International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4348-4354

Increasing the Accuracy of NEWFM using a Geometric Graph-Based Gene Selection Algorithm

Sang-Hong Lee[1], Seok-Woo Jang[2,*]

[1]Department of Computer Science & Engineering, Anyang University, Republic of Korea.
[2,]Department of Software, Anyang University, Republic of Korea.*

Abstract

In a microarray dataset with thousands of genes, the majority of the genes are irrelevant to the occurrence of disease. One of the challenges in analyzing microarray datasets is selecting a suitable number of the most relevant genes with maximum classification accuracy. A neural network with weighted fuzzy membership functions (NEWFM) is a supervised neuro-fuzzy system that reduces the size of available data to improve classification accuracy and computational efficiency. This paper proposes a new gene selection algorithm based on a geometric graph using the bounded sum of weighted fuzzy membership functions (BSWFM) that improves the accuracy of NEWFM. The proposed gene selection algorithm reduces computational load and improves accuracy by removing irrelevant genes using the Euclidean distances of the centers of gravity multiplied by the non-overlap area between two BSWFMs. Further, the results of comparative experiments conducted using the colon cancer and the leukemia microarray problem datasets indicate that NEWFM with the proposed gene selection algorithm is more accurate than NEWFM without the proposed algorithm. More specifically, 2000 genes from the colon cancer dataset and 7129 genes from the leukemia dataset used as inputs to NEWFM without the proposed algorithm resulted in accuracies of 90.3% and 58.8%, respectively. In contrast, inputs of seven minimum genes from the colon cancer dataset and nine minimum genes from the leukemia dataset to NEWFM with the proposed algorithm resulted in accuracies of 96.8% and 100%, respectively.

Keywords: Gene selection, Fuzzy neural networks, NEWFM, BSWFM, microarray.

INTRODUCTION

One of the major challenges in the handling of gene expression data is identifying genes that are relevant to a clinical diagnosis from thousands of genes in microarray datasets [1][2]. Although thousands of genes are evaluated simultaneously, most of them are irrelevant or not significant to a clinical diagnosis [3][4][5]. Dimension reduction techniques that improve computational complexity and classification accuracy by removing irrelevant and redundant genes using microarray technology have been developed to overcome this challenge [6][7]. In addition, statistical methods and machine learning have recently been widely utilized to find relevant genes [8][9]. Various statistical methods, such as t-test [10][12], correlation [11][12], regression [13], mutual information [14][15], and threshold number of misclassifications score [3], have been widely used to filter irrelevant and redundant genes. Several studies have also applied genetic algorithms (GA) to the selection of a suitable gene subset for multiclass classification [8][12][16][17][18]. Machine learning techniques, such as support vector machine (SVM) [19][20][21][22], k-nearest neighbor (k-NN) [16][23], and rough set [6], have also been applied as specific classifiers in cancer classification.

Several studies have applied fuzzy neural network (FNN) to the problem [15][24][25][26], with FNNs combining neural network [15][24] and fuzzy set theory also being proposed for learning, adaptation, and rule extraction [25]. A neural network with weighted fuzzy membership functions (NEWFM) is a supervised classification neuro-fuzzy system that uses the bounded sum of weighted fuzzy membership functions (BSWFM) [26]. This paper proposes a new gene selection algorithm based on a geometric graph that combines the Euclidean distance of centers of gravity and the non-overlap area distribution measurement between two BSWFMs to improve the accuracy of NEWFM in classifying tumor biopsies and normal biopsies from colon cancer datasets, and acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) from leukemia datasets. In the proposed algorithm, seven minimum genes and nine minimum genes with the highest classification accuracy among 2000 genes and 7129 genes from the colon cancer dataset and the leukemia dataset, respectively, are used as interpretable weighted fuzzy membership functions that preserve the disjunctive fuzzy information and characteristics [26]. As a result, a minimal set of genes with the highest classification accuracy is selected. In this study, we compared the accuracy of NEWFM without the proposed gene selection algorithm to that of NEWFM with the proposed algorithm, and also to those of existing methods [21][22]. More specifically, 2000 genes from the colon cancer problem dataset and 7129 genes from the leukemia problem dataset were used as inputs to NEWFM without the proposed gene selection algorithm, which resulted in accuracies of 90.3% and 58.8%, respectively. In contrast, inputs of seven minimum genes from the colon cancer problem dataset and nine minimum genes from the leukemia problem dataset to NEWFM with the proposed gene selection algorithm resulted in accuracies of 96.8% and 100%.

The remainder of this paper is organized as follows. In Section 2, we review the colon cancer and the leukemia problem datasets used in this study. In Section 3, we describe

the structure of NEWFM and the proposed minimum genes selection algorithm. In Section 4, we analyze the experimental results obtained by NEWFM with the gene selection algorithm proposed in this study and compare its accuracy with those of the existing methods. In Section 5, discussions and conclusion are presented.

EXPERIMENTAL DATA

In this study, the proposed gene selection algorithm was applied to data provided by the colon cancer dataset and the leukemia dataset. To demonstrate the utility of the proposed gene selection algorithm, we conducted extensive experiments using the colon cancer dataset and the leukemia dataset benchmark problems. One major advantage of the proposed gene selection algorithm is that it is simple enough to implement as a computer program without any statistical assumptions.

The colon cancer dataset

The colon cancer dataset is a collection of expression measurements from colon biopsy samples reported by Alon et al. [1]. The dataset consists of 62 samples of colon epithelial cells. These samples are divided into two variants of colon tissue: 40 colon tumor samples and 22 normal colon samples. The dataset, representing 2000 genes across 62 samples, is available at http://genomics-pubs.princeton.edu/oncology/.

The leukemia dataset

The leukemia dataset is a collection of expression measurements reported by Golub et al. [5]. It contains 72 samples split into 38 training and 34 test samples. These samples are divided into two variants of leukemia: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The dataset, representing 7129 genes across 72 samples, is available at http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43.

NEURAL NETWORK WITH WEIGHTED FUZZY MEMBERSHIP FUNCTIONS (NEWFM) AND PROPOSED GENE SELECTION ALGORITHM FOR NEWFM

NEWFM is the neuro-fuzzy system that uses the bounded sum of weighted fuzzy membership function (BSWFM) [26] that contains three *weighted fuzzy membership functions* (*WFMs)* (μ_1, μ_2, μ_3 in Fig. 1). This paper proposes a new gene selection algorithm based on a geometric graph. The proposed algorithm selects minimum genes using the Euclidean distance of centers of gravity and the non-overlap area distribution measurement between two BSWFMs (an example is shown in Fig. 1).

Fig. 2 shows a BSWFM for two classes (A, B) generated during learning. The genes shown in Fig. 2 have two BSWFMs and are obtained from the training process of NEWFM. After the training process of NEWFM is complete, all genes are interpretably formed into weighted fuzzy membership functions preserving the disjunctive fuzzy information and characteristics. The two BSWFMs graphically illustrate the gene differences between class A and class B. Using the graphical characteristics of the two BSWFMs, the proposed new gene selection algorithm is based on the fact that the greater the distance of the centers of gravity multiplied by the non-overlap area between two BSWFMs, the better the characteristics between the two classes are distinguished.

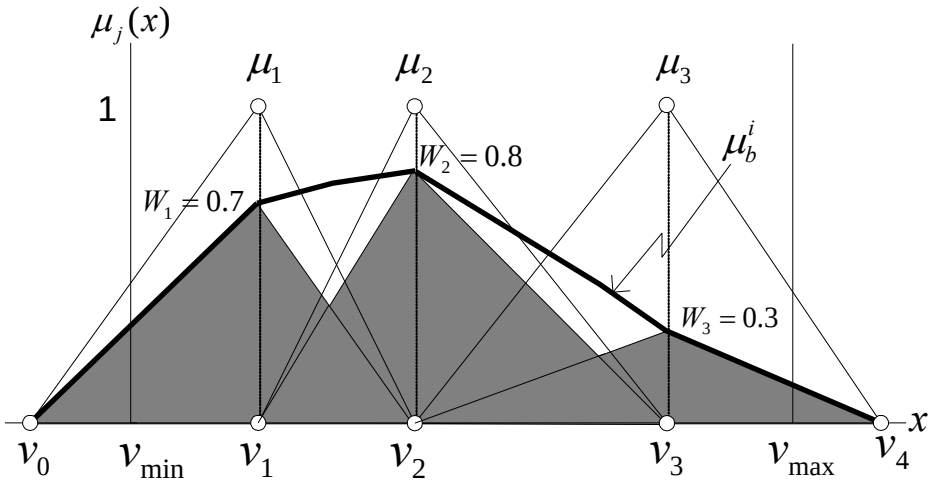

Figure 1. Bounded sum of the three weighted fuzzy membership functions (the bold line indicates the BSWFM)

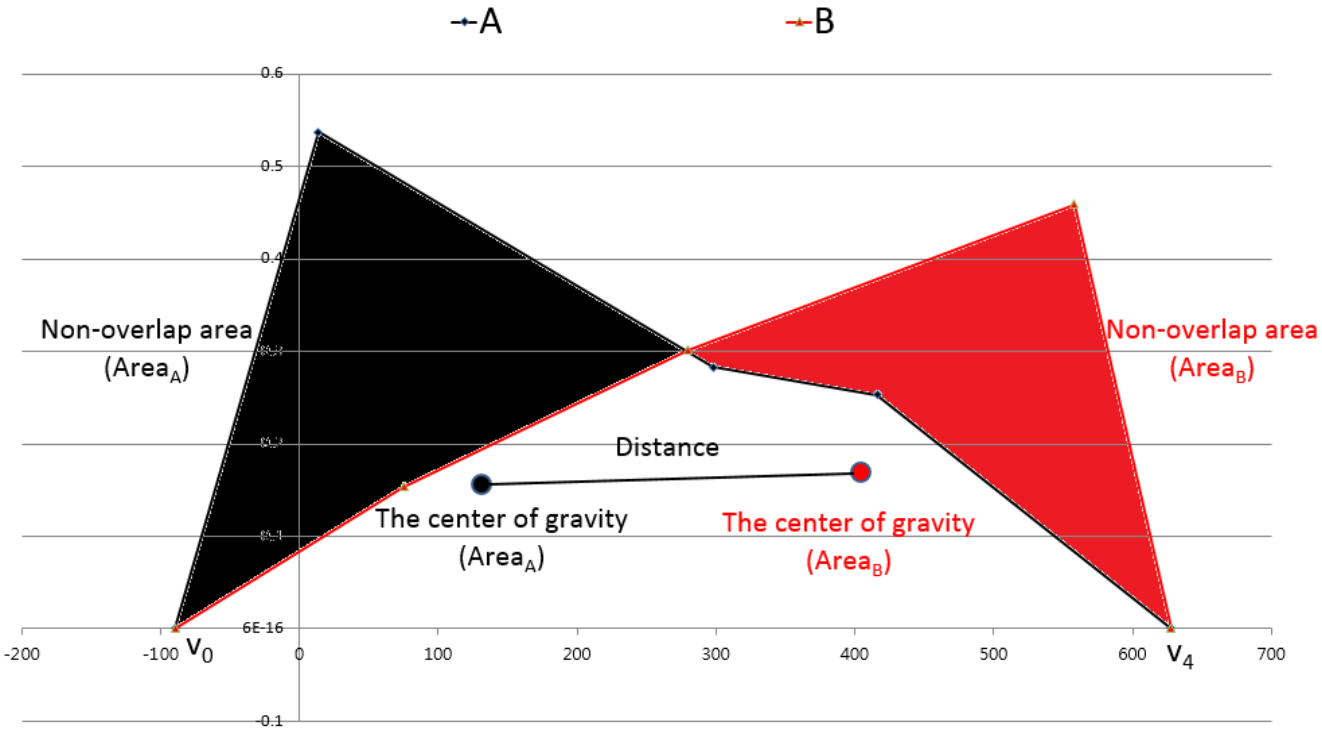

Figure 2. Example of the Euclidean distance of centers of gravity and the non-overlap area distribution measurement between two BSWFMs

The proposed algorithm uses the distance of centers of gravity multiplied by the non-overlap area between two BSWFMs to select minimum genes with maximum accuracy. The distance between the centers of gravity multiplied by the non-overlap area between two BSWFMs measures the degree of salience of the *i*th gene using the following equation:

$$Distance(i) = \sqrt{(X_A^i - X_B^i)^2 + (Y_A^i - Y_B^i)^2}$$

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

$$Value(i) = Distance(i) \times Area(i)$$

where $Area_A^i$ and $Area_B^i$ are non-overlap areas, X_A^i and X_B^i are the x-coordinates of the centers of gravity, and Y_A^i and Y_B^i are the y-coordinates of the centers of gravity of class A and class B in the *i*th gene, respectively. A larger $f(i)$ indicates better gene characteristics and good genes with good ranking. Therefore, genes with small $f(i)$ are removed one-by-one to improve classification accuracy. As a result, genes with the highest classification accuracy are selected as the minimal genes by Algorithm 1.

Algorithm 1. Gene selection algorithm.

N: Number of training for gene selection

n: Number of genes

X_n: x-coordinate of center of gravity of BSWFM // X_1, X_2, …, and $X_{2000 \text{ or } 7129}$

Y_n: y-coordinate of center of gravity of BSWFM // Y_1, Y_2, …, and $Y_{2000 \text{ or } 7129}$

x_p: x-coordinate of BSWFM // x_1, x_2, …, and x_5

y_p: y-coordinate of BSWFM // y_1, y_2, …, and y_5

Distance(n): Euclidean distance of *n*th gene

Area(n): Non-overlap area of *n*th gene

Value(n): $Distance(n) \times Area(n)$

V_n: Accumulated *Value(n)* of *n*th gene

index: Location of *n*th gene in array

01: V_1, V_2, …, and V_n are initialized to 0 // n = 2000 or 7129

02: **for** *index* = 1 **to** N // N = 1000

03: **for** i = 1 **to** n // n = 2000 or 7129

04:

$$X^i_{Class} = \frac{\sum_{p=1}^{4}((X^i_{Class,p} + X^i_{Class,p+1})(X^i_{Class,p}Y^i_{Class,p+1} - X^i_{Class,p+1}Y^i_{Class,p}))}{6\frac{1}{2}\sum_{p=1}^{4}(X^i_{Class,p}Y^i_{Class,p+1} - X^i_{Class,p+1}Y^i_{Class,p})} (Class = A, B)$$

05:

$$Y^i_{Class} = \frac{\sum_{p=1}^{4}((Y^i_{Class,p} + Y^i_{Class,p+1})(X^i_{Class,p}Y^i_{Class,p+1} - X^i_{Class,p+1}Y^i_{Class,p}))}{6\frac{1}{2}\sum_{p=1}^{4}(X^i_{Class,p}Y^i_{Class,p+1} - X^i_{Class,p+1}Y^i_{Class,p})} (Class = A, B)$$

06: $Distance(i) = \sqrt{(X^i_A - X^i_B)^2 + (Y^i_A - Y^i_B)^2}$

07: $Area(i) = \{(Area^i_A + Area^i_B)^2 \Big/ (\frac{1}{1+e^{-\left|Area^i_A - Area^i_B\right|}})\}$

08: $Value(i) = Distance(i) \times Area(i)$

09: $V_{index} = V_{index} + Value(i)$

10: **end for**

11: **end for**

12: return (V_1, …, $V_{2000\ or\ 7129}$)

EXPERIMENTAL RESULTS

In this study, tumor biopsies and normal biopsies, and AML and ALL were classified from the colon cancer dataset and the leukemia dataset, respectively. Tables 1 and 2 show the minimum genes that were finally selected. Seven minimum genes were finally selected from the 2000 genes in the colon cancer dataset and nine minimum genes from the 7129 genes in the leukemia dataset.

The accuracies of NEWFM with and without the proposed gene selection algorithm for the colon cancer dataset and the leukemia dataset are listed in Tables 3–7. As can be seen in Tables 3, 4, and 7, for the colon cancer dataset, NEWFM with the proposed gene selection algorithm outperforms NEWFM without the proposed algorithm by 6.5%. As can be seen in Tables 5–7, for the leukemia dataset, NEWFM with the proposed algorithm outperforms NEWFM without the proposed algorithm by 41.2%. Table 7 compares the classification accuracy of NEWFM with those determined by Guyon [21] and Wang [22].

For two-class problems, a true positive (TP) refers to cases where class 2 is classified as class 2, and a true negative (TN) refers to cases where class 1 is classified as class 1. A false positive (FP) refers to cases where class 2 is classified as class 1, and false negative (FN) refers to cases where class 1 is classified as class 2.

Table 1. The colon cancer dataset and selected genes

Gene	Description
H64489	238846 LEUKOCYTE ANTIGEN CD37 (Homo sapiens)
R87126	197371 MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
R36977	26045 P03001 TRANSCRIPTION FACTOR IIIA
H08393	45395 COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
M27190	Homo sapiens secretory pancreatic stone protein (PSP-S) mRNA, complete cds
R99907	201673 INTERFERON REGULATORY FACTOR 2 (Homo sapiens)
T60778	76539 MATRIX GLA-PROTEIN PRECURSOR (Rattus norvegicus)

Table 2. The leukemia dataset and selected genes

Gene	Description
M11147	FTL Ferritin, light polypeptide
M16038	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
U33822	Tax1-binding protein TXBP181 mRNA
U50136	Leukotriene C4 synthase (LTC4S) gene
X61587	ARHG Ras homolog gene family, member G (rho G)
U05681	Proto-oncogene BCL3 gene
M81695	ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)
M83652	PFC Properdin P factor, complement
X69654	RPS26 Ribosomal protein S26

Table 3. Confusion matrix of classification results without gene selection (colon cancer dataset)

Tumor biopsies samples (Class 2)	TP	FN
	36	4
Normal biopsies samples (Class 1)	FP	TN
	2	20

Table 4. Confusion matrix of classification results with gene selection (colon cancer dataset)

Tumor biopsies samples (Class 2)	TP	FN
	39	1
Normal biopsies samples (Class 1)	FP	TN
	1	21

Table 5. Confusion matrix of classification results without gene selection (leukemia dataset)

Acute myeloid leukemia (AML) (Class 2)	TP	FN
	9	5
Acute lymphoblastic leukemia (ALL) (Class 1)	FP	TN
	9	11

Table 6. Confusion matrix of classification results with gene selection (leukemia dataset)

Acute myeloid leukemia (AML) (Class 2)	TP	FN
	14	0
Acute lymphoblastic leukemia (ALL) (Class 1)	FP	TN
	0	20

Table 7. Classification accuracy of the existing methods (numbers in parentheses denote the number of selected genes)

Data set	Colon	Leukemia
Guyon et al. [21]	90.32 (8)	100 (4)
Wang et al. [22]	91.9 (3)	100 (5)
NEWFM without gene selection	90.3 (2000)	58.8 (7129)
NEWFM with gene selection	96.8 (7)	100 (9)

DISCUSSION AND CONCLUSION

NEWFM is a supervised classification neuro-fuzzy system that uses BSWFM based on a geometric graph. All genes used as inputs to NEWFM are interpretably formed into weighted fuzzy membership functions preserving the disjunctive fuzzy information and characteristics in Fig. 1. All gene differences are illustrated by the graphical characteristics of all BSWFMs in Fig. 2. In this study, a new gene selection algorithm that selects minimum genes using the graphical characteristics of BSWFM combined with Euclidean distance and the non-

overlap area distribution measurement with high accuracy was developed. This new gene selection algorithm simplifies the identification of good and bad genes by NEWFM using Euclidean distance and non-overlap area based on a geometric graph. Experiments conducted using 2000 genes from the colon cancer problem dataset and 7129 genes from the leukemia problem dataset as inputs to NEWFM without the proposed gene selection algorithm resulted in accuracies of 90.3% and 58.8%, respectively. In contrast, using seven minimum genes from the colon cancer problem dataset and nine minimum genes from the leukemia problem dataset as inputs to NEWFM with the proposed gene selection algorithm resulted in accuracies of 96.8% and 100%, respectively. These results show that the accuracy of NEWFM with the proposed gene selection algorithm is superior to that of NEWFM without the proposed algorithm.

REFERENCES

[1] Wong TT, Chen DQ (2011) A gene selection method for microarray data based on risk genes. Expert Systems with Applications 38:14065-14071.

[2] Le Cao KA, Bonnet A, Gadat S (2009) Multiclass classification and gene selection with a stochastic algorithm. Computational Statistics and Data Analysis 53:3601-3615.

[3] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000) Tissue Classification with Gene Expression Profiles. J. Computational Biology 7:559-584.

[4] Golub T, Slonim D, Tamayo P, Huard C, Caasenbeek JM, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

[5] Zhou X, Tuck DP (2007) MSVM-RFE: extensions of SVM-REF for multiclass gene selection on DNA microarray data. Bioinformatics 23:1106-1114.

[6] Maji P, Paul S (2011) Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning 52:408-426.

[7] Wong TT, Liu KL (2010) A Probabilistic mechanism based on clustering analysis and distance measure for subset gene selection. Expert Systems with Applications 37:2144-2149.

[8] Huerta EB, Duval B, Hao JK (2010) A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. Neurocomputing 73:2375-2383.

[9] Lee CP, Leu Y (2011) A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing 11:208-213.

[10] Li J, Su H, Chen H, Futscher BW (2007) Optimal search-based gene subset selection for gene array cancer classification. IEEE Transactions on Information Technology in Biomedicine 11:398-405.

[11] Van de Vijver MJ, et al (2002) A Gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347:December 19.

[12] Chen AH, Yang C (2012) The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data. Expert Systems with Applications 39: 4785-4795.

[13] Van't Veer LJ, et al (2002) Gene expression profiling predicts clinical outcome in breast cancer. Nature 415:530-536.

[14] Liu X, Krishnan A, Mondry A (2005) An entropy based gene selection method for cancer classification using microarray data. BMC Bioinformatics 6:1-14.

[15] Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:1226-1238.

[16] Li L, Weinberg CR, Darden TA, Pedersen LG (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. Bioinformatics 17:1131-1142.

[17] Zhu Z, Ong YS, Dash M (2007) Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition 40:3236-3248.

[18] Huang C, Wang C (2006) A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications 31:231–240.

[19] Huang HL, Chang FL (2007) ESVM: evolutionary support vector machine for automatic feature selection and classification of microarray data. BioSystems 90:516-528.

[20] Tapia E, Bulacio P, Angelone L (2012) Sparse and stable gene selection with consensus SVM-RFE. Pattern Recognition Letters 33:64-172.

[21] Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46:389-422.

[22] Wang Y, Makedon FS, Ford JC, Pearlman J (2005) HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21:1530-1537.

[23] Li L, Darden TA, Weinberg CR, Levine AJ, Pedersen LG (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method.

Combinatorial Chemistry & High Throughput Screening 4:727-739.

[24] Sotoca JM, Pla F (2010) Supervised feature selection by clustering using conditional mutual information-based distances. Pattern Recognition 43:2068-2081.

[25] Kabir M, Shahjahan, Murase K (2011) A new local search based hybrid genetic algorithm for feature selection. Neurocomputing 74:2914-2928.

[26] Lim JS (2009) Finding Features for Real-Time Premature Ventricular Contraction Detection Using a Fuzzy Neural Network System. IEEE Transactions on Neural Networks 20:522-527.