# Unsupervised Neuro-fuzzy Feature Selection

Jayanta Basak,
(*jayanta@sponge.riken.go.jp*)
*Lab for Information Synthesis,*
RIKEN Brain Science Institute,
*Institute of Physical and Chemical Research (RIKEN),*
*2-1 Hirosawa, Wakoshi*
*Saitama 351-01,*
*Japan.*

Rajat K. De and Sankar K. Pal
(*res9318@isical.ernet.in*) (*sankar@isical.ernet.in*)
*Machine Intelligence Unit,*
*Indian Statistical Institute,*
*Calcutta 700035,*
*India.*

**Abstract**

*This article describes a neuro-fuzzy methodology for feature selection under unsupervised training. The methodology includes connectionist minimization of a fuzzy feature evaluation index. A concept of flexible membership function incorporating weighted distance is introduced in the evaluation index to make the modeling of clusters more appropriate. A set of optimal weighting coefficients in terms of networks parameters representing individual feature importance is obtained through connectionist minimization. Besides this, another algorithm is developed for ranking different feature subsets using the aforesaid fuzzy evaluation index without neural networks. Results demonstrating the effectiveness of the algorithms for various real life data are provided.*

## 1. Introduction

FEATURE selection is a process by which a sample in an $n$-dimensional measurement space is transformed into a point in an $n'$-dimensional ($n' < n$) feature space. The problem of feature selection deals with choosing some of the features from the measurement space to constitute the feature space. The main objective of feature selection is to retain the optimum salient characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms can be devised for efficient categorization.

Fuzzy set theory enables one to deal with uncertainties in different tasks of a pattern recognition system, arising from deficiency (e.g., vagueness, incompleteness etc.) in information, in an efficient manner. Artificial Neural Networks (ANNs), having the capability of fault tolerance, adaptivity and generalization, and scope for massive parallelism, are widely used in dealing with learning and optimization tasks. There exist several methods for feature selection based on fuzzy set theory [1]–[3] and artificial neural networks (ANN) [4]–[8] in individual framework.

Neuro-fuzzy computing deals with the concept of integrating the merits of fuzzy set theory and ANN for making the systems artificially more intelligent. In the area of pattern recognition, neuro-fuzzy approaches have been attempted mostly for designing classification/clustering methodologies, not much for feature selection.

The present article is an attempt in this line and provides a neuro-fuzzy approach for feature selection under unsupervised mode of training. First of all, a fuzzy feature evaluation index for a set of features is defined in terms of membership values denoting the degree of similarity between two patterns. This does not need the information on class labels of the patterns. The similarity between two patterns is measured by an weighted distance between them. The weighting coefficients are used to denote the degree of importance of the individual features in characterizing/discriminating different clusters and to provide flexibility in modeling various clusters. The evaluation index is such that, for a set of features, the lower is its value, the higher is the importance of that set in characterizing/discriminating various clusters. A layered network is then formulated for performing the task of minimization of the evaluation index through unsupervised learning process; thereby determining the optimum weighting coefficients providing an ordering of the individual features.

J. Basak is on leave from Machine Intelligence Unit, Indian Statistical Institute, Calcutta.

R. K. De is grateful to the Department of Atomic Energy, India for providing him a Dr. K. S. Krishnan Senior Research Fellowship. The work is partly supported by the Grant No. 25(0093)/97/EMR-II of CSIR, New Delhi.

0-7803-4859-1/98 $10.00©1998 IEEE

The investigation also includes the task of ordering different subsets from a set of features. This is done by computing the evaluation index (with weighting coefficients being unity) on different subsets of features and then ordering them accordingly. The effectiveness of these algorithms is demonstrated on a speech [1], Iris [9] and a medical data [10].

## 2. Feature Evaluation Index

Let, $\mu_{pq}^O$ be the degree that both the $p$th and $q$th patterns belong to the same cluster in the $n$-dimensional original feature space, and $\mu_{pq}^T$ be that in the $n'$-dimensional ($n' < n$) transformed feature space. $\mu$ values determine how similar a pair of patterns are in the respective features spaces. Let, $s$ be the number of samples on which the feature evaluation index is computed.

The feature evaluation index for a subset ($\Omega$) of features is defined as

$$E = \frac{2}{s(s-1)} \sum_{p} \sum_{q \neq p} \frac{1}{2} [\mu_{pq}^T (1 - \mu_{pq}^O) + \mu_{pq}^O (1 - \mu_{pq}^T)]. \quad (1)$$

It has the following characteristics.
(i) If $\mu_{pq}^O = \mu_{pq}^T = 0$ or 1, the contribution of the pair of patterns to the evaluation index $E$ is zero (minimum).
(ii) If $\mu_{pq}^O = \mu_{pq}^T = 0.5$, the contribution of the pair of patterns to $E$ becomes 0.25 (maximum).
(iii) For $\mu_{pq}^O < 0.5$ as $\mu_{pq}^T \rightarrow 0$, $E$ decreases.
For $\mu_{pq}^O > 0.5$ as $\mu_{pq}^T \rightarrow 1$, $E$ decreases.

Therefore, the feature evaluation index decreases as the membership value representing the degree of belonging of $p$th and $q$th patterns to the same cluster in the transformed feature space tends to either 0 (when $\mu^O < 0.5$) or 1 (when $\mu^O > 0.5$). In other words, the feature evaluation index decreases as the decision on the similarity between a pair of patterns (*i.e.*, whether they lie in the same cluster or not) becomes more and more crisp. This means, if the intercluster/intracluster distances in the transformed space increase/decrease, the feature evaluation index of the corresponding set of features decreases. Therefore, our objective is to select those features for which the evaluation index becomes minimum; thereby optimizing the decision on the similarity of a pair of patterns with respect to their belonging to a cluster.

In order to satisfy the characteristics of $E$ (Eqn. (1)), the membership function ($\mu$) in a feature space may be defined as

$$\mu_{pq} = 1 - \frac{d_{pq} - d_{min}}{d_{max} - d_{min}}, \quad (2)$$

where $d_{pq}$ is a distance measure which provides similarity (in terms of proximity) between the $p$th and $q$th patterns in the feature space. $d_{max}$ and $d_{min}$, respectively, are the maximum and minimum values of $d_{pq}$. Note that, the higher is the value of $d_{pq}$, the lower is the similarity between $p$th and $q$th patterns, and *vice versa*. When $d_{pq} = d_{min}$ and $d_{pq} = d_{max}$, we have $\mu_{pq} = 1$ and 0, respectively. That is, when the patterns are most (least) similar, both the patterns must be in the same (different) cluster(s). If $d_{pq} = \frac{1}{2}(d_{max} + d_{min})$, $\mu_{pq} = 0.5$. That is, when the similarity between the patterns is just in between its maximum and minimum values, the difficulty in making a decision, whether both the patterns are in the same cluster or not, becomes maximum; thereby making the situation most ambiguous.

The distance $d_{pq}$ (in Eqn. (2)) can be expressed in many ways. Let us consider, for example, the Euclidian distance between the two patterns. Then,

$$d_{pq} = [\sum_{i} (x_{pi} - x_{qi})^2]^{\frac{1}{2}}, \quad (3)$$

where $x_{pi}$ and $x_{qi}$ are values of $i$th feature (in the corresponding feature space) of $p$th and $q$th patterns, respectively. $d_{max}$ is defined as

$$d_{max} = [\sum_{i} (x_{maxi} - x_{mini})^2]^{\frac{1}{2}}, \quad (4)$$

where $x_{maxi}$ and $x_{mini}$ are the maximum and minimum values of the $i$th feature in the corresponding feature space. $d_{min}$ is taken as zero as the Euclidian distance between any two identical patterns is zero in any feature space.

In the above discussion, the similarity between two patterns, in terms of proximity, is measured by $d_{pq}$ (Eqn. (3)). Since, $d_{pq}$ is an Euclidian distance, the methodology implicitly assumes that the clusters are hyperspherical. But in practice, this may not necessarily be the case. To model the practical situation we have introduced the concept of weighted distance such that

$$\begin{aligned} d_{pq} &= [\sum_{i} w_i^2 (x_{pi} - x_{qi})^2]^{\frac{1}{2}}, \\ &= [\sum_{i} w_i^2 \chi_i^2]^{\frac{1}{2}}, \quad \chi_i = (x_{pi} - x_{qi}), \end{aligned} \quad (5)$$

where $w_i \in [0, 1]$ represents weighting coefficient corresponding to $i$th feature.

The membership value $\mu_{pq}$ is now obtained by Eqns. (2), (4) and (5), and becomes dependent on $w_i$. The values of $w_i$ ($< 1$) make the $\mu_{pq}$ function of Eqn. (2) flattened along the axis of $d_{pq}$. The lower the value of $w_i$, the higher is extent of flattening. In the extreme case, when $w_i = 0$, $\forall i$, $d_{pq} = 0$ and $\mu_{pq} = 1$ for all pair

of patterns, *i.e.*, all the patterns lie on the same point making them indiscriminable.

In pattern recognition literature, the weight $w_i$ (in Eqn. (5)) can be viewed to reflect the relative importance of the feature $x_i$ in measuring the similarity (in terms of distance) of a pair of patterns. It is such that the higher the value of $w_i$, the more is the importance of $x_i$ in characterizing a cluster or discriminating various clusters. $w_i = 1$ (0) indicates most (least) importance of $x_i$.

Note that, the computation of $\mu_{pq}$ in Eqn. (2) does not require class information of the patterns, *i.e.*, the algorithm is unsupervised. In addition, it does not depend on the number of clusters present in the feature space. It is also to be noted that, the algorithm does not explicitly provide clustering of the feature space. That is, unlike the method in [3], the present algorithm is independent of the number of clusters and is able to select a set of salient features without clustering (explicitly) the feature space.

## 3. Feature Selection

In this section we describe two unsupervised algorithms for feature selection. The first one considers fuzzy feature evaluation index alone for ranking different feature subsets. The second one is based on neuro-fuzzy approach where the fuzzy feature evaluation index is minimized with a layered neural network for ranking individual features.

### 3.1 Ordering of feature subsets using $E$ (*Method 1*)

From the aforesaid discussion we see that if a particular subset ($\Omega_1$) of features is more important than another subset ($\Omega_2$) then $E$ computed over $\Omega_1$ will be less than that computed over $\Omega_2$. Therefore, the task of feature subset selection boils down to selecting the subset $\Omega$ from a given set of $n$ features for which $E$ is minimum. This is done by computing the $E$ values for different subsets of features using Eqns. (1)–(4), and ranking them accordingly. Here $\mu^O$ is computed on the $n$-dimensional original feature space, whereas $\mu^T$ is done on its various subsets. Note that, if the subset $\Omega$ contains only one feature, it provides individual feature ranking. Let us call this algorithm *Method 1* in the subsequent discussion.

### 3.2 Ordering of individual features through connectionist minimization of $E$ (*Method 2*)

In *Method 1*, we have considered Euclidian distance (Eqn. (3)) to compute $\mu$-values. Here we consider Eqn. (5) instead of Eqn. (3). Therefore, the evaluation index $E$ (Eqn. (1)) becomes a function of w ($= [w_1, w_2, \ldots w_n]$), if we consider ranking of $n$ features in a set. Here $\mu^O$ and $\mu^T$ are both computed over the original $n$-dimensional feature space. The only difference is that $\mu^O$ needs Eqns. (2)–(4), while $\mu^T$ needs Eqns. (2), (4) and (5) for their computation.

The problem of feature selection/ranking thus reduces to finding a set of $w_i$s for which $E$ becomes minimum; $w_i$s indicating the relative importance of $x_i$s. The task of minimization is performed using gradient-descent technique in a connectionist framework under unsupervised mode. Let us now describe the model.

The network (Fig. 1.) consists of an input, a hidden and an output layer. The input layer consists of a pair of nodes corresponding to each feature, *i.e.*, the number of nodes in the input layer is $2n$, for $n$-dimensional (original) feature space. The hidden layer consists of $n$ number of nodes which compute the part $\chi_i^2$ of Eqn. (5) for each pair of patterns. The output layer consists of two nodes. One of them computes $\mu^O$, and the other $\mu^T$. The feature evaluation index $E$ (Eqn. (14)) is computed from these $\mu$-values off the network.

Input nodes receive activations corresponding to feature values of each pair of patterns. A $j$th node in the hidden layer is connected only to an $i$th and $(i+n)$th input nodes via connection weights $+1$ and $-1$, respectively, where $j, i = 1, 2, \ldots, n$ and $j = i$. The output node computing $\mu^T$-values is connected to a $j$th node in the hidden layer via connection weight $W_j$ ($= w_j^2$), whereas that computing $\mu^O$-values is connected to all the nodes in the hidden layer via connection weights $+1$ each.

During training, each pair of patterns are presented at the input layer and the evaluation index is computed. The weights $W_j$s are updated using gradient-descent technique in order to minimize the index $E$. Note that, $d_{max}$ is directly computed from the unlabeled training set using Eqn. (4), and $d_{min}$ is set to zero. Both $d_{max}$ and $d_{min}$ are stored in the output nodes. When $p$th and $q$th patterns are presented to the input layer, the activation produced by $i$th ($1 \leq i \leq 2n$) input node is

$$v_i^{(0)} = u_i^{(0)} \tag{6}$$

where

$$\begin{aligned} u_i^{(0)} &= x_{pi}, \quad for\ 1 \leq i \leq n \ \ and \\ u_{i+n}^{(0)} &= x_{qi}, \quad for\ 1 \leq i \leq n, \end{aligned} \tag{7}$$

the total activations received by $i$th and $(i+n)$th ($1 \leq i \leq n$) input node, respectively. A $j$th hidden node (connecting $i$th and $(i+n)$th, $1 \leq i \leq n$, input nodes) receives a total activation

$$u_j^{(1)} = 1 \times v_i^{(0)} + (-1) \times v_{i+n}^{(0)}, \quad for\ 1 \leq i \leq n, \tag{8}$$

to produce an activation

$$v_j^{(1)} = (u_j^{(1)})^2. \tag{9}$$

The total activation received by the output node which computes $\mu^T$-values, is

$$u_T^{(2)} = \sum_j W_j v_j^{(1)}, \tag{10}$$

and that received by the other output node which computes $\mu^O$-values, is

$$u_O^{(2)} = \sum_j v_j^{(1)}. \tag{11}$$

Therefore, $u_T^{(2)}$ and $u_O^{(2)}$ represent $d_{pq}^2$ as given by Eqns. (5) and (3), respectively. The activations, $v_T^{(2)}$ and $v_O^{(2)}$, of the output nodes represent $\mu_{pq}^T$ and $\mu_{pq}^O$ for $p$th and $q$th pattern pair, respectively. Thus,

$$v_T^{(2)} = 1 - \frac{(u_T^{(2)})^{\frac{1}{2}} - d_{min}}{d_{max} - d_{min}}, \tag{12}$$

and

$$v_O^{(2)} = 1 - \frac{(u_O^{(2)})^{\frac{1}{2}} - d_{min}}{d_{max} - d_{min}}. \tag{13}$$

The evaluation index (which is computed off the network), in terms of these activations, is then written (from Eqn. (1)) as

$$E(\mathbf{W}) = \frac{2}{s(s-1)} \sum_p \sum_{q \neq p} \frac{1}{2} [v_T^{(2)}(1 - v_O^{(2)}) + v_O^{(2)}(1 - v_T^{(2)})]. \tag{14}$$

As mentioned before, the task of minimization of $E(\mathbf{W})$ (Eqn. (14)) with respect to $\mathbf{W}$ is performed using gradient-descent technique, where the change in $W_j$ ($\Delta W_j$) is computed as

$$\Delta W_j = -\eta \frac{\partial E}{\partial W_j}, \forall j, \tag{15}$$

where $\eta$ is the learning rate.

For computation of $\frac{\partial E}{\partial W_j}$, the following expressions are used.

$$\frac{\partial E(\mathbf{W})}{\partial W_j} = \frac{2}{s(s-1)} \sum_p \sum_{q \neq p} \frac{1}{2} \left[1 - 2v_O^{(2)}\right] \frac{\partial v_T^{(2)}}{\partial W_j}, \tag{16}$$

$$\frac{\partial v_T^{(2)}}{\partial W_j} = -\frac{\frac{1}{2}(u_T^{(2)})^{-\frac{1}{2}} \frac{\partial u_T^{(2)}}{\partial W_j}}{d_{max} - d_{min}}, \tag{17}$$

and

$$\frac{\partial u_T^{(2)}}{\partial W_j} = v_j^{(1)}. \tag{18}$$

After convergence, $E(\mathbf{W})$ attains a local minimum. Then the weights ($W_j = w_j^2$) of the links connecting hidden nodes and the output node computing $\mu^T$-values, indicate the order of importance of the features. Let us call this algorithm *Method 2* in the subsequent discussion. ♣

Note that, *Method 2*, which is based on neuro-fuzzy approach for individual feature ranking, finds the set of $w_i$s (for which $E$ is minimum) considering the effect of interdependence of the features, whereas in *Method 1*, each feature is considered independent of the others.

## 4. Results

The effectiveness of the aforesaid algorithms is demonstrated on a speech (vowel) data [1], Iris [9] and a medical data [10]. The speech (vowel) data consists of a set of 437 Indian Telugu vowel sounds collected by trained personnel. These were uttered in a consonant-vowel-consonant context by three male speakers in the age group of 30 to 35 years. The data set has three features, $F_1$, $F_2$ and $F_3$ corresponding to the first, second and third vowel formant frequencies obtained through spectrum analysis of the speech data, and six vowel classes ($\partial$, a, i, u, e, o). This vowel data is being extensively used for more than two decades in the area of pattern recognition.

Anderson's Iris data [9] set contains three classes, *i.e.*, three varieties of Iris flowers, namely, Iris Setosa, Iris Versicolor and Iris Virginica consisting of 50 samples each. Each sample has four features, namely, Sepal Length (SL), Sepal Width (SW), Petal Length (PL) and Petal Width (PW). Iris data has been used in many research investigations related to pattern recognition and has become a sort of benchmark-data.

The medical data consisting of 9 input features and 4 pattern classes, deals with various *Hepatobiliary disorders* [10] of 536 patient cases. The input features are the results of different biochemical tests, *viz.*, Glutamic Oxalacetic Transaminate (GOT, Karmen unit), Glutamic Pyruvic Transaminase (GPT, Karmen Unit), Lactate Dehydrase (LDH, iu/l), Gamma Glutamyl Transpeptidase (GGT, mu/ml), Blood Urea Nitrogen (BUN, mg/dl), Mean Corpuscular Volume of red blood cell (MCV, fl), Mean Corpuscular Hemoglobin (MCH, pg), Total Bilirubin (TBil, mg/dl) and Creatinine (CRTNN, mg/dl). The hepatobiliary disorders Alcoholic Liver Damage (ALD), Primary Hepatoma (PH), Liver Cirrhosis (LC) and Cholelithiasis (C), constitute the four output classes.

### 4.1 Using *Method 1*

Table I shows the ordering of different subsets for the

two types data using *Method 1*. Note that, for Iris and speech data we have computed $E$-value for all possible subsets, including the individual features, (*i.e.*, fifteen for Iris and seven for speech data) and ranked them accordingly.

It is seen from Table I that a subset of higher cardinality may not necessarily be more important than ones of lower cardinality. For speech data, it is $F_2$ which has become a member of the best four subsets. This conforms to an earlier investigation [2]. Similarly, for Iris data, PL being the best individual feature is seen to be a member of the first five best subsets. For medical data, since the number of features is large, we have, first of all, computed the $E$-value for the individual features. A few bests of them (e.g., GOT, LDH, GPT, GGT) are selected after ranking. Then we have computed $E$-value for different (20) subsets containing only these selected features.

TABLE I

IMPORTANCE OF DIFFERENT FEATURE SUBSETS USING *Method 1*. ($X > Y$ MEANS $X$ IS MORE IMPORTANT THAN $Y$.)

| Data sets | Order of importance |
|---|---|
| Speech | $\{F_2\} > \{F_1, F_2\} > \{F_2, F_3\} > \{F_1, F_2, F_3\} > \{F_3\} > \{F_1, F_3\} > \{F_1\}$ |
| Iris | $\{PL\} > \{PL, PW\} > \{SW, PL\} > \{SL, PL\} > \{SW, PL, PW\} > \{PW\} > \{SL, PL, PW\} > \{SL, SW, PL\} > \{SL, SW, PL, PW\} > \{SL, PW\} > \{SL\} > \{SL, SW, PW\} > \{SW, PW\} > \{SL, SW\} > \{SW\}$ |
| Medical | $\{GOT\} > \{GOT, GPT\} > \{LDH\} > \{GPT, LDH\} > \{GOT, LDH\} > \{GOT, GPT, LDH\} > \{GOT, GGT\} > \{GOT, GPT, GGT\} > \{LDH, GGT\} > \{GPT\} > \{GPT, LDH, GGT\} > \{GOT, LDH, GGT\} > \{GOT, GPT, LDH, GGT\} > \{GGT\} > \{GPT, GGT\} > \{CRTNN\} > \{TBil\} > \{BUN\} > \{MCV\} > \{MCH\}$ |

### 4.2 Using *Method 2*

Tables II–IV provide the degrees of importance ($w$-value) of different features corresponding to these data sets obtained by the neuro-fuzzy approach. Note that, their initial values were considered to be random numbers in $[0, 1]$. The values of $w$ were truncated to 0.0 and 1.0 during training.

In the case of the speech data, the order of importance of the features is found to be $F_2 > F_3 > F_1$ (Table II) which conforms to the order of individual features obtained by *Method 1* (Table I). For Iris data, the best two features are found to be PL and PW (Table III) which are also the best two individual features obtained by *Method 1* (Table I) and in an earlier investigation [6]. Similarly, for the medical data, the best two features are GOT and LDH (Table IV) which are also found to be the best two individual features by *Method 1* (Table I).

TABLE II

$w$-VALUES FOR SPEECH DATA USING *Method 2*.

| Feature | $w$ | Rank |
|---|---|---|
| $F_1$ | 0.045284 | 3 |
| $F_2$ | 0.888104 | 1 |
| $F_3$ | 0.600092 | 2 |

TABLE III

$w$-VALUES FOR IRIS DATA USING *Method 2*.

| Feature | $w$ | Rank |
|---|---|---|
| SL | 0.058414 | 4 |
| SW | 0.194421 | 3 |
| PL | 0.965575 | 1 |
| PW | 0.603508 | 2 |

TABLE IV

$w$-VALUES FOR THE MEDICAL DATA USING *Method 2*.

| Feature | $w$ | Order |
|---|---|---|
| GOT | 0.851015 | 1 |
| GPT | 0.665853 | 8 |
| LDH | 0.733647 | 2 |
| GGT | 0.055946 | 9 |
| BUN | 0.704469 | 6 |
| MCV | 0.704249 | 7 |
| MCH | 0.706765 | 4 |
| TBil | 0.706562 | 5 |
| CRTNN | 0.707109 | 3 |

## 5. Conclusions

The article has demonstrated how the concept of neuro-fuzzy computing can be exploited for developing a methodology for feature selection under unsupervised mode. The methodology involves connectionist minimization of a fuzzy feature evaluation index; thereby determining the ranking of various features. The algorithm considers interdependence of the original fea-

tures. Unlike the method based on fuzzy c-means algorithm [3], the method does not need the information on the number of clusters present in the feature space, and it provides ranking of individual features without clustering the feature space explicitly. The effectiveness of the method is demonstrated extensively on a 3-d speech, 4-d Iris and a 9-d medical data.

Besides the neuro-fuzzy method, we have developed another unsupervised feature selection algorithm (*Method 1*) where the aforesaid fuzzy evaluation index is used *alone* to find the best subset of features from a given set. Here the algorithm assumes, unlike the neuro-fuzzy methods, independence of the original features. Experimental results on the ordering of original features by both the algorithms conform well to those obtained using other methods [2], [6]. Although a network is used in *Method 2* for minimization of the evaluation index, one may consider other optimization techniques for this task.

## REFERENCES


[1] S. K. Pal and B. Chakraborty, "Fuzzy set theoretic measures for automatic feature evaluation," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 16, pp. 754–760, 1986.

[2] S. K. Pal, "Fuzzy set theoretic measures for automatic feature evaluation: II," *Information Sciences*, vol. 64, pp. 165–179, 1992.

[3] J. C. Bezdek and P. Castelaz, "Prototype classification and feature selection with fuzzy sets," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 7, pp. 87–92, 1977.

[4] D. W. Ruck, S. K. Rogers, and M. Kabrisky, "Feature selection using a multilayer perceptron," *Journal of Neural Network Computing*, pp. 40–48, Fall 1990.

[5] K. L. Priddy, S. K. Rogers, D. W. Ruck, G. L. Tarr, and M. Kabrisky, "Bayesian selection of important features for feedforward neural networks," *Neurocomputing*, vol. 5, pp. 91–103, 1993.

[6] J. M. Steppe and K. W. Bauer, Jr., "Improved feature screening in feedforward neural networks," *Neurocomputing*, vol. 13, pp. 47–58, 1996.

[7] M. Pregenzer, G. Pfurtscheller, and D. Flotzinger, "Automated feature selection with a distinctive sensitive learning vector quantizer," *Neurocomputing*, vol. 11, pp. 19–29, 1996.

[8] R. K. De, N. R. Pal, and S. K. Pal, "Feature analysis : Neural network and fuzzy set theoretic approaches," *Pattern Recognition*, vol. 30, pp. 1579–1590, 1997.

[9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[10] Y. Hayashi, "A neural expert system with automated extraction of fuzzy if-then rules and its application to medical diagnosis," in *Advances in Neural Information Processing Systems* (R. P. Lippmann, J. E. Moody, and D. S. Touretzky, eds.), pp. 578–584, Los Altos: Morgan Kaufmann, 1991.


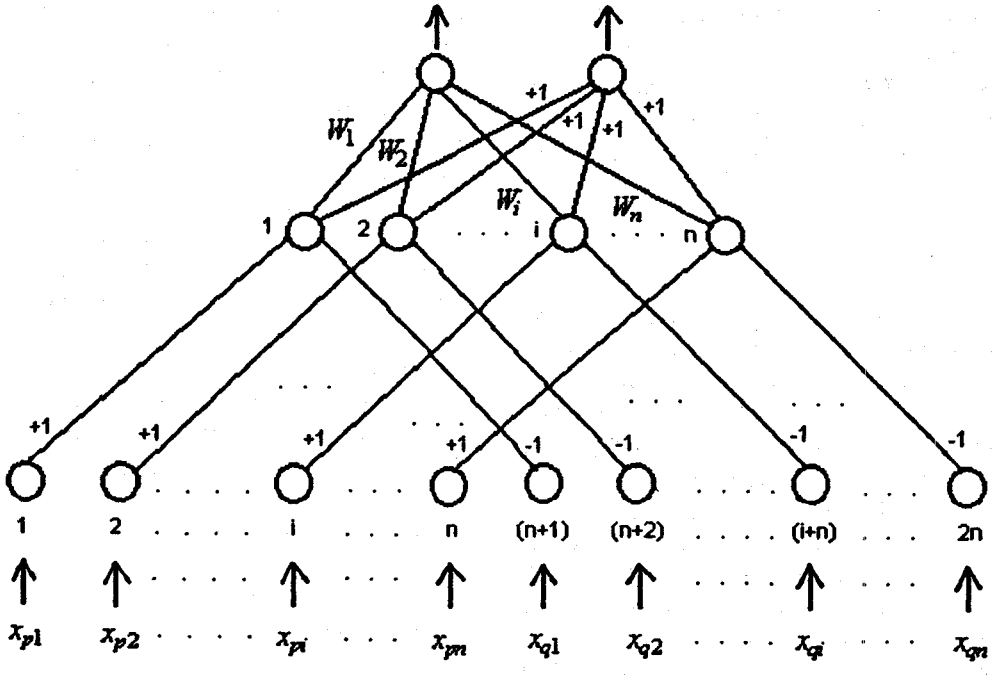


Fig. 1. A schematic diagram of the neural network model.