Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080346

Evolutionary Instance Selection Algorithm based on Takagi-Sugeno Fuzzy Model

Sang-Hong Lee[1] *and Joon S. Lim*[2,*]

[1] Department of Computer Science & Engineering, Anyang University, Republic of Korea
[2] IT College, Gachon University, Republic of Korea

Received: 17 Jun. 2013, Revised: 22 Oct. 2013, Accepted: 23 Oct. 2013
Published online: 1 May. 2014

Abstract: In this study, we propose evolutionary instance selection based on the Takagi-Sugeno (T-S) fuzzy model. The previous neural network with weighted fuzzy membership functions (NEWFM) supports feature selection; thus, it enables the selection of minimum features with the highest performance. The enhanced NEWFM supports a weighted mean defuzzification in the T-S fuzzy model with a confidence interval in the normal distribution; thus, it enables the selection of minimum instances with the highest performance. The enhanced NEWFM has two stages; feature selection is performed in the first stage, whereas instance selection is performed in the second stage. The performance of the enhanced NEWFM is compared with that of the previous NEWFM. In addition, McNemar's test reveals a significant difference between the performances of both NEWFMs ($p < 0.05$).

Keywords: Instance selection, feature selection, Takagi-Sugeno fuzzy model, McNemar's test, normal distribution

1 Introduction

In recent years, there has been an increase in the volume of data that machine learning methods are required to manage [1]. Moreover, the large amount of data that is available in any research field poses new problems for data mining and knowledge discovery methods. Data reduction is a data preprocessing task that can be applied to ease the problem of dealing with large amounts of data [10]. The best known data reduction processes are feature selection and instance selection. Too many features may result in inefficiency in terms of memory and time consumption, and they may even be inapplicable. Besides, irrelevant data may confuse algorithms, resulting in false conclusions, and hence, poor results. Feature selection is widely used for eliminating redundant or irrelevant features, and it enhances performance by improving accuracy and reducing operation costs using minimum features [4][11,12,13,14]. The objective of instance selection is to isolate the smallest set of instances that enable a data mining algorithm to determine the class of a query instance with the same quality as the initial data. Efforts to select relevant instances from initial data have stemmed from the need to reduce high storage requirements and computational load [5].

The previous neural network with weighted fuzzy membership functions (NEWFM) supports feature selection; thus, it enables the selection of minimum features with the highest performance [6,7] [15,16]. From the initial features, the minimum features that provide the highest performance are selected using a non-overlap area distribution measurement method. In this study, we propose an enhanced NEWFM that supports a weighted mean defuzzification in the Takagi-Sugeno (T-S) fuzzy model with a confidence interval in the normal distribution; thus, the enhanced NEWFM enables the selection of minimum instances with the highest performance. The remainder of this paper is organized as follows. In Section 2, we review the experimental data and related studies. In Section 3, we describe how the previous NEWFM selects minimum features using a non-overlap area distribution measurement method. In addition, we describe how the enhanced NEWFM selects minimum instances using a weighted mean defuzzification in the T-S fuzzy model with a confidence interval in the normal distribution. In Section 4, we analyze the experimental results of the instance selection algorithms proposed in this study. Finally, the conclusions are stated in Section 5.

* Corresponding author e-mail: jslim@gachon.ac.kr

2 Data Description and Related Work

2.1 Experimental data

The proposed instance selection algorithm was applied to data provided by the UCI repository of machine learning databases, known as the credit approval and heart disease databases. The credit approval and heart disease databases are mixed databases. The credit approval database has 15 features, nine of which are categorical, while six are numerical. The heart disease database has 13 features, seven of which are categorical, while six are numerical. There are 690 instances with missing feature values in the credit approval database and 303 instances with missing feature values in the heart disease database. In this study, we consider the 653 and 297 instances without missing feature values in the credit approval database and heart disease database, respectively.

2.2 Takagi-Sugeno fuzzy model

The Takagi-Sugeno (T-S) fuzzy model [9] is a powerful tool for modeling complex nonlinear systems; its consequent parts perform linear functions that can be regarded as an expansion of a piecewise linear partition.

$$R^i : IF\, x_1 is A_1^i, \ldots, x_m is A_m^i\, THEN\, y^i = a_0^i + \ldots + a_m^i x_m \quad (1)$$

$$y = \frac{\sum_{i=1}^{c} w^i y^i}{\sum_{i=1}^{c} w^i} \; where\, w^i = Min\{A_1^i(x_1), \ldots, A_m^i(x_m)\} \quad (2)$$

where $R^i (i = 1, 2, \ldots, c)$ denotes the i^{th} fuzzy rule, $x_i (i = 1, 2, \ldots, m)$ are the input variables, y^i are the rule output variables, $A_1^i, \ldots, A_m^i$ are fuzzy sets of the ith rule for x_i, and $a_0^i, \ldots, a_m^i$ are the parameter sets in the consequent part.

2.3 Normal distribution

In probability theory, the normal distribution is a continuous probability distribution that is often used as a first approximation for describing real-valued random variables that tend to cluster around a single mean value. The normal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

where μ is the mean and σ^2 is the variance. The normal distribution is a convenient choice for modeling a large variety of random variables encountered in practice, using a confidence interval that represents the area under the bell curve between $\mu - n\sigma$ and $\mu + n\sigma$ in Figure 2.1.

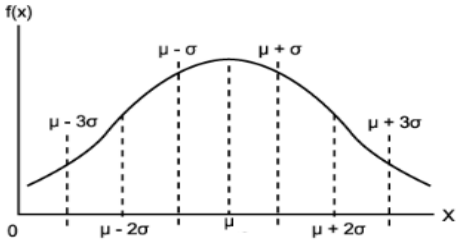


Figure 2.1: Graph of normal distribution.

2.4 Statistical significance

In this study, McNemar's test is employed to determine whether the difference between the performances of the two classification algorithms, A_1 and A_2, is statistically significant, using the same data in Table 1. The test is based on the chi-squared (χ^2) statistic; it is computed from two error matrices and given by [8]

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} \quad (4)$$

where n_{12} denotes the number of instances that are wrongly classified by algorithm A_1 but correctly classified by algorithm A_2, whereas n_{21} denotes the number of instances that are correctly classified by algorithm A_1 but wrongly classified by algorithm A_2. At the significance level of 0.05, the hypothesis H_0 that there is no difference between the performances of the two algorithms, A_1 and A_2, is rejected if χ^2 is greater than 3.84.

Table 1. McNemar's test contingency table

		A_1		
		Correct	Incorrect	Total
A_2	Correct	n_{11}	n_{12}	$n_{11}+n_{12}$
	Incorrect	n_{21}	n_{22}	$n_{21}+n_{22}$
Total		$n_{11}+n_{21}$	$n_{12}+n_{22}$	

3 Neural network with weighted fuzzy membership function (NEWFM)

A neural network with weighted fuzzy membership functions (NEWFM) is used to select minimum features in the first stage and minimum instances in the second stage to classify C_1 and C_2 into the class nodes shown in Figure 3.1. The previous NEWFM has the advantage of a non-overlap area distribution measurement method that enables the selection of minimum features [6,7] [15]. The NEWFM is a supervised classification neuro-fuzzy system that uses the bounded sum of weighted fuzzy membership functions (BSWFMs).

In this study, an enhanced NEWFM is proposed for selecting minimum instances on the basis of the Takagi-Sugeno fuzzy model. The structure of the enhanced NEWFM, shown in Figure 3.1, consists of four

layers, i.e., the input, hyperbox, class, and Takagi-Sugeno (T-S) layers. Instance selection is performed after feature selection, as shown in Figure 3.1.

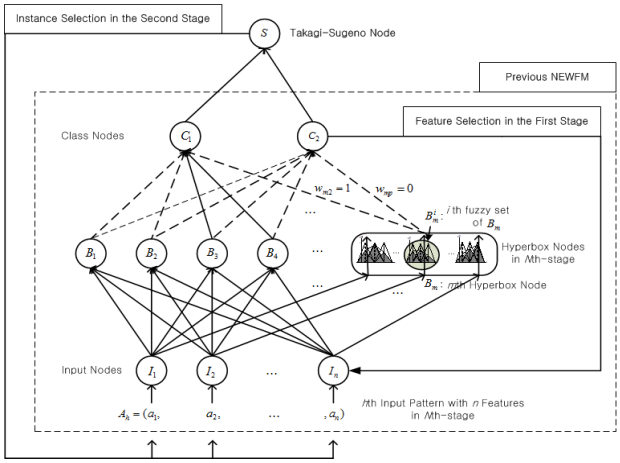


Figure 3.1: Structure of the NEWFM.

3.1 Feature selection in the first stage

The previous NEWFM provides feature selection with the highest performance using a non-overlap area distribution measurement method. The method measures the degree of salience of the *i*th feature on the basis of the non-overlap area distribution using the equation [6,7] [15].

$$f(i) = (Area_A^i + Area_B^i)^2 / \frac{1}{(1+e^{-|Area_A^i - Area_B^i|})} \quad (5)$$

where $Area_A$ and $Area_B$ are the A class superior area and the B class superior area, respectively. Examples of $Area_A$ and $Area_B$ are shown in Figure 3.2. The larger the value of $f(i)$, the stronger is the feature characteristic implied.

The features shown in Figure 3.2 have two BSWFMs, and they are obtained during the training process of the NEWFM program. The two BSWFMs graphically demonstrate the difference between class A and class B for each input feature. Figure 3.3 shows examples of good and bad candidate features selected from the initial features. Table 2 shows the minimum features selected from the initial features. 14 minimum features were finally selected from the 15 initial features of the credit approval database and 9 minimum features were finally selected from the 13 initial features of the heart disease database.

3.2 Instance selection in the second stage

The enhanced NEWFM proposed in this study provides instance selection with the highest performance by using a weighted mean defuzzification in the T-S fuzzy model with a confidence interval in the normal distribution. Instance selection involves three steps, i.e., defuzzification, normal distribution, and confidence interval selection.

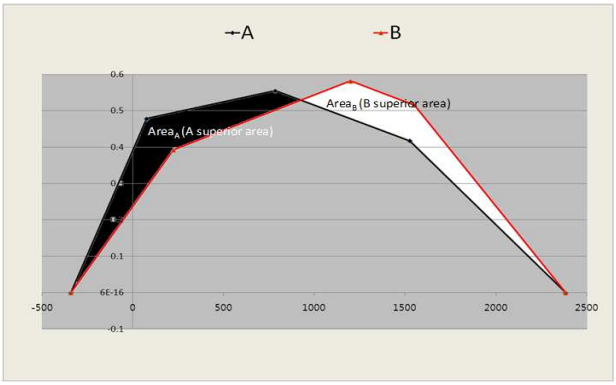
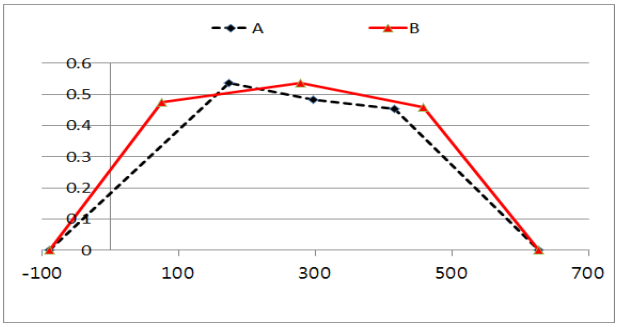
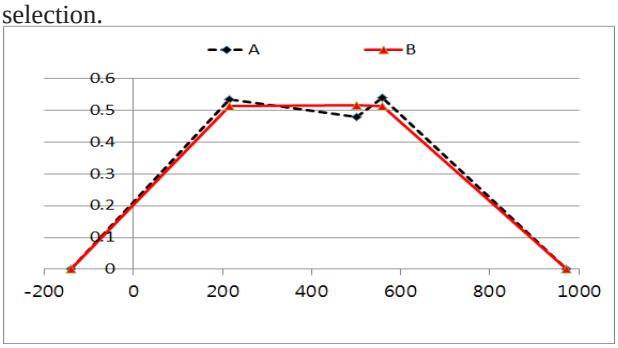


Figure 3.2: $Area_A$ (black) and $Area_B$ (white) in the non-overlap area distribution measurement method.



(a) An example of a good candidate feature for feature selection.



(b) An example of a bad candidate feature for feature selection.

Figure 3.3: Examples of good and bad candidate features for feature selection.

(A) Step 1 (Defuzzification): After feature selection is completed in the first stage, a weighted mean defuzzification is calculated using the T-S fuzzy model. Figure 3.4 shows the calculation of the weighted mean defuzzification using equation (2.2) and the BSWFM in the first stage.

(B) Step 2 (Normal distribution): Figure 3.5 shows the normal distribution of all weighted mean defuzzifications that correspond to the initial instances; the normal

distribution shows the distribution of all weighted mean defuzzifications that are correctly or incorrectly classified by feature selection in the first stage.

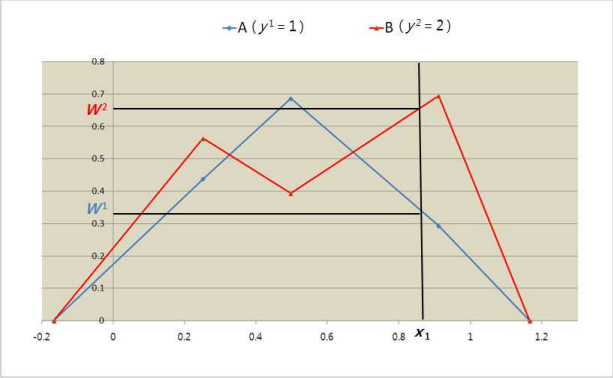


Figure 3.4: Example showing calculation of weighted mean defuzzification.

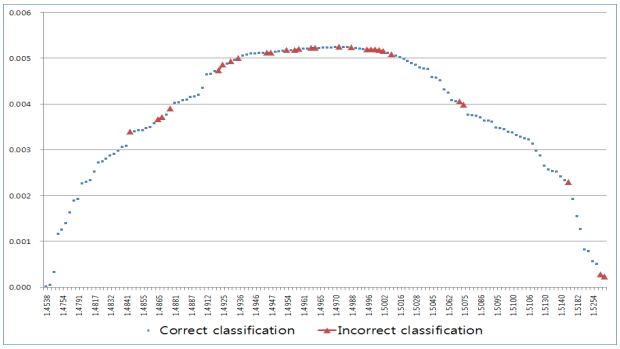


Figure 3.5: Normal distribution of weighted mean defuzzifications.

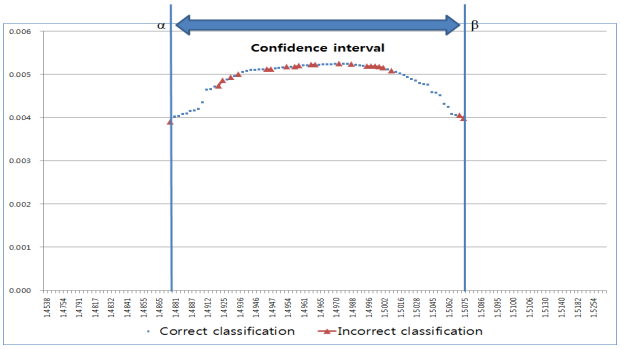


Figure 3.6: Confidence interval selection of weighted mean defuzzifications.

(C) Step 3 (Confidence interval selection): On the basis of the normal distribution shown in Figure 3.5, the weighted mean defuzzifications are selected within a confidence interval bounded by α and β, according to the weighted mean defuzzifications that are wrongly classified by feature selection, as shown in Figure 3.6. The minimum instances are selected using the equation

$$Select\ x_1, x_2, \ldots, x_i\ where\ y \geq \alpha\ and\ y \leq \beta \qquad (6)$$

where $x_i (1 \leq i \leq m)$ are the input variables selected as the minimum features in the first stage, y is a weighted mean defuzzification, and α and β are the boundary values forming the confidence interval shown in Figure 3.6. Instances that correspond to weighted mean defuzzifications within the confidence interval bounded by and are used as inputs of the NEWFM, as shown in Figure 3.1.

4 Experimental Results

The performance is evaluated after feature selection is completed in the first stage, and again, after instance selection is completed in the second stage. The performance of the enhanced NEWFM is compared with that of the previous NEWFM. In addition, McNemar's test reveals a significant difference between the performances of the two NEWFMs ($p < 0.05$).

4.1 Credit Approval

Table 3 lists the number of instances used for performance evaluation in instance selection and feature selection. As seen in Table 4, five instances are correctly classified by feature selection but wrongly classified by instance selection. However, 11 or more instances are correctly classified by instance selection because 16 instances are wrongly classified by feature selection but correctly classified by instance selection. Therefore, *instance selection* outperforms *feature selection* by 1.68% (11/653), as seen in Table 5. At the significance level of 0.05, $\chi^2 = 4.76$ in equation (2.4) is greater than 3.84; hence, the null hypothesis is rejected. Thus, the performance obtained by the proposed approach (instance selection) is significantly different from that obtained by feature selection, as seen in Table 5.

4.2 Heart Disease

Table 6 lists the number of instances used for performance evaluation in instance selection and feature selection. As seen in Table 7, one instance is correctly classified by feature selection but wrongly classified by instance selection. However, 7 or more instances are correctly classified by instance selection because 8 instances are wrongly classified by feature selection but correctly classified by instance selection. Therefore, *instance selection* outperforms *feature selection* by 2.36% (7/297), as seen in Table 8. At the significance level of 0.05, $\chi^2 = 4$ in equation (2.4) is greater than 3.84; hence, the null hypothesis is rejected. Thus, the performance obtained by the proposed approach (instance selection) is significantly different from that obtained by feature selection, as seen in Table 8.

Table 2. Description of selected features

Database name	Selected features
Credit approval database	A1, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15
Heart disease database	A2 (Sex), A6 (Fasting blood sugar), A7 (Resting ECG), A8 (Max heart rate), A9 (Exercise-induced angina), A10 (Oldpeak), A11 (Slope), A12 (Number of coloured vessels), A13 (Thal)

Table 3. Number of instances in credit approval

	Class A	Class B	Total
Instance selection	30	79	109
Feature selection	357	296	653

Table 4. McNemar's test contingency table in credit approval

		Feature selection		
		Correct	Incorrect	Total
Instance selection	Correct	74	16	90
	Incorrect	5	14	19
Total		79	30	109

Table 5. Comparisons of performance results (%) in credit approval

	Instance selection	Feature selection	Initial features	[3]
Accuracy	89.74	88.06	87.44	84.83

Table 6. Number of instances in heart disease

	Class A	Class B	Total
Instance selection	42	17	59
Feature selection	160	137	297

Table 7. McNemar's test contingency table in heart disease

		Feature selection		
		Correct	Incorrect	Total
Instance selection	Correct	42	8	50
	Incorrect	1	8	9
Total		43	16	59

Table 8. Comparisons of performance results (%) in heart disease

	Instance selection	Feature selection	Initial features	[2]
Accuracy	88.89	86.53	85.86	85.15

5 Conclusion

In this study, we proposed an enhanced NEWFM to support a weighted mean defuzzification in the T-S fuzzy model with a confidence interval in the normal distribution; this enables the selection of minimum instances with the highest performance. The performance of the enhanced NEWFM was compared with that of the previous NEWFM. In addition, McNemar's test revealed a significant difference between the performances of the two NEWFMs. The superiority of the enhanced NEWFM over the previous NEWFM was demonstrated using two experimental data sets.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology. (NRF-2012R1A1A2044134).

This research was supported by MSIP (the Ministry of Science, ICT and Future Planning), Korea, under the IT-CRSP (IT Convergence Research Support Program) (NIPA-2013-H0401-13-1001) supervised by the NIPA (National IT Industry Promotion Agency).

References

[1] Bell, G., Hey, T., and Szalay, A., "Beyond the data deluge," Science, **323**, 1297-1298 (2009).

[2] Fàtima, B.-R., Dìez, J. L., Bondia, J., "A comparative study of codification techniques for clustering heart disease database," Biomedical Signal Processing and Control, **6**, 64-69 (2011).

[3] Kim, D.-W., Lee, K., Lee, D., and Lee, K. H., "A k-populations algorithm for clustering categorical data," Pattern Recognition, **38**, 1131-1134 (2005).

[4] Kudo, M. and Sklansky, J., "Comparison of algorithms that select features for pattern classifiers," Pattern Recognition, **33**, 25-41 (2000).

[5] Kuncheva, L. I., "Editing for the k-nearest neighbors rule by a genetic algorithm," Pattern Recognition Letters, **16**, 809-814 (1995).

[6] Lim, J. S., "Finding Features for Real-Time Premature Ventricular Contraction Detection Using a Fuzzy Neural Network System," IEEE Transactions on Neural Networks, **20**, 522-527 (2009).

[7] Lee, S. H. and Lim, J. S., "Forecasting KOSPI based on a neural network with weighted fuzzy membership functions," Expert Systems with Applications, **38**, 4259-4263 (2011).

[8] McNemar, Q., "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, **12**, 153-157 (1947).

[9] Takagi, T. and Sugeno, M., "Fuzzy identification of system and its applications to modeling and control," IEEE Trans. Syst., Man, Cybern., **15**, 116-132 (1985).

[10] S-M Zhou and J. Q. Gan, ”Constructing L2-SVM-Based Fuzzy Classifiers in High-Dimensional Space With Automatic Model Selection and Fuzzy Rule Ranking,” IEEE Trans. on Fuzzy Systems, **15**, 398-409 (2007).
[11] Yi Hong, Sam Kwong, Yuchou Chang, and Qingsheng Ren, ”Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm,” Pattern Recognition, **41**, 2742-2756 (2008).
[12] Minh Hoai Nguyen and Fernando de la Torre, ”Optimal feature selection for support vector machines,” Pattern Recognition, **43**, 584-591 (2010).
[13] Josè Martìnez Sotoca and Filiberto Pla, ”Supervised feature selection by clustering using conditional mutual information-baseddistances,” Pattern Recognition, **43**, 2068-2081 (2010).
[14] Patricia E. N. Lutu and Andries P. Engelbrecht, ”A decision rule-based method for feature selection in predictive data mining,” Expert Systems with Applications, **37**, 602-609 (2010).
[15] Sang-Hong Lee and Joon S. Lim, ”Minimized Stock Forecasting Features Selection by Automatic Feature Extraction Method,” Korean Institute of Intelligent Systems, **19**, 206-211 (2009).
[16] Lee S.-H. and Lim J.S., Comparison of DBS and levodopa on resting tremor using a fuzzy neural network system, Measurement, **46**, 1995-2002 (2013).

Sang-Hong Lee received the B.S., M.S., and Ph.D. degrees in computer science from Kyungwon University, Korea in 1999, 2001, and 2012, respectively. He is currently an assistant professor in the department of computer science & engineering at Anyang University, Korea. His research focuses on neuro-fuzzy systems, stocks prediction systems, and biomedical prediction systems.

Joon S. Lim received his B.S. and M.S. degrees in computer science from Inha University, Korea, the University of Alabama at Birmingham, and Ph.D. degree was from Louisiana State University, Baton Rouge, Louisiana, in 1986, 1989, and 1994, respectively. He is currently a professor in the department of computer software at Gachon University, Korea. His research focuses on neuro-fuzzy systems, biomedical prediction systems, and human-centered systems. He has authored three textbooks on Artificial Intelligence Programming (Green Press, 2000), Javaquest (Green Press, 2003), and C# Quest (Green Press, 2006).