ELSEVIER

Pattern Recognition Letters 19 (1998) 997–1006

Pattern Recognition Letters

Unsupervised feature selection using a neuro-fuzzy approach

Jayanta Basak, Rajat K. De, Sankar K. Pal *

Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, India

Received 12 December 1997; received in revised form 19 June 1998

Abstract

A neuro-fuzzy methodology is described which involves connectionist minimization of a fuzzy feature evaluation index with unsupervised training. The concept of a flexible membership function incorporating weighed distance is introduced in the evaluation index to make the modeling of clusters more appropriate. A set of optimal weighing coefficients in terms of networks parameters representing individual feature importance is obtained through connectionist minimization. Besides, the investigation includes the development of another algorithm for ranking of 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. © 1998 Published by Elsevier Science B.V. All rights reserved.

1. Introduction

Feature selection or extraction is a process of selecting a map of the form $\boldsymbol{x}' = f(\boldsymbol{x})$ by which a sample $\boldsymbol{x}(x_1, x_2, \ldots, x_n)$ in an n-dimensional measurement space ($\mathbb{R}^n$) is transformed into a point $\boldsymbol{x}'(x'_1, x'_2, \ldots, x'_{n'})$ in an n'-dimensional ($n' < n$) feature space ($\mathbb{R}^{n'}$). The problem of feature selection deals with choosing some of the x_is from the measurement space to constitute the feature space. On the other hand, the problem of feature extraction deals with generating new x'_js (constituting the feature space) based on some x_is 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.

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. Fuzzy set theoretic approaches for feature selection are mainly based on measures of entropy and index of fuzziness (Pal and Chakraborty, 1986; Pal, 1992), fuzzy c-means and fuzzy ISODATA algorithms (Bezdek and Castelaz, 1977). Some of the recent attempts made for feature selection in the framework of ANN are mainly based on multilayer feedforward networks and self-organizing networks (Priddy et al., 1993; Steppe and Bauer, Jr., 1996; De et al., 1997; Pregenzer et al., 1996). Note that, depending on whether the class information of the samples is

[*] Corresponding author. E-mail: sankar@isical.ac.in.

0167-8655/98/$ – see front matter © 1998 Published by Elsevier Science B.V. All rights reserved.
PII: S0167-8655(98)00083-X

known or not, these methods are classified under supervised or unsupervised mode. For example, the algorithms described in (Pal and Chakraborty, 1986; Pal, 1992; Bezdek and Castelaz, 1977; Priddy et al., 1993; Steppe and Bauer, Jr., 1996; De et al., 1997) fall under the supervised category, whereas those in (Bezdek and Castelaz, 1977; Pregenzer et al., 1996) are in unsupervised mode.

Recently, attempts have been made to integrate the merits of fuzzy set theory and ANN under the heading ‘neuro-fuzzy computing’ 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 or extraction.

The present article is an attempt in this regard 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. The similarity between two patterns is measured by a weighed distance between them. The weight 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 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 by an unsupervised learning process, thereby determining the optimum weight coefficients providing an ordering of the individual features.

In another part of the investigation, the aforesaid fuzzy evaluation index is used alone to find the best subset of features. This is done by computing the evaluation index (with weight coefficients equal to 1) on different subsets of features and then ordering them accordingly. The effectiveness of these algorithms is demonstrated on four different data sets, namely, vowel (Pal and Dutta Majumder, 1986; Pal and Chakraborty, 1986), Iris (Fisher, 1936), medical (Hayashi, 1991) and mango-leaf (Pal, 1992) .

2. Feature evaluation index

In this section we first of all provide a definition of the fuzzy feature evaluation index. The membership function for its realization is then defined in terms of a distance measure and weight coefficients.

2.1. 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 ($n' < n$) 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 subset (Ω) of features is defined as

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

It has the following characteristics.

1. If $\mu_{pq}^{\mathrm{O}} = \mu_{pq}^{\mathrm{T}} = 0$ or 1, the contribution of the pair of patterns to the evaluation index E is zero (minimum).
2. If $\mu_{pq}^{\mathrm{O}} = \mu_{pq}^{\mathrm{T}} = 0.5$, the contribution of the pair of patterns to E becomes 0.25 (maximum).
3. For $\mu_{pq}^{\mathrm{O}} < 0.5$ as $\mu_{pq}^{\mathrm{T}} \to 0$, E decreases. For $\mu_{pq}^{\mathrm{O}} > 0.5$ as $\mu_{pq}^{\mathrm{T}} \to 1$, E decreases.

Therefore, the feature evaluation index decreases as the membership value representing the degree of belonging of the pth and qth patterns to the same cluster in the transformed feature space tends to either 0 (when $\mu^{\mathrm{O}} < 0.5$) or 1 (when $\mu^{\mathrm{O}} > 0.5$). 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.

2.2. Computation of membership function

In order to satisfy the characteristics of E (Eq. (1)), as stated in the previous section, the membership function (μ) in a feature space may be defined as

$$\mu_{pq} = \begin{cases} 1 - d_{pq}/D, & \text{if } d_{pq} \leqslant D, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

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 value of μ_{pq}, 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 similarity between the patterns is just in between 0 and D, the difficulty in making a decision, whether both patterns are in the same cluster or not, becomes maximum, thereby making the situation most ambiguous. We can take $D = \beta d_{\max}$ where $d_{\max}$ is the maximum separation between a pair of patterns in the entire feature space, and $0 < \beta < 1$ is a user defined constant. In one extreme case, $D = d_{\max}$ when β is chosen as 1.

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

$$d_{pq} = \left[\sum_i (x_{pi} - x_{qi})^2\right]^{1/2}, \tag{3}$$

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

$$d_{\max} = \left[\sum_i (x_{\max_i} - x_{\min_i})^2\right]^{1/2}, \tag{4}$$

where $x_{\max_i}$ and $x_{\min_i}$ are the maximum and minimum values of the ith feature in the corresponding feature space.

2.3. Incorporating weight 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} (Eq. (3)). Since, d_{pq} is a Euclidian distance, the methodology implicitly assumes that the clusters are hyperspherical. But in practice, this may not be the case. To model the practical situation we have introduced the concept of weighed distance such that

$$\begin{aligned} d_{pq} &= \left[\sum_i w_i^2 (x_{pi} - x_{qi})^2\right]^{1/2} \\ &= \left[\sum_i w_i^2 \chi_i^2\right]^{1/2}, \quad \chi_i = (x_{pi} - x_{qi}), \end{aligned} \tag{5}$$

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

The membership value μ_{pq} is now obtained by Eqs. (2)–(5), and becomes dependent on w_i. The values of w_i (< 1) make the μ_{pq} function of Eq. (2) flattened along the axis of d_{pq}. The lower the value of w_i, the higher is the extent of flattening. In the extreme case, when $w_i = 0$, $\forall i$, $d_{pq} = 0$ and $\mu_{pq} = 1$ for all pairs of patterns, i.e., all patterns lie on the same point making them indiscriminable.

In pattern recognition literature, the weight w_i (Eq. (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 one may define μ_{pq} in a different way satisfying the above mentioned characteristics. The computation of μ_{pq} in Eq. (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 (Bezdek and Castelaz, 1977), the present algorithm is independent of the number of clusters and is able to select a set of salient features without (explicitly) clustering the feature space.

3. Feature selection

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

3.1. Ordering of feature subsets using E (Method 1)

From the aforesaid discussion we see that if a particular subset (Ω_1) of features is more important than another subset (Ω_2) then E computed over Ω_1 will be less than that computed over Ω_2. Therefore, the task of feature subset selection requires selecting the subset Ω from a given set of n features for which E is minimum. This is done by computing the E values for all possible $(2^n - 1)$ subsets of features using Eqs. (1)–(4), and ranking them accordingly. Here μ^{O} is computed on the n-dimensional original feature space, whereas μ^{T} is done on its various subsets. Note that, if the subset Ω 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 the Euclidian distance (Eq. (3)) to compute μ-values. Here we consider Eq. (5) instead of Eq. (3). Therefore, the evaluation index E (Eq. (1)) becomes a function of $\boldsymbol{w}$ $(= [w_1, w_2, \ldots, w_n])$, if we consider ranking of n features in a set. Here μ^{O} and μ^{T} are both computed over the original n-dimensional feature space. The only difference is that μ^{O} needs Eqs. (2)–(4), while μ^{T} needs Eqs. (2), (4) and (5) for their computation.

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

3.3. 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 an n-dimensional (original) feature space. The hidden layer consists of n nodes which compute the part χ_i^2 of Eq. (5) for each pair of patterns. The output layer consists of two nodes. One of them computes μ^{O}, and the other μ^{T}. The feature evaluation index E (Eq. (14)) is computed from these μ-values off the network.

Input nodes receive activations corresponding to feature values of each pair of patterns. A jth node in the hidden layer is connected only to an ith 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 μ^{T}-values is connected to a jth node in the hidden layer via connection weight W_j $(= w_j^2)$, whereas that computing μ^{O}-values is connected to all the nodes in the hidden layer via connection weights $+1$ each.

During training, each pair of patterns is presented at the input layer and the evaluation index is computed. The weights W_js are updated using a gradient-descent technique in order to minimize the index E. Note that $d_{\max}$ is directly computed from the unlabeled training set. The values of $d_{\max}$ and β are stored in both the output nodes for the computation of D. When the pth and qth patterns are presented to the input layer, the activation produced by the ith $(1 \leqslant i \leqslant 2n)$ input node is

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

where

$$u_i^{(0)} = x_{pi} \quad \text{for } 1 \leqslant i \leqslant n$$

and

$$u_{i+n}^{(0)} = x_{qi} \quad \text{for } 1 \leqslant i \leqslant n, \tag{7}$$

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

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

and the activation produced by it is

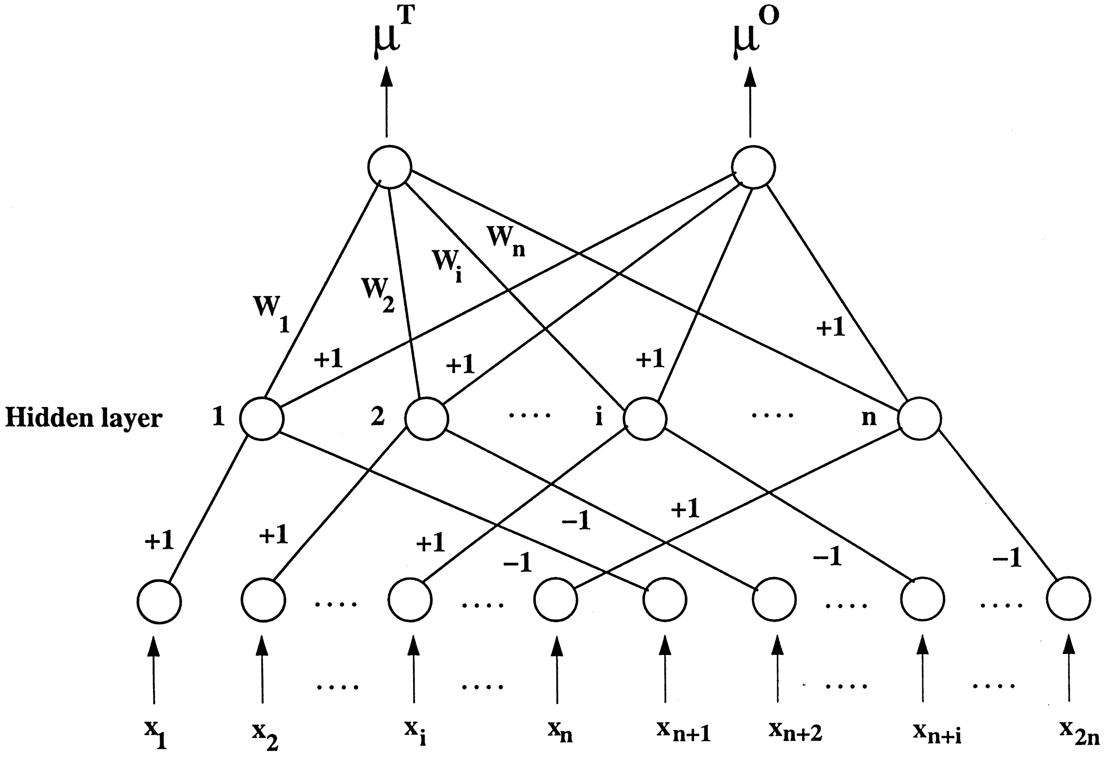

Fig. 1. A schematic diagram of the neural network model.

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

The total activation received by the output node which computes the μ^{T}-values, is

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

and that received by the other output node which computes the μ^{O}-values, is

$$u_{\mathrm{O}}^{(2)} = \sum_j v_j^{(1)}. \tag{11}$$

Therefore, $u_{\mathrm{T}}^{(2)}$ and $u_{\mathrm{O}}^{(2)}$ represent d_{pq}^2 as given by Eqs. (5) and (3), respectively. The activations, $v_{\mathrm{T}}^{(2)}$ and $v_{\mathrm{O}}^{(2)}$, of the output nodes represent μ_{pq}^{T} and μ_{pq}^{O} for the pth and qth pattern pair, respectively. Thus,

$$v_{\mathrm{T}}^{(2)} = \begin{cases} 1 - \left(u_{\mathrm{T}}^{(2)}\right)^{1/2}/D & \text{if } \left(u_{\mathrm{T}}^{(2)}\right)^{1/2} \leqslant D, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

and

$$v_{\mathrm{O}}^{(2)} = \begin{cases} 1 - \left(u_{\mathrm{O}}^{(2)}\right)^{1/2}/D & \text{if } \left(u_{\mathrm{O}}^{(2)}\right)^{1/2} \leqslant D, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

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

$$E(\boldsymbol{W}) = \frac{2}{s(s-1)} \sum_p \sum_{q \neq p} \frac{1}{2} \left[v_{\mathrm{T}}^{(2)} \left(1 - v_{\mathrm{O}}^{(2)}\right) + v_{\mathrm{O}}^{(2)} \left(1 - v_{\mathrm{T}}^{(2)}\right) \right]. \tag{14}$$

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

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

where η is the learning rate.

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

$$\frac{\partial E(\boldsymbol{W})}{\partial W_j} = \frac{2}{s(s-1)} \sum_p \sum_{q \neq p} \frac{1}{2} \left[1 - 2v_{\mathrm{O}}^{(2)}\right] \frac{\partial v_{\mathrm{T}}^{(2)}}{\partial W_j}, \qquad (16)$$

$$\frac{\partial v_{\mathrm{T}}^{(2)}}{\partial W_j} = \begin{cases} -\frac{\frac{1}{2}\left(u_{\mathrm{T}}^{(2)}\right)^{-1/2} \partial u_{\mathrm{T}}^{(2)} / \partial W_j}{D}, & \text{if } \left(u_{\mathrm{T}}^{(2)}\right)^{1/2} \leqslant D \\ 0 & \text{otherwise} \end{cases} \qquad (17)$$

and

$$\frac{\partial u_{\mathrm{T}}^{(2)}}{\partial W_j} = v_j^{(1)}. \qquad (18)$$

Algorithm

- Calculate $d_{\max}$ from the unlabeled training set. Store $d_{\max}$ and the user specified constant β in the output nodes.
- Initialize W_j with small random values in [0,1].
- Repeat until convergence, i.e., until the value of E becomes less than or equal to a certain predefined small quantity, or until the number of iterations attains a certain predefined value:
  - For each pair of patterns:
    - Present the pattern pair to the input layer.
    - Compute ΔW_j for each j using the updating rule in Eq. (15).
  - Update W_j for each j with the average value of ΔW_j.

After convergence, $E(\boldsymbol{W})$ attains a local minimum. Then the weights ($W_j = w_j^2$) of the links connecting hidden nodes and the output node computing the μ^{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 a neuro-fuzzy approach for individual feature ranking, finds the set of w_is (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

Here we demonstrate the effectiveness of the algorithms presented above on four data sets, namely, vowel data (Pal and Dutta Majumder, 1986; Pal and Chakraborty, 1986), Iris data (Fisher, 1936), medical data (Hayashi, 1991) and mango-leaf data (Pal, 1992). The 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. Fig. 2 shows a 2-D projection of the 3-D feature space of the six vowel classes (∂, a, i, u, e, o) in the F_1–F_2 plane (for ease of depiction). The details of the data and its extraction procedure are available in (Pal and Dutta Majumder, 1986). This vowel data is being extensively used for more than two decades in the area of pattern recognition.

Fisher's Iris data (Fisher, 1936) 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). This data set has been used in many research investigations related to pattern recognition and has become a sort of benchmark-data.

The medical data consisting of nine input features and four pattern classes, deals with various *Hepatobiliary disorders* (Hayashi, 1991) 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 (Pal, 1992) 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 measure-

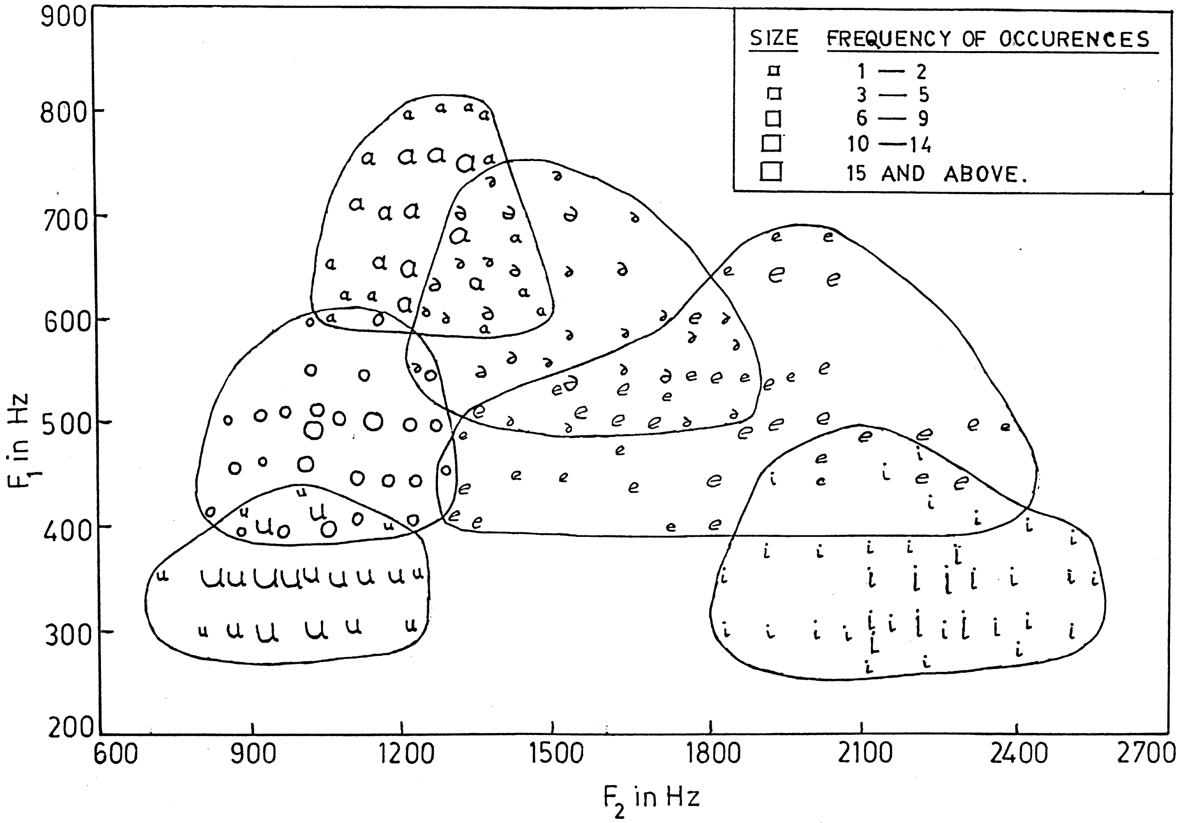

Fig. 2. Two-dimensional ($F_1 - F_2$) plot of the vowel data.

ments 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.

Although the data considered here have known classes, one may note that this class information has in no way been utilized during the experiment of feature evaluation. In other words, the methods, as described before, are entirely based on unsupervised training.

4.1. Ordering of feature subsets using Method 1

Table 1 shows the ordering of different subsets for the four types of data using Method 1. Note that for the vowel and Iris data, we have computed E-values for all possible subsets, including the individual features (i.e., seven for vowel and fifteen for Iris data), and ranked them accordingly. For medical and mango-leaf data, since the number of features is large, we have, first of all, computed the E-values for the individual features. A few of the best (e.g., GOT, LDH, GPT, GGT for medical data, and Pe, $(L + P)/B$, UM/LM, L/B, Z for mango-leaf data) are selected after ranking. Then we have computed the E-values for different subsets containing only these selected features. As a result, we have 20 subsets for the medical data and 44 subsets for the mango-leaf data. (However, for the mango-leaf data, we have shown in Table 1 the ordering of the first twenty subsets only, for brevity.)

It is seen from Table 1 that a subset of higher cardinality is not necessarily more important than

Table 1
Importance of different feature subsets using Method 1 ($X > Y$ means X is more important than Y)

Data sets	Order of importance
Vowel	$\{F_1,F_2\} > \{F_2\} > \{F_1,F_2,F_3\} > \{F_2,F_3\} > \{F_1,F_3\} > \{F_3\} > \{F_1\}$
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}
Mango-leaf	$\{\mathrm{Pe}\} > \{\mathrm{Pe}, \mathrm{UM/LM}\} > \{\mathrm{Pe}, L/B\} > \{\mathrm{Pe}, L/B, \mathrm{UM/LM}\} > \{\mathrm{Pe}, (L+P)/B\} > \{\mathrm{Pe}, (L+P)/B, \mathrm{UM/LM}\} >$ $\{\mathrm{Pe}, L/B, (L+P)/B\} > \{\mathrm{Pe}, L/B, (L+P)/B, \mathrm{UM/LM}\} >$ $\{Z, \mathrm{Pe}, \mathrm{UM/LM}\} > \{Z, \mathrm{Pe}, L/B\} > \{Z, \mathrm{Pe}, L/B, \mathrm{UM/LM}\} > \{Z, \mathrm{Pe}, (L+P)/B\} > \{Z, \mathrm{Pe}, (L+P)/B, \mathrm{UM/LM}\} >$ $\{Z, \mathrm{Pe}, L/B, (L+P)/B\} > \{Z, \mathrm{Pe}, L/B, (L+P)/B, \mathrm{UM/LM}\} >$ $\{(L+P)/B\} > \{(L+P)/B, \mathrm{UM/LM}\} > \{L/B, (L+P)/B\} > \{L/B, (L+P)/B, \mathrm{UM/LM}\} > \{\mathrm{UM/LM}\} > \cdots$

ones of lower cardinality. For vowel, F_2 being the best individual feature is seen to be a member of the best four subsets. This conforms to an earlier investigation (Pal, 1992). Similarly, for the Iris data, it is PL which has become a member of the first five best subsets. For the medical data, the best 15 subsets contain at least one of the four best individual features (GOT, LDH, GPT and GGT). Similarly, for the mango-leaf data, it is the first 10 subsets in which at least one of the five best individual features (Pe, (L + P)/B, UM/LM, L/B, Z) became a member.

In a part of the investigation, we used the k-nn classifier to study the importance of these selected features in classifying the data set in supervised mode. For this purpose, we used only the vowel data with a 50% training set and $k = 3$. It is found that the order of importance of the individual features as obtained by the k-nn classifier is $F_2 > F_3 > F_1$, which is the same as that obtained in Table 1. For the pairwise features also, the k-nn classifier and Method 1 produce the same ordering i.e., $\{F_1,F_2\} > \{F_2,F_3\} > \{F_1,F_3\}$. It may be noted that the k-nn classifier provides the best classification performance with F_1, F_2 and F_3 taken together, although this subset ranks third in case of Method 1.

4.2. Ordering of individual features using Method 2

Tables 2–5 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] while training the network. In all the cases, the value of β is taken as 2.0.

The order of importance of the features for the vowel data is found to be $F_2 > F_1 > F_3$ (Table 2)

Table 2
w-values for the vowel data using Method 2

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

Table 3
w-values for the Iris data using Method 2

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

Table 4
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

Table 5
w-values for the mango-leaf data using Method 2

Feature	w	Order
Z	0.021467	17
A	0.0	18
Pe	1.0	1
L	0.657581	14
B	0.704104	8
P	0.700014	10
K	0.707102	2
S	0.707097	3
SI	0.656674	15
$L+P$	0.621514	16
L/P	0.674567	12
L/B	0.703800	9
$(L+P)/B$	0.673539	13
A/L	0.705429	6
A/B	0.680706	11
A/Pe	0.706753	4
UM/LM	0.704375	7
UPe/LPe	0.706594	5

which conforms to that obtained in an earlier investigation (Pal, 1992). For the Iris data, the best two features are found to be PL and PW (Table 3) which are also the best two individual features obtained by Method 1 (Table 1) and in an earlier investigation (Steppe and Bauer, Jr., 1996). Similarly for the medical data, the best two features are GOT and LDH (Table 4) which are also the best two individual features found by Method 1 (Table 1). However, for the mango-leaf data, only the best feature (Pe) obtained by Method 2 (Table 5) matches with that of Method 1.

5. Conclusions

In this article we have demonstrated how the concept of neuro-fuzzy computing can be exploited for developing a methodology for feature selection in unsupervised mode. The methodology developed involves connectionist optimization of a fuzzy feature evaluation index, thereby determining the ranking of various features. The algorithm considers interdependence of the original features. Unlike the method based on the fuzzy c-means algorithm (Bezdek and Castelaz, 1977), the algorithm provides a ranking of the individual features, without clustering the feature space explicitly. The effectiveness of the method is demonstrated extensively on 3-d speech (vowel) data, 4-d Iris data, 9-d medical data and a 18-d mango-leaf data set.

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 algorithms conform well to those obtained using other methods (Pal, 1992; Steppe and Bauer, Jr., 1996). Although a network is used in Method 2 for minimization of the evaluation index, one may consider other optimization techniques also for this task.

We have also provided a comparison of the feature subset selection algorithm (Method 1) with the k-nn classifier. However, one may note that the former is based on unsupervised partitioning, whereas the later is a supervised classification method.

Acknowledgements

Mr. Rajat K. De is grateful to the Department of Atomic Energy, Government of India for providing him a Dr. K.S. Krishnan Senior Research Fellowship. The work is partly supported by Grant No. 25(0093)/97/EMR-II of CSIR, New Delhi. The work was partly done when Jayanta Basak was in RIKEN Brain Science Institute, Wakoshi, Saitama, Japan.

References

Bezdek, J.C., Castelaz, P., 1977. Prototype classification and feature selection with fuzzy sets. IEEE Trans. on Systems, Man and Cybernetics 7, 87–92.

Pal, S.K., 1992. Fuzzy set theoretic measures for automatic feature evaluation: II. Information Sciences 64, 165–179.

Pal, S.K., Chakraborty, B., 1986. Fuzzy set theoretic measures for automatic feature evaluation. IEEE Trans. on Systems, Man and Cybernetics 16, 754–760.

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

Steppe, J.M., Bauer Jr., K.W., 1996. Improved feature screening in feedforward neural networks. Neurocomputing 13, 47-58.

De, R.K., Pal, N.R., Pal, S.K., 1997. Feature analysis: neural network and fuzzy set theoretic approaches. Pattern Recognition 30, 1579–1590.

Pregenzer, M., Pfurtscheller, G., Flotzinger, D., 1996. Automated feature selection with a distinctive sensitive learning vector quantizer. Neurocomputing 11, 19–29.

Pal, S.K., Dutta Majumder, D., 1986. Fuzzy Mathematical Approach to Pattern Recognition. Wiley (Halsted Press), New York.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

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