Indian Journal of Science and Technology, *Vol 8(26), DOI: 10.17485/ijst/2015/v8i26/80982, October 2015*
*ISSN (Print) : 0974-6846*
*ISSN (Online) : 0974-5645*

# Minimum Gene Selection using BSWFM

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

[1]Department of Computer Science and Engineering, Anyang University, Anyang-si - 708113, Republic of Korea; shleedosa@gmail.com
[2]IT College, Gachon University, Seongnam-si, Republic of Korea; jslim@gachon.ac.kr

## Abstract

In this paper, we propose a supervised gene selection method to classify tumor and normal samples based on the Bounded Sum of Weighted Fuzzy Membership Functions (BSWFM). This study compares the performance of a Neural Network with a Weighted Fuzzy Membership Function (NEWFM) with and without the proposed gene selection method. The superiority of the NEWFM with gene selection over the one without gene selection was demonstrated using a colon cancer dataset. Two thousand genes were used as inputs for the NEWFM without gene selection, and these resulted in accuracy, specificity, and sensitivity of 79%, 59.1% and 90%, respectively. A minimum of 19 genes were used as inputs for the NEWFM with gene selection, and these resulted in accuracy, specificity, and sensitivity of 87.4%, 72.7% and 95%, respectively. The results show that the NEWFM with gene selection performed better than the one without gene selection.

**Keywords:** BSWFM, Distance, Fuzzy Neural Network, Gene Selection, NEWFM

## 1. Introduction

Fuzzy Neural Network (FNN), an adaptive decision support tool that combines a neural network with fuzzy set theories for pattern classification, diagnosis, and prediction, has been previously proposed[1,2]. FNNs with different structures have been presented together with the algorithms for learning, adaptation, and rule extraction[7,8]. In order to extract knowledge from a given series of learning data, FNNs based on self-organizing systems have been developed[11,12]. In addition, in the field of artificial intelligence, Bayesian networks have been introduced as an important method to handle uncertainty. The Bayesian network is a model that expresses the problems of the real world with a combined probability distribution and has the advantage of being able to reflect experts' knowledge well[5,19].

Recently, in all areas, particularly the field of processing genetic data, the amount of data needed to perform data mining or machine learning has rapidly increased[13–15]. In general, it is believed that when the amount of input data is high, a certain fact can be classified or judged more efficiently; however, excessive data or inputs may trigger inefficiency in terms of memory and time. In addition, data with little cross-relevance may generate wrong results. Therefore, research must focus on decreasing the amount of data used for learning by existing FNNs or the number of features used for inputs[16]. Feature selection is used as a method to reduce the number of features, and instance selection and prototype selection are used as methods to decrease the values of such features. Feature selection removes overlapping or irrelevant features, thereby improving classification performance; utilizes minimal features; and reduces operating costs, thereby enhancing performance[4]. Instance selection is an algorithm that induces good learning by selecting good values from the features, thereby improving classification performance[17]. Moreover, prototype selection is an efficient algorithm that may reduce feature values while maintaining classification performance[18]. Bayesian networks reflect a causal relationship among features using experts' knowledge, thereby not only selecting minimal features

*Author for correspondence

and instances but also being able to select minimal features and instances based on association among features. In particular, recent research on gene selection classified methods to improve classification performance into the filter, wrapper, and embedded techniques. Moreover, an ensemble technique that simultaneously utilizes different learning algorithms was recently reported[6,10].

A Neural Network with a Weighted Fuzzy Membership Function (NEWFM) is a supervised neuro-fuzzy classification system that uses the Bounded Sum of Weighted Fuzzy Membership Functions (BSWFM)[3,9]. After the training process of the NEWFM is complete, all features are interpretably formed into weighted fuzzy membership functions that preserve the disjunctive fuzzy information and features, and all feature differences are illustrated by the graphical features in all BSWFMs. In this study, we propose a gene selection BSWFM combined with distance in order to decrease the computational load and improve accuracy by removing irrelevant genes. This enables the selection of minimal genes while achieving the highest classification performance. In the results, 2000 genes were used as inputs for the NEWFM without gene selection, and these resulted in accuracy, specificity, and sensitivity of 79%, 59.1% and 90%, respectively. A minimum of 19 genes were used as inputs for the NEWFM with gene selection, and these resulted in accuracy, specificity, and sensitivity of 87.4%, 72.7% and 95%, respectively. The results show that the NEWFM with gene selection performed better than the one without gene selection.

The remainder of this study is organized as follows. In Section 2, we review the experimental data used in this study and describe the structure of the NEWFM, the learning process of the NEWFM, and the BSWFM. In Section 3, we analyze the experimental results of the gene selection algorithms proposed in this study. Finally, the conclusion is presented in Section 4.

# 2. Materials and Methods

## 2.1 Experimental Dataset

Descriptions of the colon cancer dataset studied are as follows. The colon cancer dataset involves the comparison of tumor and normal samples of the same tissue. The dataset consists of 62 samples of colon epithelial cells. These samples are divided into two variants of colon tissue: 40 tumor colon samples and 22 normal colon samples. The dataset, representing 2000 genes across 62 samples, is available at http://genomics-pubs.princeton.edu/oncology/.

## 2.2 Neural Network with Weighted Fuzzy Membership Function (NEWFM)

An NEWFM is an FNN of supervised learning that performs classification using a learned BSWFM[3,9]. The BSWFM synthesizes the bounded sums of three fuzzy membership functions with weights—namely, large, middle, and small—into one fuzzy membership function. The structure of the NEWFM, illustrated in Figure 1, is composed of three layers—namely, input, hyperbox, and class. The input layer comprises $n$ input nodes, and each node receives the input of one feature. The hyperbox class consists of $m$ hyperbox nodes, and the $l$th hyperbox node $B_l$ is connected to a single class node and has $n$ fuzzy sets. An NEWFM that has undergone learning may be used as a fuzzy rule set to classify input patterns. In each fuzzy set of hyperbox nodes, three Weighted Fuzzy Memberships (WFMs) are created. The BSWFM is expressed as $\mu_b^i(x)$. The bold line in Figure 2 is defined as the following formula and integrates fuzzy features of the three WFMs in Figure 2. The learned BSWFM $\mu_b^i(x)$ becomes the fuzzy rule of the $i$th input.

$$\mu_b^i(x) = \sum_{j=1}^{3} B_l^i(\mu_j(x)). \tag{1}$$

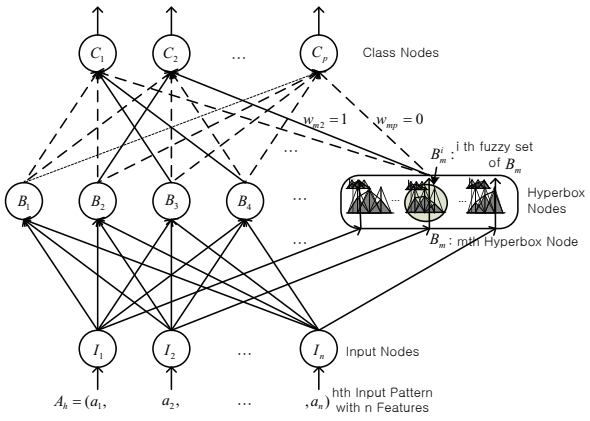


**Figure 1.** NEWFM structure.

## 2.3 Algorithm for Gene Selection based on Distances between Weights of the BSWFM

Feature selection enables reduction of calculation costs or the improvement of performance by removing redundant

or noisy features[4]. This paper proposes a measure that takes the same time to derive excellent, minimal features in classification performance regardless of the number of NEWFM features by removing features of little importance. Figure 2 shows the algorithm based on distances between weights of the BSWFM. As the Figure explains, this paper intends to propose minimal gene selection using these distances.

In this paper, learning is performed using an NEWFM from initial genes. Using the distances between weights of the BSWFM generated through such learning, gene selection was performed in the following four steps.

[Step 1] $v_0$ and $v_4$ of the BSWFM in Figure 2, which were generated during the process of learning, are normalized to zero and 100, respectively.

[Step 2] Using the normalized BSWFM, the BSWFM of tumor colon samples and the center of gravity for the BSWFM of normal colon samples are derived.

[Step 3] The distance between the two BSWFM centers of gravity is derived, and the distances between the centers of gravity about the initial genes are aligned in order to rank them.

[Step 4] Classification performance is compared by removing genes one by one with the shortest average of distance between the centers of gravity.

The reason behind the normalization step (step 1) is that initial genes used for NEWFM inputs have different ranges of values. For example, some initial genes may have a value between zero and one; others may have a value between zero and 10,000. Therefore, after learning is complete, the range of the $x$-axis of the BSWFM in Figure 2 may have a value of 1 or 10,000 according to the genes. If normalization is not performed, genes with a large $x$-axis range will have a larger distance between the centers of gravity than genes with a small $x$-axis range. In order to remove such error, the process of normalization is necessary. In step 2, the center of gravity for the BSWFM of the normal colon sample and that of the BSWFM for the tumor colon sample, as shown in Figure 3, are calculated using the BSWFM normalized to a value between zero and 100. In step 3, the distance between the centers of gravity in Figure 3 is derived using the center of gravity derived similar to that in step 2. In addition, the average value, standard deviation, maximal value, and minimal value are derived relating to the distance of the centers of gravity about the initial genes. In step four, classification performance is derived by removing genes with the lowest average value one by one based on the result derived in step 3.

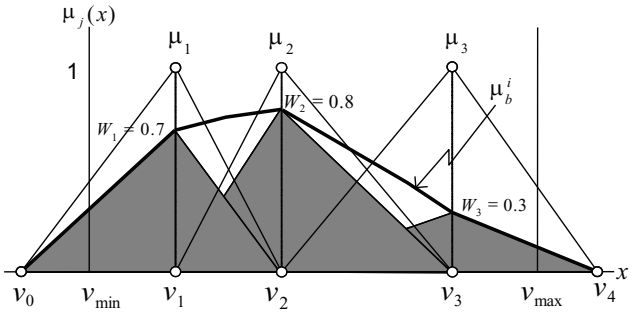


**Figure 2.** Example of three BSWFMs.

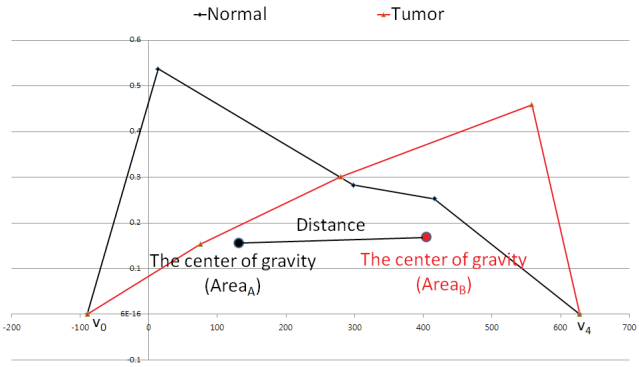


**Figure 3.** Distances between the centers of gravity of BSWFM.

## 3. Experimental Results

In this study, tumor biopsies and normal biopsies were classified from colon cancer datasets. Table 1 shows 19 minimum genes that were finally selected from the 2000 initial genes of the colon cancer datasets. Differences between tumor biopsies and normal biopsies can be visualized, and related genes can be analyzed from the 19 minimum genes in Figure 4.

In Equation (2), TP (True Positive) indicates the cases in which a tumor colon sample was identified as a tumor colon sample, and TN (True Negative) indicates the cases in which a normal colon sample was identified as a normal colon sample. Conversely, FP (False Positive) denotes the cases in which a tumor colon sample was identified as a normal colon sample, and FN (False Negative) denotes the cases in which a normal colon sample was identified as tumor colon sample. The classification performances of the NEWFM with and without

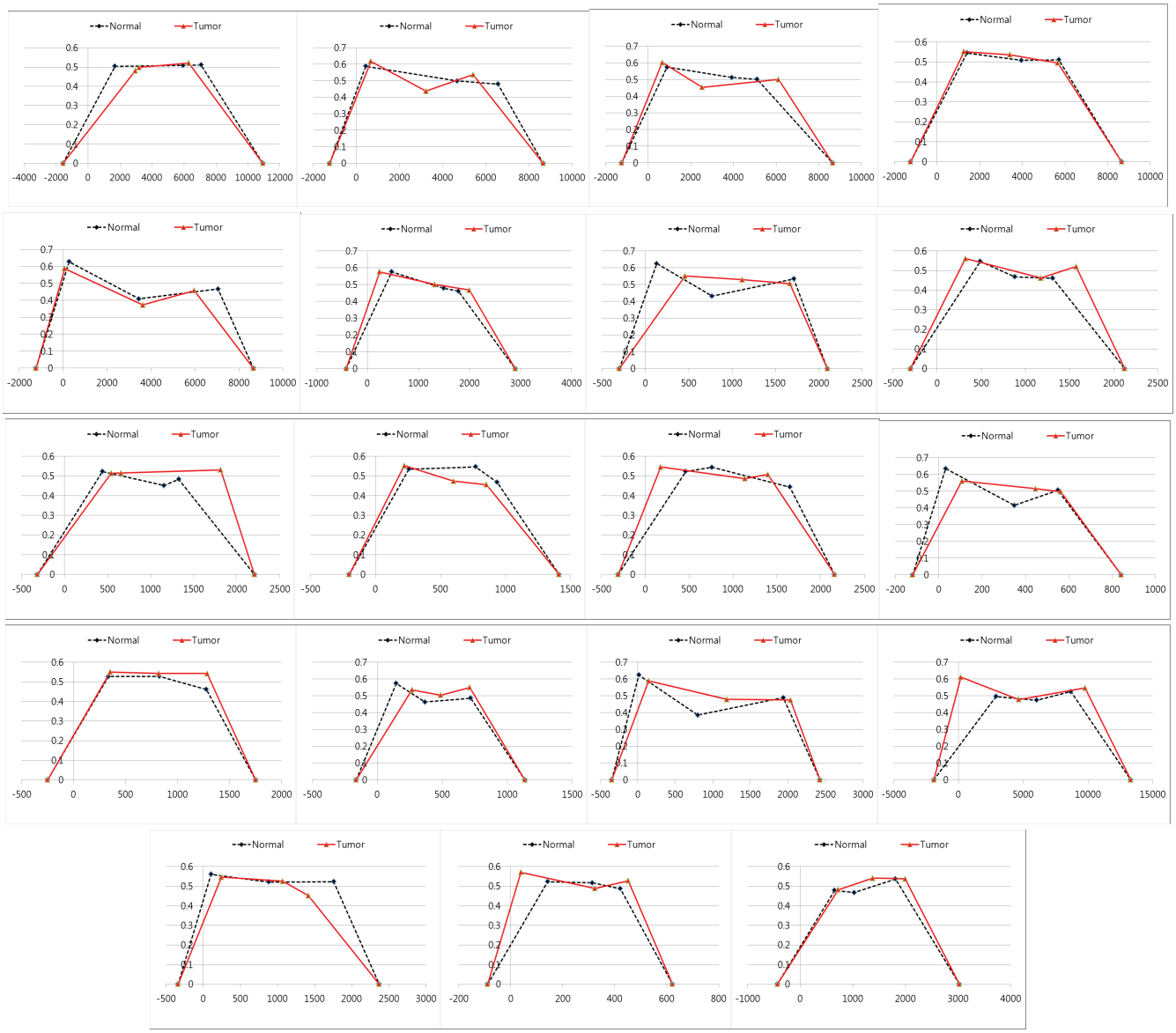


**Figure 4.** BSWFMs of the 19 genes.

gene selection are listed in Tables 2-4. As can be seen in Table 4, gene selection outperforms no gene selection by 8.1%.

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \qquad (2)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \times 100$$

## 4. Conclusions

In this study, the use of an NEWFM was proposed to select genes using a BSWFM combined with distance; this gene selection process enables the selection of a few genes with the highest performance. The superiority of the NEWFM with gene selection over the one without gene selection was demonstrated using a colon cancer dataset. The NEWFM obtained BSWFMs for a minimum of 19 (colon cancer) genes to identify the fuzzy membership functions for minimal genes. These fuzzy membership functions were used to classify tumor biopsies and normal biopsies in the colon cancer dataset.

**Table 1.** Description of 19 minimum genes selected from a total of 2000 colon cancer data points

| Gene | Description |
|---|---|
| M27190 | Homo Sapiens Secretory Pancreatic Stone Protein (PSP-S) mRNA, Complete CDs |
| R62945 | Complement Decay-Accelerating Factor 1 Precursor (Homo Sapiens) |
| R99907 | Interferon Regulatory Factor 2 (Homo Sapiens) |
| Control (260$^{th}$) | None |
| U09367 | Human Zinc Finger Protein ZNF136 |
| Control (261$^{st}$) | None |
| T49941 | Putative Insulin-like Growth Factor II Associated (Human) |
| H72965 | 26S Protease Regulatory Subunit 7 (Homo Sapiens) |
| L20859 | Human Leukemia Virus Receptor 1 (GLVR1) mRNA, Complete CDs |
| Y00062 | Human mRNA for T200 Leukocyte Common Antigen (CD45, LC-A) |
| R62549 | Putative Serine/Threonine-Protein Kinase B0464.5 in Chromosome III (Caenorhabditis elegans) |
| Control (263$^{rd}$) | None |
| T48014 | Hemoglobin Alpha Chain (Human) |
| M26383 | Human Monocyte-Derived Neutrophil-Activating Protein (MONAP) mRNA, Complete CDs |
| Control (262$^{nd}$) | None |
| R53455 | Serine Carboxypeptidase I Precursor (Hordeum Vulgare) |
| M86934 | Human GS1 (protein of unknown function) mRNA, Complete CDs |
| M28373 | Homo Sapiens Amyloid Protein A4 Precursor mRNA, 3' End of CDs |
| X86693 | H. sapiens mRNA for Hevin like Protein |

**Table 2.** Confusion matrix of classification results without gene selection

| Tumor colon samples | TP | FN |
|---|---|---|
| | 36 | 4 |
| Normal colon samples | FP | TN |
| | 9 | 13 |

**Table 3.** Confusion matrix of classification results with gene selection

| Tumor colon samples | TP | FN |
|---|---|---|
| | 38 | 2 |
| Normal colon samples | FP | TN |
| | 6 | 16 |

**Table 4.** Classification results using NEWFM

| | Accuracy | Specificity | Sensitivity |
|---|---|---|---|
| Without gene selection (%) | 79 | 59.1 | 90 |
| With gene selection (%) | 87.1 | 72.7 | 95 |

## 5. Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2054293).

## 6. References

1. Carpenter GA, Grossberg S, Reynolds J. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks, 1991; 4:565–88.
2. Jang R. ANFIS: Adaptive network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics. 1993; 23:665–85.
3. Lee SH, Lim JS. Parkinson's disease classification using gait characteristics and wavelet-based feature extraction. Expert Systems with Applications. 2012; 39:7338–44.
4. Zhou S-M, Gan JQ. Constructing L2-SVM-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking. IEEE Transactions on Fuzzy Systems. 2007; 15(3):398–409.

5. Heckerman D. Bayesian networks for data mining. Data Mining and Knowledge Discovery. 1997; 1(1):79–119.
6. Yang P, et al. A review of ensemble methods in bioinformatics. Current Bioinformatics. 2010; 5(4):296–308.
7. Ishibuchi H, Nakashima T. Voting in fuzzy rule-based systems for pattern classification problems. Fuzzy Sets and Systems. 1999; 103:223–38.
8. Kasabov N. Foundation of neural networks, fuzzy systems and knowledge engineering. Cambridge, MA: The MIT Press; 1996.
9. Lim JS. Finding features for real-time premature ventricular contraction detection using a fuzzy neural network system. IEEE Transactions on Neural Networks. 2009; 20(3):522–7.
10. Kim K-J, Cho S-B. An evolutionary algorithm approach to optimal ensemble classifiers for DNA microarray data analysis. IEEE Transactions on Evolutionary Computation. 2008; 12(3):377–88.
11. Juang CF, Lin CT. An on-line self-constructing neural fuzzy inference network and its applications. IEEE Transactions on Fuzzy Systems. 1998; 6(1): 12–32.
12. Tanaka K, Sano M, Watanabe H. Modeling and control of carbon monoxide concentration using a neuro-fuzzy technique. IEEE Transactions on Fuzzy Systems. 1995; 3:271–9.
13. Chen H-L, Huang C-C, Yu X-G, Xu X, Sun X, Wang G, Wang S-J. An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Systems with Applications. 2013; 40:263–71.
14. Saeys Y, et al. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19):2507–17.
15. Bell G, Hey T, Szalay A. Beyond the data deluge. Science. 2009; 323:1297–8.
16. Zhou S-M, Gan JQ. Constructing L2-SVM-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking. IEEE Transactions on Fuzzy Systems. 2007; 15(3):398–409.
17. Kuncheva LI. Editing for the k-nearest neighbors rule by a genetic algorithm. Pattern Recognition Letters. 1995; 16:809–14.
18. Chen H-L, Huang C-C, Yu X-G, Xu X, Sun X, Wang G, Wang S-J. An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Systems with Applications. 2013; 40:263–71.
19. Neapolitan RE. Learning bayesian networks. New Jersey: Pearson Prentice Hall; 2004.