Information Gain and Adaptive Neuro-Fuzzy Inference System for Breast Cancer Diagnoses

M. Ashraf
ISE, University of Canberra
Bruce ACT 2601 AUSTRALIA
Ashraf.ali@uni.canberra.edu.au

Kim Le
ISE, University of Canberra
Bruce ACT 2601 AUSTRALIA
Kim.le@canberra.edu.au

Xu Huang
ISE, University of Canberra
Bruce ACT 2601 AUSTRALIA
Xu.Huang@canberra.edu.au

Abstract- **This paper presents a new approach for breast cancer diagnosis using a combination of an Adaptive Network based Fuzzy Inference System (ANFIS) and the Information Gain method. In this approach, the ANFIS is to build an input-output mapping using both human knowledge and machine learning ability and the information gain method is to reduce the number of input features to ANFIS. An experimental result shows 98.23% accuracy which underlines the capability of the proposed algorithm.**

Keywords- Breast cancer diagnoses, Adaptive Neuro-Fuzzy Inference Systems, Information Gain, Sugeno Inference System.

I. INTRODUCTION

Yearly around the world, millions of ladies suffer from breast cancer, making it the second common non-skin cancer after lung cancer, and the fifth cause of death among cancer diseases in the world [1]. Discovering the disease in its early stages may reduce the breast cancer tragedy. Computing and machine learning tools can be used to assist physicians in diagnosing and predicting the disease so they can provide necessary treatment and prevent the impacts, including the possibility of death.

With advances in computer technologies, medical expert systems and Computer Aided Diagnosis (CAD) tools become one of the foremost research areas in the field of medical diagnoses. The aim is to design a complete expert system that combines the human expertise and the technology intelligence to achieve more accurate diagnosis. In addition, it may speed up the diagnoses, and reduce the errors and mistakes occurred by human being [2].

Übeyli [3] aimed to integrate adaptive neural fuzzy inference system (ANFIS) for breast cancer diagnoses. The author used a database of patients with known diagnosis (i.e. supervised learning). An ANFIS classifier was trained with a set of records for nine examined features for breast cancer, and then was used to diagnose new cases. The system joined the neural network learning ability and the fuzzy modeling approach. The performance of Übeyli's ANFIS-based model has capability to diagnose the disease with 98% accuracy. Song et al. [4] presented a system for automatic breast cancer diagnosis with improved computational performance due to input data selection. They focused on obtaining higher computational performance in collaboration between Genetic Algorithm (GA) and Fuzzy Neural Network. They also showed that input reduction can be used for many other problems which have high complexity and strong non-linearity with huge data to be analysed.

Arulampalam and Bouzerdoum [5] have proposed a method for diagnosing breast cancer and diabetes called Shunting Inhibitory Artificial Neural Networks (SIANNs). SIANN is a neural network stimulated by human biological networks in which the neurons interact among each other's via a nonlinear mechanism called shunting inhibition. The feed forward SIANNs have been applied to several medical diagnosis problems and the results were more favorable than those obtained using Multilayer Perceptrons (MLPs). In addition, a reduction in the number of inputs was investigated. Setiono [6] has explained how the pre-processing of data set can improve the accuracy of the neural network and the accuracy of the rules because some rules may be extracted from human experience, and may be erroneous. The data pre-processing involves the selection of significant attributes and the elimination of records with missed attribute values. The rules generated by Setiono's method were more brief and accurate than those generated by other methods mentioned in the literature. Meesad and Yen [7] have proposed a Hybrid Intelligent System (HIS) which integrates the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules have been determined based on knowledge embedded in the trained ILFN or been extracted from real experts. In addition, the method also utilized Genetic Algorithm (GA) to reduce the number of the linguistic rules that sustain high accuracy and consistency. The proposed method has been evaluated using Wisconsin Breast Cancer Diagnosis (WBCD) data set. The results have shown that the proposed HIS perform better than some well-known methods.

II. Adaptive Neural Fuzzy Inference System (ANFIS)

A. ANFIS Structure

Adaptive Neural Fuzzy Inference System (ANFIS), proposed by Jang in 1993 [9], is a combination of two machine learning approaches: Neural Network (NN) and Fuzzy Inference System (FIS). ANFIS compromises the advantages of NN and FIS by combining the human expert knowledge (FIS rules) and the ability to adapt and learn (NN). For simple illustration, suppose the fuzzy system contains two Sugeno fuzzy rules:

$$Rule1: IF\ x\ is\ A_1 AND\ y\ is\ B_1, THEN\ f = p_1 x + q_1 y + r_1$$

$$Rule2: IF\ x\ is\ A_2 AND\ y\ is\ B_2, THEN\ f = p_2 x + q_2 y + r_2$$

Fig. 1 (a) shows the fuzzy reasoning and Fig.1 (b) shows the corresponding structure of ANFIS. In Fig. 1(b), the node function in each layer is as the follows:

Layer1: Each node i (represented by a square) in this layer accepts input and computes the membership $\mu_{A_i}(x)$.

$$O_i^1 = \mu_{A_i}(x) \quad (1)$$

where x is the input to node i, and A_i is the linguistic label (small, large, etc.) associated with this node. In other words, O_i^1 is the membership function of A_i and it specifies the degree to which the given x satisfies the quantifier A_i .Usually we choose $\mu_{A_i}(x)$ to be bell-shaped with values between 0 and 1, such as the generalized bell function:

$$\mu_{A_i}(x) = \exp\left[-\left(\frac{x-c_i}{a_i}\right)^2\right] \quad (2)$$

where a_i and c_i are two parameters called premises.

Layer2: Every node in this layer (represented by a circle) takes the corresponding outputs from Layer 1 and multiplies them to generate a weight:

$$\overline{w} = \mu_{A_i}(x) \times \mu_{B_i}(x),\ i=1,2. \quad (3)$$

The output of this node represents the firing strength of the rule.

Layer3: Every node in this layer is a circle node labelled *N*. This layer normalize the weight of a certain node in compare to the sum of other nodes weights (The ration of weight) then compute the implication of each output member function.

$$\overline{w}_i = \frac{w_i}{\sum_j w_j},\ i=1,2.\ j=2. \quad (4)$$

Layer 4: Every node in this layer is illustrated with a square. Based on Sugeno inference system, the output of a rule can be written on the following linear format:

$$O_i^4 = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i) \quad (5)$$

where p_i, q_i, are the consequent parameters and r_i is the bias.

Layer 5: This layer called the aggregation layer, which computes the summation of rules, the proposed algorithm produce a single output (centroid):

$$O_i^5 = final\ output = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (6)$$

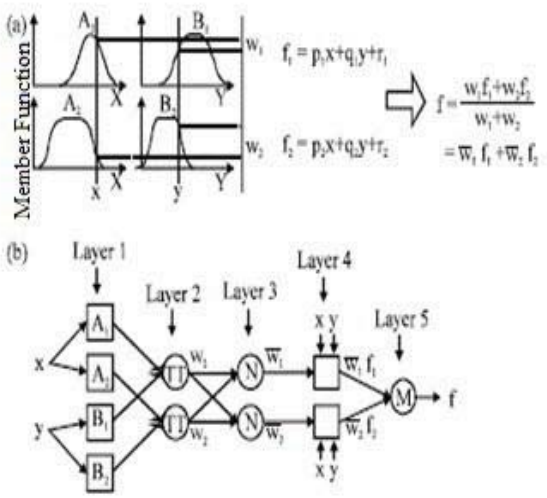

Fig. 1. (a) Fuzzy reasoning (b) equivalent ANFIS Structure

B. ANFIS learning

The method to train ANFIS is the hybrid learning algorithm which uses the gradient descent method and Least Square Estimate (LSE). Each cycle of the hybrid learning consists of a forward pass and a backward pass. In the forward pass the signal travels forward until Layer 4 and the consequent parameters are identified using the LSE method. In the backward pass the errors are propagated backward and the premise parameters are updated by gradient descent. The process repeated until achieving the lowest error [9].

III. Information Gain & Waikato Environment For Knowledge Analysis (WEKA)

Machine can learn about a certain problem from a set of prior information (database or dataset) based on the importance and the rank for each attribute in the dataset [8]. The information gain method was proposed to approximate quality of each attributes using the entropy by estimating the difference between the prior entropy and the post entropy [8]:
when C and V are discrete variables then the prior entropy of C is:

$$Ent(C) = -\sum_C P(C)\log_2 P(C) \quad (7)$$

where $P(C)$ is the probability function of variable C. The conditional entropy of C given V (post entropy):

$$Ent(C|V) = \sum_V P(V)Ent(C|V) \quad (8)$$

$$= -\sum_V P(V)\ \sum_C P(C|V)log_2\, P(C|V) \quad (9)$$

The information gain $IG(C;V)$ is:

$$IG(C;V)\ = Ent\ (C) -\ Ent\ (C|V) \quad (10)$$

$$IG(C;V) = -\sum_C P(C)\ log_2\, P(C) - \sum_V(-P(V) \times \sum_C P(C|V)log_2\, P(C|V)) \quad (11)$$

WEKA provides the environment to calculate the information gain. WEKA is an open source machine learning software written in JAVA language. WEKA contains some

data mining and machine learning methods for data pre-processing, classification, regression, clustering, association rules, and visualization [10].

IV. The Proposed Method

Our proposed approach is to combine the information gain method and ANFIS method for diagnosing diseases such as breast cancer. The information gain will be used for selecting the quality of attributes. The output of applying the information gain method is a set of features with high ranking values, the set of high ranked features will be the input for ANFIS. The selected features will be applied to ANFIS to train and test the proposed approach. The structure of the proposed approach is shown in Fig. 2 where $X = \{x_1, x_2, .., x_n\}$ are the original features in dataset, $Y = \{y_1, y_2, .., y_k\}$ are the features after applying the information gain (features selections), and Z denotaed to the final output after applying Y on ANFIS (the diagnose).

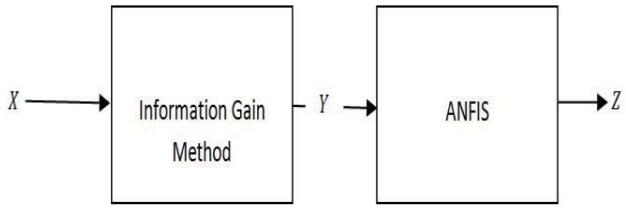

Fig. 2. The general structure for the proposed approach

V. The Experimintal Results

The database, Wisconsin Breast Cancer Diagnosis (WBCD), have been created by William Wolberg *et al.* [13] from the University of Wisconsin-Madison, USA. The database attributes were collected from digital fine needle aspirate (FNA) of breast mass.
WBCD contains 699 records. Each record consists of 9 features plus the class attribute. Table I shows the statistical details of the data.

In our experiment, the database was divided into training and testing datasets. 341 records used for training and 342 records for testing. We have ignored the records which contain the missing values (16 records). The class attributes have been normalized to [0=Benign, 1=Malignant]. The information gain method has been used to select the quality of attributes. Table II shows the ranking of attributes after applying the attribute evaluator *InfoGainAttributeVal* and the searching method *Ranker-T-1* using WEKA on WBCD dataset.

TABLE I
WBCD Statistical Details

#	Attribute	Value Range
1	Clump Thickness	1 – 10
2	Uniformity of Cell Size	1 - 10
3	Uniformity of Cell Shape	1 - 10
4	Marginal Adhesion	1 - 10
5	Single Epithelial Cell Size	1 - 10
6	Bare Nuclei	1 - 10
7	Bland Chromatin	1 - 10
8	Normal Nucleoli	1 - 10
9	Mitoses	1 - 10
10	Class	Benign, Malignant

TABLE II
Information Gain Ranking Using WEKA On WBCD

Attribute	Rank
Uniformity of Cell Size (UCSize)	0.636
Uniformity of Cell Shape (UCSshape)	0.633
Normal Nucleoli (NN)	0.555
Bare Nuclei (BN)	0.538
Single Epithelial Cell Size (SECS)	0.421
Clump Thickness (CT)	0.411
Marginal Adhesion (MA)	0.394
Bland Chromatin (BC)	0.316
Mitoses(MI)	0.278

Fig. 3 shows the graph of table II**.** It shows the most significant change in the graph (the slope point) which gave us an indication to choose the first four top ranking features located above the slope point as the recommended number of features to be used later as inputs to ANFIS. At this stage, the attributes have been deducted and the recommended number of features has been set to 4 features.

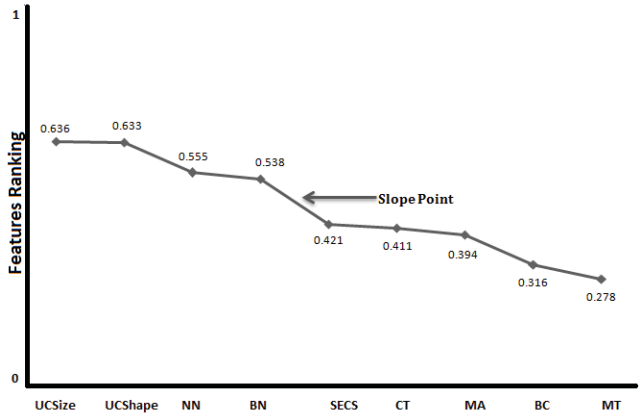

Fig. 3. Information Gain Ranking

The second stage is to construct the fuzzy inference system (FIS). In our proposed approach, we have used Sugeno Fuzzy Inference system that maps feature to feature membership functions, feature membership function to rules, rules to a set of output, output to output membership functions, and the output membership function to a single-valued output as shown in fig 4. The membership function maps input with a membership values as shown in fig. 5.

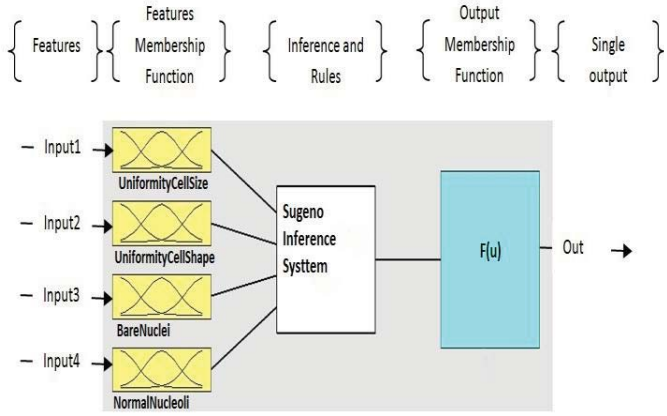

Fig. 4. Sugeno Fuzzy Inference System with 4 features input

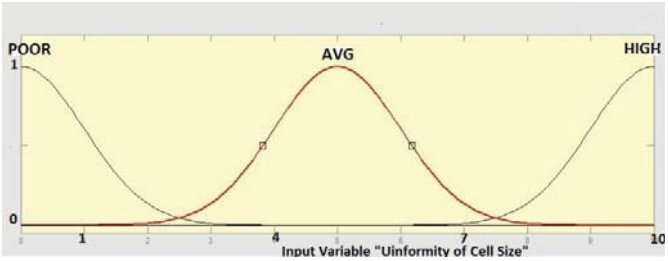

Fig. 5. Input Membership Function for "Uniformity of Cell Size"

In addition to the membership function, FIS contains the rules that add human reasoning capabilities to machine intelligences, which are usually based on Boolean logic. In our proposed approach, the rules have been defined from the real data. The rules express the weight of each feature by giving higher priority for features that have the highest rank. The proposed approach contains 81 rules (Number of rules = x^y where x is the Number of member functions and y is the number of features i.e.3^4). The following are examples of some rules used in the proposed approach:

IF AND (*UniformityCellSize is poor, UniformityCellShape is Avg, BareNuclei is poor, NormalNucleoli is poor*) THEN *(output is OK*)

IF AND (*UniformityCellSize is poor, UniformityCellShape is high, BareNuclei is poor, NormalNucleoli is avg*) THEN *(output is NOT_OK*)

In the final stage, the constructed Fuzzy Inference System loaded to ANFIS which will train and test the proposed approach as shown in Fig. 6. The structure of ANFIS on MATLAB is shown in Fig. 7.

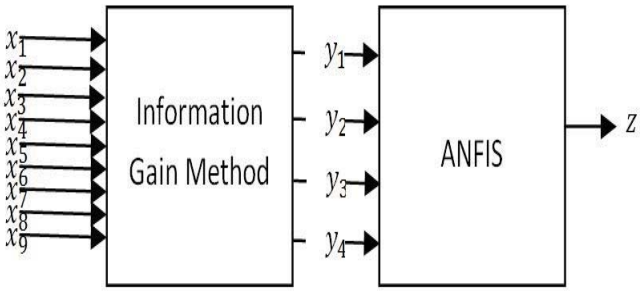

Fig. 6. The structure for the proposed approach

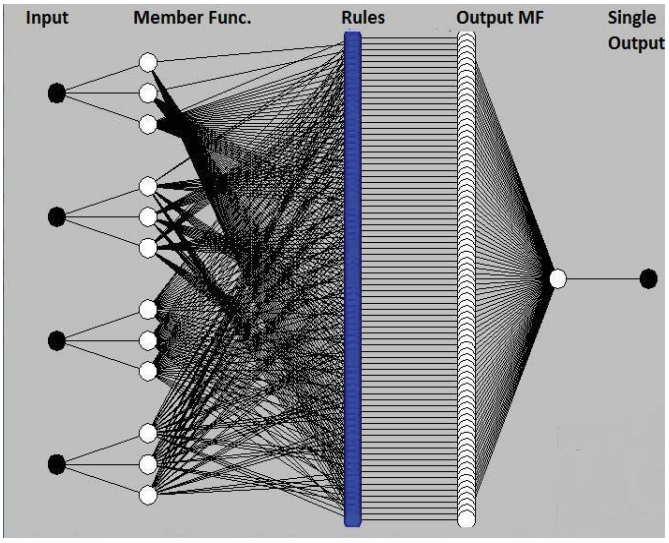

Fig. 7. ANFIS Structure on MATLAB

The result of applying ANFIS on the features selected using the information gain on WBCD dataset shows 98.24% accuracy. The results of previous work (using the same dataset) are shown in Table III:

TABLE III
COMPARISON WITH PREVIOUS WORK

The approach	Accuracy	Ref
AdaBoost	57.60%	[11]
ANFIS	59.90%	[11]
SANFIS	96.07%	[4]
FUZZY	96.71%	[7]
FUZZY-GENETIC	97.07%	[12]
ILFN	97.23%	[7]
NNs	97.95%	[6]
ILFN&FUZZY	98.13%	[7]
IGANFIS (our method)	98.24%	
SIANN	100.00%	[5]

VI. DISCUSSION

We have proposed a new approach for diagnosing the breast cancer by reducing the number of features to the optimal number using the information gain and then apply the new dataset to the Adaptive Neuro Fussy system (ANFIS). We found that the accuracy for the proposed approach is 98.24% compared with other methods. In our future work, we will concentrate on the computation time for our proposed approach and apply it on more databases.

REFERENCES

[1] International Agency for Research on Cancer, "Mammography Screening Can Reduce Deaths from Breast Cancer," *International Agency for Research on Cancer.* Press Release No. 139, March 2002. Available: http://www.iarc.fr/en/media-centre/pr/2002/pr139.html, [Accessed: Mar. 20, 2010].

[2] J. Giarratano and G. Riley, *Expert Systems Principles and Programming*, 2nd ed., vol. 1. Boston: MA, 1994.

[3] E. Übeyliand. "Adaptive Neuro-Fuzzy systems for Automatic Detection of Breast Cancer," *Journal of Medical Systems*, vol. 33, No. 5, pp. 353 – 358, October 2009.

[4] H. Song, S. Lee, D. Kim and G. Park. "New Methodology of Computer Aided Diagnostic system on Breast Cancer," in *Second International Symposium on Neural Networks*, 2005, pp. 780 -789.

[5] G. Arulampalam and A. Bouzerdoum. "Application of Shunting Inhibitory Artificial Neural Networks to Medical Diagnosis," in *Seventh Australian and New Zealand Intelligent Information Systems Conference*, 2001, pp. 89 -94.

[6] R. Setiono. "Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis," in *Artificial Intelligence in Medicine*, 2000.

[7] P. Meesad and G. Yen. "Combined Numerical and Linguistic Knowledge Representation and Its Application to Medical Diagnosis," in *Component and Systems Diagnostics, Prognostics, and Health Management II*, 2003.

[8] I. Kononenko. "Estimating attributes: Analysis and extensions of RELIEF," in *Machine Learning ECML-94*, 1994.

[9] J. Shing and R. Jang. "ANFIS: Adaptive-Network-based Fuzzy Inference system", *IEEE transactions on systems*, vol. 23, no. 3, pp. 665 – 685, June 1993.

[10] H. Lan and F. Eibe, *Data mining: practical machine learning tools and techniques*, 2nd ed., San Francisco: Diane Cerra, 2005.

[11] W. Land and E. Veheggen. "Experiments Using an Evolutionary Programmed Neural Network with Adaptive Boosting for Computer Aided Diagnosis of Breast Cancer," in *IEEE International Workshop on Soft Computing in Industrial Application*, 2003.

[12] C. Pena-Reyes and M. Sipper. "Designing Breast Cancer Diagnostic System via a Hybrid Fuzzy-Genetic Methodology," in *IEEE International Fuzzy Systems Conference Proceeding*, 1999.

[13] O. Mangasarian and W. Wolberg. "Cancer diagnosis via linear programming", SIAM News, vol. 23, no. 5, September 1990.