Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM

Fatemeh Saiti[1], Afsaneh Alavi Naini[2]
[1]Biomedical Engineering Group, Electrical Engineering Department, K. N. Toosi University of Technology, Tehran, Iran
Fatemeh.saiti@gmail.com, Afsaneh_alavi@yahoo.com

Mahdi Aliyari shoorehdeli[1], Mohammad Teshnehlab[1]
[2]Mechatronics Department, Science and research branch Islamic Azad University, Tehran, Iran
M_aliyari@eetd.kntu.ac.ir, Teshnehlab@eetd.kntu.ac.ir

Abstract—**Thyroid gland produces thyroid hormones to help the regulation of the body's metabolism. The abnormalities of producing thyroid hormones are divided into two categories. Hypothyroidism which is related to production of insufficient thyroid hormone and hyperthyroidism related to production of excessive thyroid hormone. Separating these two diseases is very important for thyroid diagnosis. Therefore Support Vector Machines and Probabilistic Neural Network are proposed to classification. These methods rely mostly on powerful classification algorithms to deal with redundant and irrelevant features. In this paper feature selection is argued as an important problem via diagnosis and demonstrate that GAs provide a simple, general and powerful framework for selecting good subsets of features leading to improved diagnosis rates. Thyroid disease datasets are taken from UCI machine learning dataset.**

Keywords- Genetic algorithms; Probabilistic Neural Network; Support Vector Machine; Thyroid disease diagnosis.

I. INTRODUCTION

Thyroid gland produces thyroid hormones to help the regulation of the body's metabolism. It produces two active hormones, levothyroxine (abbreviated T4) and triodothyroine (abbreviated T3). These hormones are important in the production of proteins, in the regulation of the body temperature, and in overall energy production and regulation. The seriousness of thyroid disorders should not be underestimated as thyroid storm (an episode of severe hyperthyroidism) and myxedema coma (the end stage of untreated hypothyroidism) may lead to death in a significant number of cases [1].

Hyperthyroidism, or an overactive thyroid, may also be caused by inflammation of the thyroid, various kinds of medications, and lack of the most common causes is Graves' disease. Graves' disease happens when the body makes proteins that constantly tell the thyroid to make more thyroid hormones [1]. Most thyroid problems can be treated successfully. In this case thyroid function diagnosis is an important classification problem [1, 2]. Various new methods, such as pattern recognition techniques, fuzzy classifiers, artificial immune recognition system, neural networks, neuro fuzzy, etc, have been used to diagnose thyroid disease [1, 2, 7]

This paper is organized as follows, in section II, thyroid datasets are introduced. In section III, feature selection is explained by genetic algorithms. In section IV, SVM and PNN are introduced as two kinds of classifiers. Then we discuss about the results of comparing them in section V. Finally, section VI presents our conclusion and plans for future.

II. THYROID DATASETS

In order to perform the research reported in this manuscript, two sets of thyroid datasets taken from the UCI machine learning respiratory were used [9].The first dataset includes 215 samples and 5 features, the second datasets includes 7200 samples and 21 features. Both two datasets consist of 3 classes which are normal, hyperthyroidism and hypothyroidism.

III. FEATURE SELECTION WITH GENETIC ALGORITHM

GA is a class of optimization procedures inspired by the biological mechanism of reproduction. In the past, they have been used to solve various problems including target recognition [12], object recognition [15], face recognition [15], and face detection/verification [15]. GAs operate iteratively on a population of structures, each on of which represents a candidate solution to the problem at hand, properly encoded as a string of symbols(e.g. binary). A randomly generated set of such string forms the initial population from which the GA starts its search. Three basic genetic operators guide this search: selection, crossover, and mutation. The genetic search process is iterative: evaluating, selection, and recombining string in the population during each iteration (generation) until reaching some termination condition. Evaluation of each string is based on a fitness function that is problem-dependent. It determines which of the candidate solutions are better. This corresponds to the environmental determination of survivability in national selection. Selection of a string, which represents a point in the search space, depends on the string's fitness relative to those of other strings in the population, those points that have relatively low fitness. Mutation, as in natural systems, is a very low probability operator and just flips bit. Mutation plays the rule of restoring lost genetic material. Crossover in contrast is applied with high probability. It is a randomized yet structured operator that allows information exchange between points. Its goal is to preserve the fittest individual without introducing any new value. The goal of feature subset selection is to use less features to achieve the same or better performance. Therefore the fitness evaluation

978-1-4244-2902-8/09/$25.00 ©2009 IEEE

contains two terms: (1) accuracy and (2) the number of features selected.

We used the fitness function shown below to combine the two terms:

$$Fitness = error + \alpha \times ones \quad (1)$$

Where error corresponds to the classification error and ones corresponds to the number of features selected (i.e., ones in the chromosome). In this function α is considered between (0, 1) and the higher α results in less selected features. In this paper $\alpha = 0.01$ is chosen.

In Fig.1, GA is used to find an optimal binary vector, where each bit is associated with a feature. If the i^{th} bit of this vector equals to 1, then the i^{th} feature is allowed to participate in classification; if the bit is a 0, then the corresponding feature does not participate.

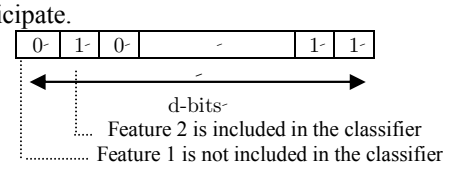

Figure 1. A d-dimensional binary vector, comprising a single member of the GA population for GA-based feature selection.

IV. PROBABILISTIC NEURAL NETWORK

The probabilistic neural network (PNN) was introduced by Specht [4]. It is a supervised neural network that is widely used in the area of pattern recognition, nonlinear mapping, and estimation of the probability of class membership and likelihood ratios. It is closely related to Bayes classification rule and Parzen nonparametric probability density function estimation theory [3]. The fact that PNNs offer a way to interpret the network structure in terms of probability density functions is an important merit of this type of networks. This characteristic renders PNNs faster to train compared to feed forward neural networks. The structure of a PNN is similar to that of FNNs, although the architecture of a PNN is limited to four layers; the input layer, the pattern layer, the summation layer, and the output layer, as illustrated in Fig 2.

An input vector $X = (x_1,...,x_n)^T \in \Re^n$, is applied to the n input neurons and is passed to the pattern layer. The neurons of the pattern layer are divided into groups, one for each class. The i^{th} pattern neuron in the k^{th} group computed its output using a Gaussian kernel of the form.

$$F_{k,i}(X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\| X - X_{k,i} \|^2}{2\sigma^2}\right) \quad (2)$$

Where $X_{k,i} \in \Re^n$ is the center of the kernel, and σ, also known as the *spread (smoothing) parameter*, determines the size of the receptive field of the kernel. The summation layer of the network computes the approximation of the conditional class probability functions through a combination of the previously computed densities [3].

$$G_k(X) = \sum_{i=1}^{M_k} w_{ki} F_{k,i}(X), \qquad k \in \{1,.....,k\} \quad (3)$$

Where M_k is the number of pattern neurons of class k, and w_{ki} are positive coefficients satisfying $\sum_{i=1}^{M_k} w_{ki} = 1$. Pattern vector X is classified to belong to the class that corresponds to the summation unit with the maximum output:

$$C(X) = \arg \max_{1 \subseteq k \subseteq K} (G_k) \quad (4)$$

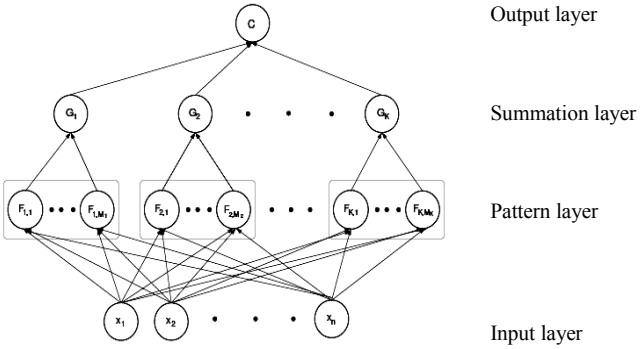

Figure 2. A Probabilistic Neural Network[3].

V. SVM CLASSIFIER

A. Linear SVM Classifier

In this section, we briefly describe the basic SVM concepts for typical two-class classification problems. We begin with the simplest case, in which the training patterns are linearly separable. That is, there exists a linear function of the form:

$$f(x) = w^T x + b \quad (5)$$

Such that for each training sample, the function yields $f(x_i) \geq 0$ for $y_i = +1$, and $f(x_i) < 0$ for $y_i = -1$. In other words, training examples from the two different classes are separated by the hyperplane $f(x) = w^T x + b = 0$. For a given training set, while there may exist many hyperplanes that separate the two classes, the SVM classifier is based on the hyperplane that maximizes the separating margin between the two classes as illustrated in Fig. 3 [5,14].

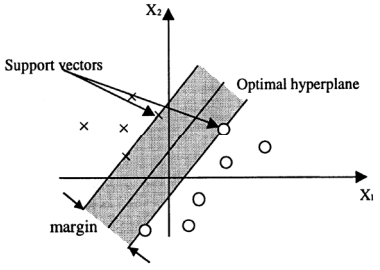

Figure 3. SVM classification with a hyperplane that maximizes the separating margin between the two classes (indicated by data points marked by "X "s and "O "s). Support vectors are elements of the training set that lie on the boundary hyperplanes of the two classes.

In other words, SVM classifier finds the hyperplane that causes the largest separation between the decision function values for the "borderline" samples from the two classes.

Mathematically, this hyperplane can be found by minimizing the following cost function:

$$J(w)=\frac{1}{2}w^T w=\frac{1}{2}\|w\|^2 \quad (6)$$

Subject to the separability constraints

$$w^T x_i+b\geq +1, \qquad \text{for } y_i=+1 \quad (7)$$

Or

$$w^T x_i+b\leq -1 \qquad \text{for } y_i=-1;\ i=1,2,\ldots,l. \quad (8)$$

Equivalently, these constraints can be written more compactly as:

$$y_i\left(w^T x_i+b\right)\geq 1; \qquad i=1,2,\ldots,l. \quad (9)$$

This specific problem formulation may not be useful in practice because the training data may not be completely separable by a hyperplane. In this case, slack variables, denoted by ξ_i , can be introduced to relax the separability constraints in (9) as follows:

$$y_i\left(w^T x_i+b\right)\geq 1-\xi_i \qquad \xi_i\geq 0; i=1,2,\ldots,l. \quad (10)$$

The cost function in (6) can be modified as follows:

$$J(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i \quad (11)$$

where is a user-specified, positive, regularization parameter. In (11), the variable ξ is a vector containing all the slack variables $\xi_i, i=1,2,\ldots,l$.

The modified cost function in (11) constitutes the so-called *structural risk*, which balances the *empirical risk* (i.e., the training errors reflected by the second term); with model complexity (the first term).The regularization parameter C controls this trade-off. The purpose of using model complexity to limit the optimization of empirical risk is to avoid *over fitting*, a situation in which the decision boundary too precisely corresponds to the training data, and thereby fails to perform well on data outside the training set.

B. Nonlinear SVM Classifier

The linear SVM can be readily extended to a nonlinear classifier by first using a nonlinear operator $\Phi(\cdot)$ to map the input pattern x into a higher dimensional space H . The nonlinear SVM classifier obtained from this procedure is defined as:

$$f(x)=w^T\Phi(x)+b \quad (12)$$

which is linear in terms of the transformed data $\Phi(x)$, but nonlinear in terms of the original data $x\in R^n$. Following nonlinear transformation, the parameters of decision function $f(x)$ are determined by the following equation:

$$\min\ J(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i \quad (13)$$

Subject to

$$y_i\left(w^T\Phi(x_i)+b\right)\geq 1-\xi_i, \qquad \xi_i\geq 0; i=1,2,...,l. \quad (14)$$

The mapping $\Phi(\cdot)$ enters the problem only implicitly through the kernel function $K(\cdot,\cdot)$, thus, it is only necessary to define $K(\cdot,\cdot)$, which implicitly defines $\Phi(\cdot)$. In this paper, we consider Gaussian RBF as our kernel function which is defined as follow:

$$K(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (15)$$

where $\sigma>0$ is a constant that defines the kernel width.

VI. Simulation and Result

The classification accuracies obtained by this and other studies for two of thyroid disease datasets were presented in Table I and Table II. In this paper, sensitivity and specificity are computed in this way. Sensitivity is the number of true positive decisions/number of actually positive cases, and specificity is the number of true negative decisions/number of actually negative cases. A true positive decision occurs when the positive detection of the classifier coincides with a positive detection of the physician. A true negative decision occurs when both the classifier and the physician suggest the absence of a positive detection. In this study we use the classification accuracy as performance measures:

$$\text{Classification accuracy } (N)=\frac{\sum_{i=1}^{|N|}calculate(n_i)}{|N|}, n_i\in N \quad (16)$$

$$calculate(n)=\begin{cases}1 & if\ classify(n)=nc\\ 0 & otherwise\end{cases} \quad (17)$$

where N is the set of data items to be classified (the test set), $n\in N$, nc is the class of the item n, *classify (n)* returns the classification of n by PNN and SVM. So the comparison between this study and other studies can be made easily without any problem. The results of our approach are shown in Table I, II. As they illustrate the second dataset SVM classifier has given more accurate results than other methods.

The first dataset includes 215 samples and 5 features which consist of 3 classes which are normal, hyperthyroidism and hypothyroidism. The results of second dataset are shown in Table II.

TABLE I. THE RESULTS OF CLASSIFICATION ACCURACIES, SENSITIVITY AND SPECIFICITY FOR THE FIRST THYROID DISEASE DATASET

study	*Method*	*Classification accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*
Ozylmaz, Yildirim (2002)	MLNN[a] MLNN RBF[b] CSFNN[c]	86.33 89.80 79.08 91.14	-	-
Polat (2007)	AIRS[d]	81.00	-	-
Feyzulla Temurts (2007)	MLNN with LM[e] Learning vector quantization	93.19 90.05	-	-
Proposed method	PNN with GA feature selection	100	100	100
	SVM with GA feature selection	100	100	100

a. Multilayer neural network

b. Radial basis function

c. conic section function neural network

d. Artificial immune recognition system

e. Levenberg marquardt

TABLE II. THE RESULTS OF CLASSIFICATION ACCURACIES, SENSITIVITY AND SPECIFICITY FOR THE SECOND THYROID DISEASE DATASET AND COMPARISON WITH OTHER STUDIES.

study	*Method*	*Classification accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*
A.Sierra, A.Echeverría (2006)	EDA[a] WEDA[b]	98.06 98.00	-	-
Włodzisław Duch (2000)	Feature space mapping	97,9	-	-
Shigeo Abe(1997)	A Fuzzy Classifier with Ellipsoidal Regions	93.34	-	-
Michael L. Raymer (2003)	Hybrid Bayes Classifier	97.20	-	-
A. Sierra (2002)	HOFDA[c]	97.73	-	-
Proposed method	PNN with GA feature selection	96.8	84.2	97.9
	SVM with GA feature selection	99.02	93.33	99.7

a. Evolutionary discriminate analysis

b. Wrapped Evolutionary discriminate analysis

c. High order Fisher discriminate analysis

The second dataset includes 7200 samples and 21 features which consist of 3 classes which are normal, hyperthyroidism and hypothyroidism. The features selected by genetic algorithm in this study are the same as clinical features to diagnose thyroid disease.

VII. CONCLUSION

This paper presents a comparative study on thyroid disease diagnosis by using two kinds of classifiers; Support vector Machines and Probabilistic Neural Network. The results were also compared with the results of the previous studies. The classification accuracies obtained by this study is better than those obtained by previous works. As it seen in the resulting tables, both PNN and SVM have given the same efficient results, but according to the second dataset, it is observed that SVM has performed better.

REFERENCES

[1] L. Ozilmaz, T. Yildirim, "Diagnosis of thyroid disease using artificial neural network method," ICONIP'029th international conference on neural Information, pp. 2033-2036, 2002.

[2] K. Polat, S. Sahan ,S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis," Selcuk University, Expert System with application pp. 1141-1147 ,2007.

[3] L. Georgiou, N.G. Pavlidis, K.E. Parsopoulos, "Optimizing the Performance of Probabilistic Neural Networks in a Bioinformatics Task," University of Patras Artificial Intelligence Research center, 2003.

[4] D.F. Specht, "Probabilistic neural networks," Neural Networks, 1(3), pp. 109-118, 1990.

[5] S. Theodoridis, "Pattern Recognition" second edition Academic Press An imprint Of Elsevier Science, ISBN 0-12-685875-6, 2003.

[6] M.L. Ramer, "Knowledge Discovery in Biological Dataset Using a Hybrid Bayes classifier/Evolutionary Algorithm," IEEE Trans, on Bioinformatics and Bioengineering, 2003.

[7] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," Sakarya University, Expert System with Application, 2007.

[8] S. Abe and R. Thawonmas, "A fuzzy classifier with Ellipsoidal regions," IEEE Trans. Fuzzy Syst, Vol.5, No.3, Aug 1997.

[9] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/thyroid

[10] A. Sierra, "High order Fisher's discriminants," Pattern Recogn., vol. 35, pp. 1291-1302, 2002.

[11] W. Duch, R. Adamczack, and K. Grabczewski, "A new methodology of extraction, optimization and application of crisp and fuzzy logical rules," IEEE Trans. Neural Netw., vol. 12, pp. 277-306, 2001.

[12] A. Katz, P. Thrift, "Generation image filters for target recognition by genetic learning," IEEE Trans. Pattern Anal. Mach. Intell. 16, 1994.

[13] A. Sierra and A. Echeverria, "Evolutionary Discriminant Analysis," IEEE Trans. Evolutionary computation, vol. 10, no. 1, Feb.2006.

[14] C. J. Burges, "A tutorial on support vector machines for pattern recognition on," Knowledge Discovery and Data Mining, vol. 2, pp. 121–167, June 1998.

[15] Z. Sun, G. Bebis, R. Miller, "Object detection using feature subset selection ,"Pattern recognition 372165, 2004.