Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach

Sankar K. Pal, *Fellow, IEEE*, Rajat K. De, *Member, IEEE*, and Jayanta Basak, *Senior Member, IEEE*

***Abstract*—The present article demonstrates a way of formulating neuro-fuzzy approaches for both feature selection and extraction under unsupervised learning. A fuzzy feature evaluation index for a set of features is defined in terms of degree of similarity between two patterns in both the original and transformed feature spaces. A concept of flexible membership function incorporating weighted distance is introduced for computing membership values in the transformed space. Two new layered networks are designed. The tasks of membership computation and minimization of the evaluation index, through unsupervised learning process, are embedded into them without requiring the information on the number of clusters in the feature space. The network for feature selection results in an optimal order of individual importance of the features. The other one extracts a set of optimum transformed features, by projecting n-dimensional original space directly to n'-dimensional ($n' < n$) transformed space, along with their relative importance. The superiority of the networks to some related ones is established experimentally.**

***Index Terms*—Feature selection/extraction, fuzzy feature evaluation index, layered networks, soft computing.**

I. Introduction

FEATURE selection or extraction is a process of selecting a map of the form $\boldsymbol{x}' = f(\boldsymbol{x})$ by which a sample $\boldsymbol{x}(x_1, x_2, \cdots, x_n)$ in an n-dimensional measurement space $(\Re^n)$ is transformed into a point $\boldsymbol{x}'(x'_1, x'_2, \cdots, x'_{n'})$ in an n'-dimensional $(n' < n)$ feature space $(\Re^{n'})$. The problem of feature selection deals with choosing some of $x_i s$ from the measurement space to constitute the feature space. On the other hand, the problem of feature extraction deals with generating new $x'_j s$ (constituting the feature space) based on some $x_i s$ in the measurement space. The main objective of these processes 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.

Different useful classical techniques for feature selection and extraction are available in [1]. Some of the recent attempts made for these tasks in the framework of artificial neural networks (ANN's) are mainly based on multilayer feedforward networks and self-organizing networks [2]–[11]. Depending on whether the class information of the samples is known or not, these methods are classified under supervised [2], [3], [5], [7] or unsupervised [9]–[11] modes.

Recently, attempts are being made to integrate the merits of fuzzy set theory and ANN under the heading "neuro-fuzzy computing" [12] with an aim of making the systems artificially more intelligent. Incorporation of fuzzy set theory enables one to deal with uncertainties in different tasks of pattern recognition system, arising from deficiency (e.g., vagueness, incompleteness, etc.) in information, in an efficient manner. ANN's, having the capability of fault tolerance, adaptivity, and generalization, and scope for massive parallelism, are widely used in dealing with learning and optimization tasks. In the area of pattern recognition, neuro-fuzzy approaches have been attempted mostly for designing classification/clustering methodologies; the problem of feature evaluation, particularly under unsupervised mode of learning, has not been addressed much.

The present article is an attempt in this regard and provides a neuro-fuzzy approach for both feature selection and extraction under unsupervised learning. The methodology involves formulation of two different layered networks for minimization of a fuzzy feature evaluation index. The fuzzy index for a set of features is defined in terms of membership values denoting the degree of similarity between two patterns both in the original and the transformed spaces. 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.

For the task of feature selection, a set of weighting coefficients is 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 similarity between two patterns in the transformed space, obtained by incorporating these weighting factors in the original feature space, is measured by an weighted distance between them. Minimization of the evaluation index through unsupervised learning of the network determines the optimum weighting coefficients providing an ordering of the importance of features individually.

For feature extraction, the transformed space is obtained through a set of linear transformations. The similarity between two patterns in the transformed space is computed, as in the case of feature selection, using a set of weighting coefficients. An optimum transformed space along with the degrees of individual importance of the transformed (extracted) features is obtained through connectionist minimization. The network is such that the number of nodes in its second hidden layer determines the desired number of extracted features.

Manuscript received October 27, 1998; revised November 2, 1999 and January 4, 2000. The work of R. K. De was supported by a Dr. K. S. Krishnan Senior Research Fellowship from the Department of Atomic Energy, Government of India.

The authors are with the Machine Intelligence Unit, Indian Statistical Institute, Calcutta, 700035 India (e-mail: {sankar; rajat; jayanta}@isical.ac.in).

Publisher Item Identifier S 1045-9227(00)02995-7.

The effectiveness of the algorithms is demonstrated on four different real-life data sets, namely, Iris [13], vowel [14], [15], medical [16], [17], and mango-leaf [18]. The validity of the feature selection algorithm and the superior discrimination ability of the extracted features over the original ones are established using k-NN classifier for different values of k. The method for feature extraction is also compared with a principal component analysis network (PCAN) [11].

II. Feature Evaluation Index

A. Definition

Let, μ_{pq}^O be the degree that both the pth and qth patterns belong to the same cluster in the n-dimensional original feature space, and μ_{pq}^T be that in the n'-dimensional ($n' \leq n$) transformed feature space. μ values determine how similar a pair of patterns are in the respective features spaces. That is, μ may be interpreted as the membership value of a pair of patterns belonging to the fuzzy set "similar." Let s be the number of samples on which the feature evaluation index is computed.

The feature evaluation index for a set (Ω) of transformed 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: 1) for $\mu_{pq}^O < 0.5$ as $\mu_{pq}^T \to 0$, E decreases. For $\mu_{pq}^O > 0.5$ as $\mu_{pq}^T \to 1$, E decreases. In both the cases, the contribution of the pair of patterns to the evaluation index E becomes minimum (=0) when $\mu_{pq}^O = \mu_{pq}^T = 0$ or 1; 2) for $\mu_{pq}^O < 0.5$ as $\mu_{pq}^T \to 1$, E increases. For $\mu_{pq}^O > 0.5$ as $\mu_{pq}^T \to 0$, E increases. In both the cases, the contribution of the pair of patterns to E becomes maximum (=0.5) when $\mu_{pq}^O = 0$ and $\mu_{pq}^T = 1$, or $\mu_{pq}^O = 1$ and $\mu_{pq}^T = 0$; and 3) if $\mu_{pq}^O = 0.5$, the contribution of the pair of patterns to E becomes constant (=0.25), i.e., independent of μ_{pq}^T.

The characteristics 1) and 2) can be verified as follows. From (1) we have

$$\frac{\partial E}{\partial \mu_{pq}^T} = \frac{1}{s(s-1)}(1 - 2\mu_{pq}^O). \quad (2)$$

For $\mu_{pq}^O < 0.5$, $(\partial E/\partial \mu_{pq}^T) > 0$. This signifies that E decreases (increases) with decrease (increase) in μ_{pq}^T. For $\mu_{pq}^O > 0.5$, $(\partial E/\partial \mu_{pq}^T) < 0$. This signifies that E decreases (increases) with increase (decrease) in μ_{pq}^T. Since $\mu_{pq}^T \in [0, 1]$, E decreases (increases) as $\mu_{pq}^T \to 0(1)$ in the former case, and $\mu_{pq}^T \to 1(0)$ in the latter. ♣

Therefore, the feature evaluation index decreases as the membership value representing the degree of belonging of pth and qth patterns to the same cluster in the transformed feature space tends to either zero (when $\mu^O < 0.5$) or one (when $\mu^O > 0.5$), and becomes minimum for $\mu_{pq}^O = \mu_{pq}^T = 0$ or 1. In other words, the index decreases as the similarity (dissimilarity) between two patterns belonging to the same cluster (different clusters) in the original feature space, increases; thereby making the decision regarding belongingness of patterns to a cluster 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/extract 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.

Characteristic 2) implies that E increases when similar (dissimilar) patterns in the original space becomes dissimilar (similar) in the transformed space. That is, any occurrence of such a situation will be automatically protected by the process of minimizing E. Similarly, when $\mu_{pq}^O = 0.5$ [characteristic 3)], i.e., decision regarding the similarity between a pair of patterns whether they lie in the same cluster or not, is most ambiguous, the contribution of the pattern pair to E does not have any impact on the minimization process.

B. Computation of Membership Function

In order to satisfy the characteristics of E (1), as stated in Section II-A, the membership function (μ) in a feature space may be defined as

$$\begin{aligned} \mu_{pq} &= 1 - \frac{d_{pq}}{D} \quad \text{if } d_{pq} \leq D \\ &= 0, \quad \text{otherwise.} \end{aligned} \quad (3)$$

d_{pq} is a distance measure which provides similarity (in terms of proximity) between the pth and qth patterns in the feature space. Note that, the higher the value of d_{pq}, the lower is the similarity between pth and qth patterns, and *vice versa*. D is a parameter which reflects the minimum separation between a pair of patterns belonging to two different clusters. When $d_{pq} = 0$ and $d_{pq} = D$, we have $\mu_{pq} = 1$ and 0, respectively. If $d_{pq} = D/2$, $\mu_{pq} = 0.5$. That is, when the distance between the patterns is just half the value of D, 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 term D [in, (3)] may be expressed as

$$D = \beta\, d_{\max} \quad (4)$$

where $d_{\max}$ is the maximum separation between a pair of patterns in the entire feature space, and $0 < \beta \leq 1$ is a user defined constant. β determines the degree of flattening of the membership function (3). The higher the value of β, more will be the degree, and *vice versa*.

The distance d_{pq} (3) can be defined in many ways. Considering Euclidian distance, we have

$$d_{pq} = \left[\sum_i (x_{pi} - x_{qi})^2\right]^{1/2} \quad (5)$$

where x_{pi} and x_{qi} are values of ith feature (in the corresponding feature space) of pth and qth patterns, respectively. $d_{\max}$ is defined as

$$d_{\max} = \left[\sum_i (x_{\max i} - x_{\min i})^2\right]^{1/2} \quad (6)$$

where $x_{\max i}$ and $x_{\min i}$ are the maximum and minimum values of the ith feature in the corresponding feature space.

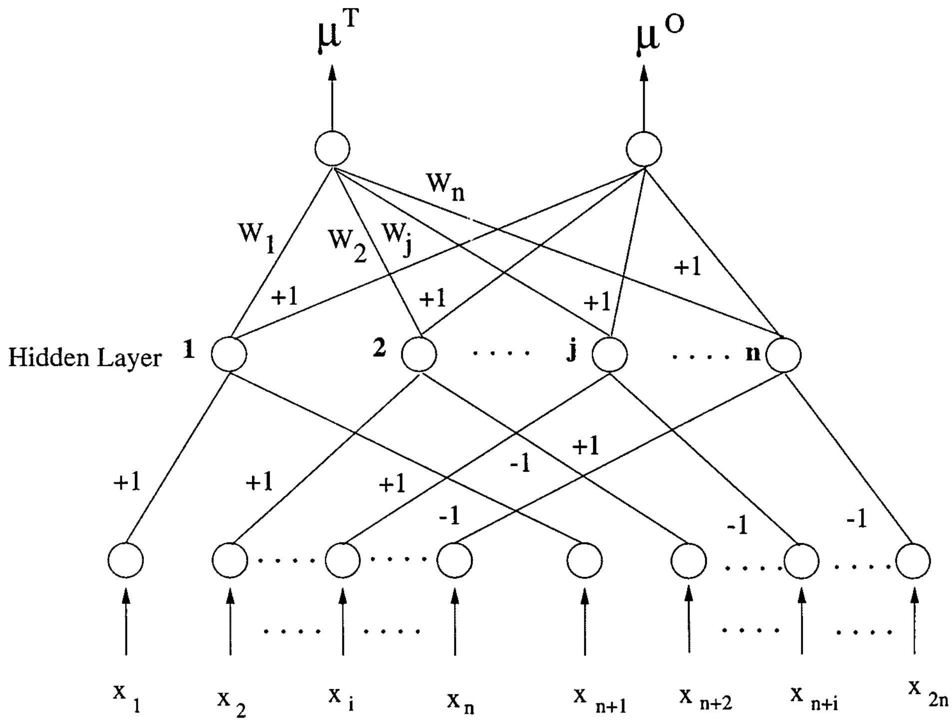


Fig. 1. Neural-network model for feature selection.

Incorporating Weighting Coefficients: In the above discussion, we have measured the similarity between two patterns in terms of proximity, as conveyed by the expression for d_{pq} (5). Since, d_{pq} is an Euclidian distance, the methodology implicitly assumes that the clusters are hyperspherical. In practice, this may not necessarily be the case. To model the practical situations we have introduced the concept of weighted distance such that

$$d_{pq} = \left[\sum_i w_i^2 (x_{pi} - x_{qi})^2\right]^{1/2} = \left[\sum_i w_i^2 \chi_i^2\right]^{1/2}, \qquad \chi_i = (x_{pi} - x_{qi}) \tag{7}$$

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

The membership value μ_{pq} is now obtained by (3), (4), (6), and (7), and becomes dependent on w_i. The values of w_i (<1) make the μ_{pq} function of (3) 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.

The weight w_i [in (7)] reflects the relative importance of the feature x_i in measuring the similarity (in terms of distance) of a pair of patterns. 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, one may define μ_{pq} in a different way satisfying the above mentioned characteristics. The computation of μ_{pq} in (3) does not require the information on class label of the patterns.

III. Feature Selection

As mentioned in Section II-A, our objective is to minimize the evaluation index E (1) which involves the terms μ^O and μ^T. Note that the n-dimensional transformed space is obtained by introducing $\boldsymbol{w}$ (=$[w_1, w_2, \cdots, w_n]$) on the n-dimensional original space. The computation of μ^O requires (3)–(6), while μ^T needs (3), (4), (6), and (7). Therefore, E (1) becomes a function of $\boldsymbol{w}$, if we consider ranking of n features in a set.

The problem of feature selection/ranking thus reduces to finding a set of w_is for which E becomes minimum; w_is indicating the relative importance of x_is. The task of minimization is performed using gradient-descent technique under unsupervised mode. The network designed for performing all these operations (i.e., computation of μ^O and μ^T, and minimization) is described below.

Connectionist 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 χ_i^2 of (7). The output layer consists of two nodes. One of them computes μ^O, and the other μ^T. The index E (12) is computed from these μ-values off the network.

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

During learning, each pair of patterns are presented at the input layer and the evaluation index is computed. The weights

W_js are updated using gradient-descent technique in order to minimize the index E. Note that, the values of $d_{\max}$ (computed from the unlabeled training set) and β (user specified) are stored in both the output nodes for the computation of D. When pth and qth patterns are presented to the input layer, the activation produced by ith ($1 \le i \le 2n$) input node is $v_i^{(0)} = u_i^{(0)}$ where

$$u_i^{(0)} = x_{pi}, \qquad \text{for } 1 \le i \le n$$

and

$$u_{i+n}^{(0)} = x_{qi}, \qquad \text{for } 1 \le i \le n \tag{8}$$

are the total activations received by ith and $(i+n)$th ($1 \le i \le n$) input node, respectively. The activation received by jth hidden node is given by

$$u_j^{(1)} = 1 \times v_i^{(0)} + (-1) \times v_{i+n}^{(0)}, \qquad \text{for } 1 \le i \le n \tag{9}$$

and that produced by it is $v_j^{(1)} = (u_j^{(1)})^2$. The total activation received by the output node which computes μ^T-values, is $u_T^{(2)} = \sum_j W_j v_j^{(1)}$, and that received by the other, is $u_O^{(2)} = \sum_j v_j^{(1)}$.

Therefore, $u_T^{(2)}$ and $u_O^{(2)}$ represent d_{pq}^2 as given by (7) and (5), respectively. The activations, $v_T^{(2)}$ and $v_O^{(2)}$, of the output nodes represent μ_{pq}^T and μ_{pq}^O for pth and qth pattern pair, respectively. Thus,

$$\begin{aligned} v_T^{(2)} &= 1 - \frac{(u_T^{(2)})^{1/2}}{D}, \qquad \text{if } (u_T^{(2)})^{1/2} \le D \\ &= 0, \qquad \text{otherwise} \end{aligned} \tag{10}$$

and

$$\begin{aligned} v_O^{(2)} &= 1 - \frac{(u_O^{(2)})^{1/2}}{D}, \qquad \text{if } (u_O^{(2)})^{1/2} \le D \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{11}$$

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

$$\begin{aligned} E(\boldsymbol{W}) = \frac{2}{s(s-1)} \sum_p \sum_{q \ne p} \frac{1}{2} \\ \cdot \left[v_T^{(2)}(1 - v_O^{(2)}) + v_O^{(2)}(1 - v_T^{(2)}) \right]. \end{aligned} \tag{12}$$

As mentioned before, the task of minimization of $E(\boldsymbol{W})$ (12) with respect to $\boldsymbol{W}$ is performed using gradient-descent technique, where the change in W_j ($\triangle W_j$) is computed as

$$\triangle W_j = -\eta \frac{\partial E}{\partial W_j}, \; \forall j \tag{13}$$

where η is the learning rate.

For computation of $(\partial E / \partial W_j)$ the following expressions are used:

$$\frac{\partial E(\boldsymbol{W})}{\partial W_j} = \frac{1}{2} \left[1 - 2 v_O^{(2)} \right] \frac{\partial v_T^{(2)}}{\partial W_j} \tag{14}$$

and

$$\begin{aligned} \frac{\partial v_T^{(2)}}{\partial W_j} &= - \frac{\frac{1}{2}(u_T^{(2)})^{-(1/2)} \frac{\partial u_T^{(2)}}{\partial W_j}}{D}, \qquad \text{if } (u_T^{(2)})^{1/2} \le D \\ &= 0, \qquad \text{otherwise} \end{aligned} \tag{15}$$

and

$$\frac{\partial u_T^{(2)}}{\partial W_j} = v_j^{(1)}. \tag{16}$$

After minimization, i.e., when $E(\boldsymbol{W})$ attains a local minimum, the weights ($W_j = w_j^2$) of the links connecting hidden nodes and the output node computing μ^T-values, indicate the order of importance of the features. Note that this unsupervised method performs the task of feature selection without clustering the feature space explicitly and does not need to know the number of clusters present in the feature space.

IV. Feature Extraction

In the case of feature extraction, the original feature space ($\boldsymbol{x}$) is transformed to $\boldsymbol{x}'$ by a matrix $\boldsymbol{\alpha}$ ($=[\alpha_{ji}]_{n' \times n}$), i.e.,

$$\boldsymbol{x} \xrightarrow{\alpha} \boldsymbol{x}'.$$

The jth transformed feature is, therefore

$$x_j' = \sum_i \alpha_{ji} x_i \tag{17}$$

where α_{ji} ($j = 1, 2, \cdots, n'$, $i = 1, 2, \cdots, n$, and $n > n'$) is a set of coefficients. Then the distance d_{pq} between pth and qth patterns in the transformed space is

$$\begin{aligned} d_{pq} &= \left[\sum_j w_j^2 \left(\sum_i \alpha_{ji}(x_{pi} - x_{qi}) \right)^2 \right]^{1/2}, \\ &= \left[\sum_j w_j^2 \left(\sum_i \alpha_{ji} \chi_i \right)^2 \right]^{1/2}, \; \chi_i = x_{pi} - x_{qi}, \\ &= \left[\sum_j w_j^2 \psi_j^2 \right]^{1/2}, \; \psi_j = \sum_i \alpha_{ji}(x_{pi} - x_{qi}) \end{aligned} \tag{18}$$

and the maximum distance $d_{\max}$

$$\begin{aligned} d_{\max} &= \left[\sum_j \left(\sum_i |\alpha_{ji}|(x_{\max i} - x_{\min i}) \right)^2 \right]^{1/2}, \\ &= \left[\sum_j \phi_j^2 \right]^{1/2}, \; \phi_j = \sum_i |\alpha_{ji}|(x_{\max i} - x_{\min i}). \end{aligned} \tag{19}$$

Weighting coefficients (w_j) representing the importance of the transformed features, make the shape of clusters in the transformed space hyperellipsoidal.

The membership μ^T is computed using (3), (4), (18) and (19), while μ^0, as in Section III, is done by (3)–(6). Therefore, the

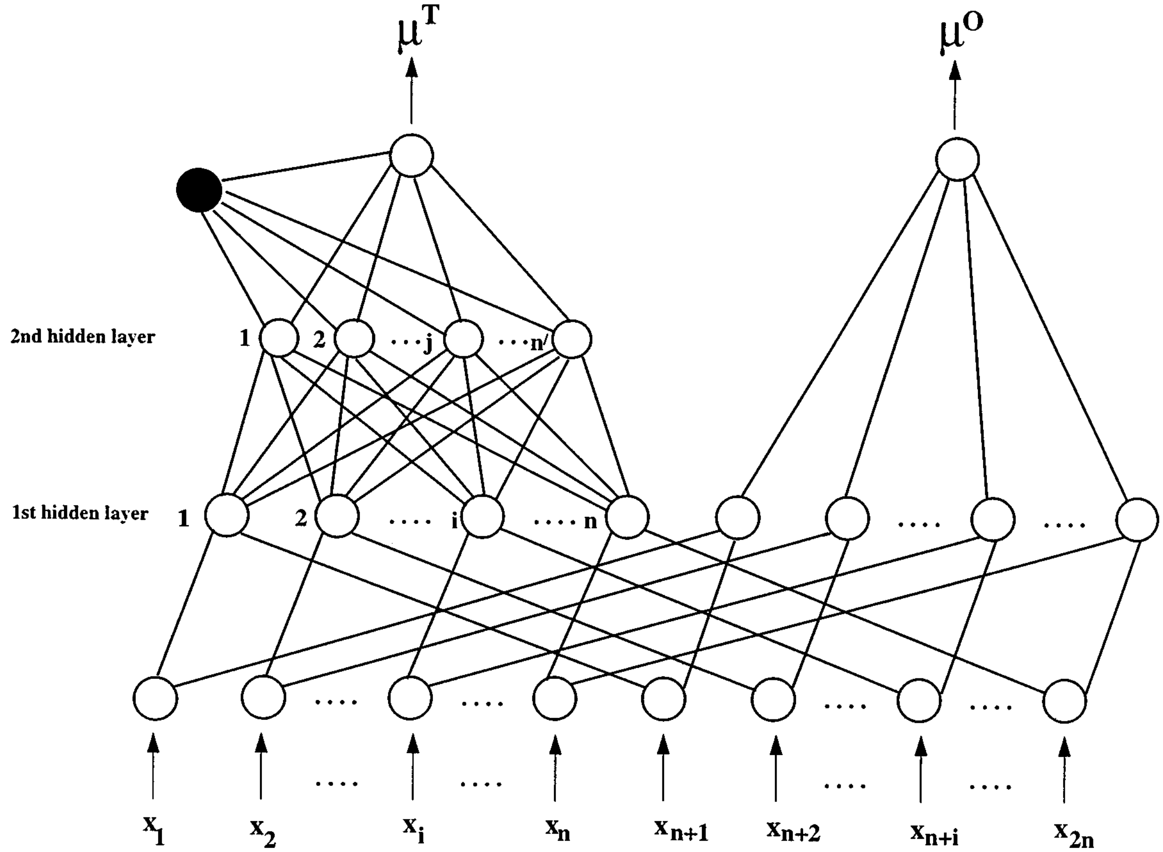


Fig. 2. Neural-network model for feature extraction.

evaluation index E (1) becomes a function of α and $\boldsymbol{w}$. The problem of feature extraction thus reduces to finding a set of α_{ji} and ω_j for which E (1) becomes a minimum. The task of minimization has been performed by gradient-descent technique under unsupervised learning. Like feature selection method, all these operations (i.e., computation of μ^O and μ^T, and minimization for learning α and $\boldsymbol{w}$) are performed in a single network. This is described below.

Connectionist Model: The network (Fig. 2) consists of an input, two hidden and an output layers. The input layer consists of a pair of nodes corresponding to each feature. The first hidden layer consists of $2n$ (for n-dimensional original feature space) number of nodes. Each of the first n nodes computes the part χ_i of (18) and the rest compute χ_i^2. The value of $(x_{\max\, i} - x_{\min\, i})$ is stored in each of the first n nodes. The number of nodes in the second hidden layer is taken as n', in order to extract n' number of features. Each of these nodes has two parts; one of which computing ψ_j^2 of (18) and the other ϕ_j^2 of (19). The output layer consists of two nodes which compute μ^T and μ^O values. There is a node (represented by black circle) in between the output node computing μ^T-values and the second hidden layer. This node computes $d_{\max}$ (19) in the transformed feature space and sends it to the output node for computing μ^T. The value of β is stored in both the output nodes. The feature evaluation index E (25) is computed from these μ-values off the network.

Input nodes receive activations corresponding to feature values of each pair of patterns. A j_1th node in the first hidden layer is connected to ith $(1 \le i \le n)$ and $(i+n)$th $(1 \le i \le n)$ input nodes via weights $+1$ and -1, respectively. A j_2th node in the second hidden layer is connected to a j_1th node in the first hidden layer via weight $\alpha_{j_2 j_1}$. The output node computing μ^T-values is connected to a j_2th node in the second hidden layer via weight W_{j_2} $(=w_{j_2}^2)$, and that computing μ^O-values is connected to a j_1th $(n+1 \le j_1 \le 2n)$ node in the first hidden layer via weights $+1$ each. The node represented by the black circle is connected via weights $+1$ with the second hidden layer and also with the output node computing μ^T-values.

During learning, each pair of patterns are presented to the input layer and the evaluation index is computed. The weights $\alpha_{j_2 j_1}$ and W_{j_2}s are updated using gradient-descent technique in order to minimize the index E. When pth and qth patterns are presented to the input layer, the activation produced by ith $(1 \le i \le 2n)$ input node is $v_i^{(0)} = u_i^{(0)}$ where

$$\begin{aligned} u_i^{(0)} &= x_{pi}, \quad \text{for } 1 \le i \le n \text{ and} \\ u_{(i+n)}^{(0)} &= x_{qi}, \quad \text{for } 1 \le i \le n. \end{aligned} \tag{20}$$

$u_i^{(0)}$ $(1 \le i \le 2n)$ is the total activation received by an ith input node. The activation received by j_1th node in the first hidden layer is given by

$$u_{j_1}^{(1)} = 1 \times v_i^{(0)} + (-1) \times v_{i+n}^{(0)}, \quad \text{for } 1 \le i \le n \tag{21}$$

and that produced by it is

$$\begin{aligned} v_{j_1}^{(1)} &= (u_{j_1}^{(1)}), \quad \text{for } 1 \le j_1 \le n \\ &= (u_{j_1}^{(1)})^2, \quad \text{for } n+1 \le j_1 \le 2n. \end{aligned} \tag{22}$$

The activation received by j_2th node in the second hidden layer is $u_{j_2}^{(2)} = \sum_{j_1=1}^{n} \alpha_{j_2 j_1} v_{j_1}^{(1)}$, and that produced by it is $v_{j_2}^{(2)} = (u_{j_2}^{(2)})^2$. The total activation received by the output node which computes μ^T-values is $u_T^{(3)} = \sum_{j_2} W_{j_2} v_{j_2}^{(2)}$, and that received by the other, is $u_O^{(3)} = \sum_{j_1=n+1}^{2n} v_{j_1}^{(1)}$.

Therefore, $u_T^{(3)}$ and $u_O^{(3)}$ represent d_{pq}^2 as given by (18) and (5), respectively. The activations, $v_T^{(3)}$ and $v_O^{(3)}$, of the output nodes represent μ_{pq}^T and μ_{pq}^O for pth and qth pattern pair, respectively. Thus,

$$\begin{aligned} v_T^{(3)} &= 1 - \frac{(u_T^{(3)})^{1/2}}{D}, \quad \text{if } (u_T^{(3)})^{1/2} \le D, \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{23}$$

and

$$\begin{aligned} v_O^{(3)} &= 1 - \frac{(u_O^{(3)})^{1/2}}{D}, \quad \text{if } (u_O^{(3)})^{1/2} \le D \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{24}$$

The evaluation index, in terms of these activations, can then be expressed [from (1)] as

$$\begin{aligned} E(\alpha, \boldsymbol{W}) \\ = \frac{2}{s(s-1)} \sum_p \sum_{q \ne p} \frac{1}{2} [v_T^{(3)}(1 - v_O^{(3)}) + v_O^{(3)}(1 - v_T^{(3)})]. \end{aligned} \tag{25}$$

The task of minimization of $E(\alpha, \boldsymbol{W})$ [(25)] with respect to $\alpha_{j_2 j_1}$ and W_{j_2}, for all j_1 and j_2 is performed using simple gradient-descent technique where the changes in $\alpha_{j_2 j_1}$ ($\triangle \alpha_{j_2 j_1}$) and W_{j_2} ($\triangle W_{j_2}$) are computed as

$$\triangle \alpha_{j_2 j_1} = -\eta_1 \frac{\partial E}{\partial \alpha_{j_2 j_1}}, \forall j_1, j_2, \text{ and} \tag{26}$$

$$\triangle W_{j_2} = -\eta_2 \frac{\partial E}{\partial W_{j_2}}, \forall j_2 \tag{27}$$

where η_1 and η_2 are the learning rates.

For computation of $\partial E / \partial \alpha_{j_2 j_1}$ and $\partial E / \partial w_{j_2}$ the following expressions are used:

$$\frac{\partial E}{\partial \alpha_{j_2 j_1}} = \frac{1}{2} [1 - 2 v_O^{(3)}] \frac{\partial v_T^{(3)}}{\partial \alpha_{j_2 j_1}} \tag{28}$$

$$\begin{aligned} \frac{\partial v_T^{(3)}}{\partial \alpha_{j_2 j_1}} &= -\frac{Y}{D^2}, \quad \text{if } (u_T^{(3)})^{1/2} \le D \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{29}$$

where

$$Y = D \frac{1}{2} (u_T^{(3)})^{-1/2} \frac{\partial u_T^{(3)}}{\partial \alpha_{j_2 j_1}} - (u_T^{(3)})^{1/2} \frac{\partial D}{\partial \alpha_{j_2 j_1}},$$

$$\frac{\partial u_T^{(3)}}{\partial \alpha_{j_2 j_1}} = W_{j_2} \frac{\partial v_{j_2}^{(2)}}{\partial \alpha_{j_2 j_1}} \tag{30}$$

$$\frac{\partial v_{j_2}^{(2)}}{\partial \alpha_{j_2 j_1}} = 2 u_{j_2}^{(2)} \frac{\partial u_{j_2}^{(2)}}{\partial \alpha_{j_2 j_1}} \tag{31}$$

and

$$\frac{\partial u_{j_2}^{(2)}}{\partial \alpha_{j_2 j_1}} = v_{j_1}^{(1)}, \quad \text{for } 1 \le j_1 \le n \tag{32}$$

$$\frac{\partial D}{\partial \alpha_{j_2 j_1}} = \frac{\beta (x_{\max j_1} - x_{\min j_1})}{d_{\max}} \sum_i |\alpha_{j_2 i}| (x_{\max i} - x_{\min i}) \tag{33}$$

$$\frac{\partial E}{\partial W_{j_2}} = \frac{1}{2} [1 - 2 v_O^{(3)}] \frac{\partial v_T^{(3)}}{\partial W_{j_2}} \tag{34}$$

$$\begin{aligned} \frac{\partial v_T^{(3)}}{\partial W_{j_2}} &= - \frac{\frac{1}{2} (u_T^{(3)})^{-1/2} \frac{\partial u_T^{(3)}}{\partial W_{j_2}}}{D}, \quad \text{if } (u_T^{(3)})^{1/2} \le D \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{35}$$

and

$$\frac{\partial u_T^{(3)}}{\partial W_{j_2}} = v_{j_2}. \tag{36}$$

After minimization, i.e., when $E(\alpha, \boldsymbol{W})$ attains a local minimum, the extracted features are obtained by (17) using the optimum α-values. The weights of the links, connecting the output node computing μ^T-values and the nodes in the second hidden layer, indicate the order of importance of the extracted features. Like feature selection, the method does not need to know the number of clusters in the feature space, and provides feature extraction without clustering the feature space explicitly.

V. RESULTS

Here we demonstrate the effectiveness of the above mentioned algorithms on four data sets, namely, Iris [13], vowel [14], [15], [19], medical [16], [17], and mango-leaf [18]. Anderson's Iris data [13] 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 vowel data [14], [15], [19] consists of a set of 871 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 30–35 yr. 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 containing six overlapping vowel classes (∂, a, i, u, e, o). The details of the data and its extraction procedure are available in [14]. This vowel data is being extensively used for more than two decades in the area of pattern recognition.

The medical data consisting of nine input features and four pattern classes, deals with various *Hepatobiliary disorders* [16], [17] 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.

The mango-leaf data [18], on the other hand, provides information on different kinds of mango-leaf with 18 features, (i.e., 18-dimensional data) for 166 patterns. It has three classes representing three kinds of mango. The feature set consists of measurements like Z-value (Z), area (A), perimeter (Pe), maximum length (L), maximum breadth (B), petiole (P), K-value (K), S-value (S), shape index (SI), L+P, L/P, L/B, (L+P)/B, A/L, A/B, A/Pe, upper midrib/lower midrib (UM/LM) and perimeter upper half/perimeter lower half (UPe/LPe). The terms "upper" and "lower" are used with respect to maximum breadth position.

A. Feature Selection/Ordering of Individual Features

Tables I and II provide the degrees of importance (w-value) of different features corresponding to Iris and vowel data obtained by the neuro-fuzzy approach. Note that, their initial values were considered to be random numbers in [0, 1] while training the network. The order of importance of the features for the vowel data is found, from Table II, to be $F_2 > F_1 > F_3$, where $x > y$ means feature x is more important than y. This conforms to those obtained in several earlier investigations based on both feature evaluation and classification under supervised mode [14], [15], [18], [19]. For Iris data, the best two features are found to be PL and PW (Table I). This is also in agreement with those obtained using supervised neural [2] and neuro-fuzzy [19] methods.

In the case of medical data, the proposed unsupervised method results in the order of importance of the nine features as $GOT > LDH > CRTNN > MCH > TBil > BUN > MCV > GPT > GGT$, whereas it is $MCV > GOT > GPT > LDH > GGT > MCH > TBil > CRTNN > BUN$, obtained by an earlier investigation using supervised learning [19]. From these results, it is interesting to note that the relative importance of five features, e.g., $GOT > LDH > MCH > TBil > BUN$, remains the same in both the approaches, although their individual ranks are different. Further, the features GOT and LDH have come out as members of the sets of best four features by both the methods. Similarly for mango-leaf data, such common features are found to be K and A/Pe out of best four by these methods. (To restrict the size of the article, tables for these data sets have not been included.)

One may note that the recognition scores obtained by k-NN classifier using $\{GOT, LDH, MCH, CRTNN\}$ are 44.22, 41.98, and 47.57 for $k = 1, 3$ and 5, respectively, whereas the corresponding figures are 44.40, 48.51, and 47.76 for the set $\{GOT, GPT, LDH, MCV\}$. On the other hand, for the mango-leaf data, these results are 77.71, 69.88, and 69.88 using the set $\{Pe, K, S, A/Pe\}$, and 61.90, 67.86, and 64.29 by $\{K, A/L, A/Pe, UPe/LPe\}$. These demonstrate that the proposed algorithm, although being unsupervised, performs comparable/superior to the method under supervised learning.

TABLE I
w-VALUES FOR IRIS DATA

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

TABLE II
w-VALUES FOR VOWEL DATA

Feature	w	Rank
F_1	0.590065	2
F_2	0.896044	1
F_3	0.120944	3

TABLE III
RECOGNITION SCORE WITH k-NN CLASSIFIER FOR IRIS DATA

Feature	% classification		
	$k=1$	$k=3$	$k=5$
SL	48.67	66.67	67.33
SW	55.33	52.67	52.67
PL	93.33	95.33	95.33
PW	89.33	96.00	96.00

In order to show the validity of these orders of importance, we consider both scatter plots and k-NN classifier for $k = 1, 3$, and 5. The results are shown only for Iris data. From the results of k-NN classifier (Table III), PL and PW are found, as in Table I, to be the best two individual features. PL is seen to be better than PW for $k = 1$, and it is the reverse for $k = 3$ and 5, although the difference is not significant. From Tables I and III it is seen that both w-values and recognition scores corresponding to features PL and PW are much higher than those of SL and SW. This signifies the larger difference in importance of PL, PW over SL, SW. All the six scatter plots, given in [19] also reflect the similar observation that the feature pair $\{PL, PW\}$ is the best of all. Further, it is hard to discriminate the relative importance of PL and PW (Fig. 3).

Note that, there is no unsupervised connectionist feature selection approach available in literature, to our knowledge. For this reason, we have compared our results only with those of the related supervised methods.

B. Feature Extraction

As mentioned in Section IV, the number of nodes in the second hidden layer determines the desired number of extracted features. That is, in order to extract n' number of features, one needs to employ exactly n' nodes in the second hidden layer. For each data set, we performed experiments for different number of nodes in the second hidden layer for finding different sets of extracted features. The particular set for which E-value

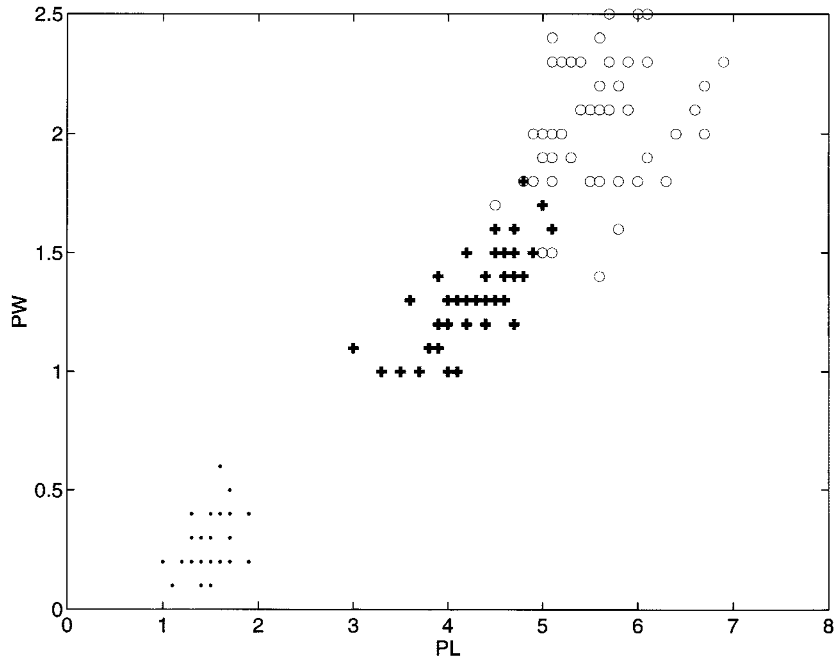


Fig. 3. Scatter plot $PL - PW$ of Iris data. Here "·," "+," and "o" represent classes Iris Setosa, Iris Versicolor, and Iris Virginica, respectively.

TABLE IV
VALUES OF α AND E FOR EXTRACTED FEATURE SETS OF IRIS DATA

Extracted feature set containing	Coefficients (α) of				E (Eqn. (1))
	SL	SW	PL	PW	
one feature	0.072	-0.029	0.195	0.140	0.1024
two features	0.001	0.001	-0.003	-0.051	0.1016
	0.009	0.024	-0.269	-0.066	
three features	-0.017	0.005	-0.123	-0.153	
	-0.004	-0.025	-0.006	-0.084	0.1048
	0.024	-0.004	0.237	0.200	

TABLE V
VALUES OF α AND w FOR THE BEST EXTRACTED FEATURE SET OF VOWEL DATA

Extracted Features	Coefficients (α) of			w	Rank
	F_1	F_2	F_3		
V_1	-0.0057	0.0507	0.0006	0.7101	2
V_2	0.0008	-0.1598	0.0009	0.7376	1

is minimum in a fixed number of iterations is considered to be the best.

Let us consider the case of Iris data. Table IV shows the values of α_{ji} [in (17)] for different sets of extracted features along with their E-values. The extracted features are obtained by (17). Note that, the set containing two extracted features results in minimum E-value, and therefore, is considered to be the best of all. The expressions for these two extracted features are then written, from (17), as

$$I_1 = 0.001 * SL + 0.001 * SW - 0.003 * PL - 0.051 * PW$$

and

$$I_2 = 0.009 * SL + 0.024 * SW - 0.269 * PL - 0.066 * PW.$$

w-values representing the importance of the features I_1 and I_2 are found to be 0.712 669 and 0.889 967, respectively.

Similarly, the dimension of the best extracted feature space is found to be two for vowel data (Table V), and eight for both medical and mango-leaf data. (For medical and mango-leaf data, tables are not included to restrict the size of the article.) In order to demonstrate the effectiveness of the feature extraction method, we have compared the discriminating capability of the extracted features with that of the original ones, using k-NN classifier for $k = 1, 3$, and 5. For Iris and vowel data, Tables III, VI, and VII demonstrate the percentage classification using the extracted feature set and all possible subsets of the original feature set. In the case of Iris data, the recognition score using the extracted feature set is found to be greater than or equal to that obtained using any set of the original features, except for one case (e.g., the set $\{SL, SW, PL, PW\}$ with $k = 5$). Similar is the case with the vowel data, where the extracted feature pair performs better than any other set of original features, except the set $\{F_1, F_2, F_3\}$. For medical and mango-leaf data, comparison is made only between the extracted feature set and the entire original feature set (Tables VIII–IX). Table IX shows that the classification performance in the eight-dimensional extracted feature space of mango-leaf data is much better than that of its 18-dimensional original feature space for all values of k. Similar finding is obtained in the case of medical data, except for $k = 1$ (Table VIII).

In a part of the experiment, the neuro-fuzzy method for feature extraction is compared with the well-known principal component analysis in a connectionist framework, called principal component analysis network (PCAN) [11]. Here, we provide the comparison for Iris data only. Scatter plots in Figs. 4 and 5 show

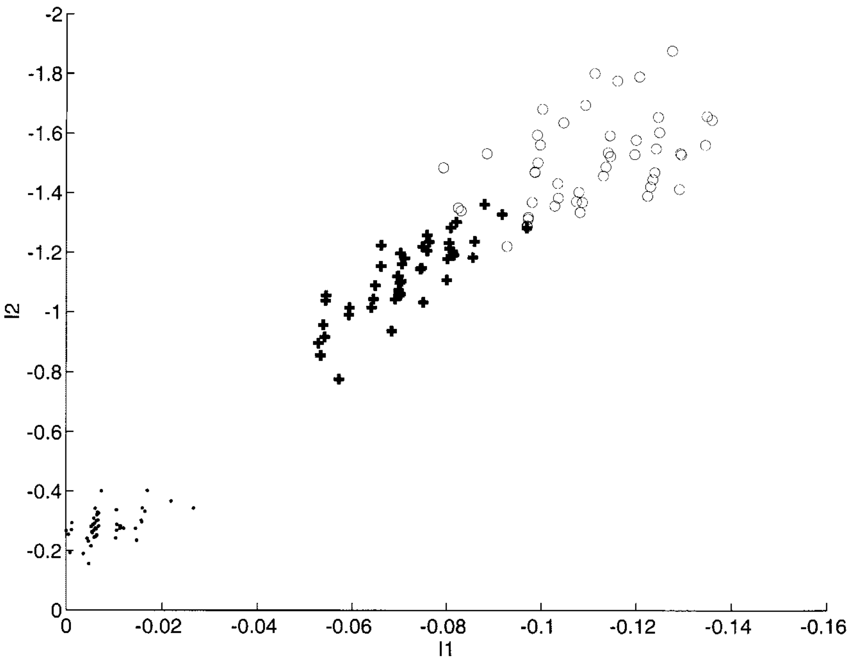


Fig. 4. Scatter plot $I_1 - I_2$, in the extracted plane obtained by the neuro-fuzzy method, of Iris data. Here "·," "+," and "o" represent classes Iris Setosa, Iris Versicolor, and Iris Virginica, respectively.

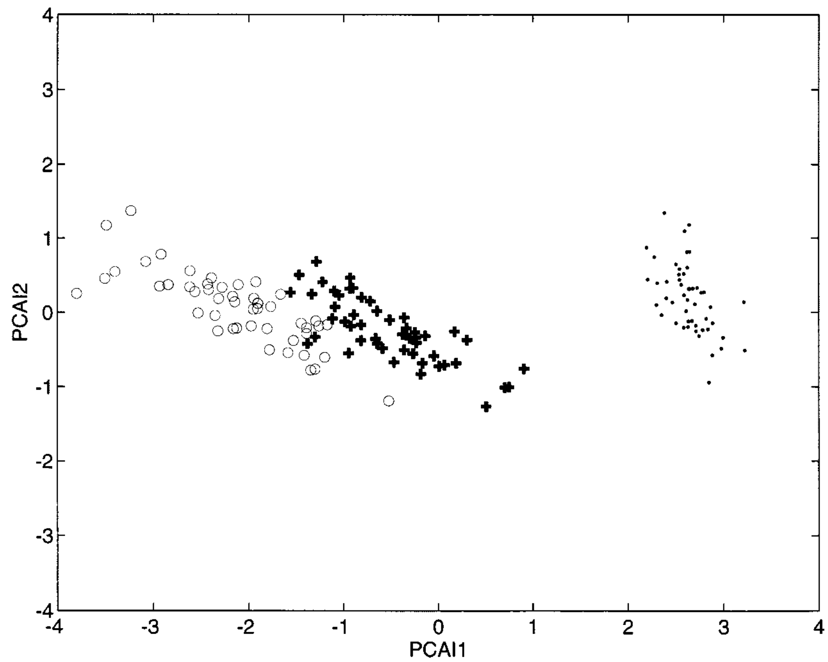


Fig. 5. Scatter plot $PCA_1 - PCA_2$, in the extracted plane obtained by PCAN, of Iris data. Here "·," "+," and "o" represent classes Iris Setosa, Iris Versicolor, and Iris Virginica, respectively.

the class structures in the two-dimensional extracted planes obtained by our neuro-fuzzy method and the PCAN, respectively. The number of samples lying in the overlapping region is seen to be more for the latter case. This is also verified from the results of fuzzy c-means clustering algorithm (for $c = 3$), where the number of misclassified samples (lying in other regions) is 14 for the former case, as compared to 17 in the latter.

In order to compare the said class structures of the extracted planes (Figs. 4 and 5) with that of the original feature space, one may refer to scatter plot for $PL - PW$ (Fig. 3). Note that $\{PL, PW\}$ is found, in Section V-A, to be the best feature pair for Iris data. The extracted feature plane $I_1 - I_2$ (Fig. 4) is seen to have more resemblance with that in Fig. 3, as compared to Fig. 5.

VI. Conclusions

In this article we have demonstrated how the concept of neuro-fuzzy computing can be exploited for developing a

TABLE VI
RECOGNITION SCORE WITH k-NN CLASSIFIER (CONTINUED FROM TABLE III) FOR IRIS DATA

Data set	Feature set	% classification		
		$k=1$	$k=3$	$k=5$
Original	$\{SL, SW\}$	74.67	76.67	76.00
	$\{SL, PL\}$	95.33	93.33	95.33
	$\{SL, PW\}$	94.67	94.00	94.00
	$\{SW, PL\}$	94.67	92.00	93.33
	$\{SW, PW\}$	90.67	94.00	94.67
	$\{PL, PW\}$	93.33	96.00	96.00
	$\{SL, SW, PL\}$	94.00	94.00	94.00
	$\{SL, SW, PW\}$	93.33	93.33	92.00
	$\{SL, PL, PW\}$	96.00	96.67	96.00
	$\{SW, PL, PW\}$	94.00	96.67	95.33
	$\{SL, SW, PL, PW\}$	95.33	96.00	96.67
Extracted	$\{I_1, I_2\}$	96.00	96.67	96.00

TABLE VII
RECOGNITION SCORE WITH k-NN CLASSIFIER FOR VOWEL DATA

Data set	Feature set	% classification		
		$k=1$	$k=3$	$k=5$
Original	$\{F_1\}$	26.52	27.21	27.21
	$\{F_2\}$	38.58	38.23	47.76
	$\{F_3\}$	26.06	33.41	33.87
	$\{F_1, F_2\}$	56.37	68.20	76.35
	$\{F_1, F_3\}$	44.32	46.84	55.80
	$\{F_2, F_3\}$	58.21	63.03	63.95
	$\{F_1, F_2, F_3\}$	78.42	81.29	82.43
Extracted	$\{V_1, V_2\}$	74.63	75.78	76.35

TABLE VIII
RECOGNITION SCORE WITH k-NN CLASSIFIER FOR MEDICAL DATA

Feature set	% classification		
	$k=1$	$k=3$	$k=5$
Extracted	53.92	56.34	59.89
Original	55.22	56.16	59.14

TABLE IX
RECOGNITION SCORE WITH k-NN CLASSIFIER FOR MANGO-LEAF DATA

Feature set	% classification		
	$k=1$	$k=3$	$k=5$
Extracted	85.71	88.10	92.86
Original	71.69	68.67	70.48

methodology for both feature selection and extraction under unsupervised mode. Two different layered networks are designed. Various tasks like membership computation in both original and transformed spaces, and minimization of feature evaluation index are embedded into them. Both the algorithms consider interdependence of the original features.

Since there is no unsupervised connectionist feature selection method available in the literature, to our knowledge, we have compared our results with those of many related supervised algorithms [2], [14], [15], [18], [19]. Interestingly, our unsupervised method has performed like supervised ones. Its validity is demonstrated in terms of both classification performance and class structures with the help of k-NN classifier and scatter plots.

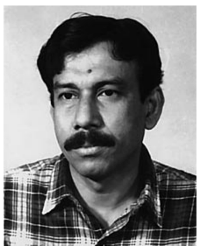
The extracted feature space has been able to provide better classification performance than the original ones for all the data sets. The extent of overlapping region in the extracted plane of the neuro-fuzzy method is less (as found by the scatter plots and fuzzy c-means algorithm) than that of the PCAN. Moreover, the neuro-fuzzy feature extraction preserves the data structure, cluster shape and inter pattern distances better than the PCAN. Here we mention that the scatter plots obtained by the PCAN and Sammon's nonlinear discriminant analysis (NDA) network [6] are alike.

Note that, the task of feature extraction by the neuro-fuzzy method involves projection of an n-dimensional original space directly to an n'-dimensional transformed space. On the other hand, in the case of PCAN, this task involves projection of an n-dimensional original space to an n-dimensional transformed space, followed by selection of best n' number of transformed components.

REFERENCES

[1] P. A. Devijver and J. Kittler, *Pattern Recognition, A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1982.
[2] J. M. Steppe and K. W. Bauer Jr., "Improved feature screening in feedforward neural networks," *Neurocomputing*, vol. 13, pp. 47–58, 1996.
[3] 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.
[4] M. Pregenzer, G. Pfurtscheller, and D. Flotzinger, "Automated feature selection with a distinctive sensitive learning vector quantizer," *Neurocomputing*, vol. 11, pp. 19–29, 1996.
[5] D. Lowe and A. R. Webb, "Optimized feature extraction and Bayes decision in feedforward classifier networks," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 355–364, 1991.
[6] J. Mao and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Trans. Neural Networks*, vol. 6, pp. 296–317, 1995.
[7] E. Saund, "Dimensionality-reduction using connectionist networks," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 304–314, 1989.
[8] W. A. C. Schmidt and J. P. Davis, "Pattern recognition properties of various feature spaces for higher order neural networks," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 795–801, 1993.
[9] J. Lampinen and E. Oja, "Distortion tolerant pattern recognition based on self-organizing feature extraction," *IEEE Trans. Neural Networks*, vol. 6, pp. 539–547, 1995.
[10] M. A. Kraaijveld, J. Mao, and A. K. Jain, "A nonlinear projection method based on Kohonen's topology preserving maps," *IEEE Trans. Neural Networks*, vol. 6, pp. 548–559, 1995.
[11] J. Rubner and P. Tavan, "A self-organizing network for principal component analysis," *Europhys. Lett.*, vol. 10, pp. 693–698, 1989.
[12] S. K. Pal and S. Mitra, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*. New York: Wiley, 1999.
[13] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.
[14] S. K. Pal and D. Dutta Majumder, *Fuzzy Mathematical Approach to Pattern Recognition*. New York: Wiley (Halsted), 1986.
[15] S. K. Pal and B. Chakraborty, "Fuzzy set theoretic measures for automatic feature evaluation," *IEEE Trans. Syst., Man., Cybern.*, vol. 16, pp. 754–760, 1986.

[16] Y. Hayashi, "A neural expert system with automated extraction of fuzzy if-then rules and its application to medical diagnosis," in *Advances in Neural Inform. Processing Syst.*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds. Los Altos, CA: Morgan Kaufmann, 1991, pp. 578–584.
[17] S. Mitra, "Fuzzy MLP-based expert system for medical diagnosis," *Fuzzy Sets Syst.*, vol. 65, pp. 285–296, 1994.
[18] S. K. Pal, "Fuzzy set theoretic measures for automatic feature evaluation: II," *Inform. Sci.*, vol. 64, pp. 165–179, 1992.
[19] R. K. De, J. Basak, and S. K. Pal, "Neuro-fuzzy feature evaluation with theoretical analysis," *Neural Networks*, vol. 12, pp. 1429–1455, 1999.



Sankar K. Pal (M'81–SM'84–F'93) received the M.Tech. and Ph.D. degrees in radio physics and electronics in 1974 and 1979, respectively, from the University of Calcutta, India. In 1982, he received the Ph.D. degree in electrical engineering along with the D.I.C. degree from the Imperial College, University of London, U.K.

He is a Distinguished Scientist and Founding Head of the Machine Intelligence Unit at the Indian Statistical Institute, Calcutta. He worked at the University of California, Berkeley, and the University of Maryland, College Park, from 1986 to 1987 as a Fulbright Postdoctoral Visiting Fellow; at the NASA Johnson Space Center, Houston, Texas from 1990 to 1992 and in 1994 as a Guest Investigator under the NRC-NASA Senior Research Associateship program; and at the Hong Kong Polytechnic University, Hong Kong in 1999 as a Visiting Professor. He served as a Distinguished Visitor of IEEE Computer Society (USA) for the Asia-Pacific Region from 1997 to 1999. His research interests include pattern recognition, image processing, soft computing, neural nets, genetic algorithms, and fuzzy systems. He is a Coauthor of six books including *Fuzzy Mathematical Approach to Pattern Recognition* (New York: Wiley, 1986) and *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing* (New York: Wiley, 1999).

Dr. Pal is a Fellow of the Third World Academy of Sciences, Italy, and all the four National Academies for Science/Engineering in India. He received the 1990 S. S. Bhatnagar Prize, the 1993 Jawaharlal Nehru Fellowship, the 1993 Vikram Sarabhai Research Award, the 1993 NASA Tech Brief Award, the 1994 IEEE Transactions Neural Networks Outstanding Paper Award, the 1995 NASA Patent Application Award, the 1997 IETE–Ram Lal Wadhwa Gold Medal, the 1998 Om Bhasin Foundation Award, and the 1999 G. D. Birla Award for Scientific Research. He was an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS from 1994 to 1998, *Pattern Recognition Letters*, *Neurocomputing*, *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, and *Fundamenta Informaticae*; a Member of the Executive Advisory Editorial Board of IEEE TRANSACTIONS ON FUZZY SYSTEMS and the *International Journal of Approximate Reasoning*; and a Guest Editor of many journals including the *IEEE Computer*.

Rajat K. De (M'00) received the B.Tech. degree in computer science and engineering from the University of Calcutta, India, in 1991, and the Master of Computer Science and Engineering degree from Jadavpur University, India, in 1993.

He is a Computer Engineer in the Machine Intelligence Unit of the Indian Statistical Institute, Calcutta, India. He was a Dr. K. S. Krishnan Senior Research Fellow, sponsored by the Department of Atomic Energy, Government of India, from 1993 to 1998 and a Research Associate, sponsored by the Council for Scientific and Industrial Research, Government of India, from 1998 to 1999. His research interests include pattern recognition, image processing, fuzzy sets, neural networks, soft computing, and case-based reasoning.

Jayanta Basak (M'95–SM'99) received the Bachelor's degree in electronics and telecommunication engineering from Jadavpur University, Calcutta, in 1987; the Master's degree in computer science and engineering from the Indian Institute of Science (IISc), Bangalore, in 1989; and the Ph.D. degree from the Indian Statistical Institute (ISI), Calcutta, in 1995.

He served as a Computer Engineer in the Knowledge Based Computer Systems Project of ISI, Calcutta, from 1989 to 1992. In 1992, he joined as a faculty in the Electronics and Communication Sciences Unit of ISI, Calcutta. Since 1996, he has been Associate Professor in the Machine Intelligence Unit of ISI, Calcutta. He was a Researcher in the RIKEN Brain Science Institute, Saitama, Japan from 1997 to 1998, and a Visiting Scientist in the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA, from 1991 to 1992. His research interests include neural networks, pattern recognition, image analysis, and fuzzy sets.

Dr. Basak received a Gold Medal from Jadavpur University, Calcutta, in 1987, the Young Scientist Award in Engineering Sciences from the Indian National Science Academy (INSA) in 1996, and the Junior Scientist Award in Computer Science from the Indian Science Congress Association in 1994.