ELSEVIER Journal of Information Sciences 105 (1998) 173–188

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

Fuzzy feature evaluation index and connectionist realization

Sankar K. Pal *, Jayanta Basak [1], Rajat K. De [2]

Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India

Received 22 January 1996; accepted 10 October 1997

Abstract

A new feature evaluation index based on fuzzy set theory and a connectionist model for its evaluation are provided. A concept of flexible membership function incorporating weighting factors, is introduced which makes the modeling of the class structures more appropriate. A neuro-fuzzy algorithm is developed for determining the optimum weighting coefficients representing the feature importance. The overall importance of the features is evaluated both individually and in a group considering their dependence as well as independence. Effectiveness of the algorithms along with comparison is demonstrated on speech and Iris data. © 1998 Elsevier Science Inc. All rights reserved.

1. Introduction

The process of selecting the necessary information to present to the decision rule is called *feature selection*. Its main objective 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 classification.

The criterion of a good feature is that it should be unchanging with any other possible variation within a class, while emphasizing differences that are im-

* Corresponding author. E-mail: sankar@isical.ernet.in.

[1] E-mail: jayanta@isical.ernet.in.

[2] E-mail: res9318@isical.ernet.in.

PII: S0020-0255(97)10023-8

portant in discriminating between patterns of different types. One of the useful techniques to achieve this is clustering transformation [1], which maximizes/minimizes the interset/intraset distance using a diagonal transformation, such that smaller weights are given to features having larger variance (less reliable). Other separability measures based on information theoretic approach include divergence, Bhattacharya coefficient, and the Kolmogorov variational distance [1–3]. Several methods based on fuzzy set theory [4–6] and Artificial Neural Network (ANN) [7–11] have also been reported. Incorporation of fuzzy set theory enables one to deal with uncertainties in a system, arising from vagueness, incompleteness in information etc., in an efficient manner. ANNs, having the capability of fault tolerance, adaptivity and generalization, and scope for massive parallelism, are widely used in dealing with optimization tasks. Recently, attempts are being made to integrate the merits of fuzzy set theory and ANN under the heading "neuro-fuzzy computing" for making the systems artificially more intelligent.

The present article is an attempt in this line, and has two parts. In the first part a new fuzzy set theoretic feature evaluation index, in terms of individual class membership, is defined and its performance with an existing one [4,5] is compared for ranking the features (or subsets of features). Its relation with Mahalanobis distance and divergence measure is experimentally demonstrated. The second part provides a neuro-fuzzy approach where a new connectionist model has been designed in order to perform the task of optimizing a modified version of the aforesaid fuzzy evaluation index which incorporates weighted distance for computing class membership values. This optimization process results in a set of weighting coefficients representing the importance of the individual features. These weighting coefficients lead to a transformation of the feature space for flexible modeling of class structures.

The effectiveness of the algorithms is demonstrated on two different data sets, namely, vowel and Iris data.

2. Evaluation index and feature subset selection

Let the pth pattern vector (pattern) be represented as $\mathbf{f}^{(p)} = [f_1^{(p)}, f_2^{(p)}, \ldots, f_i^{(p)}, \ldots, f_n^{(p)}]$, where n is the number of features in M (set of measurable quantities) and $f_i^{(p)}$ is the ith component of the vector. Let prob_k and $d_k(\mathbf{f}^{(p)})$ stand for the a priori probability for the class C_k and the distance of the pattern $\mathbf{f}^{(p)}$ from the kth mean vector $\mathbf{m}_k(= [m_{k_1}, m_{k_2}, \ldots, m_{k_i}, \ldots, m_{k_n}])$, respectively. m_{k_i} indicates the ith component of the vector $\mathbf{m}_k$.

The feature evaluation index for a subset (Ω) containing few of these n features is defined as,

$$E = \sum_{\mathbf{f}^{(p)} \in C_k} \sum_{k} \frac{s_k(\mathbf{f}^{(p)})}{\sum_{k' \neq k} s_{kk'}(\mathbf{f}^{(p)})} \times \alpha_k, \tag{1}$$

where $\mathbf{f}^{(p)}$ is constituted by the features of Ω only.

$$s_k(\mathbf{f}^{(p)}) = \mu_{C_k}(\mathbf{f}^{(p)}) \times \left(1 - \mu_{C_k}(\mathbf{f}^{(p)})\right) \tag{2}$$

and

$$\begin{aligned} s_{kk'}(\mathbf{f}^{(p)}) = {} & \frac{1}{2}\left[\mu_{C_k}(\mathbf{f}^{(p)}) \times \left(1 - \mu_{C_{k'}}(\mathbf{f}^{(p)})\right)\right] \\ & + \frac{1}{2}\left[\mu_{C_{k'}}(\mathbf{f}^{(p)}) \times \left(1 - \mu_{C_k}(\mathbf{f}^{(p)})\right)\right]. \end{aligned} \tag{3}$$

$\mu_{C_k}(\mathbf{f}^{(p)})$ and $\mu_{C_{k'}}(\mathbf{f}^{(p)})$ are the membership values of the pattern $\mathbf{f}^{(p)}$ in classes C_k and $C_{k'}$, respectively. α_k is the normalizing constant for class C_k which takes care of the effect of relative sizes of the classes.

Note that, s_k is zero (minimum) if $\mu_{C_k} = 1$ or 0, and is 0.25 (maximum) if $\mu_{C_k} = 0.5$. On the other hand, $s_{kk'}$ is zero (minimum) when $\mu_{C_k} = \mu_{C_{k'}} = 1$ or 0, and is 0.5 (maximum) for $\mu_{C_k} = 1, \mu_{C_{k'}} = 0$ or vice-versa.

Therefore, the term $s_k / \sum_{k' \neq k} s_{kk'}$ is minimum if $\mu_{C_k} = 1$ and $\mu_{C_{k'}} = 0$ for all $k' \neq k$ i.e., if the ambiguity in the belongingness of a pattern $\mathbf{f}^{(p)}$ to classes C_k and $C_{k'}$ $\forall k' \neq k$ is minimum (the pattern belongs to only one class). It is maximum when $\mu_{C_k} = 0.5$ for all k. In other words, the value of E decreases as the belongingness of the patterns increases to only one class (i.e., compactness of individual classes increases) and at the same time decreases for other classes (i.e., separation between classes increases). E increases when the patterns tend to lie at the boundaries between classes (i.e., $\mu \to 0.5$). Our objective is, therefore, to select those features for which the value of E is minimum.

In order to achieve this, the membership $(\mu_{C_k}(\mathbf{f}^{(p)}))$ of a pattern $\mathbf{f}^{(p)}$ to a class C_k is defined, with a multi-dimensional π-function [12] which is given by,

$$\begin{aligned} \mu_{C_k}(\mathbf{f}^{(p)}) &= 1 - 2d_k^2(\mathbf{f}^{(p)}), \quad 0 \leqslant d_k(\mathbf{f}^{(p)}) < \frac{1}{2}, \\ &= 2\left[1 - d_k(\mathbf{f}^{(p)})\right]^2, \quad \frac{1}{2} \leqslant d_k(\mathbf{f}^{(p)}) < 1, \\ &= 0, \quad \text{otherwise}. \end{aligned} \tag{4}$$

The distance $d_k(\mathbf{f}^{(p)})$ of the pattern $\mathbf{f}^{(p)}$ from $\mathbf{m}_k$ (the center of class C_k) is defined as,

$$d_k(\mathbf{f}^{(p)}) = \left[\sum_i \left(\frac{f_i^{(p)} - m_{k_i}}{\lambda_{k_i}}\right)^2\right]^{1/2}, \tag{5}$$

where

$$\lambda_{k_i} = 2 \max_p \left[|f_i^{(p)} - m_{k_i}|\right], \tag{6}$$

and

$$m_{k_i} = \frac{\sum_{p \in C_k} f_i^{(p)}}{|C_k|}. \tag{7}$$

Eqs. (4)-(7) are such that the membership $\mu_{C_k}(\mathbf{f}^{(p)})$ of a pattern $\mathbf{f}^{(p)}$ is 1 if it is located at the mean of C_k, and 0.5 if it is at the boundary (i.e., ambiguous region) for a symmetric class structure.

Let us now explain the role of α_k. In Eq. (1), E is computed over all the samples in the feature space irrespective of the size of the classes. Therefore, it is expected that the contribution of a class of bigger size (i.e. with larger number of samples) will be more in the computation of E. As a result, the index value will be more biased by the bigger classes; which might affect the process of feature selection. In order to overcome this i.e., to normalize this effect of the size of the classes, a factor α_k, corresponding to the class C_k, is introduced. In the present investigation, we have chosen $\alpha_k = 1 - \text{prob}_k$. However, other expressions like $\alpha_k = 1/|C_k|$ or $\alpha_k = 1/\text{prob}_k$ could also have been used.

The feature evaluation index (E in Eq. (1)) provides an aggregated measure of compactness of individual classes and separation between different classes. If a particular subset (F_1) of features is more important than another subset (F_2) in characterizing/discriminating the classes/between classes then the value of E computed over F_1 will be less than that computed over F_2. In that case, both individual class compactness and between class separation would be more in the feature space constituted by F_1 than that of F_2. Therefore, the task of feature subset selection boils down to selecting the subset (F) among all possible combinations of a given set (M) of n features for which E is minimum. In the case of individual feature ranking, the subset F contains only one feature.

3. Weighted membership function and feature ranking

It is clear from Eqs. (4)-(7) that the class structures are modeled using a set of predefined membership functions which are kept fixed throughout the computation. Instead of modeling the class structures rigidly, a flexible (adaptive) membership function is defined by introducing a set of weighting coefficients such that the feature space can suitably be transformed depending on these weighting coefficients. The incorporation of weighting factors also makes the method of modeling the class structures more generalized.

The new weighted membership function is expressed by Eq. (4) where $d_k(\mathbf{f}^{(p)})$ is defined as

$$d_k(\mathbf{f}^{(p)}) = \left[\sum_i w_i^2 \left(\frac{f_i^{(p)} - m_{k_i}}{\lambda_{k_i}}\right)^2\right]^{1/2}, \quad w_i \in [0, 1]. \tag{8}$$

Therefore, the membership values (μ) of the sample points of a class become dependent on w_i. $w_i = 1$, for all i, corresponds to Eq. (4). Other values of $w_i(< 1)$ make the function of Eq. (4) flattened along the axis of f_i. The lower the value of w_i, the higher is the extent of flattening. In the extreme case, when $w_i = 0$, for all $i, d_k = 0$ and $\mu_{C_k} = 1$ for all the patterns. Therefore, the incorporation of the weighting factors adds flexibility to the expanse of the modeled class structures. The extent to which the modeled class structures needs to be expanded, depends on the amount of overlap between the adjacent classes. In other words, w_is should be such that both compactness of individual classes and separation between classes increase. This is essentially being guided by the feature evaluation index E based on the weighted distance measure.

In pattern recognition literature, the weight w_i (Eq. (8)) can be viewed to reflect the relative importance of the feature f_i in measuring the similarity (in terms of distance) of a pattern to a class. It is such that the higher the value of w_i, the more is the importance of f_i in characterizing (discriminating) a class (between classes). $w_i = 1\,(0)$ indicates that f_i is most (least) important.

Therefore, the compactness of the individual classes and the separation between the classes as measured by E (Eq. (1)) is now essentially a function of $\mathbf{w}$ $(= [w_i, w_2, \ldots, w_n])$. 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 f_is in characterizing/ discriminating classes. The task of minimization may be performed with various techniques [13,14]. Here, we have adopted gradient descent technique in a connectionist framework (because of its massive parallelism, fault tolerance etc.) for minimizing E. A new connectionist model is developed for this purpose. This is described in Section 4.

Note that, the method of individual feature ranking, explained in Section 2 considers each feature individually independent of others. On the other hand, the method described in this section finds the set of w_is (for which E is minimum) considering the effect of inter-dependencies of the features.

4. Neural network model for feature evaluation

The network (Fig. 1) consists of two layers, namely, input and output. The input layer represents the set of all features in M and the output layer corresponds to the pattern classes. Input nodes accept activations corresponding to the feature values of the input patterns. The output nodes produce the membership values of the input patterns corresponding to the respective pattern classes. With each output node, an auxiliary node is connected which controls the activation of the output node through modulatory links. An output node can be activated from the input layer only when the corresponding auxiliary node remains active. Input nodes are connected to the auxiliary nodes through

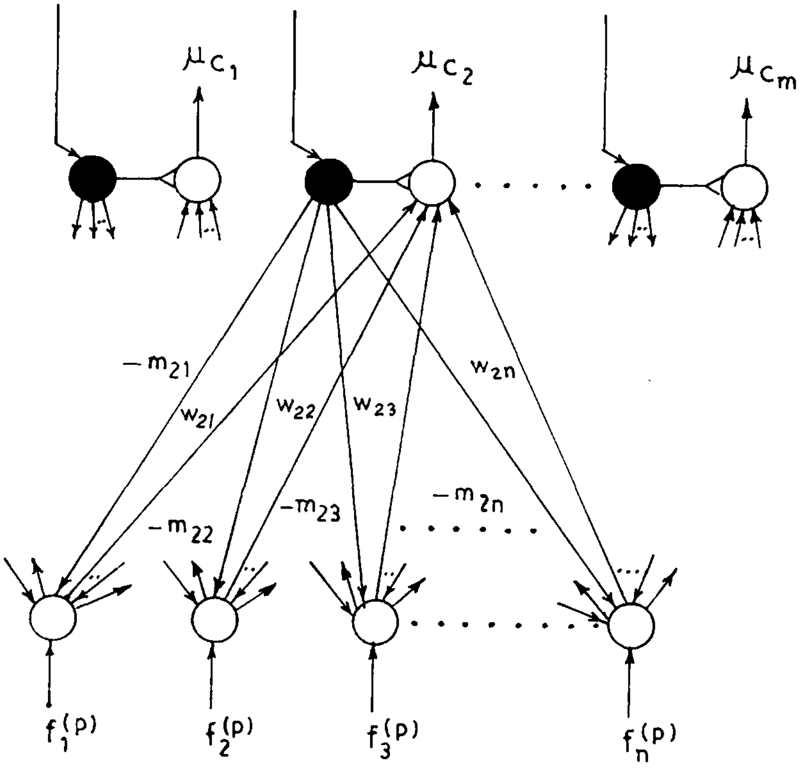

Fig. 1. A schematic diagram of the proposed neural network model. Black circles represent the auxiliary nodes, and white circles represent input and output nodes. Small triangles attached to the output nodes represent the modulatory connections from the respective auxiliary nodes.

feedback links. The weight of the feedback link from the auxiliary node, connected to the kth output node (corresponding to the class C_k), to the ith input node (corresponding to the feature f_i) is equated to $-m_{k_i}$. The weight of the feedforward link from the ith input node to the kth output node provides the degree of importance of the feature f_i, and is given by,

$$W_{k_i} = \left(\frac{w_i}{\lambda_{k_i}}\right)^2. \tag{9}$$

During training, the patterns are presented at the input layer and the membership values are computed at the output layer. The feature evaluation index for these membership values is computed (Eq. (14)) and the values of w_is are updated in order to minimize this index. Note that, λ_{k_i}s and m_{k_i}s are directly computed from the training set and kept fixed during updation of w_is. The auxiliary nodes are activated (i.e., activation values are equated to unity) one at a time while the others are made inactive (i.e., activation values are fixed at 0). Thus, during training, at a time only one output node is allowed to get activated.

When the kth auxiliary node is activated, input node i has an activation value as,

$$u_{ik}^{(p)} = \left(x_{ik}^{(p)}\right)^2, \tag{10}$$

where $x_{ik}^{(p)}$ is the total activation received by the ith input node for the pattern $\mathbf{f}^{(p)}$, when the auxiliary node k is active. $x_{ik}^{(p)}$ is given by,

$$x_{ik}^{(p)} = f_i^{(p)} - m_{k_i}. \tag{11}$$

$f_i^{(p)}$ is the external input (value of the ith feature for the pattern $\mathbf{f}^{(p)}$) and $-m_{k_i}$ is the feedback activation from the kth auxiliary node to the ith input node. The activation value of the kth output node is given by,

$$v_k^{(p)} = g\left(y_k^{(p)}\right), \tag{12}$$

where $g(.)$, the activation function of each output node, is a π-function as given in Eq. (4). $y_k^{(p)}$, the total activation received by the kth output node for the pattern $\mathbf{f}^{(p)}$, is given by

$$y_k^{(p)} = \left(\sum_i u_{ik}^{(p)} \times \left(\frac{w_i}{\lambda_{k_i}}\right)^2\right)^{1/2}. \tag{13}$$

Note that, $y_k^{(p)}$ is the same as d_k (Eq. (8)) for the given input pattern $\mathbf{f}^{(p)}$, and $v_k^{(p)}$ is equal to the membership value of the input pattern $\mathbf{f}^{(p)}$ in the class C_k.

The expression for $E(\mathbf{w})$ (from Eq. (1)), in terms of the output node activations, is given by

$$E(\mathbf{w}) = \sum_{\mathbf{f}^{(p)} \in C_k} \sum_k \frac{v_k^{(p)}\left(1 - v_k^{(p)}\right)}{\sum_{k' \neq k} \frac{1}{2}\left[v_k^{(p)}\left(1 - v_{k'}^{(p)}\right) + v_{k'}^{(p)}\left(1 - v_k^{(p)}\right)\right]} \times \alpha_k. \tag{14}$$

The training phase of the network takes care of the task of minimization of $E(\mathbf{w})$ (Eq. (14)) with respect to $\mathbf{w}$ which is performed using gradient-descent technique. The change in w_i (Δw_i) is computed as,

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

where η is the learning rate.

For the computation of $\partial E/\partial w_i$, the following expressions are used.

$$\frac{\partial s_{kk'}(\mathbf{f}^{(p)})}{\partial w_i} = \frac{1}{2}\left[\left[1 - 2v_{k'}^{(p)}\right]\frac{\partial v_k^{(p)}}{\partial w_i} + \left[1 - 2v_k^{(p)}\right]\frac{\partial v_{k'}^{(p)}}{\partial w_i}\right], \tag{16}$$

$$\frac{\partial s_k(\mathbf{f}^{(p)})}{\partial w_i} = \left[1 - 2v_{k'}^{(p)}\right]\frac{\partial v_k^{(p)}}{\partial w_i}, \tag{17}$$

$$\begin{aligned}\frac{\partial v_k^{(p)}}{\partial w_i} &= -4y_k^{(p)}\frac{\partial y_k^{(p)}}{\partial w_i}, \quad 0 \leqslant y_k^{(p)} < \frac{1}{2},\\ &= -4\left[1 - y_k^{(p)}\right]\frac{\partial y_k^{(p)}}{\partial w_i}, \quad \frac{1}{2} \leqslant y_k^{(p)} < 1,\\ &= 0, \quad \text{otherwise},\end{aligned} \tag{18}$$

and

$$\frac{\partial y_k^{(p)}}{\partial w_i} = \frac{w_i}{y_k^{(p)}}\left(\frac{f_i^{(p)} - m_{k_i}}{\lambda_{k_i}}\right)^2. \tag{19}$$

The steps involved in the training phase of the network are as follows:

- Calculate the mean vectors ($\mathbf{m}_k$) of all the classes from the data set and equate the weight of the feedback link from the auxiliary node corresponding to the class C_k to the input node i as $-m_{k_i}$ (for all i and k).
- Get the values of λ_{k_i} s (bandwidths in Eq. (6)) from the data set and initialize the weight of the feedforward link from ith input node to kth output (for all values of i and k) node.
- For each input pattern:

  Present the pattern vector to the input layer of the network.

  Activate only one auxiliary node at a time.

  Whenever an auxiliary node is activated, it sends the feedback to the input layer. The input nodes in turn send the resultant activations to the output nodes. The activation of the output node (connected to the active auxiliary node) provides the membership value of the input pattern to the corresponding class. Thus, the membership values of the input pattern corresponding to all the classes are computed by sequentially activating the auxiliary nodes one at a time.

  Compute the desired change in w_is to be made using the updating rule given in Eq. (15).
- Compute total change in w_i for each i, over the entire set of patterns. Update w_i (for all i) with the average value of Δw_i.
- Repeat the whole process until convergence, i.e., the change in E becomes less than certain predefined small quantity.

After convergence, $E(\mathbf{w})$ attains a local minima. In that case, the values of w_is indicate the order of importance of the features.

5. Results

The effectiveness of the above-mentioned algorithms was tested on two types of data sets, namely, vowel data [3] and Iris data [15]. 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 30–35 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 [3]. This vowel data is being extensively used for two decades in the area of pattern recognition.

Anderson's Iris data [15] 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, Sepal Width, Petal Length and Petal Width corresponding to f_1, f_2, f_3 and f_4, respectively. Iris data has been used in many research investigation related to pattern recognition and has become a sort of benchmark-data.

5.1. Results obtained using fuzzy feature evaluation index

Table 1 indicates the order of different subsets of features of vowel data based on the values of E (Eq. (1)). This order is also compared with that obtained by Pal et al. [4,5]. Table 1 shows that the subset $\{f_2\}$ is the best and $\{f_1,f_2\}$ is the second best using Eq. (1), while the corresponding order is $\{f_1,f_2\}$ and $\{f_2\}$ in the case of Pal et al. However, in both the methods, the difference in index values for the subsets $\{f_2\}$ and $\{f_1,f_2\}$ is insignificant. f_3 stands at the bottom of the order list, in both the cases. Note also that, the inclusion of f_2 in a subset improves its characterization/discrimination ability. This further justifies the significant importance of f_2 in characterizing vowel classes. These results conform to the earlier findings [3] on speech recognition (from the point of correct rate of classification of vowel sounds).

Table 2 provides the order of different subsets of features of the Iris data. Although, the order obtained using Eq. (1) differs from that obtained with the FEI of Pal et al., like vowel sounds, the successive difference of the index values between these subsets are found to be small. Among the individual flower features, the ranking done by Eq. (1) and FEI of Pal et al. [4,5] is f_4, f_3, f_1, f_2 and f_3, f_4, f_2, f_1, respectively. This shows that f_3 and f_4 are more important than f_1 and f_2. This conforms to the earlier finding using fuzzy set theoretic [16] and neural approaches [17].

The relation of FEI with Mahalanobis distance and divergence measure is graphically depicted in Figs. 3 and 4 (for vowel data), and Figs. 5 and 6 (for

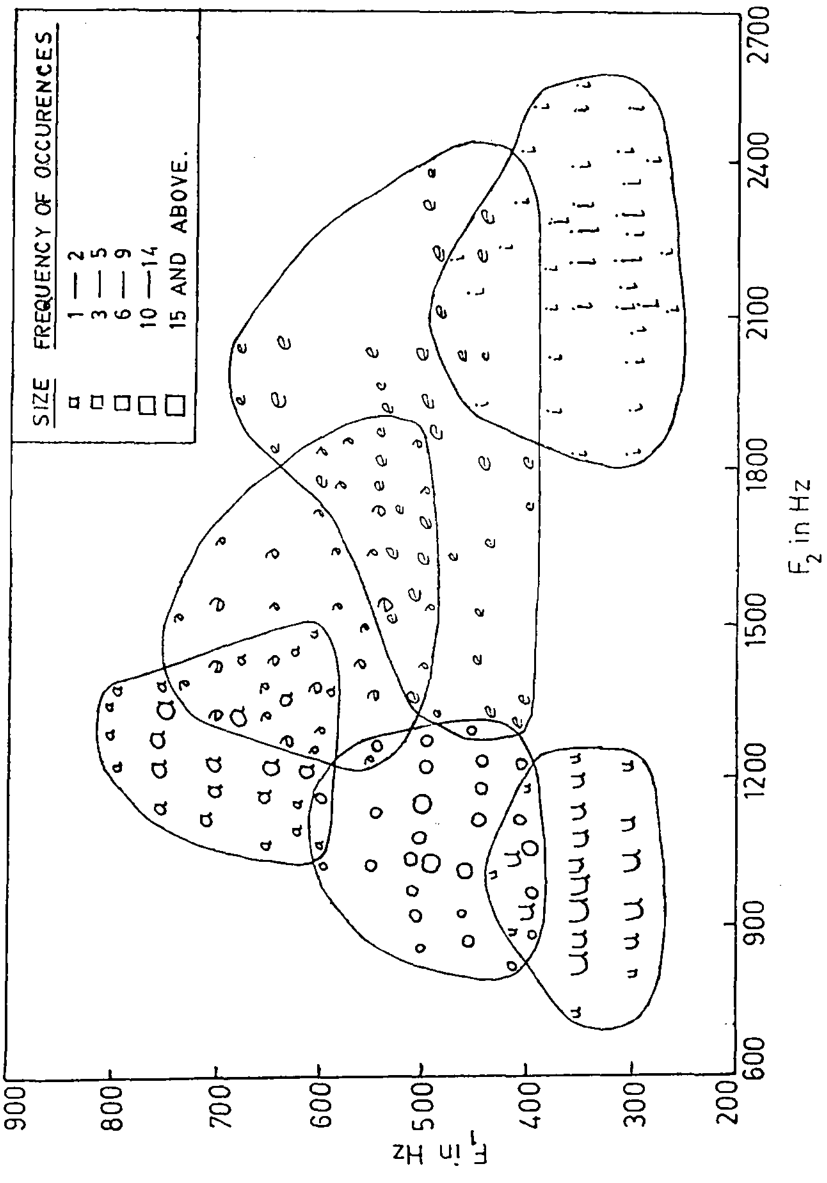

Fig. 2. Two-dimensional ($f_1 - f_2$) plot of the vowel data.

Table 1
Values of FEI for every feature subset of vowel data

Feature subset	Order obtained using	
	Eq. (1)	*FEI* of Pal et al. [4,5]
$\{f_1\}$	5	3
$\{f_2\}$	1	2
$\{f_3\}$	7	6
$\{f_1,f_2\}$	2	1
$\{f_1,f_3\}$	6	7
$\{f_2,f_3\}$	4	4
$\{f_1,f_2,f_3\}$	3	5

Iris data). They are computed over every pair of classes. As expected, Figs. 3–6 show a decrease in feature evaluation index with increase in Mahalanobis distance and divergence measure between the classes.

5.2. *Results obtained with neural network*

Tables 3 and 4 provide the degrees of importance (w) of different features corresponding to the vowel and Iris data respectively, obtained by the neural network method described in Section 4. Three different initializations of $\mathbf{w}$ were used in order to train the network.

Table 2
Values of FEI for every feature subset of Iris data

Feature subset	Order obtained using	
	Eq. (1)	*FEI* of Pal et al. [4,5]
$\{f_1\}$	14	14
$\{f_2\}$	15	13
$\{f_3\}$	3	1
$\{f_4\}$	1	4
$\{f_1,f_2\}$	13	15
$\{f_1,f_3\}$	9	8
$\{f_1,f_4\}$	6	12
$\{f_2,f_3\}$	8	2
$\{f_2,f_4\}$	4	7
$\{f_3,f_4\}$	2	3
$\{f_1,f_2,f_3\}$	12	10
$\{f_1,f_2,f_4\}$	10	11
$\{f_1,f_3,f_4\}$	7	9
$\{f_2,f_3,f_4\}$	5	5
$\{f_1,f_2,f_3,f_4\}$	11	6

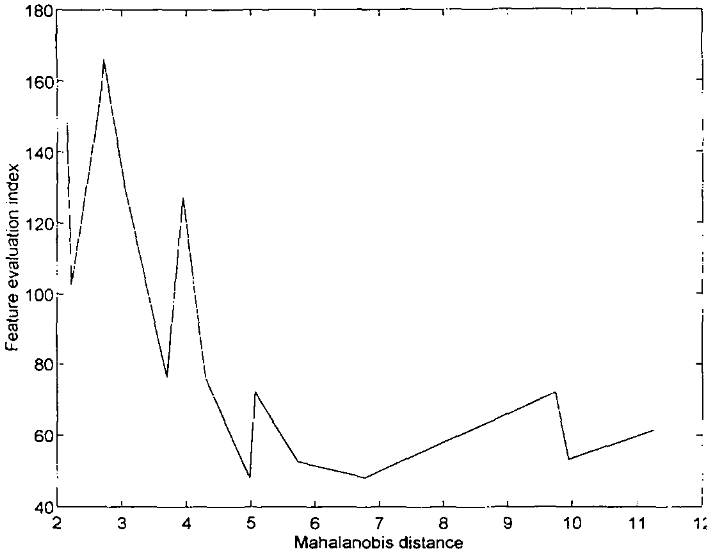

Fig. 3. Graphical representation of the relationship between feature evaluation index and Mahalanobis distance for the vowel data.

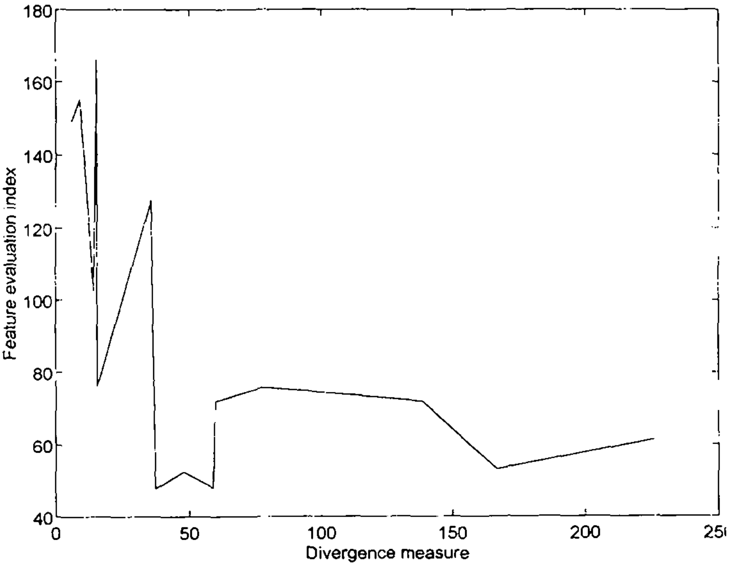

Fig. 4. Graphical representation of the relationship between feature evaluation index and divergence measure for the vowel data.

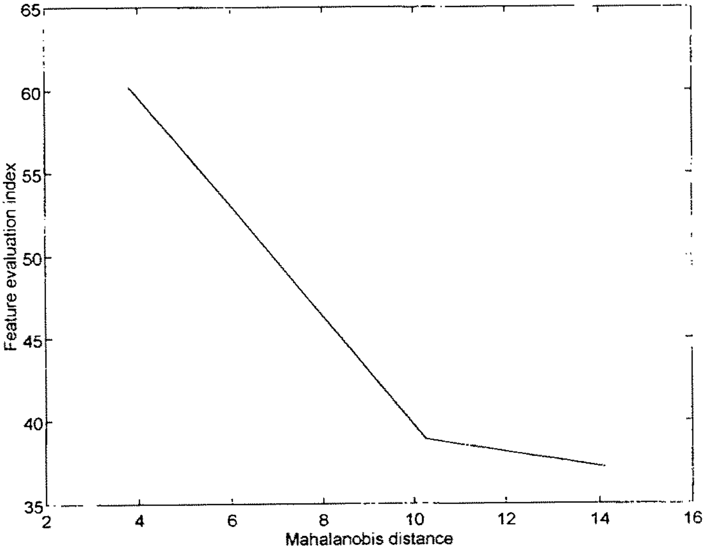

Fig. 5. Graphical representation of the relationship between feature evaluation index and Mahalanobis distance for Iris data.

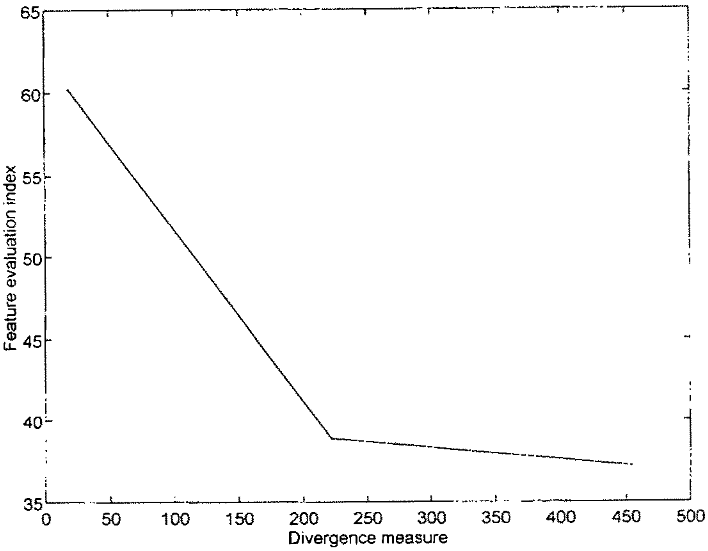

Fig. 6. Graphical representation of the relationship between feature evaluation index and divergence measure for Iris data.

Table 3
Degrees of importance of different features of vowel data

Feature	Initial w					
	$= 1.0$		in $[0,1]$		$= 0.5 \pm \epsilon$	
	w	Rank	w	Rank	w	Rank
f_1	0.001194	3	0.000048	3	0.001037	3
f_2	0.342003	1	0.337536	1	0.342621	1
f_3	0.192297	2	0.001745	2	0.092156	2

Table 4
Values of degrees of importance of different features of Iris data

Feature	Initial w					
	$= 1.0$		in $[0,1]$		$= 0.5 \pm \epsilon$	
	w	Rank	w	Rank	w	Rank
f_1	0.029140	3	0.003230	3	0.029066	3
f_2	0.090552	2	0.102529	2	0.074984	2
f_3	0.320185	1	0.322186	1	0.320367	1
f_4	0.002404	4	0.002027	4	0.002833	4

These are:

(i) $w_i = 1$, for all i, i.e., all the features are considered to be equally most important,

(ii) $w_i \in [0, 1]$, for all i, i.e., the network starts searching for a sub-optimal set of weights from an arbitrary point in the search space, and

(iii) $w_i = 0.5 \pm \epsilon$, for all i, $\epsilon \in [0, 0.01]$. In this case the features are considered to be almost equally but not fully important. Note that, $w_i = 1$ means the feature f_i is most important. That is, its presence is a must for characterizing the pattern classes. Similarly, $w_i = 0$ means f_i has no importance and therefore, its presence in the feature vector is not required. $w_i = 0.5$ indicates an ambiguous situation about such presence of f_i. ϵ adds a small perturbation to the degree of presence/importance.

It is found from Table 3 that the order of features of the vowel data, in all the cases, is f_2, f_3, f_1 whereas it is f_2, f_1, f_3 in Table 1. Similarly, for Iris data (Table 4), the order is seen to be f_3, f_2, f_1, f_4 unlike f_4, f_3, f_1, f_2 in Table 2. This discrepancy may be because of the fact that the neural network based method considers interdependence among the features, whereas, the other method assumes features to be independent of the others. It has been observed experimentally that the network converges much slower with the initial

$w_i = 1$, for all i, as compared to the others. For example, the number of iterations required to converge the network corresponding to the initializations $w_i = 1$, [0,1] and $0.5 \pm \epsilon$ are 152, 49 and 60 for vowel data, and 269, 154 and 134 for the Iris data.

6. Conclusions

In this article, we have presented a new feature evaluation index based on fuzzy set theory and a neuro-fuzzy approach for feature evaluation. The index is defined based on the aggregated measure of compactness of the individual classes and the separation between the classes in terms of class membership functions. The index value decreases with the increase in both the compactness of individual classes and the separation between the classes. Using this index, the best subset from a given set of features can be selected. As Mahalanobis distance and divergence between the classes increase, the feature evaluation index decreases.

The incorporation of feature importance as weighting factors into membership functions gives rise to a transformation of the feature space which provides a generalized framework for modeling class structures. A new connectionist model is designed in order to perform the task of minimizing this index. Note that, this neural network based minimization procedure considers all the features simultaneously, in order to find the relative importance of the features. In other words, the interdependencies of the features have been taken into account. Whereas, the other method (without considering the weighting factors and neural network), considers each feature or subset of features independently.

Results obtained by the feature evaluation index (Eq. (1)) is seen from Tables 1 and 2 to be comparable with that defined in [4,5]. However, in [4,5], the separation between two classes is measured by pooling the classes together, and modeling them with a single membership function. Therefore, for an m-class problem, the number of membership functions required is $m + \binom{m}{2}$; where the first and the second terms correspond to individual class and pairwise class membership functions, respectively. In other words, one needs $m(m + 1)$ parameters for computing the *FEI* [4,5]. On the other hand, for computing the evaluation index of Eq. (1), one needs to compute only m individual class membership functions i.e., $2m$ parameters.

In the neuro-fuzzy approach, the class means and bandwidths are determined directly from the training data (under supervised mode). However, the method may be suitably modified in order to adaptively determine the class means and bandwidths under unsupervised mode so that it can give rise to a versatile self-organizing neural network model for feature evaluation.

Acknowledgements

Mr. Jayanta Basak is grateful to Ms. Basabi Chakraborty for the helpful discussions with her. 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.

References

[1] J.T. Tou, R. Gonzalez, Pattern Recognition Principles, Addison-Wesley, Reading, MA, 1974.

[2] P.A. Devijver, J. Kittler, Pattern Recognition, A Statistical Approach, Prentice-Hall, London, 1982.

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

[4] S.K. Pal, B. Chakraborty, Fuzzy set theoretic measures for automatic feature evaluation, IEEE Trans. on Systems, Man and Cybernetics, vol. SMC-16, no. 5, 1986, pp. 754 760.

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

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

[7] Y.H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley, New York, 1989.

[8] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. on Neural Networks, vol. 5, no. 4, 1994, pp. 537 550.

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

[10] A. Kowalczyk, H.L. Ferra, Developing higher-order neural networks with empirically selected units, IEEE Trans. on Neural Networks, vol. 5, no. 5, 1994, pp. 698–711.

[11] L.M. Belue, K.W. Bauer Jr., Determining input features for multilayer perceptrons, Neurocomputing, vol. 7, no. 2, 1995, pp. 111–121.

[12] S.K. Pal, P.K. Pramanik, Fuzzy measures in determining seed points in clustering, Pattern Recognition Letter, vol. 4, 1986, pp. 159 164.

[13] D.M. Himmelblau, Applied Nonlinear Programming, McGraw-Hill, New York, 1972.

[14] L. Davis (Ed.), Genetic Algorithms and Simulated Annealing, Pitman Publishing, London, 1987.

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

[16] B. Chakraborty, On Some Fuzzy Set Theoretic Measures and Knowledge based Approach for Feature Selection in a Pattern Recognition System, Ph.D. Thesis, Calcutta University, India, 1994.

[17] J.M. Keller, D.J. Hunt, Incorporating fuzzy membership functions into the perceptron algorithm, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-7, no. 6, 1985, pp. 693–699.