Effective Liver Cancer Diagnosis Method based on Machine Learning Algorithm

Sangman Kim / Seungpyo Jung / Youngju Park / Jihoon Lee / Jusung Park
Dept. of Electronics and Electrical Engineering,
Pusan National University
Pusan, Korea

Abstract—**In this paper, we introduce a method to find useful markers from sensor arrays which have massive sensing points and diagnose liver cancer based on machine learning algorithms which are neural network and fuzzy neural network. We obtain reliable results by using a learning ability and n-fold cross validation. For the verification of the proposed method, raw data of serums from 314 normal and 81 patients reacted to 1,142 aptamers are used. According to the results, we can detect liver cancer with the accuracy of 99.19 % by average use of 132 aptamers based on neural network and 98.19 % by average use of 226 aptamers based on fuzzy neural network.**

Keywords-component; neural network; fuzzy neural network; machine learning; diagnosis, select-drop, feature

I. INTRODUCTION

Research for more accurate and easier way of disease diagnosis has been studied from the past to the present. Generally, disease diagnosis is performed by reacting to limited number of markers. However, we cannot guarantee the accurate disease diagnosis. Especially, it is more difficult in the first stage with the limited number of markers. To resolve the problem, disease diagnosis using microarray has been actively studied in the recent years. Microarray consists of many unspecified markers and serum is reacted to them for disease diagnosis. We use the reaction between serum and markers for disease diagnosis. Generally, the microarray has many arrays from several hundred to 10 thousands sensing points. The reason that we need so many markers is that we don't know the exact markers for a certain disease. [1][2]. More markers we use in diagnosis, more resource and time are required. Since the neural network (NN) or fuzzy neural network (FNN) that are used to diagnosis disease are a kinds of multi-dimensional none linear function, some markers may contain un-useful information for disease diagnosis and it decreases the accuracy of disease diagnosis. Therefore, we need to get rid of those un-useful substances of features in microarray to enhance accuracy, and to find a way to select efficient features for diagnosis. In this paper, we proposed a useful feature selection and disease diagnosis method by using artificial intelligence of NN and FNN.

This paper is organized as follow. In the section II, the basic concept of NN and FNN, which are widely used to make a decision for uncertain fact such as disease diagnosis, and feature selection method based on NN and FNN will be explained. In section III, the result of proposed algorithm applied to real clinical data will be introduced and discussion will be given to the result. Finally, conclusion will be made in section IV.

II. PROPOSED ALGORITHM

Disease diagnosis and feature selection method based on NN and FNN which are widely used for machine learning is developed with this work [3]~[18].

A. *Neural Network*

We design NN with the most basic structure for algorithm that this paper proposed. The NN we used comes with 3-stage and 10 hidden layers as Fig. 1.

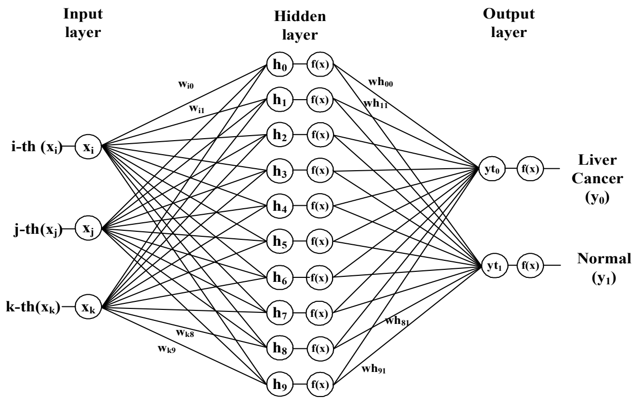

Figure 1. Three-layer neural network structure

The NN use back propagation algorithm to adjust weight of neural network. NN, which is used to verify the proposed algorithm in this paper, uses marker's measurement value as input. The output has two different types, normal and liver cancer. Output of liver cancer (y_0) and normal (y_1) goes through the following steps. Firstly, multiply the hidden layer of 10 nodes by linked weight to get input x_i. Set h_k for the K-th hidden node and then, the result is as (1). w_{ik} means weight which is linked between input layer and hidden layer.

$$h_k = \sum_{i=0}^{n} x_i \cdot w_{ik} \tag{1}$$

h_k values are multiplied by wh_{km} which is the linked weight between hidden layer and output layer. After that, all the values linked to output nodes are added to get the final output value (2). Hereafter, m presents output node and has 0 or 1.

$$yt_m = \sum_{k=0}^{9} h_k \cdot wh_{km} \quad (m = 0, 1) \tag{2}$$

Values by activation function needs to prevent overflow hidden layer and output layer, the function we used is sigmoid function of (3).

$$f(x) = \frac{1}{1+exp^{(-x)}} \tag{3}$$

NN we need to use during the test make use of error back propagation to control weight between each layer. There are two kinds of the weight. One is weight between input layer and hidden layer, and the other is weight between hidden layer and output layer. Since Error back-propagation inversely controls weight from the output, we get weight between output layer and hidden layer from (4) at first. t means weight we used for conducting algorithm, t+1 is newly updated weight.

$$wh_{km}(t+1) = wh_{km}(t) + \Delta wh_{km} \tag{4}$$

Δ wh_{km} in (4) can be expressed as (5)

$$\Delta wh_{km} = \rho \cdot \delta_m \cdot y_m \tag{5}$$

ρ is learning rate, y_m is output from the output node. wh_{km} is error and follows (6).

$$\delta_m = y_m \cdot (1 - y_m) \cdot (y_t - y_m) \tag{6}$$

y_t is an expected output when we enter input. If we use liver cancer data for the input, we have values as y_0 =1, y_1 =0. If we use normal data instead, it becomes y_0 =0, y_1 =1. When the weights between output layers and hidden layers are controlled, hidden layer and input layer will be conducted as (7).

$$w_{km}(t+1) = w_{km}(t) + \Delta w_{km} \tag{7}$$

where, Δw_{km} can be obtained from (8). ρ and y_m are applied as the same value as we used for weight control between output layer and hidden layer.

$$\Delta w_{km} = \rho \cdot \delta_k \cdot y_m \tag{8}$$

The error of k-node in hidden layer is the same as (9).

$$\delta_k = \sum_{m-0}^{1} y_m \cdot (1 - y_m) \cdot (\delta_m \cdot wh_{km}) \tag{9}$$

δ_m is the error value between output layer and hidden layer which we gained in the previous step. NN is possible to make right decision through repeated update of weight; it called learning, with various inputs.

B. Fuzzy Neural Network

FNN is an intelligence model hybrid fuzzy theory and NN. The model we used is a basic linear fuzzy neural network. Fig. 2 shows a structure for an input of i-th marker of used FNN structure. If the designed FNN gets marker's measurement value as input, it changes the value to fuzzy membership value by fuzzification function. The final output y_0 and y_1 can be gathered from sum of each weight connected to learning network.

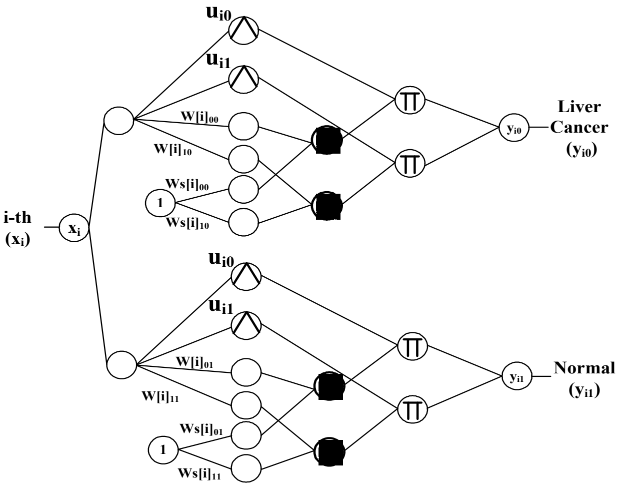

Figure 2. Basic linear FNN structure

If fuzzification function for i-th marker's liver cancer is u_{i1}, i-th marker belongs to liver cancer sample value among all samples will be used to get variation, standard deviation for u_{i0}. Then, getting variation, standard deviation for u_{i1} by using i-th marker values belongs to normal sample. Outputs y_{i0}, y_{i1} for each marker can get gathered from (14). The final output values, y_{i0}, y_{i1}, are sum of all output values of used markers and can be expressed as (15).

$$\begin{aligned} y_{im} = {} & u_{i0}(x_i) \cdot (w[i]_{0m} + ws[i]_{0m}) \\ & + u_{i1}(x_i) \cdot (w[i]_{1m} + ws[i]_{1m}) \quad (m = \; 1, 0) \end{aligned} \tag{14}$$

$$y_m = \sum_{i=0}^{k} y_{im} \quad (\, m = 0, 1) \tag{15}$$

FNN is also an algorithm based on NN and the weight adjustment by learning follows back propagation method. The weights adjusted by learning are w, ws expressed as (16), (17).

$$\mathrm{w}(t+1) = w(t) + \Delta w \tag{16}$$

$$\mathrm{ws}(t+1) = ws(t) + \Delta ws \tag{17}$$

Δw and Δws can be expressed as (18), (19).

$$\Delta \mathrm{w}[\mathrm{i}]_{km} = 2 \cdot \rho \cdot (y_m - y_t) \cdot u_{im}(x_t) \cdot x_t + \alpha \cdot \left(w[i]_{km}(t) - w[i]_{km}(t-1)\right) \quad (k = 0, 1) \tag{18}$$

$$\Delta \mathrm{ws}[\mathrm{i}]_{km} = 2 \cdot \rho \cdot (y_m - y_t) \cdot u_{im}(x_t) + \alpha \cdot \left(ws[i]_{km}(t) - ws[i]_{km}(t-1)\right) \quad (k = 0, 1) \tag{19}$$

where, y_m is expressed in (15) and y_t is an expected output value for each input and is the same as in NN. ρ is learning rate. The 2nd term of (19) is related to learning time and learning improvement. FNN is possible to make right decision through repeated update of weight; it called learning, with various inputs.

C. Feature selection method

Feature selection method that we propose use of learning ability of NN and FNN to prevent features from making bad effect on accuracy of diagnosis. The basic concept is to add up marker one by one and check the result. Finally, we select the best set of marker as feature. The proposed method consists of three parts: pre-screening, select-drop and determination.

1) Pre-screening: It is not efficient way to add uncertain microarray of hundreds of thousand data at random. Also it is hard to get the optimized result from random selection of data. Thus, pre-screening of features, which are likely to be helpful for the disease diagnosis, is necessary to make a effient feature slection method. For this purpose, we use one-way analysis of varaince(ANOVA)[18]. ANOVA is a collection of statistical models used to analyze the differences between group means and their associated procedures (such as "variation" among and between groups). As the result of the one way ANOVA, we can get the p-value of each marker and sort them in acsending order based on p-value.

2) Select-Drop : The basic concept of select-drop step is that we select feature which increase diagnosis accuracy and drop which lower the accuracy, when adding up features based on learning as Fig. 3. In this step, we use the accuracy calculated by using n-fold cross validation(CV) [19]. At the first step we use the 1st marker of pre-screen list to learn through (n-1) group and calculate accuracy 1 group, and save the accuracy as previous result. We add the 2nd marker and carry out the same procedure as in the 1st marker and save accuracy as present result. If the present result is better than or equal to the previous one the 2nd maker is selected as feature, if not, it is rejcted. The selsect-drop steps are carried out as many as the desired number of markers in the pre-screen list. NN and FNN algorithms consist of non-linear functions which were introduced in the section II. Thus, best accuracy cannot be obtained by simply adding the markers with low *p*-value. It is not always true that adding up markers result in the better accuracy. In some cases, accuracy drops although more markers are added. To resolve this problem, after Select-Drop sequence for all the markers, perform this sequence once again for dropped markers. If some markers don't make good result during the early stage Select-Drop sequence, we can possibly expect the good result by combining features in the later stage. N-fold CV results depends on the grouping, so generalization process is needed. For this process, a set of data included in each group selected randomly and large number of the n fold-CV is performed. After gen*e*ralization, feature consist of marker candidates could be gathered.

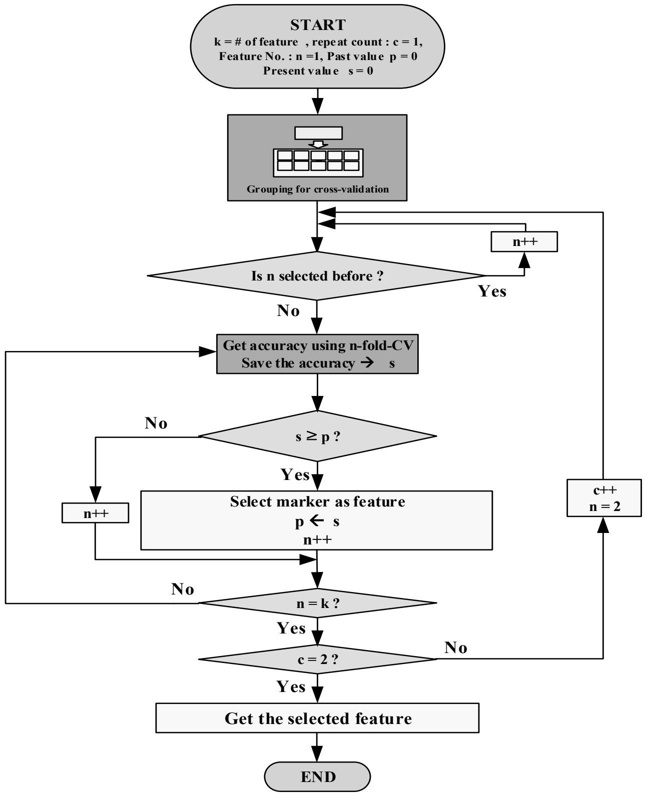

Figure 3. Select-Drop algorithm

3) Determinaiton : After select-drop step, we can get re-ordered feature set. NN and FNN are complex systems, which consist of none linear functions. It means that increasing the number of markers does not guarantee monotonically increased accuracy. Therefore, in order to obtain the best feature we need trial-and-error method experimentally. In this step, we will do rough search to find which number of markers have the best accuracy at first, and then precisely adjusting the number of markers to find the best marker set.

III. Experiment

For verification of the proposed method, raw data of serums from 314 normal and 81 liver cancer patients reacted to 1,142 aptamers with containing a fluorescent material, are used. The aptamer is oligonucleic acid or peptide molecules that bind to a specific target molecule. At the select-drop step, we used 10 fold CV that each group has 31 normal and 8 liver cancer patients as shown in Fig. 4.

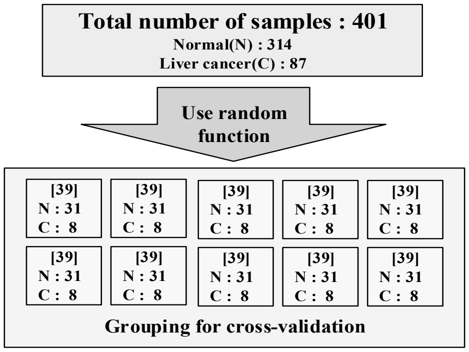

Figure 4. Grouping for 10 fold cross validation

Select-drop algorithm which is shown in Fig. 3 was performed 1,000 times for generalization. The whole procedure was carried out 10 times with different grouping for 10 fold-CV, and the results are shown in table I. Since the time to carry out the above calculation steps is astronomically required, we used MPI(Message Passing Interface) which is kind of parallel programming method and super computer of KIST (Korea Institute of Science and Technology). The super computer was equipped with Intel Xeon 2.5Ghz CPU-144 nodes(threads), Intel Xeon 2.6Ghz CPU-512 nodes(threads), Intel Xeon 2.53Ghz CPU-488 nodes(threads).

Table I. Accuracy of the proposed method

Run	Neural Network		Fuzzy Neural Network	
	Number of aptamers	Accuracy	Number of aptamers	Accuracy
1	136	0.991	236	0.983
2	157	0.993	220	0.982
3	151	0.992	205	0.981
4	142	0.993	232	0.982
5	105	0.991	182	0.982
6	160	0.992	254	0.982
7	111	0.991	252	0.983
8	88	0.992	235	0.982
9	126	0.992	201	0.981
10	141	0.992	245	0.982
Average	131.7	0.9919	226.2	0.9820

As the result, we can detect liver cancer with the accuracy of 99.19 % by average use of 132 aptamers based on NN and 98.19 % by average use of 226 aptamers based on FNN. These results mean that about 80% of 1,142 aptamers are not necessary to detect liver cancer and through the proposed feature selection step, we can find more helpful feature for liver cancer detection.

IV. Conclusion

In this paper, we propose the method that selects useful makers from several hundred markers and detects liver cancer with accuracy level of 98~99%. The feature selection method we proposed can dramatically reduce number of markers without making a sacrifice of diagnosis accuracy. It is meaningful that we can get the same diagnosis accuracy with small numbers of marker because we can save time and make a cost down in disease diagnosis. And also the computational complexity of NN and FNN is normally in proportion to the number of inputs, small numbers of markers lead to lower power computation of POCT (Point –of- Care- Test) equipment. The method is reliable because it is verified by the 400 clinical data of liver cancer and the super computer. The concept of the proposed method is applicable to select feature of other diseases and used to diagnose a various diseases.

Acknowledgment

This work was supported by PLSI supercomputing resources of Korea Institute of Science and Technology Information, Ministry of Science, ICT and Future Planning and IDEC Platform Center (IPC) and BK21PLUS, Creative Human Resource Development Program for IT Convergence.

References

[1] Churchill, G. A., ”Fundamentals of experimental design for cDNA microarrays,” Nature Genetics, Suppl. 32, pp.490-495, 2002.

[2] Leung, Y.F., and Cavalieri D., “Fundamentals of cDNA microarray data analysis”, Trends in Genetics, , 19(11), pp.649-659, 2003.

[3] W.Pedrycz, “An identification algorithm in fuzzy relational system”, Fuzzy Sets Syst., Vol. 13, pp. 153-167, 1984.

[4] T. Takagi, M. Sugeno, “Fuzzy Identification of Systems and Its Applications to Modeling and control”, IEEE SMC-15, No.1, pp.116-132, 1984.

[5] Baxt, W. G., ”Use of an artificial neural network for data analysis in clinical decision making : The diagnosis of acute coronary occlusion,” Neural Computing, Vol. 2, No. 4, pp.480-489, 1990.

[6] E. Rich and K. Knight, “Artificial intelligence”, McGraw-Hill, New York, 1991.

[7] H. Takagi, N. Suzuki, T. Kouda, Y. Kojima, “Neural networks Designed on Approximate Reasoning Architecture and Their Applications”, IEEE trnas. Neural Networks, Vol. 3, No. 5, pp. 752-760, 1992.

[8] S. Horikawa, T. Furuhashi, Y. Uchigawa, “On Fuzzy Modeling Using Fuzzy neural Networks with the Back Propagation Algorithm”, IEEE trans. Neural Networks, Vol. 3, No. 5, pp.752-760, 1992.

[9] J.R. Jang, “ANFIS:Adaptive-network-based fuzzy inference system”, IEEE Trans. Syst., Man, Cybern., 23, 665-685, 1993.

[10] Chin-Teng Lin, C.S. George Lee, “Neural Fuzzy Systems”, Prentice Hall, 1996.

[11] L.P. Wang, “On competitive learng”, IEEE Transactions on Neural Networks, vol. 8, no.5, pp. 1214-1217, 1997.

[12] Y. Frayman and L.P. Wang, “Data mining using dynamically constructed recurrent fuzzy neural networks”, Research and Development in Knowledge Discovery and Data Mining, vol. 1394, pp. 122-131, 1998.

[13] V. Kecman, Learning and Soft Computing, Support Vector mchines, Neural Networks and Fuzzy Loic Models, The MIT Press, Cambraige, MA, 2001.

[14] Sung-Kwun Oh, Witold Pedrycz, T. C. Ahn, “Self-organizing neural networkswith fuzzy polynmial neurons”, Applied Soft Computing, Vol. 1, pp.1-10, Aug. 2002.

[15] Byoung-Jun Park, Witold Pedrycz, Sung-Kwun Oh, "Fuzzy Polynomial Neural Networks:Hybrid Architectures of Fuzzy modeling", IEEE Trans.on Fuzzy Systems, Vol. 10, No. 5, Oct. 2002.

[16] Y. Frayman and L.P. Wang, " A Dynamically-constructed fuzzy nural controller for direct model reference adapive control of multi-input-ouput nonlinear processes", Soft Computing, vol.6, pp.244-253, 2002.

[17] L.P. Wang and X.J. Fu, Data Mining with Computational Intelligence, Springer, Berlin, 2005.

[18] Bing Liu, Chunru Wan, and L.P. Wang, "An efficient semi-unsupervised gene selection method vi spectral biclustering", IEEE Transactions on Nano-Bioscience, vol.5, no.2, pp.110-114, Jun, 2006.

[19] Viv Bewick, Liz Cheek, Jonathan Ball, "Statistics review 9: One-way analysis of variance," , Critical Care, Vol.8, No. 2, April, 2004.

[20] Kun Yang, Haipeng Wang, Guojun Dai, Sanging Hu, Yanbin Zhang, Jing Xu, "Determining the repeat number of cross-validation," Biomedical Engneering and Informatics 4th international conference, Vol. 3,pp. 1957-1960, 2011.