0.0161	0.0143	0.0322	0.0231	0.0194	0.0217
50	0.0162	0.0136	0.0324	0.028	0.0220	0.0204
60	0.0094	0.0132	0.0178	0.0172	0.0143	0.0121
70	0.0138	0.0089	0.0142	0.0161	0.0152	0.0127

In Fig. 4, the FNN learning results of using domain external expansion technology is better than that of the FNN without expansion. When the learning data set is smaller, the efficiency in reducing RMSE is clearer. When the size of learning data is five, the degree of improvement for prediction accuracy is the largest. As the size of learning data increases, the RMSE values between the values without expansion and the values with external expansion are getting closer. When the data size is approaching 30, it's large enough to indicate the actual domain range.

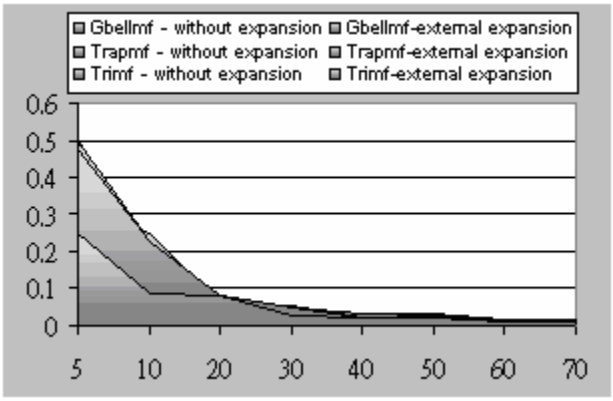

Fig. 4. The RMSE curves comparison using chaotic data.

6.3 Variability of the learning results

In the calculations above, only one set of data is used in learning. However, learning results can be variable and unsteady on randomly selected data sets, especially in the very small learning data set size.

Figure 5 is the RMSE values of various training data sizes using chaotic data. For each data size, ten sets of the training data were chosen randomly from the original data. For each set of data, both methods (with and without external expansion) were applied in the FNN learning. Fig. 5 illustrates the curves of RMSE means and Fig. 6 presents the curves of standard deviation. In Fig. 5, while data set size is increasing, the RMSE mean is decreasing, and when data size is smaller, the efficiency of the system with external expansion is better than without. In Fig. 6, it's clear that the standard deviation of RMSE value is decreasing while the data size is increasing. In general, although the accuracies vary in small data set learning, the proposed method can decrease the learning errors.

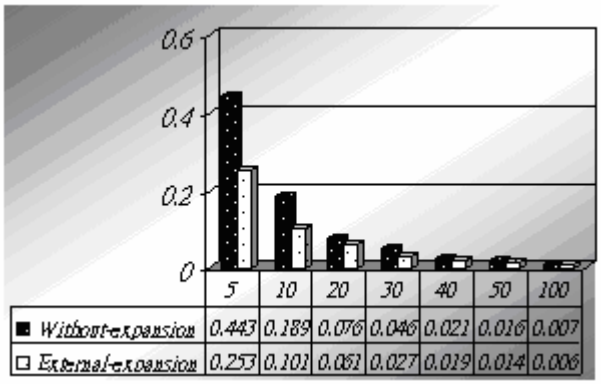

Fig. 5. RMSE mean curves.

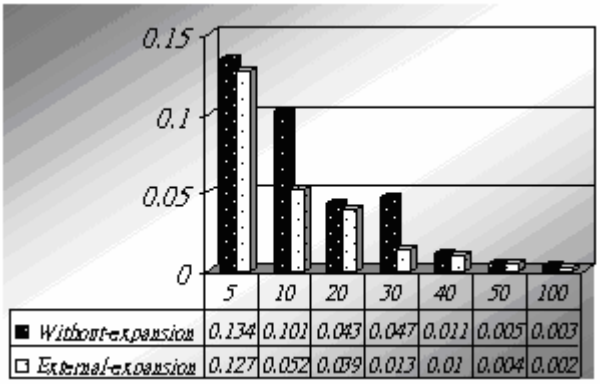

Fig. 6. Standard deviation of RMSE values.

6.4 Comparisons with other methods

In this subsection, 15 data are chosen for learning, and different kinds of classification methods are applied. The result is shown in Fig. 7. SVM method has the largest RMSE value and Mega-fuzzification method has the lowest.

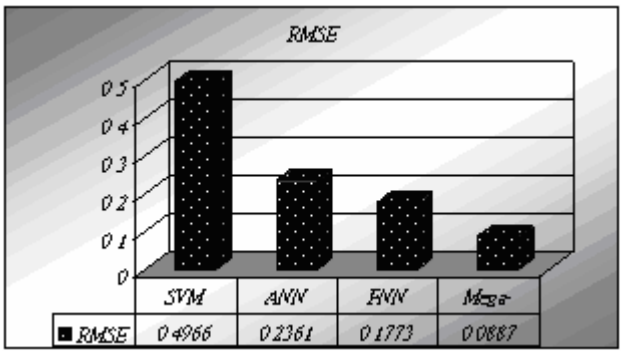

Fig. 7. The comparison of classification methods.

6.5 A challenge of number of attribute

In our experiment, data with more than six input attributes may become a challenge to neuro-fuzzy or mega-fuzzification learning. Therefore, it presents a challenge in small data set learning using mega-fuzzificaiton. Next subsection, a credit card data

with 15 input attributes are presented for learning comparison without neuro-fuzzy and mega-fuzzification.

6.6 Learning property and credit card data

In previous subsection, some classification methods can not used to compare because they can only be used for nominal or integer output classes. For this reason, another data are used in the following subsection. The data is credit approval data download from University of California. There are total 690 instances. Each instance has 15 input and one output attributes. However, a new problem is raised. Too many input attributes are fail to perform FNN in the Mega-fuzzificaiton. In next subsection, some classification methods are compared but without FNN and Mega-fuzzification.

6.7 Comparisons using credit card data

In this subsection, Bayesian, C4.5, SVM, and ANN methods are compared using the credit card data. Fig. 8 to 10 are the comparison results of accuracy, learning time, and RMSE.

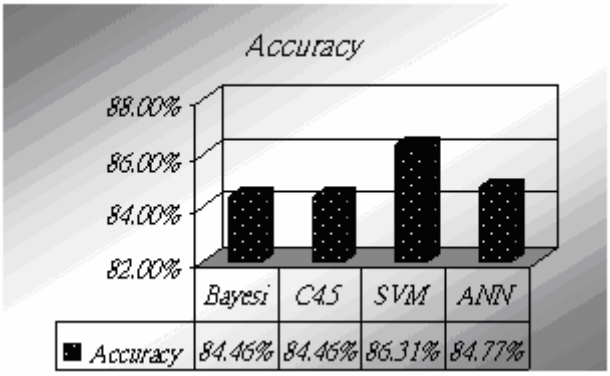

Fig. 8. The comparison of prediction accuracy.

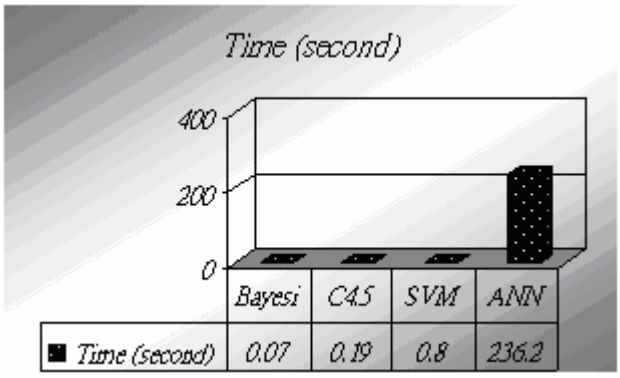

Fig. 9. The comparison of learning time period.

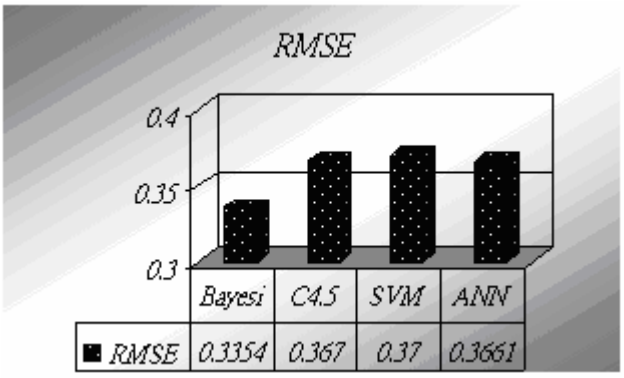

Fig. 10. The comparison of root mean squared error.

6.8 Monk3, Hayes-Roth, and Nbus data

Two data sets from Department of Information and Computer Sciences, University of California Irvine [18] and one from Rose2 software [19-21] for rough set are used in this study: Monk's problems, Hayes-Roth database, and Nbus database. Monk3 data are one of the Monk's problem data donated by Sebastian Thrun of Carnegie Mellon University at Pittsburgh. They have 432 instances, 6 inputs and 1 output attributes, and are good data for machine learning tests [19]. Hayes-Roth data are created by Barbara and Frederick Hayes-Roth [21] which have 132 instances, 4 inputs and 1 output attributes. As well as Nbus data have 76 instances, 8 input and 1 output attributes.

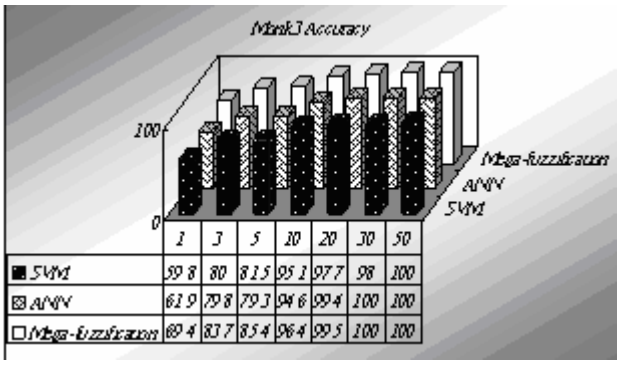

Fig. 11. The comparison of Monk3 accuracy.

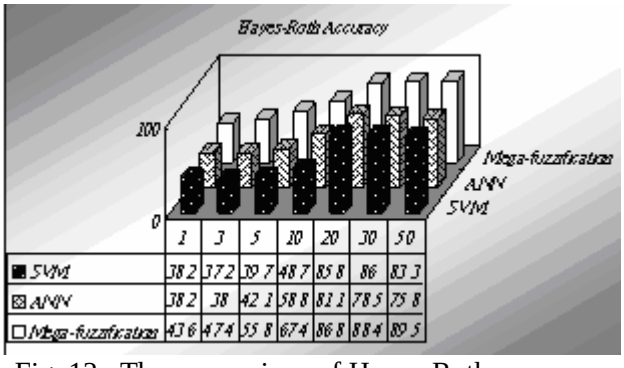

Fig. 12. The comparison of Hayes-Roth accuracy.

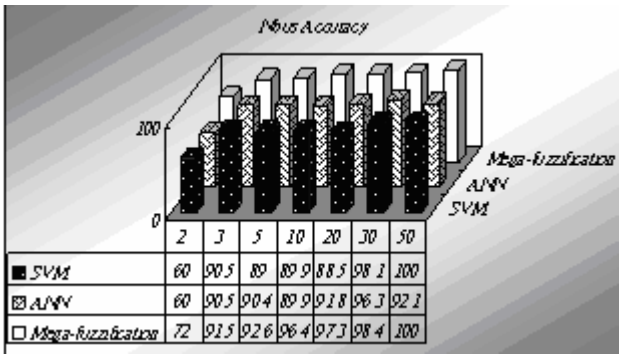

Fig. 13. The comparison of Nbus accuracy.

7 Further Study

A large number of attributes becomes a block to neuro-fuzzy and mega-fuzzificaiton. Unfortunately, most of the data in the real world have a large number of attributes and methods for attribute reduction become a challenging study in the further study.

8 Conclusions

When data size on hand is small, some informations are missing and the predict accuracy is low. In order to increase the learning accuracy, the mega-fuzzification method was proposed to solve the problem. The Mega-fuzzification method indeed can increase the prediction accuracy than other method. After testing different kinds of data bases, the results of this study indicate that when data size increases, the prediction accuracy increases, RMSE value decreases, and the variability of learning results decreases. However, the fuzzy based methods, including neuro-fuzzy and mega-fuzzificaton can not performed well with data have a large number of attributes. This is a limit for the fuzzy based methods. In this study, several learning methods with different data are compared. It indicates that the mega-fuzzification method has better accuracy than other methods. Unfortunately, when data have a large number of attributes, other methods are used to instead of mega-fuzzification. The study of attributes reduction becomes a challenging topic in the further. Therefore, the Mega-fuzzification can increase the predict accuracy but still has some limit.

Acknowledgement

Thanks are due to the support in part by the National Science Council of Taiwan under Grant No. NSC 95-2416-H-272-002.

References:

[1] D. C. Li, L. S. Chen, and Y. S. Lin, "Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments," *International Journal of Production Research*, vol. 41, no. 17, pp. 4011-4024, 2003.

[2] D. C. Li, C. Wu, and F. M. Chang, "Using Data-fuzzification Technology in Small Data Set Learning to Improve FMS Scheduling Accuracy," *International Journal of Advanced Manufacturing Technology*, Vol. 27, No. 3-4, pp. 321-328, 2005.

[3] F. M. Chang and M. Y. Chiu, "A Method of Small Data Set Learning for Early Knowledge Acquisition," *WSEAS Transactions on Information Science and Applications*, Vol. 2, No. 2, pp.89-94, 2005

[4] F. M. Chang, "An intelligent method for knowledge derived from limited data," *2005 Proceeding - IEEE International Conference on Systems, Man, and Cybernetics*, The Big Island, Hawaii, USA, Oct. 10-12, pp.566-571, 2005.

[5] F. M. Chang, "Determination of the Economic Prediction in Small Data Set Learning," *WSEAS Transactions on Computers*, Vol. 5, No. 11, pp.2743-2750, 2006.

[6] D. C. Li, C. Wu, and F. M. Chang," Using data continualization and expansion to improve small data set learning accuracy for early FMS scheduling,," *International Journal of Production Research*, Vol. 44, No. 21, pp.4491-4509, 2006.

[7] F. M. Chang and Y. C. Chen, "A Frequency Assessment Expert System of Piezoelectric Transducers in Paucity of data," *Expert Systems with Applications*, Vol. 36, No. 2**,** to be published in December 2008.

[8] M. Y. Kiang, "A comparative assessment of classification methods," *Decision Support Systems*, Vol. 35, pp.441-454, 2003.

[9] M. Anthony, N. Biggs, *Computational Learning Theory*. Cambridge University Press, 1997.

[10] P. Niyogi, F. Girosi, P. Tomaso, "Incorporating prior information in machine learning by creating virtual examples," *Proceeding of the IEEE*, pp. 275-298, 1998.

[11] C. Huang, C. Moraga, "A diffusion-neural-network for learning from small samples," *International Journal of Approximate Reasoning*, vol. 35, pp. 137-161, 2004.

[12] C. F. Huang, "Principle of information diffusion," *Fuzzy Sets and Systems*, vol. 91, pp. 69-90, 1997.

[13] E. J. M. Lauría, J. Duchessi, "A methodology for developing Bayesian networks: An application to information technology (IT) implementation," *European Journal of Operational Research* Vol. 179, No.1, pp.234-252, 2007.

[14] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys* Vol. 28, No. 1, pp.71-72, 1986.

[15] K. Seo, "An application of one-class support vector machines in content-based image retrieval," *Expert Systems With Applications* Vol. 33, No. 2, pp.491-498, 2007.

[16] J.-S. R. Jang, "ANFIS: Adaptive-Network-based Fuzzy Inference Systems," *IEEE Transactions on System, Man, and Cybernetics*, vol. 23, no.3, pp. 665-685, 1993.

[17] M. C. Mackey and L. Glass, "Oscillation and chaos in physiological control systems," *Science*, vol. 197, pp. 287-289, 1977.

[18] UCI Machine Learning Repository, http://mlearn.ics.uci.edu/MLRepository.html

[19] B. Predki, R. Slowinski, J. Stefanowski, R. Susmaga, and Sz. Wilk, "ROSE - Software Implementation of the Rough Set Theory," In: L. Polkowski, A. Skowron, eds, "Rough Sets and Current Trends in Computing," *Lecture Notes in Artificial Intelligence*, vol. 1424, pp. 605-608,. 1998.

[20] B. Predki and Sz.Wilk, "Rough Set Based Data Exploration Using ROSE System," In: Z. W. Ras, A. Skowron, eds, "Foundations of Intelligent Systems," *Lecture Notes in Artificial Intelligence*, vol. 1609, pp.172-180, 1999.

[21] B. Hayes-Roth, and F. Hayes-Roth, "Concept learning and the recognition and classification of exemplars," *Journal of Verbal Learning and Verbal Behavior*, vol. 16, pp. 321-338, 1977.