Contributed article

Neuro-fuzzy feature evaluation with theoretical analysis

R.K. De, J. Basak, S.K. Pal*

Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700035, India

Received 18 December 1996; received in revised form 5 July 1999; accepted 5 July 1999

Abstract

The article provides a fuzzy set theoretic feature evaluation index and a connectionist model for its evaluation along with their theoretical analysis. A concept of weighted membership function 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. It is shown theoretically that the evaluation index has a fixed upper bound and a varying lower bound, and it monotonically increases with the lower bound. A relation between the evaluation index, interclass distance and weighting coefficients is established. Effectiveness of the algorithms for evaluating features both individually and in a group (considering their independence and dependency) is demonstrated along with comparisons on speech, Iris, medical and mango-leaf data. The results are also validated using scatter diagram and *k*-NN classifier.

Keywords: Fuzzy sets; Neural networks; Pattern recognition; Feature evaluation index; Softcomputing; Weighted membership function

1. Introduction

Feature selection or extraction is a process of selecting a map of the form $\mathbf{x}' = f(\mathbf{x})$ by which a sample $\mathbf{x}(x_1, x_2, \ldots, x_n)$ in an n-dimensional measurement space ($\mathbb{R}^n$) is transformed into a point $\mathbf{x}'(x'_1, x'_2, \ldots, x'_q)$ in a q-dimensional ($q < n$) feature space ($\mathbb{R}^q$). The main objective of this problem 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.

In general, in the feature selection/extraction process, the features considered to have optimal saliencies (usefulness) are that for which interclass/intraclass distances are maximized/minimized. 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 important in discriminating between patterns of different types. Different useful classical techniques to achieve this are based on diagonal transformation, Mahalanobis distance, divergence, Bhattacharya coefficient, and the Kolomogorov variational distance (Devijver & Kittler, 1982; Tou & Gonzalez, 1974).

There also exist several methods based on fuzzy set theory (Bezdek, 1981; Bezdek & Castelaz, 1977; Pal, 1992; Pal & Chakraborty, 1986) and Artificial neural networks (ANN) (Belue & Bauer, 1995; Kowalczyk & Ferra, 1994; Kraaijveld, Mao & Jain, 1995; Lampinen & Oja, 1995; Lowe & Webb, 1991; Mao & Jain, 1995; Priddy, Rogers, Ruck, Rogers & Kabrisky, 1990; Ruck, Tarr & Kabrisky, 1993; Saund, 1989; Schmidt & Davis, 1993). Fuzzy set theoretic approaches for feature selection are mainly based on measures of entropy and index of fuzziness (Pal, 1992; Pal & Chakraborty, 1986), fuzzy *c*-means (Bezdek, 1981) and fuzzy ISODATA (Bezdek & Castelaz, 1977) algorithms, etc. Some of the recent attempts made for feature selection/extraction in the framework of ANN are mainly based on multilayer feedforward networks (Belue & Bauer, 1995; Kowalczyk & Ferra, 1994; Lowe & Webb, 1991; Mao & Jain, 1995; Priddy et al., 1993; Ruck et al., 1990; Saund, 1989; Schmidt & Davis, 1993) and self-organizing networks (Kraaijveld et al., 1995; Lampinen & Oja, 1995; Mao & Jain, 1995). The methods based on multilayer feedforward networks include, among others, determination of saliency of input features (Priddy et al., 1993), development of Sammon's nonlinear discriminant analysis (NDA) network, linear discriminant analysis (LDA) network (Mao & Jain, 1995), whereas those based on self-organizing networks include development of nonlinear projection (NP-SOM) based Kohonen's self-organizing feature map (Mao & Jain, 1995), distortion tolerant Gabor transformations followed by minimum distortion clustering by multilayer self-organizing maps (Lampinen & Oja, 1995), a non-linear projection method based on Kohonen's topology preserving maps (Kraaijveld et al., 1995).

Incorporation of fuzzy set theory enables one to deal with uncertainties in a system, arising from deficiency (e.g. vagueness, incompleteness, etc.) in information, 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 provides a neuro-fuzzy approach for feature evaluation and a theoretical analysis of its performance. First of all, a new fuzzy set theoretic evaluation index is defined in terms of individual class membership. Its performance with an existing one (Pal, 1992; Pal & Chakraborty, 1986) is compared for ranking the features (or subsets of features). Its relation with Mahalanobis distance and divergence measure is demonstrated. Then, we provide a new connectionist model to perform the task of optimizing 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 modeling better the class structures. Finally, the performance of the system is theoretically analyzed. This includes derivation of upper and lower bounds of the evaluation index, and determining its relation with interclass distance and weighting coefficient. The effectiveness of the algorithms, along with extensive comparisons, is demonstrated on four different data sets, namely, three-dimensional 6-class vowel data, four-dimensional 3-class Iris data, nine-dimensional 4-class medical data and 18-dimensional 3-class mango-leaf data. The validity of the experimental results is analyzed independently with scatter plots and k-NN classifier for different values of k.

The article is organized as follows. Section 2 provides the description of a new feature evaluation index and weighted membership function. Section 3 describes the connectionist model for the evaluation of the feature evaluation index. Theoretical analysis of the feature evaluation index is provided in Section 4. The effectiveness of the methods is established with experimental results in Section 5. Finally, the paper is concluded in Section 6.

2. Fuzzy feature evaluation index and weighted membership function

Let us consider an n-dimensional feature space Ω containing $x_1, x_2, x_3, \ldots, x_i, \ldots, x_n$ features (components). Let there be M classes $C_1, C_2, C_3, \ldots, C_k, \ldots, C_M$. The feature evaluation index for a subset ($\Omega_{\mathbf{x}}$) containing few of these n features is defined as

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

where $\mathbf{x}$ is constituted by the features of $\Omega_{\mathbf{x}}$ only and

$$s_k(\mathbf{x}) = \mu_{C_k}(\mathbf{x}) \times (1 - \mu_{C_k}(\mathbf{x})), \tag{2}$$

$$s_{kk'}(\mathbf{x}) = \tfrac{1}{2}[\mu_{C_k}(\mathbf{x}) \times (1 - \mu_{C_{k'}}(\mathbf{x}))] + \tfrac{1}{2}[\mu_{C_{k'}}(\mathbf{x}) \times (1 - \mu_{C_k}(\mathbf{x}))], \tag{3}$$

with $\mu_{C_k}(\mathbf{x})$ and $\mu_{C_{k'}}(\mathbf{x})$ being the membership values of the pattern $\mathbf{x}$ in classes C_k and $C_{k'}$, respectively. Here 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 $\mathbf{x}$ 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 for 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). The value of 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. Here 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 $\alpha_k = 1 - P_k$, where P_k is a priori probability for class C_k. However, other expressions like $\alpha_k = (1/|C_k|)$ or $\alpha_k = (1/P_k)$ could also have been used.

The membership ($\mu_{C_k}(\mathbf{x})$) of a pattern $\mathbf{x}$ to a class C_k is defined with a multi-dimensional π-function (Pal & Pramanik, 1986) which is given by,

$$\mu_{C_k}(\mathbf{x}) = \begin{cases} 1 - 2d_k^2(\mathbf{x}) & 0 \leq d_k(\mathbf{x}) < \frac{1}{2}, \\ 2[1 - d_k(\mathbf{x})]^2 & \frac{1}{2} \leq d_k(\mathbf{x}) < 1, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where $d_k(\mathbf{x})$ is the distance of the pattern $\mathbf{x}$ from $\mathbf{m}_k$ (the

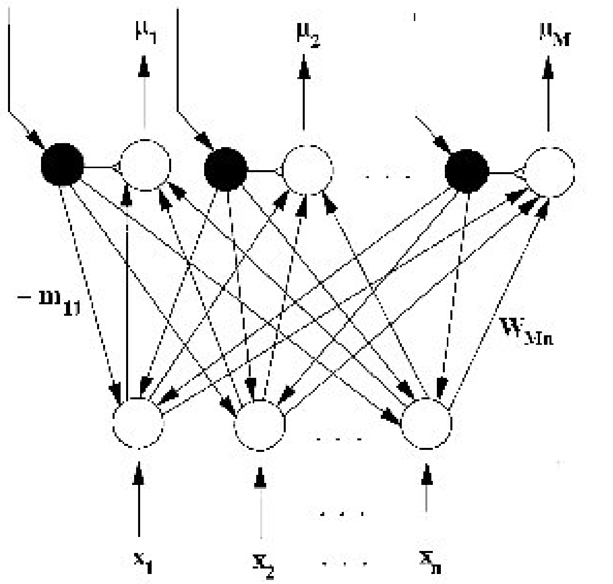

Fig. 1. A schematic diagram of the proposed neural network model. Black circles represent the auxiliary nodes, and white circles represent input and output nodes. Small triangles attached to the output nodes represent the modulatory connections from the respective auxiliary nodes.

center of class C_k). It can be defined as,

$$d_k(\mathbf{x}) = \left[\sum_i \left(\frac{x_i - m_{ki}}{\lambda_{ki}} \right)^{r_k} \right]^{1/r_k}, \quad r_k > 0, \tag{5}$$

where

$$\lambda_{ki} = 2 \max_{\mathbf{x} \in C_k} [|x_i - m_{ki}|], \tag{6}$$

and

$$m_{ki} = \frac{\sum_{\mathbf{x} \in C_k} x_i}{|C_k|}. \tag{7}$$

Eqs. (4)–(7) are such that the membership $\mu_{C_k}(\mathbf{x})$ of a pattern $\mathbf{x}$ 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.

In practice, the class structure may not be symmetric. In that case, the membership values of some patterns at the boundary of the class will be greater than 0.5. Also, some patterns of other classes may have membership values greater than 0.5 for the class under consideration. For handling this undesirable situation, the membership function corresponding to a class needs to be transformed so that it can model the real life class structures appropriately. For this purpose, we have incorporated a weighting factor corresponding to a feature, which transforms the feature space in such a way that the transformed membership functions model the class structures appropriately. Note that, this incorporation of weighting factors makes the method of modeling the class structures more generalized; a symmetric class structure being a special case.

For this purpose, we define weighted distance from Eq. (5) as

$$d_k(\mathbf{x}) = \left[\sum_i w_i^{r_k} \left(\frac{x_i - m_{ki}}{\lambda_{ki}} \right)^{r_k} \right]^{1/r_k}, \quad w_i \in [0, 1]. \tag{8}$$

The membership values (μ) of the sample points of a class become dependent on w_i. The values of w_i (< 1) make the function of Eq. (4) flattened along the axis of x_i. The lower the value of w_i, the higher is the extent of flattening. In the extreme case, when $w_i = 0$, $d_k = 0$ and $\mu_{C_k} = 1$ for all the patterns.

In pattern recognition literature, the weight w_i (Eq. (8)) can be viewed to reflect the relative importance of the feature x_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 x_i in characterizing/discriminating a class/between classes. $w_i = 1(0)$ indicates that x_i is the 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 $\mathbf{w}$ ($= [w_1, w_2, \ldots, w_n]$), if we consider all the n features together. The problem of feature selection/ranking thus reduces to finding a set of w_is for which E becomes minimum; w_is indicate the relative importance of x_is in characterizing/discriminating classes. The task of minimization may be performed with various techniques (Davis, 1987; Himmelblau, 1972). 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 the next section.

Note that, the method of individual feature ranking is not identical to that described in this section. The later one finds the set of w_is (for which E is minimum) considering the effect of inter-dependencies of the features, whereas in the case of former one, each feature is considered individually independent of other.

3. Neural network model for fuzzy 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 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 x_i) is equated to $-m_{ki}$. The weight of the feedforward link from the ith input node to the kth output node provides the degree of importance of the feature x_i, and is given by

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

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. (4)) and the values of w_is are updated in order to minimize this index. Note that, λ_{ki}s and m_{ki}s are directly computed from the training set and kept fixed during updating 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. the activation values are fixed at 0). Thus during training, at a time, only one output node is allowed to be activated.

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

$$u_{ik} = (I_{ik})^{r_k}, \tag{10}$$

where I_{ik} is the total activation received by the ith input node for the pattern $\mathbf{x}$, when the auxiliary node k is active, which is given by

$$I_{ik} = x_i - m_{ki}, \tag{11}$$

with x_i being the external input (value of the ith feature for the pattern $\mathbf{x}$) and $-m_{ki}$ the feedback activation from the kth auxiliary node to the ith input node. The activation value of the kth output node is given by

$$v_k = g(y_k), \tag{12}$$

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

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

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

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

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

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 simple gradient-descent technique. The change in w_i (Δw_i) is computed as

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}, \ \forall_i, \tag{15}$$

where η is the learning rate.

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

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

$$\frac{\partial s_k(\mathbf{x})}{\partial w_i} = [1-2v_k]\frac{\partial v_k}{\partial w_i}, \tag{17}$$

$$\frac{\partial v_k}{\partial w_i} = \begin{cases} -4y_k \dfrac{\partial y_k}{\partial w_i}, & 0 \le y_k < \frac{1}{2}, \\ -4[1-y_k]\dfrac{\partial y_k}{\partial w_i}, & \frac{1}{2} \le y_k < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{18}$$

and

$$\frac{\partial y_k}{\partial w_i} = \left(\frac{w_i}{y_k}\right)^{r_k-1}\left(\frac{x_i - m_{ki}}{\lambda_{ki}}\right)^{r_k}. \tag{19}$$

Alternately, we can also express E as a function of W_{ki}, where $W_{ki} = (w_i/\lambda_{ki})^{r_k}$, and then minimize E with respect to W_{ki}. In this case, during training phase, the values of W_{ki}s can be updated using the same gradient-descent technique. After training, the degree of importance of ith feature can be computed as $w_i = W_{ki}^{1/r_k} \times \lambda_{ki}$.

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. Set the weight of the feedback link from the auxiliary node corresponding to the class C_k to the input node i as $-m_{ki}$ (for all i and k).
- Compute λ_{ki}s from Eq. (6) and initialize the weight of the feedforward link from ith input node to kth output (for all values of i and k) node. Set the values of r_ks (in Eq. (8)) so that the membership values of all the patterns of the kth class are at least 0.5 for that class.
- 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 weights of the

feedforward links 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 minimum. In that case, the weights of the feedforward links indicate the order of importance of the features. In the following section, the convergence of E is theoretically established, and the validity of the ordering of features in terms of network parameters is demonstrated for some well-defined class structures.

4. Theoretical analysis

Here, we analyze mathematically the characteristics of the feature evaluation index (E) and the significance of weighting coefficients (w_i). For this purpose we proceed as follows.

- A fixed upper bound and a varying lower bound of E (Eq. (1)) are derived. The variation of E with respect to the lower bound is studied.
- A relation between E, w_i and interclass distance is derived.

4.1. Upper bound and lower bound of E

We can write E (Eq. (1)) as

$$E = \sum_k \sum_{\mathbf{x}\in C_k} \frac{\mu_k \times (1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]} \quad (20)$$

where $\mu_k = \mu_{C_k}(\mathbf{x})$ and $\mu_{k'} = \mu_{C_{k'}}(\mathbf{x})$. Let, $E = \sum_k E_k = \sum_k \sum_{\mathbf{x}\in C_k} E_k(\mathbf{x}|\mathbf{x}\in C_k)$ where

$$E_k = \sum_{\mathbf{x}\in C_k} \frac{\mu_k \times (1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]} \quad (21)$$

and

$$E_k(\mathbf{x}|\mathbf{x}\in C_k) = \frac{\mu_k \times (1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]}. \quad (22)$$

That is, E_k is the value of the evaluation index corresponding to a class C_k, and $E_k(\mathbf{x}|\mathbf{x}\in C_k)$ is contribution of a pattern $\mathbf{x}$ in class C_k, to E_k.

For a pattern $\mathbf{x}$ in class C_k,

$$\frac{1}{2}\sum_{k'\neq k}[\mu_k(1-\mu_{k'}) + \mu_{k'}(1-\mu_k)]$$

$$= \frac{1}{2}\sum_{k'\neq k}[\mu_k(1-\mu_k) + (\mu_k-\mu_{k'})^2 + \mu_{k'}(1-\mu_{k'})].$$

Since $[(\mu_k-\mu_{k'})^2 + \mu_{k'}(1-\mu_{k'})] \geq 0$,

$$\frac{1}{2}\sum_{k'\neq k}[\mu_k(1-\mu_{k'}) + \mu_{k'}(1-\mu_k)] \geq \frac{M-1}{2}\mu_k(1-\mu_k),$$

where M is the number of classes. Since, $0 < \alpha_k < 1$, we can write

$$E_k(\mathbf{x}|\mathbf{x}\in C_k) \leq \frac{2}{(M-1)}. \quad (23)$$

Therefore,

$$\mathscr{E}(E) \leq \frac{2M}{M-1}, \quad (24)$$

where $\mathscr{E}$ denotes the 'mathematical expectation' operator. ♣

Again, for a pattern $\mathbf{x}$ in class C_k, $\mu_k, \mu_{k'} \in [0,1]$, we can write

$$\tfrac{1}{2}[\mu_k(1-\mu_{k'}) + \mu_{k'}(1-\mu_k)] \leq \tfrac{1}{2},$$

$$\sum_{k'\neq k}\frac{1}{2}[\mu_k(1-\mu_{k'}) + \mu_{k'}(1-\mu_k)] \leq \frac{1}{2}(M-1),$$

$$\frac{1}{\sum_{k'\neq k}\frac{1}{2}[\mu_k(1-\mu_{k'}) + \mu_{k'}(1-\mu_k)]} \geq \frac{2}{(M-1)},$$

$$\sum_k \frac{\mu_k(1-\mu_k)\alpha_k}{\sum_{k'\neq k}\frac{1}{2}[\mu_k(1-\mu_{k'}) + \mu_{k'}(1-\mu_k)]}$$

$$\geq \frac{2}{(M-1)}\sum_k \mu_k(1-\mu_k)\alpha_k.$$

Thus

$$E_k(\mathbf{x}|\mathbf{x}\in C_k) \geq \frac{2}{(M-1)}\mu_k(1-\mu_k)\alpha_k.$$

That is

$$\mathscr{E}(E) \geq \frac{2}{(M-1)}\mathscr{E}(\sum_k \mu_k(1-\mu_k)\alpha_k). \quad (25)$$

Therefore,

$$\frac{2}{(M-1)}\mathscr{E}(\sum_k \mu_k(1-\mu_k)\alpha_k) \leq \mathscr{E}(E) \leq \frac{2M}{(M-1)}. \quad (26)$$

Note that, the upper bound of $\mathscr{E}(E)$ is fixed, whereas the lower bound is varying with $[2/(M-1)]\mathscr{E}(\sum_k \mu_k(1-\mu_k)\alpha_k)$. ♣

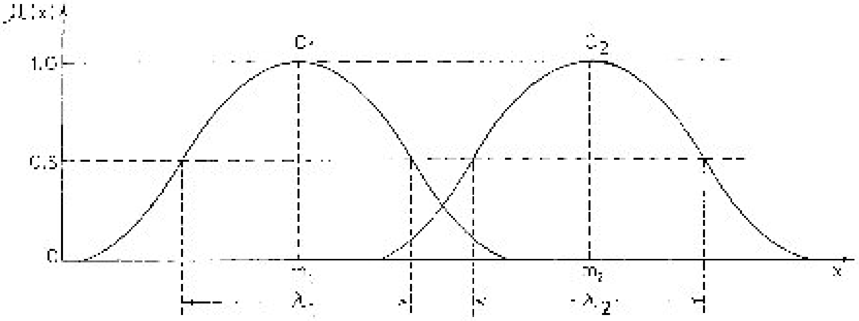

Fig. 2. Non-overlapping pattern classes modeled with π-function.

Let us now analyze the behavior of $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ with respect to $\mu_k(1-\mu_k)$. For this purpose, we substitute $\mu_k(1-\mu_k)$ by h_k in Eq. (22). In that case,

$$\frac{\mathrm{d}E_k(\mathbf{x}|\mathbf{x} \in C_k)}{\mathrm{d}h_k} = \frac{\alpha_k[\sum_{k' \neq k}[\mu_{k'}(1-\mu_k)+\mu_k(1-\mu_{k'})](1-2\mu_k) - \mu_k(1-\mu_k)\sum_{k' \neq k}(1-2\mu_{k'})]}{\frac{1}{2}[\sum_{k' \neq k}[\mu_{k'}(1-\mu_k)+\mu_k(1-\mu_{k'})]]^2(1-2\mu_k)}$$

$$= \frac{\nu_k \alpha_k}{\frac{1}{2}[\sum_{k' \neq k}[\mu_{k'}(1-\mu_k)+\mu_k(1-\mu_{k'})]]^2}, \tag{27}$$

where

$$\nu_k = \frac{\sum_{k' \neq k}[\mu_{k'}(1-\mu_k)+\mu_k(1-\mu_{k'})](1-2\mu_k) - \mu_k(1-\mu_k)\sum_{k' \neq k}(1-2\mu_{k'})}{(1-2\mu_k)}. \tag{28}$$

It is clear from Eq. (27) that $(\mathrm{d}E_k(\mathbf{x}|\mathbf{x} \in C_k)/\mathrm{d}h_k)$ is positive/negative if ν_k is positive/negative. In other words, $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ increases/decreases monotonically with $\mu_k(1-\mu_k)$ if ν_k is positive/negative. Simplifying the expression on the right-hand side of Eq. (28) we get

$$\nu_k = \sum_{k' \neq k}\mu_{k'} - \frac{\mu_k^2 \sum_{k' \neq k}(1-2\mu_{k'})}{(1-2\mu_k)}. \tag{29}$$

In order to show that $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ monotonically increases with $\mu_k(1-\mu_k)$ for both *non-overlapping* and *overlapping* class structures, we consider the following cases.

Case 1 (Non-overlapping (Fig. 2)). Here, for a pattern $\mathbf{x}$, if $|x_i - m_{ki}| \leq (\lambda_{ki}/2)$ holds for all values of i, $\mu_k \geq 0.5$ and $\mu_{k'} < 0.5$, $\forall k' \neq k$. Therefore, $\nu_k > 0$ (Eq. (29)), and as a result $(\mathrm{d}E_k(\mathbf{x}|\mathbf{x} \in C_k)/\mathrm{d}h_k) > 0$. This indicates $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ is monotonically increasing with $\mu_k(1-\mu_k)$.

Case 2 (Overlapping (Fig. 3)). In this case, for a pattern $\mathbf{x}$, if $|x_i - m_{ki}| \leq (\lambda_{ki}/2)$ holds for all values of i, $\mu_k \geq 0.5$ and $\mu_{k'} \leqq 0.5$, $\forall k' \neq k$. Since the classes are overlapped, we consider two different possibilities: $\mathbf{x}$ lying outside the

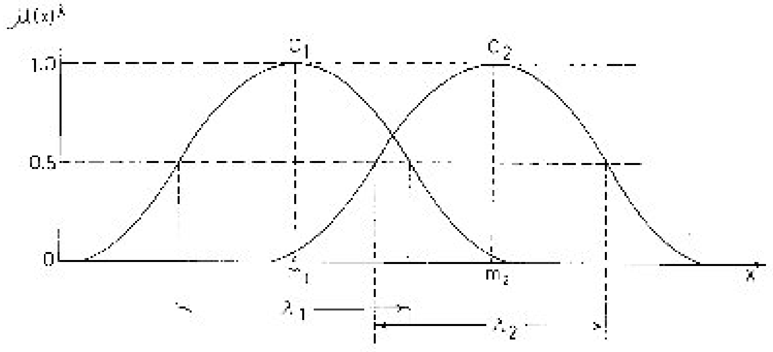

Fig. 3. Overlapping pattern classes modeled with π-function.

overlapping zone (i.e. $|x_i - m_{ki}| \le (\lambda_{ki}/2), \forall i$ and $|x_i - m_{k'i}| > (\lambda_{k'i}/2)$) and $\mathbf{x}$ lying within the overlapping zone (i.e. $|x_i - m_{ki}| < \lambda_{ki}/2, \forall i$ and $|x_i - m_{k'i}| < (\lambda_{k'i}/2), \forall i$).

If the pattern $\mathbf{x}$ lies outside the overlapping zone, then $\mu_{k'} < 0.5$ and thereby $\nu_k > 0$ (Eq. (29)). This indicates $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ monotonically increases with $\mu_k(1 - \mu_k)$.

If $\mathbf{x}$ lies within the overlapping zone, both $\mu_k, \mu_{k'} > 0.5$. Then we have three possibilities: (a) $\mu_k > \mu_{k'}$; (b) $\mu_k \approx \mu_{k'}$; and (c) $\mu_k < \mu_{k'}$.

(a) $\mu_k > \mu_{k'}$. Let $\mu_{k'} = \mu_k - \epsilon_{kk'}$ where $\epsilon_{kk'} > 0$. Therefore, from Eq. (29) we get

$$\nu_k = \sum_{k' \neq k} (\mu_k - \epsilon_{kk'}) - \frac{\mu_k^2 \sum_{k' \neq k} (1 - 2\mu_k + 2\epsilon_{kk'})}{(1 - 2\mu_k)}, \qquad (30)$$

i.e.

$$\nu_k = (M-1)\mu_k - \sum_{k' \neq k} \epsilon_{kk'} - \frac{2\mu_k^2 \sum_{k' \neq k} \epsilon_{kk'} - \mu_k^2(2\mu_k - 1)(M-1)}{(1 - 2\mu_k)}. \qquad (31)$$

Thus, $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ increases monotonically with $\mu_k(1 - \mu_k)$ if

$$(M-1)\mu_k - \sum_{k' \neq k} \epsilon_{kk'} - \frac{2\mu_k^2 \sum_{k' \neq k} \epsilon_{kk'} - \mu_k^2(2\mu_k - 1)(M-1)}{(1 - 2\mu_k)} > 0, \qquad (32)$$

i.e. if

$$\frac{1}{M-1} \sum_{k' \neq k} \epsilon_{kk'} > -\frac{\mu_k(1-\mu_k)(2\mu_k - 1)}{(1-\mu_k)^2 + \mu_k^2} \qquad (33)$$

Since, $\epsilon_{kk'} > 0$, the above inequality always holds, and therefore, in such cases, $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ always increases monotonically with $\mu_k(1 - \mu_k)$.

(b) $\mu_k \approx \mu_{k'}$. In this case, $\epsilon_{kk'} \approx 0$, and therefore, inequality (33) always holds. Thus, in this case also, we get a monotonic increasing nature of $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ with respect to $\mu_k(1 - \mu_k)$.

(c) $\mu_k < \mu_{k'}$. In this case, $\epsilon_{kk'} < 0$. Let us replace $\epsilon_{kk'}$ by $-\epsilon_{kk'}$, i.e. $\mu_{k'} = \mu_k + \epsilon_{kk'}$. Then, the condition for $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ being monotonically increasing function with respect to $\mu_k(1 - \mu_k)$ becomes

$$\frac{1}{M-1} \sum_{k' \neq k} \epsilon_{kk'} < \frac{\mu_k(1-\mu_k)(2\mu_k - 1)}{(1-\mu_k)^2 + \mu_k^2}. \qquad (34)$$

This condition provides an upper bound on the average value of $\epsilon_{kk'}$ (hence on the average value of $\mu_{k'}$) that can be allowed in order to get a monotonic increasing behavior of $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ with respect to $\mu_k(1 - \mu_k)$.

First of all, the chance of $\mu_k < \mu_{k'}$ is low for a pattern in class C_k. Even if this happens (say, for overlapping case), the chance of happening

$$\frac{1}{M-1} \sum_{k' \neq k} \epsilon_{kk'} > (\mu_k(1-\mu_k)(2\mu_k - 1))/((1-\mu_k)^2 + \mu_k^2)$$

is very low (as illustrated in the following two examples). Therefore, $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ is most likely monotonically increasing with $\mu_k(1 - \mu_k)$.

Example 1. Let, $\mu_1 = 0.6$ for a pattern $\mathbf{x}$ lying within the region $\|\mathbf{x} - \mathbf{m}_1\| < \lambda_1/2$ in class C_1. Then, the condition (34) becomes

$$\frac{1}{M-1} \sum_{k' \neq k} \epsilon_{kk'} < 0.1.$$

In order to violate this condition, the average membership value of $\mathbf{x}$ (say, μ_2) to classes other than C_1 should be at least 0.7. It can also be seen that whatever be the value of μ_1 (> 0.5), the value of μ_2 should be greater than μ_1. This is unusual. Thus, we can say that in this case the inequality (34) will be satisfied and thereby, we can expect a monotonic increasing behavior of $E_1(\mathbf{x}|\mathbf{x} \in C_1)$ with respect to $\mu_1(1 - \mu_1)$.

Example 2. Let, $\mu_1 = 0.5$. In that case, condition (34) becomes

$$\frac{1}{M-1} \sum_{k' \neq k} \epsilon_{kk'} < 0.$$

That is, the average membership value of $\mathbf{x}$ to classes other than C_1 should be greater than or equal to 0.5. This situation occurs when the classes are highly overlapped. In other words, if there is high amount of overlap, the behavior of $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ becomes unpredictable for ambiguous patterns. ♣

Thus, we can say that almost in all the cases, $E_k(\mathbf{x}|\mathbf{x} \in C_k)$ is monotonically increasing with $\mu_k(1 - \mu_k)$. Therefore, we can expect that $E_k (= \sum_{\mathbf{x} \in C_k} E_k(\mathbf{x}|\mathbf{x} \in C_k))$ increases monotonically with $\sum_k \mu_k(1 - \mu_k)$. In other words, almost in all the cases $\mathscr{E}(E)$ is a monotonically increasing function of $\mathscr{E}(\sum_k (\mu_k(1 - \mu_k)\alpha_k)$, as α_ks are positive constants.

4.2. Relation between E, interclass distance and w_i

Let us now derive a relation of the lower bound of $\mathscr{E}(E)$ with interclass distance and weighting coefficients for some well-defined class structures.

- Let us assume that the classes $C_1, C_2, \ldots, C_k, \ldots, C_M$ have independent, identical Gaussian distributions with

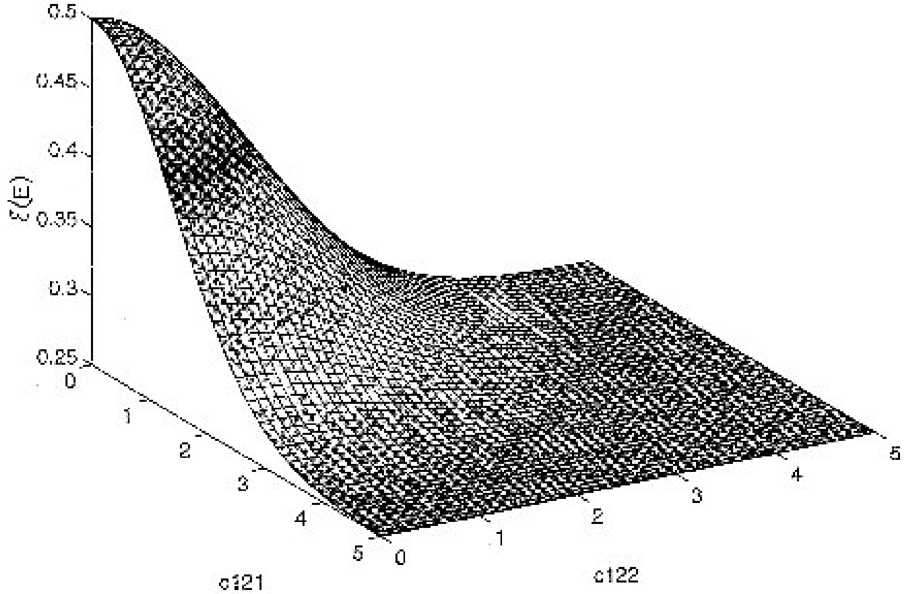

Fig. 4. Graphical representation of $\mathscr{E}(E)$ with respect to c_{121} and c_{122} with $w_1 = w_2 = 1.0$.

respective means $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_k, \ldots, \mathbf{m}_M$ and with the same variance σ^2. Let $\wp(\mathbf{x}|C_k)$ be the class-conditional probability density function for class C_k. Then

$$\wp(\mathbf{x}|C_k) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\sum_i \frac{(x_i - m_{ki})^2}{2\sigma^2}\right) \tag{35}$$

- Let the membership of a pattern $\mathbf{x}$ in a class C_k be given by

$$\mu_k = \mu_k(\mathbf{x}) = \exp\left(-\sum_i \frac{(x_i - m_{ki})^2 w_i^2}{2\lambda^2}\right) \tag{36}$$

where λ is the bandwidth of the class C_k, and is the same for all the classes.

$\mathscr{E}(E)$ is given by

$$\mathscr{E}(E) = \int_{\mathbf{x}} E\wp(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{37}$$

where

$$\wp(\mathbf{x}) = \sum_k P_k \wp(\mathbf{x}|C_k); \tag{38}$$

with P_k being a priori probability of class C_k. Evaluating the right-hand side of Eq. (37) (see Appendix A), we have

$$\mathscr{E}(E) \approx \sum_k \frac{\alpha_k P_k}{M-1}\frac{\sum_i w_i^2}{2\rho^2} \times \left(1 + \sum_{k' \neq k} \exp\left[-\sum_i \frac{c_{kk'i}^2}{2\sigma^2\left(1 + \frac{\rho^2}{w_i^2}\right)}\right]\Bigg]\right), \tag{39}$$

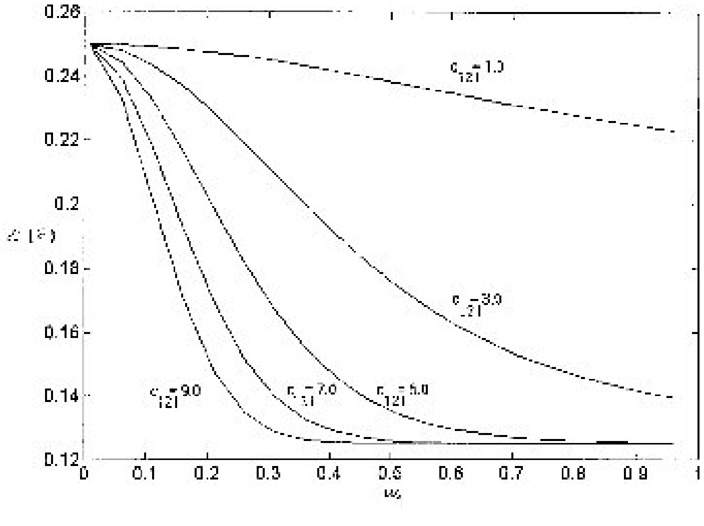

Fig. 5. Graphical representation of $\mathscr{E}(E)$ with respect to w_1 for different values of c_{121}, with $c_{122} = 0$ and $\sum_{i=1}^{2} w_i^2 = 1$.

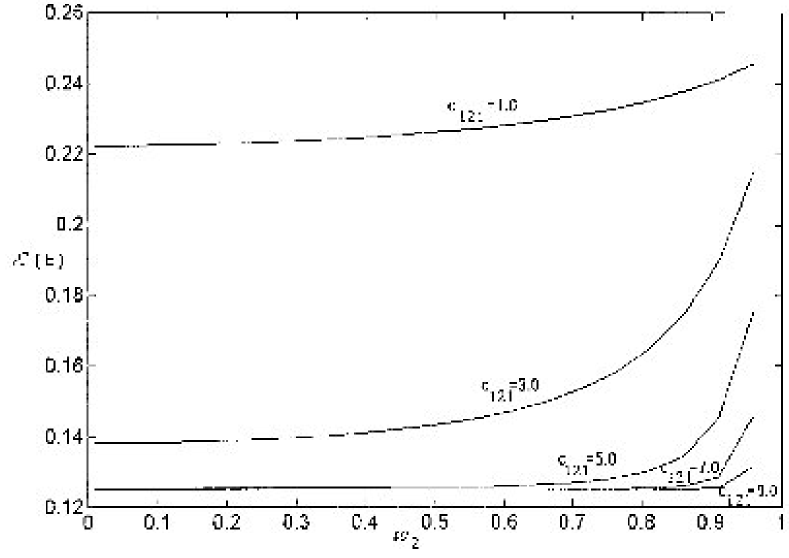

Fig. 6. Graphical representation of $\mathscr{E}(E)$ with respect to w_2 for different values of c_{121}, with $c_{122} = 0$ and $\sum_{i=1}^{2} w_i^2 = 1$.

where $\rho = (\lambda/\sigma)$ and $c_{kk'i} = m_{ki} - m_{k'i}$ is a measure of interclass distance between the classes C_k and $C_{k'}$ along the feature axis x_i.

Let us consider two classes C_1 and C_2, with two features x_1 and x_2. Let, C_1 and C_2 have unit normal distribution, i.e. $\sigma = 1.0$. Let, $\lambda = 1.0$ and $P_k = \alpha_k = 0.5$ $(\forall k)$. c_{121} and c_{122} are the interclass distances between class C_1 and class C_2 along the feature axes x_1 and x_2, respectively. We now demonstrate graphically the variation of $\mathscr{E}$ (E) with respect to c_{121} and c_{122}, and w_1 and w_2.

Fig. 4 shows the variation of $\mathscr{E}$ (E) with respect to c_{121} and c_{122} with $w_1 = w_2 = 1$. $\mathscr{E}(E)$ is maximum when $c_{121} = c_{122} = 0$, i.e. when the two classes completely overlap. Here $\mathscr{E}(E)$ decreases with the increase in c_{121} and c_{122}. This variation is symmetric with respect to both c_{121} and c_{122}. The rate of decrease in $\mathscr{E}(E)$ also decreases as c_{121} (and c_{122}) increases. Finally, after a certain value of c_{121} (and c_{122}) the rate of decrease in $\mathscr{E}(\sum_k \mu_k(1 - \mu_k)\alpha_k)$ becomes infinitesimally small. This is also evident from the way of computing μ-value where μ_2 of a pattern $\mathbf{x}$ with fixed μ_1 decreases with increase in interclass distance. If the interclass distance exceeds a certain value, μ_2 becomes very small. Thus, the contribution of the pattern to the evaluation index does not get affected further by the extent of the class separation.

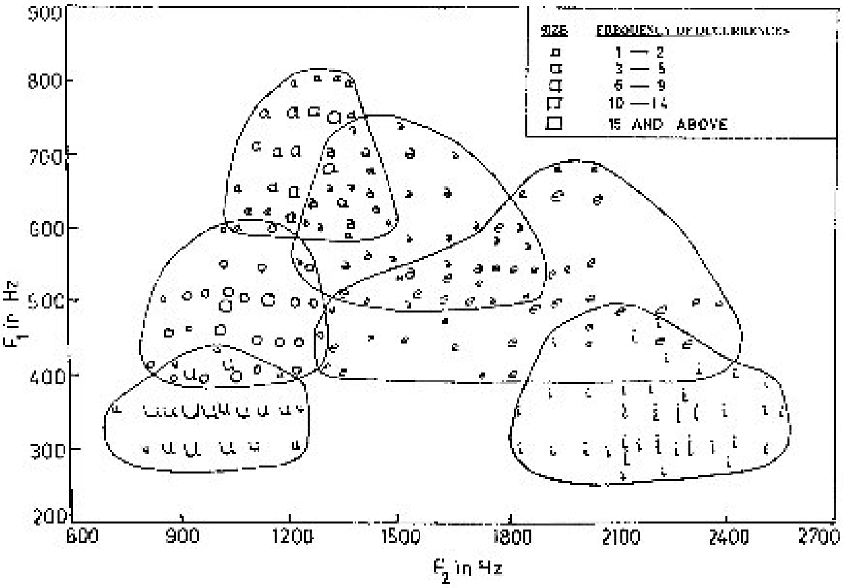

Fig. 7. Two-dimensional (F_1–F_2) plot of the vowel data. This figure is the same as Fig. 22. The only difference is that here approximate boundary of the classes are drawn.

Table 1
Importance of different feature subsets. $X > Y$ means feature subset X is more important than Y. Since the number of subsets for medical and mango-leaf data is large, only first 15 are shown

Data sets	Order of importance using	
	E (Eq. (1))	FEI of Pal (1992), Pal and Chakraborty (1986)
Vowel	{F_2} > {F_1} > {F_1,F_2} > {F_2,F_3} > {F_1,F_2,F_3} > {F_1,F_3} > {F_3}	{F_1,F_2} > {F_2} > {F_1} > {F_2,F_3} > {F_1,F_2,F_3} > {F_3} > {F_1,F_1,F_3}
Iris	{PW} > {PL} > {PL,PW} > {SW,PW} > {SW,PL} > {SL,PL} > {SL,PW} > {SW,PL,PW} > {SL} > {SL,SW,PW} > {SL,SW,PL} > {SL,PL,PW} {SL,SW,PL,PW} > {SL,SW} > {SW}	{PL} > {SW,PL} > {PL,PW} > {PW} > {SW,PL,PW} > {SL,SW,PL,PW} > {SW,PW} > {SL,PL} > {SL,PL,PW} > {SL,SW,PL} > {SL,SW,PW} > {SL,PW} > {SW} > {SL} > {SL,SW}
Medical	{MCV} > {LDH,MCV} > {MCH} > {MCV,MCH} > {MCV,TBil} > {LDH,MCV,TBil} > {LDH,MCV,MCH} > {LDH,MCH} > {BUN,MCV} > {LDH} > {MCH,TBil} > {LDH,BUN,MCV} > {BUN,MCV,MCH} > {BUN,MCV,Tbil} > {LDH,BUN,MCV,MCH} > …	{MCV,MCH,TBil} > {TBil} > {MCV,TBil} > {MCH} > {BUN,MCV,MCH} > {BUN,MCV} > {MCH,TBil} > {BUN,MCV,TBil} > {BUN,MCV,MCH,TBil} > {BUN,MCH} > {MCV,MCH} > {BUN,Tbil} > {BUN} > {BUN,MCH,TBil} > {MCV} > …
Mango-leaf	{L/B} > {L/B,UPe/LPe} > {SI,L/B} > {SI} > {SI,L/B,UPe/LPe} > {SI,UPe/Lpe} > {SI,L/B,(L + P)/B} > {B,L/B} > {B,SI,L/B} > {SI,(L + P)/B} > {B,L/B,(L + P)/B} > {B,SI} > {L/B,(L + P)/B,UPe/LPe} > {(L + P)/B,UPe/LPe} > …	{B} > {L/B} > {B,UPe/LPe} > {Pe} > {(L + P)/B} > {A/L} > {B,L/B} > {B,L/B,UPe/LPe] > {P} > {A} > {L + P > S} > {SI,L/B} > {SI,L/B,UPe/LPe} > {L/B,(L + P)/B,UPe/LPe} > …

Figs. 5 and 6 show the variation of $\mathscr{E}(E)$ with respect to w_1 and w_2 for different interclass distances when $\sum_{i=1}^{2} w_i^2 = 1$. Here we have considered $c_{122} = 0$ throughout whereas c_{121} is considered to be 1.0, 3.0, 5.0, 7.0 and 9.0, respectively. It is seen from the figures that E decreases with w_1 (or increases with w_2) and attains a maximum (or minimum) when $w_1 = 0$ (or when $w_2 = 0$). This is due to the fact that the feature x_2 has no discriminating power as $c_{122} = 0$. On the other hand, the feature x_1 is necessary for classification as there is a separation ($c_{121} \neq 0$) between the classes along its axis. Note also from Figs. 5 and 6 that for higher values of c_{121}, the decrease (or increase) of E is more sharp. This indicates that the rate of convergence of the network to a

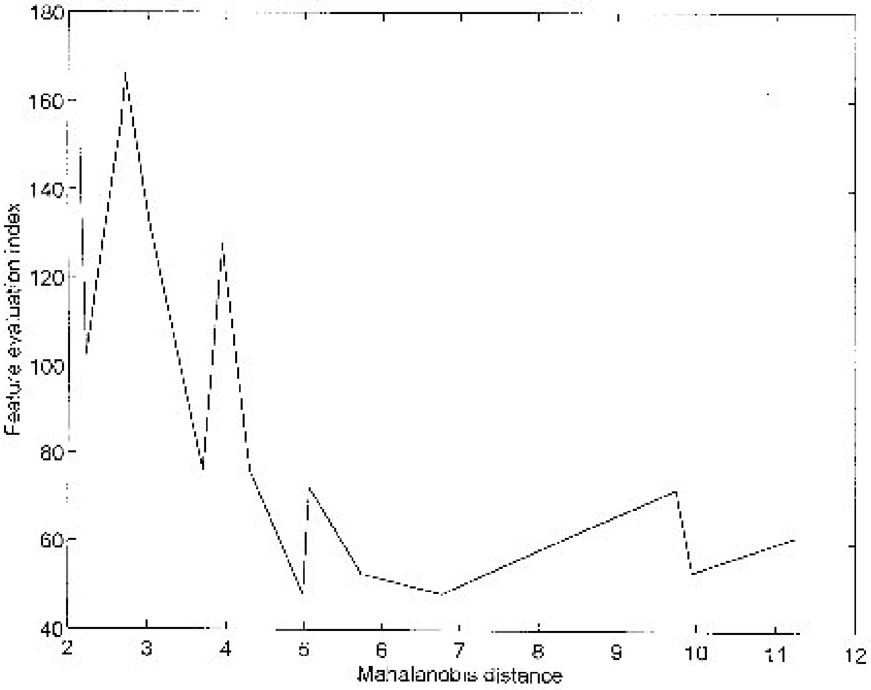

Fig. 8. Graphical representation of the relationship between feature evaluation index and Mahalanobis distance for the vowel data.

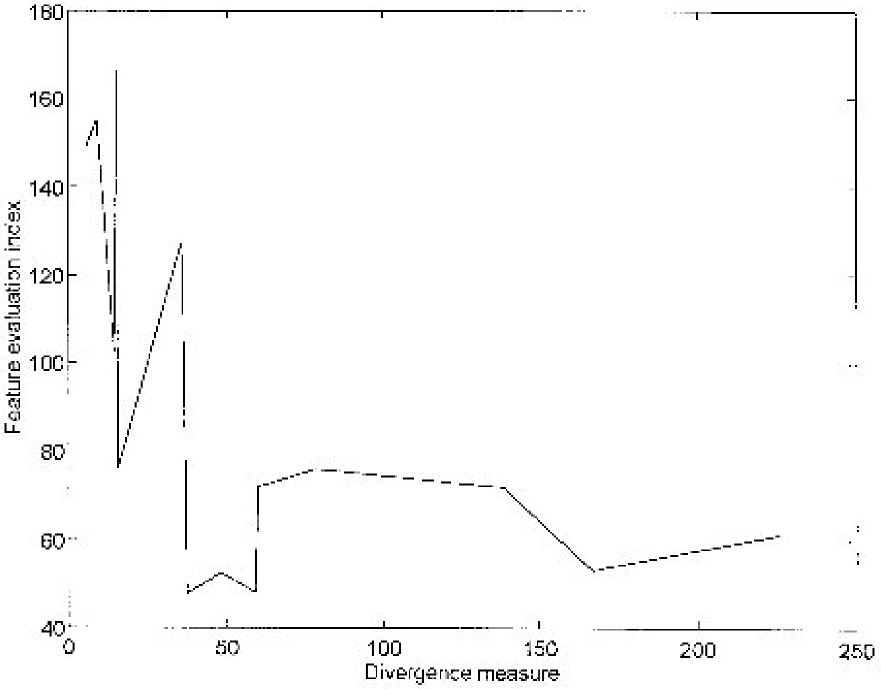

Fig. 9. Graphical representation of the relationship between feature evaluation index and divergence measure for the vowel data.

local minimum increases, as expected, with the decrease in overlap between the classes.

5. Results

The effectiveness of the above-mentioned algorithms was tested on four data sets, namely, vowel data (Pal & Dutta Majumder, 1986), Iris data (Fisher, 1936), medical data (Hayashi, 1991) and mango-leaf data (Bhattacharjee, 1986). The vowel data 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

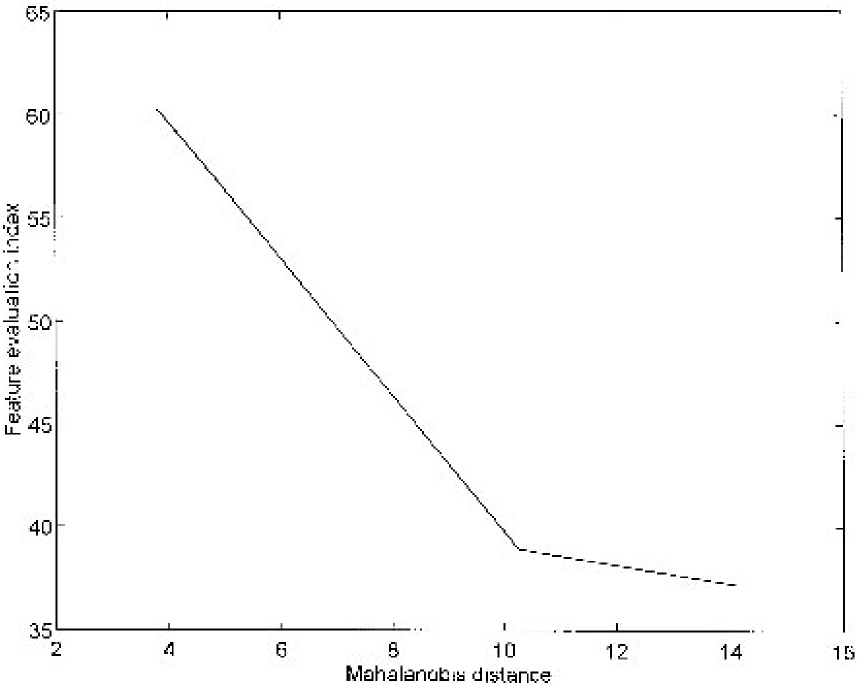

Fig. 10. Graphical representation of the relationship between feature evaluation index and Mahalanobis distance for Iris data.

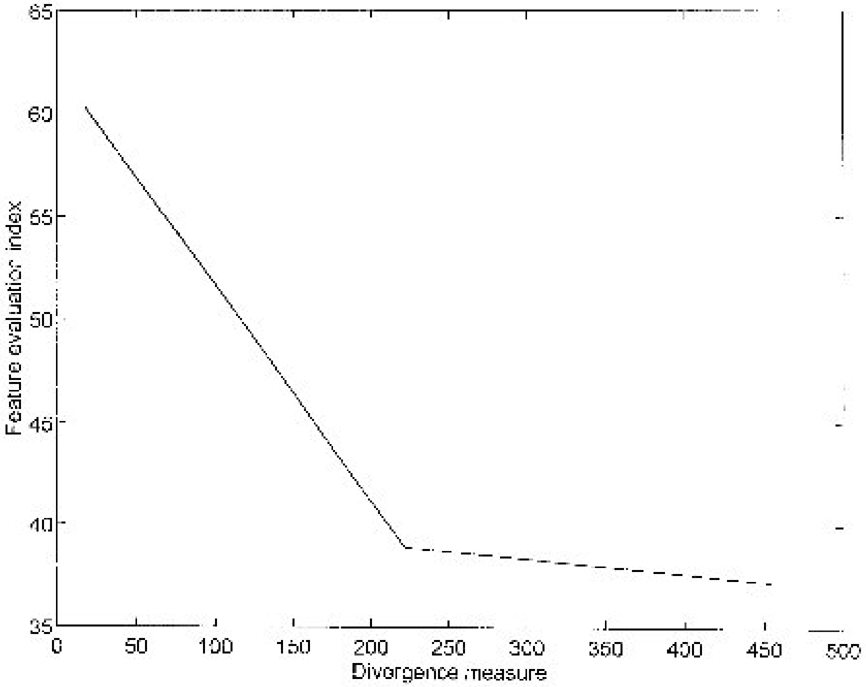

Fig. 11. Graphical representation of the relationship between feature evaluation index and divergence measure for Iris data.

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 format frequencies obtained through spectrum analysis of the speech data. Fig. 7 shows a two-dimensional projection of the three-dimensional 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 & Dutta Majumder, 1986). This vowel data is being extensively used for two decades in the area of pattern recognition.

Anderson'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

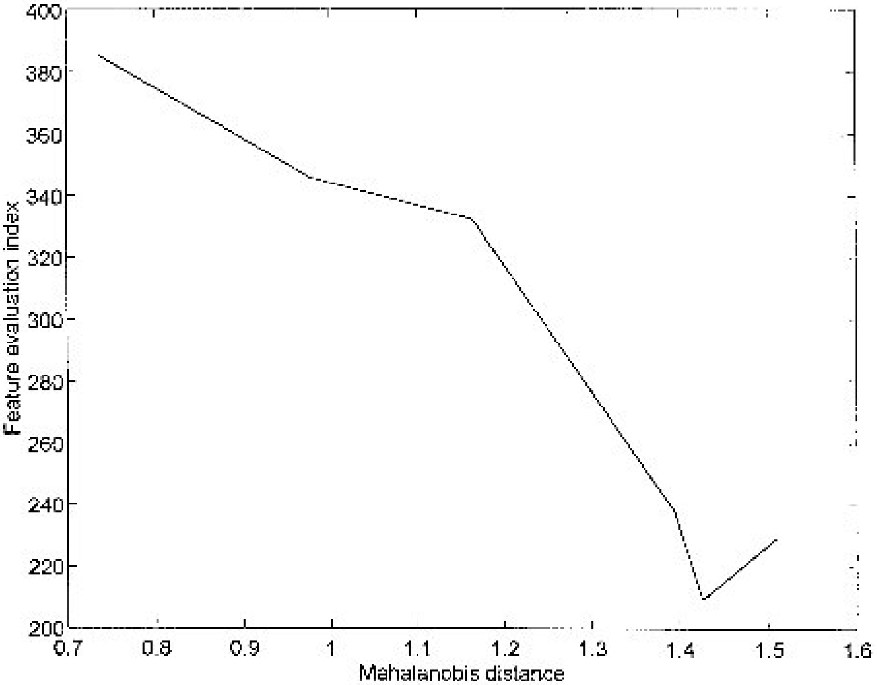

Fig. 12. Graphical representation of the relationship between feature evaluation index and Mahalanobis distance for the medical data.

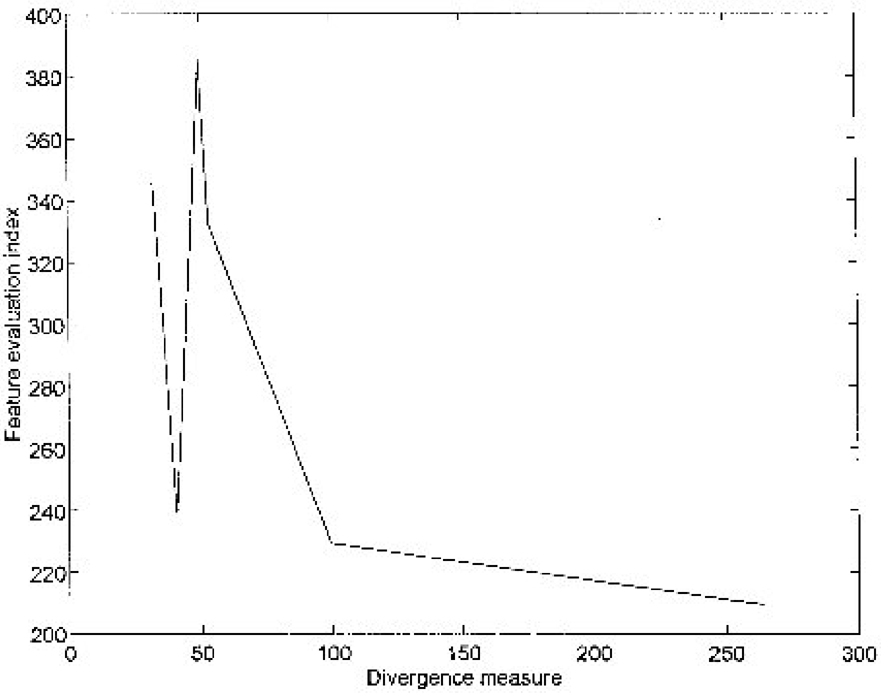

Fig. 13. Graphical representation of the relationship between feature evaluation index and divergence measure for the medical data.

Length (SL), Sepal Width (SW), Petal Length (PL) and Petal Width (PW). Iris data has been used in many research investigation 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,

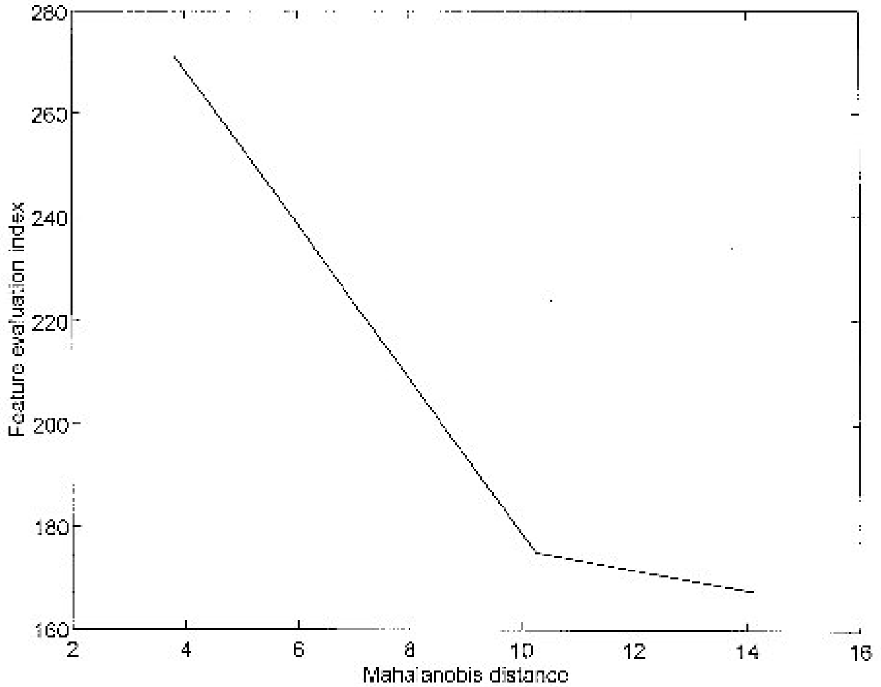

Fig. 14. Graphical representation of the relationship between feature evaluation index and Mahalanobis distance for mango-leaf data.

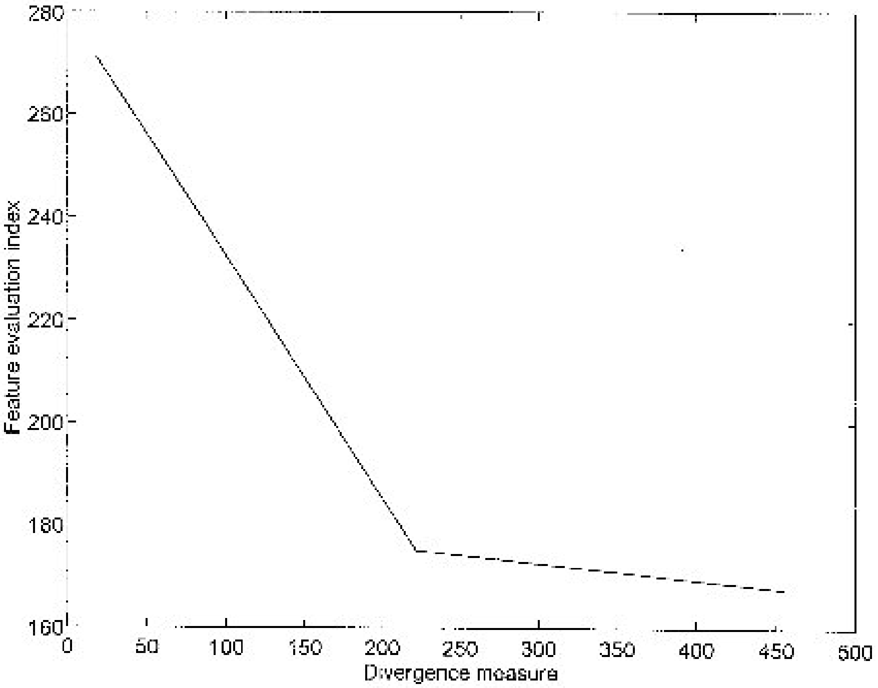

Fig. 15. Graphical representation of the relationship between feature evaluation index and divergence measure for mango-leaf data.

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.

Mango-leaf data (Bhattacharjee, 1986), on the other hand, is a data set on different kinds of mango-leaf with 18 features, (i.e. 18-dimensional data) with 166 data points. 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,

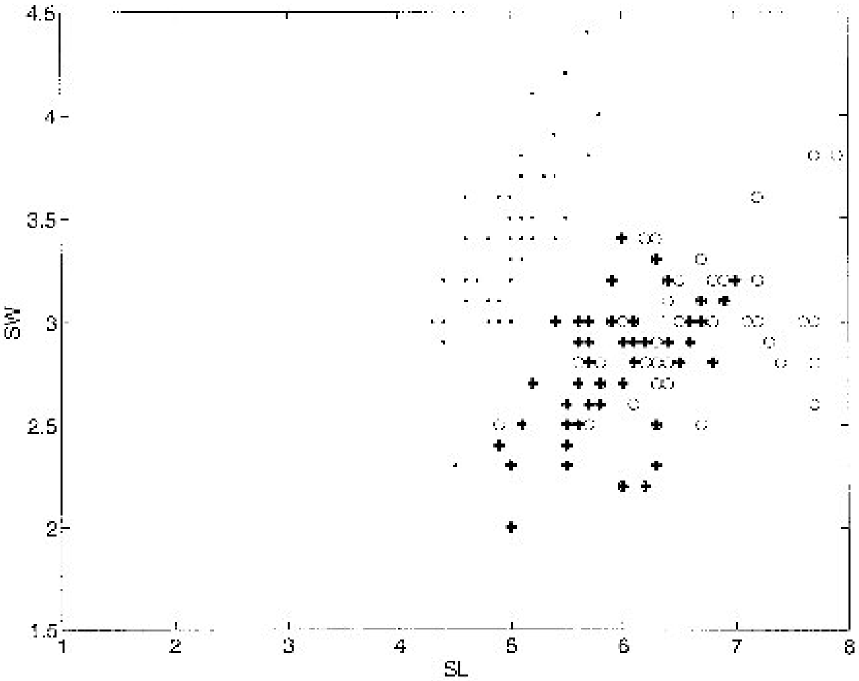

Fig. 16. Scatter plot SL–SW of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

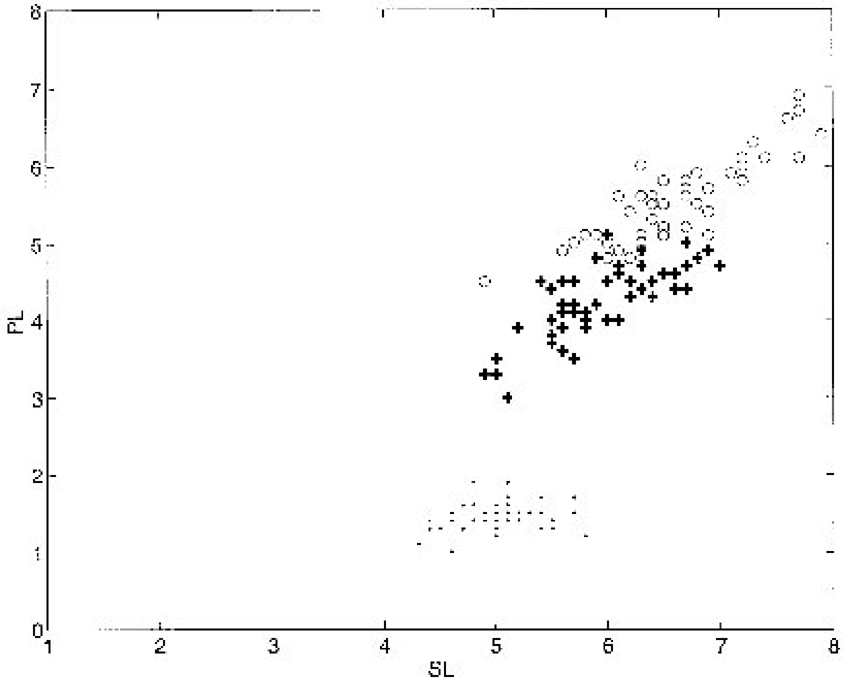

Fig. 17. Scatter plot SL–PL of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

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.

In the following experiments the values of r_k in Eqs. (5) and (8) are so chosen that the membership values of all the patterns of a class are at least 0.5 for that class. For 6-class vowel data the values of r_k are found to be 28.8, 78.5, 21.4, 74.0, 20.4 and 47.8 corresponding to its classes. Similarly, these values are 71.7, 241.3 and 193.9 for 3-class Iris data, 65.0, 38.5, 12.8 and 163.2 for 4-class medical data, and 133.8, 71.2 and 225.2 for 3-class mango-leaf data.

5.1. Using feature evaluation indices

The evaluation index, E (Eq. (1)), was computed for various subsets of features of all the data sets described

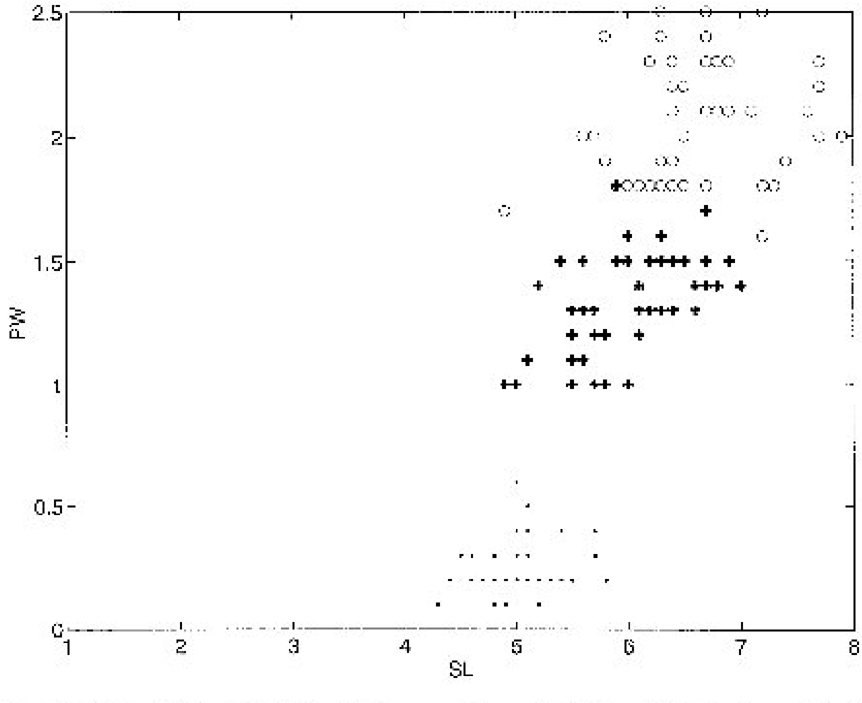

Fig. 18. Scatter plot SL–PW of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

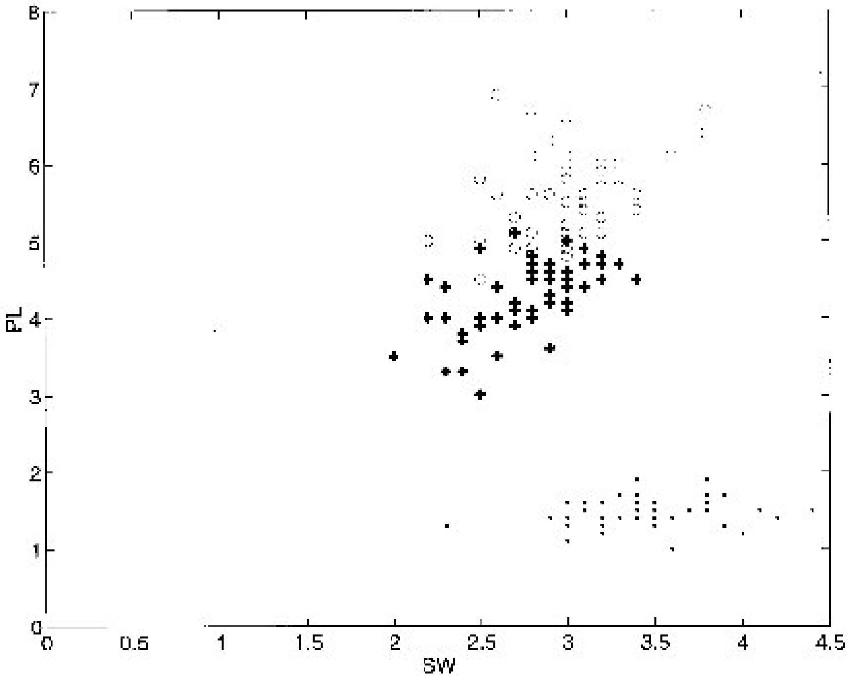

Fig. 19. Scatter plot SW–PL of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

before. The order of importance of these subsets was compared with that obtained by the feature evaluation index (FEI) used by Pal (1992), Pal and Chakraborty (1986).

In the case of vowel data, the order of importance of the subsets of features is

$$\{F_2\} > \{F_1\} > \{F_1,F_2\} > \{F_2,F_3\} > \{F_1,F_2,F_3\}$$
$$> \{F_1,F_3\} > \{F_3\}$$

according to E of Eq. (1), and

$$\{F_1,F_2\} > \{F_2\} > \{F_1\} > \{F_2,F_3\}$$
$$> \{F_1,F_2,F_3\} > \{F_3\} > \{F_1,F_3\}$$

according to the FEI of Pal (1992), Pal and Chakraborty (1986). Here $x > y$ indicates that the importance of feature x is greater than that of feature y. For both the methods, three best subsets are found to be the same. Similarly, in the case of Iris data (Table 1), the subsets {PW}, {PL} and {SW,PW} are found to be the first, second and third best subsets by E (Eq. (1)), whereas the corresponding subsets are {PL}, {SW,PL} and {PL,PW} by the index of Pal (1992), Pal and Chakraborty (1986). Note that, *SL* has not come out as a member of these subsets by either method.

In the case of medical data, since the number of features is nine, we have computed the evaluation indices for individual features (i.e. for the nine subsets), and for all the subsets containing elements of the best four individual features obtained by the respective indices. Note that, these four features are found to be MCV, MCH, LDH and TBil by Eq. (1), and TBil, MCH, BUN and MCV by FEI of Pal (1992), Pal and Chakraborty (1986). Therefore we consider five features LDH, BUN, MCV, MCH and TBil to constitute these subsets. The total number of subsets thus considered including the nine individual features becomes 35. Among all these, the order of importance of the best five subsets, as seen from Table 1, is

$$\{\text{MCV}\} > \{\text{LDH, MCV}\} > \{\text{MCH}\} > \{\text{MCV,MCH}\}$$
$$> \{\text{MCV, TBil}\}$$

according to E of Eq. (1), and

$$\{\text{MCV, MCH, TBil}\} > \{\text{TBil}\} > \{\text{MCV, TBil}\} > \{\text{MCH}\}$$
$$> \{\text{BUN, MCV, MCH}\}$$

according to the FEI of Pal (1992), Pal and Chakraborty (1986). Note that, the features MCV and/or MCH are present in all these subsets obtained by E (Eq. (1)), whereas it is MCH and/or TBil which are present in all the best five subsets obtained by the index of Pal (1992), Pal and Chakraborty (1986). This conforms to the ranking order obtained for individual features where MCV and MCH are found to be the best two features using Eq. (1), and TBil and MCH are those as obtained by the algorithm in Pal (1992), Pal and Chakraborty (1986).

Similarly, in the case of mango-leaf data, since the number of features is 18, we have computed the evaluation indices for individual features (i.e. for the 18 subsets), and for all the subsets containing elements of the best four individual features obtained by the respective indices. Here, the best four features obtained by these two indices are found to be the elements of {Pe, B, SI, L/B, (L + P)/B, UPe/LPe};

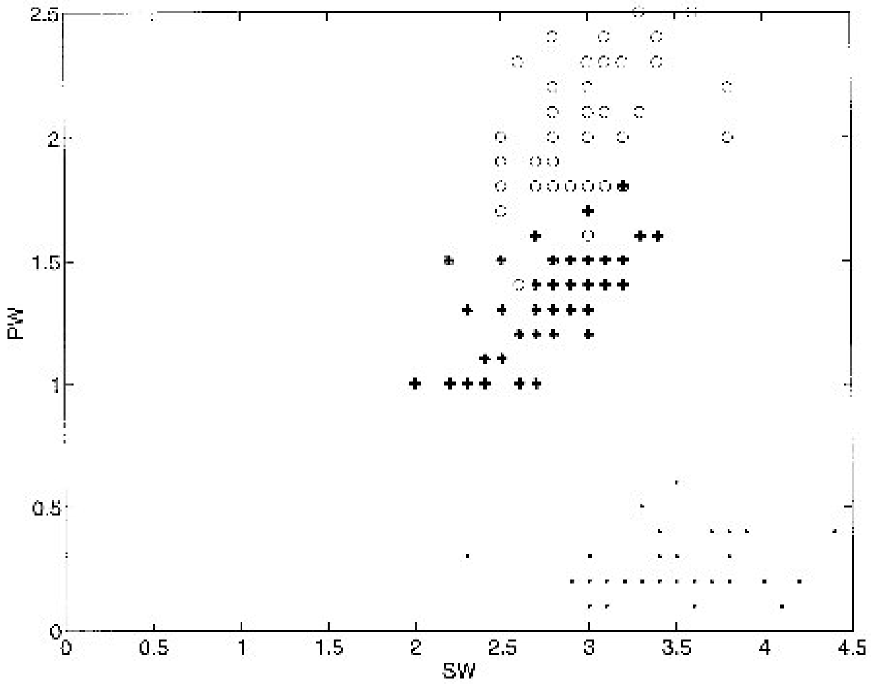

Fig. 20. Scatter plot SW–PW of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

thereby making a total of 75 subsets. Among them, the best five subsets as obtained with *E* (Eq. (1)) and the FEI of Pal (1992), Pal & Chakraborty (1986) are (Table 1)

{L/B} > {L/B, UPe/LPe} > {SI, L/B} > {SI} > {Pe, L/B}

and

{B} > {L/B} > {B, UPe/LPe} > {Pe} > {(L + P)/B}

respectively. Note that the features L/B and/or SI are present in all these five subsets obtained by E (Eq. (1)). This conforms to the ranking order obtained for individual feature where L/B and SI are found to be the best two features using Eq. (1). On the other hand, for FEI (Pal, 1992; Pal & Chakraborty, 1986) the best two individual features, e.g. B and L/B are seen to be present only in the first three subsets.

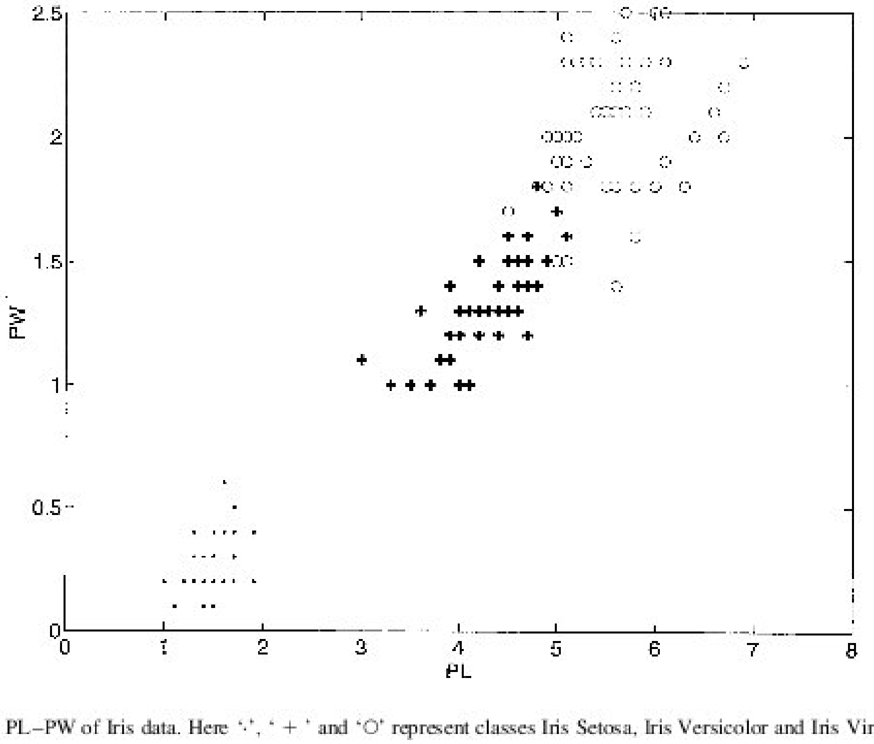

Fig. 21. Scatter plot PL–PW of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

Table 2
Recognition score with k-NN classifier for individual and pairwise features of Iris data

Feature	Classification (%)				
Subset	$k = 1$	$k = 2$	$k = 3$	$k = 5$	$k = 9$
{SL}	48.67	64.00	66.67	67.33	66.67
{SW}	55.33	55.33	52.67	52.67	54.67
{PL}	93.33	89.33	95.33	95.33	95.33
{PW}	89.33	89.33	96.00	96.00	94.67
{SL,SW}	74.67	76.67	76.67	76.00	78.00
{SL,PL}	95.33	92.00	93.33	95.33	96.00
{SL,PW}	94.67	94.67	94.00	94.00	91.33
{SW,PL}	94.67	90.67	92.00	93.33	95.33
{SW,PW}	90.67	92.67	94.00	94.67	94.00
{PL,PW}	93.33	94.00	96.00	96.00	96.67

In order to show the validity of these orders of importance, we consider both scatter plots and k-NN classifier for $k = 1, 2, 3, 5$ and $\sqrt{S}$; S being the number of samples in the training set. The results are shown only for Iris and vowel data. In the case of Iris data, it is seen from Figs. 16–21 that the order of importance (in terms of class structures) of the feature pairs conforms to those (Table 1) obtained by the evaluation index E (Eq. (1)). Among all the feature pairs, {PL,PW} is the best. In other words, the result obtained by FEI of Pal (1992), Pal & Chakraborty (1986), that the subset {SW,PL} is more important than {PL,PW}, does not get reflected by the scatter plots. Although, the order of importance of PW and PL, individually, is found to be different for E and FEI, according to Fig. 21, they are seen to have more or less the same importance.

From the results of k-NN classifier (Table 2), PW is seen to be better than PL for most of the values of k, although the difference is not significant. In fact, the ranking PW > PL > SL > SW as obtained by E for individual features is seen to be exactly reflected in Table 2. As in the case of scatter plots, {PL,PW} is seen here to be the best of all such pairs. In other words, the order obtained by FEI of Pal (1992), Pal & Chakraborty (1986), that {SW, PL} > {PL, PW} does not get supported by the k-NN classifier. The subset {SW,PW} is also found to be more important (in terms of classification performance) than {SW,PL} for all the cases except $k = 9$. These signify the superiority of the measure E over FEI considering the ranking within both individual features and pairwise features.

Table 3
Recognition score with k-NN classifier for individual and pairwise features of vowel data

Feature	Classification (%)				
Subset	$k = 1$	$k = 2$	$k = 3$	$k = 5$	$k = 21$
{F_1}	26.52	18.25	27.21	27.21	31.92
{F_2}	38.58	36.28	38.23	47.76	60.28
{F_3}	26.06	26.41	33.41	33.87	26.75
{F_1,F_2}	56.37	55.68	68.20	76.35	77.73
{F_1,F_3}	44.32	45.58	46.84	55.80	54.65
{F_2,F_3}	58.21	56.14	63.03	63.95	65.10

Table 4
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	0.640382	2	0.257358	2	0.213647	2
F_2	0.759389	1	0.437536	1	0.342621	1
F_3	0.435496	3	0.154319	3	0.123651	3

In the case of overlapping vowel data, it is seen from Figs. 22–24 that $\{F_1, F_2\}$ is the best feature pair, and this conforms to that obtained by both the indices. The order of importance of the feature pairs, $\{F_1, F_2\} > \{F_2, F_3\} > \{F_1, F_3\}$, as obtained by both the indices, is also in conformity to the results obtained by k-NN classifier. However, unlike E, the relative importance of the best three subsets obtained by FEI is seen to be maintained in the results of k-NN classifier.

Finally, the relation of feature evaluation index, E (Eq. (1)) with Mahalanobis distance and divergence measure is graphically depicted in Figs. 8 and 9 (for vowel data), in Figs. 10 and 11 (for Iris data), in Figs. 12 and 13 (for the medical data) and in Figs. 14 and 15 (for mango-leaf data). They are computed over every pair of classes. As expected, Figs. 8–15 show a decrease in feature evaluation index with increase in Mahalanobis distance and divergence measure between the classes.

5.2. Using the neural network model

Tables 4–7 provide the degrees of importance (w) of individual features, obtained by the neural network-based method (Section 3), corresponding to the vowel, Iris, medical and mango-leaf data. Three different initializations of **w** were used in order to train the network. 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

Table 5
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
SL	0.480797	4	0.203230	4	0.229066	4
SW	0.572347	3	0.302529	3	0.374984	3
PL	0.617570	1	0.422186	1	0.420367	1
PW	0.617173	2	0.402027	2	0.402833	2

Table 6
Importance of different features of the medical data

Feature	Initial w					
	$= 1.0$		$\in [0, 1]$		$= 0.5 \pm \epsilon$	
	w	Rank	w	Rank	w	Rank
GOT	0.576090	2	0.601643	2	0.613058	2
GPT	0.300417	3	0.529896	3	0.534147	3
LDH	0.181370	4	0.341677	4	0.322765	4
GGT	0.133649	5	0.300638	5	0.235711	6
BUN	0.070480	9	0.142536	8	0.123007	9
MCV	0.735713	1	0.748205	1	0.747224	1
MCH	0.128931	6	0.101046	7	0.300428	5
Tbil	0.123402	7	0.204479	6	0.201762	7
CRTNN	0.103465	8	0.125008	9	0.149290	8

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 x_i is most important. That is, its presence is a must for characterizing the pattern classes. Similarly, $w_i = 0$ means x_i has no importance and therefore, its presence in the feature vector is not required. $w_i = 0.05$ indicates an ambiguous situation about such presence of x_i. ϵ adds a small perturbation to the degree of presence/importance.

It is found from Table 4 that the order of importance of individual features for the vowel data, under all initializations of $\mathbf{w}$, is $F_2 > F_1 > F_3$ which is the same as obtained by both E (Eq. (1)) and FEI (Pal, 1992; Pal & Chakraborty, 1986). For Iris data (Table 5), like both E (Eq. (1)) and FEI (Pal, 1992; Pal & Chakraborty, 1986), PL and PW are found to be the best two features. As established in Section 5.1 by the scatter plots (Figs. 16–21) and the results of k-NN classifier (Table 2), {PL,PW} is the best feature pair. Within them it is hard to find the edge of one over the other. This justifies the interchangeable order as obtained by E (Eq. (1)) and FEI (Pal, 1992; Pal & Chakraborty, 1986) between PW and PL.

In the case of medical data (Table 6), the order of the best four features as obtained by neuro-fuzzy approach is MCV > GOT > GPT > LDH, whereas this is MCV > MCH > LDH > TBil by Eq. (1). Note that, MCV has come out as the best individual feature in both the cases. Table 8 shows that the results of k-NN classifier using these feature sets. Here, the neuro-fuzzy method is seen to perform better than E (Eq. (1)) (with respect to classification performance) for all values of k. On the other hand, for mango-leaf data, the set of best four features obtained by the neuro-fuzzy approach (Table 7) is found to perform poorer (Table 9). In this connection we mention here that the neuro-fuzzy method considers interdependence among the features, whereas the other method assumes features to be independent of each other.

Table 7
Importance of different features of mango-leaf data

Feature	Initial w					
	$= 1.0$		$\in [0, 1]$		$= 0.5 \pm \epsilon$	
	w	Rank	w	Rank	w	Rank
Z	0.398839	13	0.096816	10	0.007504	17
A	0.509456	9	0.080296	12	0.121824	13
Pe	0.451312	12	0.080145	13	0.209411	11
L	0.507300	11	0.070094	14	0.007141	18
B	0.598589	5	0.426404	4	0.445410	5
P	0.273254	17	0.012582	15	0.300251	9
K	0.600539	4	0.411154	5	0.457997	4
S	0.535693	7	0.186507	9	0.328927	6
SI	0.313462	15	0.008756	16	0.201877	12
L + P	0.508099	10	0.300547	7	0.233489	10
L/P	0.191838	18	0.096777	11	0.111012	15
L/B	0.588887	6	0.213001	8	0.310926	7
(L + P)/B	0.293149	16	0.007061	18	0.116798	14
A/L	0.625549	3	0.500711	3	0.529431	3
A/B	0.523274	8	0.401327	6	0.309092	8
A/Pe	0.643935	2	0.600085	2	0.714805	2
UM/LM	0.322303	14	0.007913	17	0.095220	16
UPe/LPe	1.0	1	0.768731	1	0.720648	1

As mentioned in Section 2, the transformed feature space is obtained by multiplying the original feature values with their respective (optimum) weighting coefficients as obtained by the ANN model. As typical illustrations, Figs. 25–27 depict three scatter plots in the two-dimensional transformed spaces for Iris data. Note that, the scales along both the transformed axes are kept identical to those of the original ones, for the sake of comparison. From Figs. 16–21 and 25–27 it is seen that the classes in the transformed feature spaces are more compact than those in the original spaces; thereby validating one of the objectives of the algorithm. In order to support this finding, k-NN classifier was also used on the transformed spaces. It was found, for example, for the pair {PL,PW} that k-NN classifier results in 94, 94, 96, 96.67 and 97.33% in the transformed space as compared to 93.33, 94, 96, 96 and 96.67% in the original one for $k = 1, 2, 3, 5$ and 9, respectively. Similarly, for overlapping vowel classes, the classification performance is seen to improve in the transformed space for lower values of k. For example, for the feature pairs $\{F_1, F_2\}$, $\{F_1, F_3\}$ and $\{F_2, F_3\}$ in the transformed space, k-NN classifier results in 59.01, 55.34 and 62.80% for $k = 1$, and 57.98, 52.81 and 60.05% for $k = 2$. In contrast to that the figures are (Table 3) 56.37, 44.32 and 58.21% for $k = 1$, and 55.68, 45.58 and 56.14% for $k = 2$ in the original space.

It has been observed experimentally that the network converges much slower with the initialization $w_i = 1$, for all i, as compared to the other values. For example, the number of iterations required to converge the network corresponding to the initializations 1, [0,1] and $0.5 \pm \epsilon$ are 17 300, 10 000 and 11 500 for vowel data, 9400, 7000

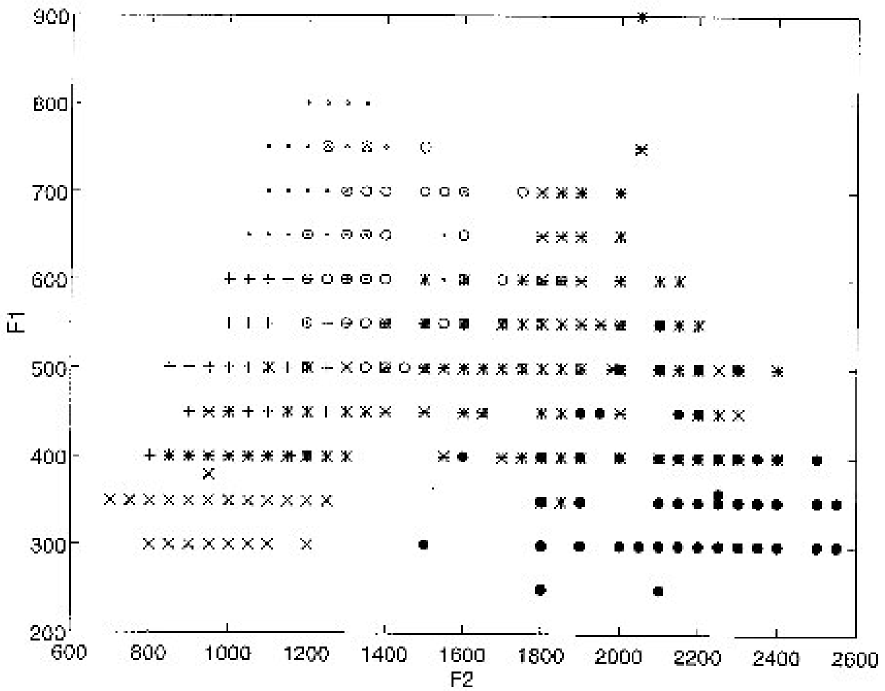

Fig. 22. Scatter plot F$_2$–F$_1$ of vowel data. Here '○', '·', '●', '×', '✳' and '+' represent classes ∂, a, i, u, e and o, respectively.

and 5600 for the Iris data, 4700, 3000 and 1900 for medical data, and 1700, 1200 and 900 for mango-leaf data.

6. Conclusions

In this article, we have presented a neuro-fuzzy model for feature evaluation along with its theoretical analysis and experimental performance on speech (vowel) data, Iris data, medical data and mango-leaf data (having dimension three, four, nine and eighteen respectively). First, a feature evaluation 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.

Weighting factors representing feature importance are then introduced into membership functions. Incorporation of these weighting factors into membership function 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

Table 8
Recognition score for medical data with k-NN classifier corresponding to four best individual features, obtained by the neuro-fuzzy method and E

Feature	Classification (%)									
Subset	$k = 1$	Rank	$k = 2$	Rank	$k = 3$	Rank	$k = 5$	Rank	$k = 16$	Rank
{GOT,GPT,LDH,MCV}	44.40	1	45.90	1	48.51	1	47.76	1	48.88	1
{LDH,MCV, MCH,TBil}	43.66	2	38.06	2	40.67	2	45.90	2	45.15	2

Table 9
Recognition score for mango-leaf data with k-NN classifier corresponding to four best individual features, obtained by the neuro-fuzzy method and E

Feature	Classification (%)									
Subset	$k = 1$	Rank	$k = 2$	Rank	$k = 3$	Rank	$k = 5$	Rank	$k = 9$	Rank
{K,A/L,A/Pe,UPe/LPe}	61.90	2	67.86	2	67.86	2	64.29	2	70.24	2
B,SI,L/B,UPe/LPe	76.19	1	80.95	1	78.57	1	77.38	1	77.38	1

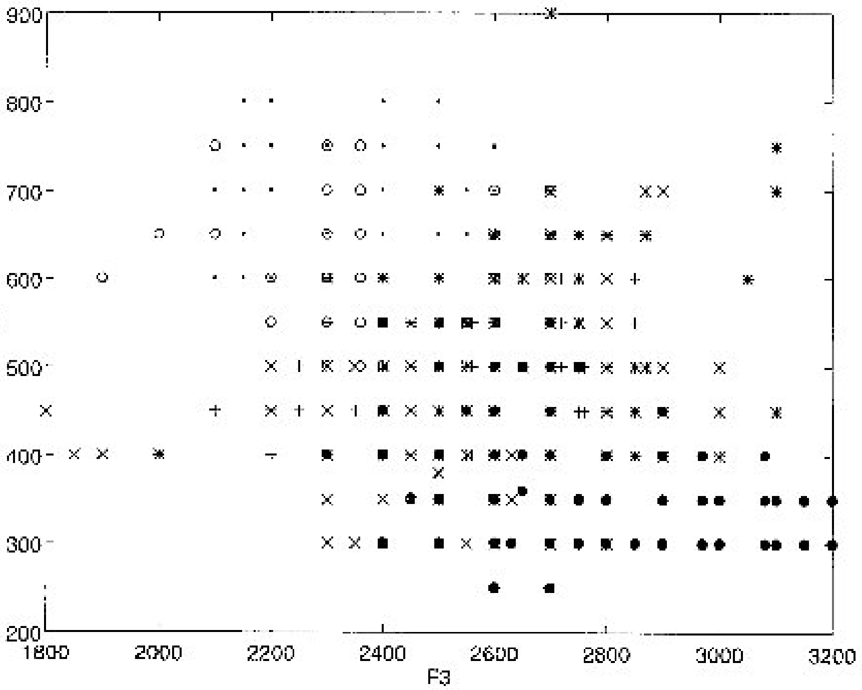

Fig. 23. Scatter plot F_3–F_1 of vowel data. Here '○', '·', '●', '×', '✳' and '+' represent classes ∂, a, i, u, e and o, respectively.

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.

It is shown theoretically that the evaluation index has a fixed upper bound and a varying lower bound. The monotonic increasing behavior of the evaluation index with respect to the lower bound is established for different cases. A relation of the evaluation index, interclass distance and weighting coefficients is derived. It is also shown that

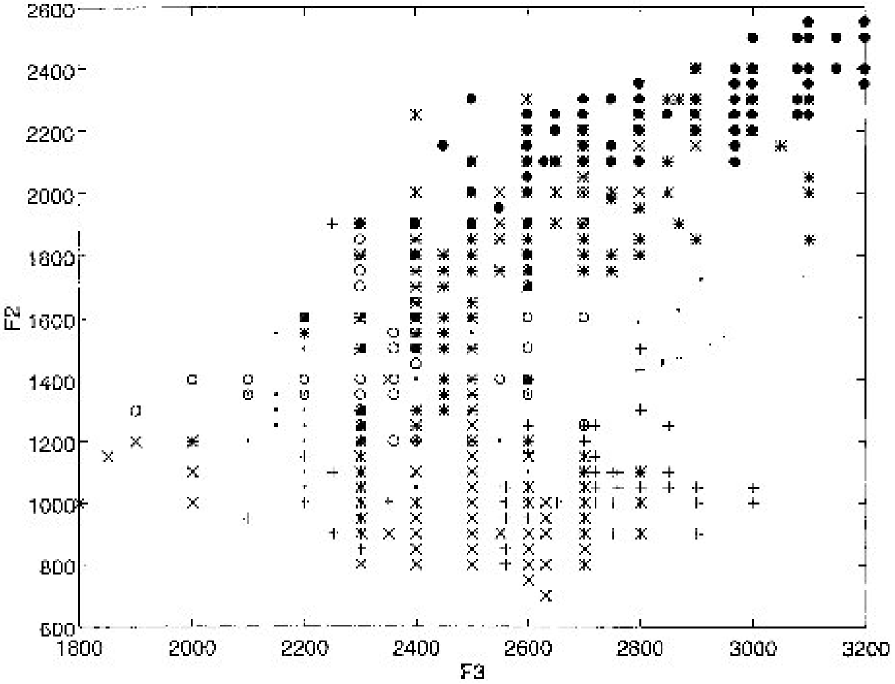

Fig. 24. Scatter plot F_3–F_2 of vowel data. Here '○', '·', '●', '×', '✳' and '+' represent classes ∂, a, i, u, e and o, respectively.

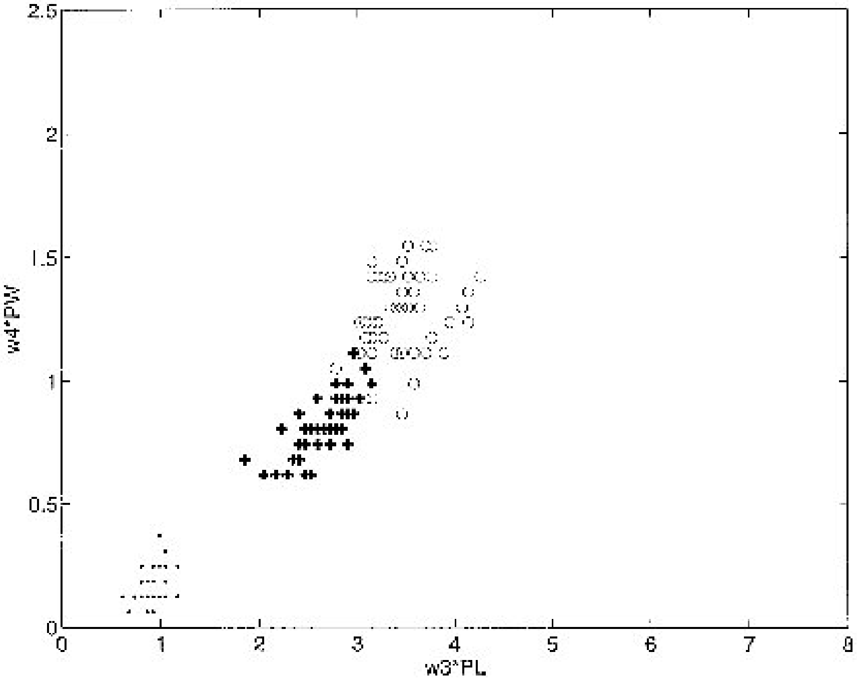

Fig. 25. Scatter plot PL–PW, in the transformed space, of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

the higher the interclass distances, the greater is the chance of the network in getting converged into local minima.

Results obtained by the feature evaluation index E of Eq. (1) is seen to be superior to that of FEI of Pal (1992), Pal and Chakraborty (1986). This is validated by both scatter plots (i.e. in terms of class structures) and k-NN classifier (i.e. in terms of classification performance). Moreover, in the index FEI, 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 $M + {}^{M}C_2$; where the first term and the

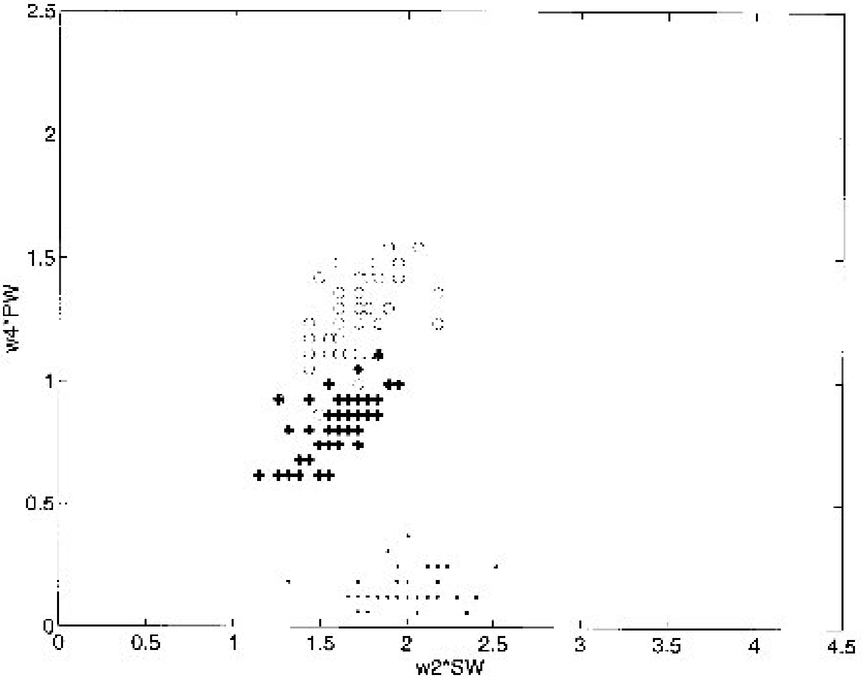

Fig. 26. Scatter plot SW–PW, in the transformed space, of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

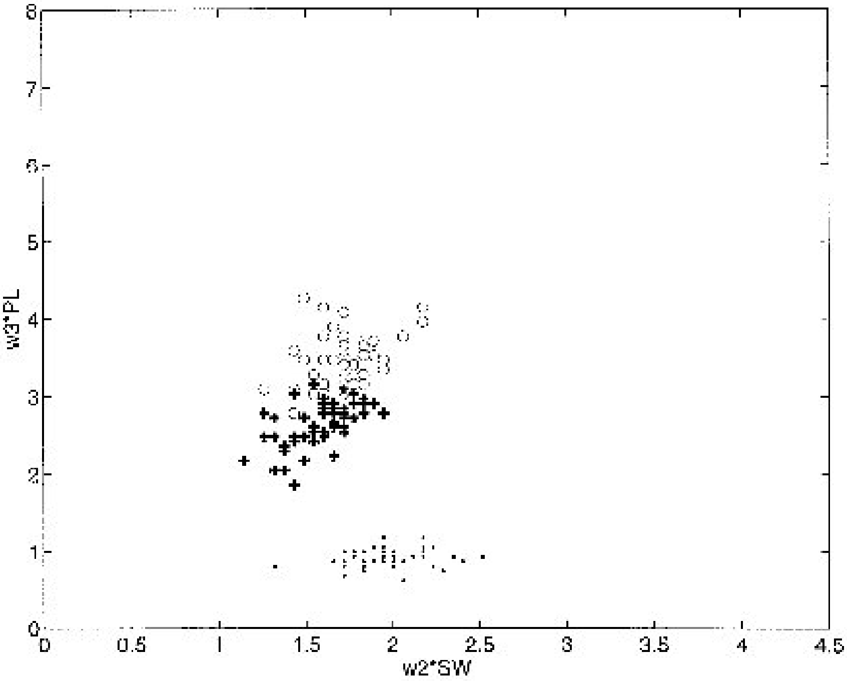

Fig. 27. Scatter plot SW–PL, in the transformed space, of Iris data. Here '·', ' + ' and 'O' represent classes Iris Setosa, Iris Versicolor and Iris Virginica, respectively.

second term correspond to individual class and pairwise class membership functions, respectively. In other words, one needs $M(M+1)$ parameters for computing the FEI. On the other hand, for computing the evaluation index E, we need to compute only M individual class membership functions, i.e. $2M$ parameters. Individual ranking, as obtained by neuro-fuzzy method, conforms well to those obtained by E (Eq. (1)) for both vowel and Iris data. For medical data the former method is seen to perform better as per the k-NN classifier is concerned, whereas it is the reverse for the mango-leaf data.

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 determine, adaptively, 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.

Appendix A. Derivation of Eq. (39)

For a pattern $\mathbf{x} \in C_k$,

$$\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k\times(1-\mu_{k'})+\mu_{k'}\times(1-\mu_k)]}$$

$$=\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k+\mu_{k'}-2\mu_k\mu_{k'}]}$$

$$=\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\mu_k\sum_{k'\neq k}\left[1-\left(2-\frac{1}{\mu_k}\right)\mu_{k'}\right]}$$

$$=\frac{(1-\mu_k)\alpha_k}{\frac{1}{2}[(M-1)-\left(2-\frac{1}{\mu_k}\right)\sum_{k'\neq k}\mu_{k'}]}$$

$$=\frac{(1-\mu_k)\alpha_k}{\frac{1}{2}(M-1)\left[1-\left(2-\frac{1}{\mu_k}\right)\frac{\sum_{k'\neq k}\mu_{k'}}{M-1}\right]}$$

$$\approx\frac{2(1-\mu_k)\alpha_k}{(M-1)}\left[1+\left(2-\frac{1}{\mu_k}\right)\frac{\sum_{k'\neq k}\mu_{k'}}{M-1}\right],$$

as

$$\left(2 - \frac{1}{\mu_k}\right)\frac{\sum_{k\neq k'} \mu_{k'}}{M-1} < 1.$$

Thus,

$$\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]}$$

$$= \frac{2\alpha_k}{M-1}\left(1 - \mu_k + \left(3 - 2\mu_k - \frac{1}{\mu_k}\right)\frac{\sum_{k'\neq k}\mu_{k'}}{M-1}\right)$$

Therefore, using Eq. (38), $\mathscr{E}(\mu_k(1-\mu_k)\alpha_k/\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)])$ is given by

$$\mathscr{E}\left(\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]}\right)$$

$$= \int_{\mathbf{x}\in C_k} \frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]}\wp(\mathbf{x})\mathrm{d}\mathbf{x}$$

$$\approx \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\frac{2\alpha_k}{M-1}\Bigg(1-\mu_k$$

$$+\left(3-2\mu_k-\frac{1}{\mu_k}\right)\frac{\sum_{k'\neq k}\mu_{k'}}{M-1}\Bigg)P_k\wp(\mathbf{x}|C_k)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

Let,

$$J_k = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}(1-\mu_k)P_k\frac{1}{(\sqrt{2\pi}\sigma)^n}$$

$$\times\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$= P_k - P_k\int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\mu_k\frac{1}{(\sqrt{2\pi}\sigma)^n}$$

$$\times\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n = P_k - P_kJ_{k1},$$

where

$$J_{k1} = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\mu_k\frac{1}{(\sqrt{2\pi}\sigma)^n}$$

$$\times\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

Also let,

$$J_{k2} = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\left(3-2\mu_k-\frac{1}{\mu_k}\right)$$

$$\times\sum_{k'\neq k}\mu_{k'}P_k\frac{1}{(\sqrt{2\pi}\sigma)}\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$= P_k\sum_{k'\neq k}3J_{kk'1} - 2J_{kk'2} - J_{kk'3}),$$

where

$$J_{kk'1} = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\frac{\mu_{k'}}{M-1}\frac{1}{(\sqrt{2\pi}\sigma)^n}$$

$$\times\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

$$J_{kk'2} = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\mu_k\frac{\mu_{k'}}{M-1}\frac{1}{(\sqrt{2\pi}\sigma)^n}$$

$$\times\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

$$J_{kk'3} = \int_{x_1=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty}\frac{1}{\mu_k}\frac{\mu_{k'}}{M-1}\frac{1}{(\sqrt{2\pi}\sigma)^n}$$

$$\times\exp\left(-\sum_i\frac{(x_i-m_{ki})^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

Therefore,

$$\mathscr{E}\left(\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]}\right)$$

$$= \frac{\alpha_k}{M-1}(J_K + J_{K2}). \qquad \text{(A1)}$$

Let us also assume that,

$$J_{k1i} = \int_{x_i=-\infty}^{\infty}$$

$$-\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-m_{ki})^2}{2\sigma^2}-\frac{(x_i-m_{ki})^2w_i^2}{2\lambda^2}\right]\mathrm{d}x_i,$$

$$J_{kk'1i} = \int_{x_i=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}$$

$$\times\exp\left[-\frac{(x_i-m_{ki})^2}{2\sigma^2}-\frac{(x_i-m_{k'i})^2w_i^2}{\lambda^2}\right]\mathrm{d}x_i,$$

$$J_{kk'2i} = \int_{x_i=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\Bigg[-\frac{(x_i-m_{ki})^2}{2\sigma^2}$$

$$-\frac{(x_i-m_{k'i})^2w_i^2}{2\lambda^2}-\frac{(x_i-m_{ki})^2w_i^2}{2\lambda^2}\Bigg]\mathrm{d}x_i,$$

$$J_{kk'3i} = \int_{x_i=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - m_{ki})^2}{2\sigma^2} - \frac{(x_i - m_{k'i})^2 w_i^2}{\lambda^2} + \frac{(x_i - m_{ki})^2 w_i^2}{\lambda^2} \right] \mathrm{d}x_i,$$

so that,

$$J_k = P_k(1 - J_{k1}) = P_k\left(1 - \prod_i J_{k1i}\right)$$

and

$$J_{k2} = P_k \sum_{k' \neq k} \left(3\prod_i J_{kk'1i} - 2\prod_i J_{kk'2i} - \prod_i J_{kk'3i}\right).$$

Therefore, from Eq. (A1) we have,

$$\mathcal{E}\left(\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k \times (1-\mu_{k'}) + \mu_{k'} \times (1-\mu_k)]}\right) = \frac{\alpha_k}{M-1}\left(J_k + \frac{J_{k2}}{M-1}\right). \qquad \text{(A2)}$$

For evaluating the integrals J_{k1i}, $J_{kk'1i}$, $J_{kk'2i}$, and $J_{kk'3i}$ we use the result of the following integral,

$$J = \int_{-\infty}^{\infty} \exp\left[-(\alpha x^2 + \beta x + \gamma)\right] \mathrm{d}x.$$

Now,

$$J = \int_{-\infty}^{\infty} \exp\left[-\alpha\left(x^2 + 2x\frac{\beta}{2\alpha} + \frac{\beta^2}{4\alpha^2}\right) + \left(\frac{\beta^2}{4\alpha} - \gamma\right)\right] \mathrm{d}x$$

$$= \exp\left(\frac{\beta^2}{4\alpha} - \gamma\right) \int_{-\infty}^{\infty} \exp\left[-\alpha\left(x + \frac{\beta}{2\alpha}\right)^2\right] \mathrm{d}x$$

$$= \exp\left(\frac{\beta^2}{4\alpha} - \gamma\right) \int_{-\infty}^{\infty} \exp\left[-\alpha y^2\right] \mathrm{d}y$$

where

$$y = x + \frac{\beta}{2\alpha}.$$

Therefore,

$$J = 2\exp\left(\frac{\beta^2 - 4\alpha\gamma}{4\alpha}\right) \int_0^{\infty} \exp(-\alpha y^2)\, \mathrm{d}y$$

$$= 2\exp\left(\frac{\beta^2 - 4\alpha\gamma}{4\alpha}\right) \int_0^{\infty} \frac{1}{2\sqrt{\alpha}} \exp(-z)\, z^{-1/2}\, \mathrm{d}z$$

where

$$z = \alpha y^2.$$

Hence,

$$J = \frac{\exp\left(\dfrac{\beta^2 - 4\alpha\gamma}{4\alpha}\right)\sqrt{\pi}}{\sqrt{\alpha}}. \qquad \text{(A3)}$$

We use the following transformation for evaluating J_{k1i}, $J_{kk'1i}$, $J_{kk'2i}$ and $J_{kk'3i}$.

$$y_i = \left(\frac{x_i - m_{ki}}{\sqrt{2}\lambda}\right) w_i,$$

$$\mathrm{d}x_i = \frac{\sqrt{2}\lambda}{w_i}\, \mathrm{d}y_i.$$

Then we can write,

$$\left[\frac{(x_i - m_{ki})^2 w_i^2}{2\lambda^2} + \frac{(x_i - m_{ki})^2}{2\sigma^2}\right] = y_i^2 + \frac{\rho^2}{w_i^2} y_i^2$$

$$= \left(1 + \frac{\rho^2}{w_i^2}\right) y_i^2,$$

$$\left[\frac{(x_i - m_{k'i})^2 w_i^2}{2\lambda^2} + \frac{(x_i - m_{ki})^2}{2\sigma^2}\right]$$

$$= y_i^2 + \frac{\sqrt{2} w_i c_{kk'i}}{\lambda} y_i + \frac{c_{kk'i}^2 w_i^2}{2\lambda^2} \frac{\rho^2}{w_i^2} y_i^2$$

$$= \left(1 + \frac{\rho^2}{w_i^2}\right) y_i^2 + \frac{\sqrt{2} w_i c_{kk'i}}{\lambda} y_i + \frac{c_{kk'i}^2 w_i^2}{2\lambda^2},$$

$$\left[\frac{(x_i - m_{ki})^2 w_i^2}{2\lambda^2} + \frac{(x_i - m_{k'i})^2 w_i^2}{2\lambda^2} + \frac{(x_i - m_{ki})^2}{2\sigma^2}\right]$$

$$= \left(2 + \frac{\rho^2}{w_i^2}\right) y_i^2 + \frac{\sqrt{2} w_i c_{kk'i}}{\lambda} y_i + \frac{c_{kk'i}^2 w_i^2}{2\lambda^2},$$

$$\left[-\frac{(x_i - m_{ki})^2 w_i^2}{2\lambda^2} + \frac{(x_i - m_{k'i})^2 w_i^2}{2\lambda^2} + \frac{(x_i - m_{ki})^2}{2\sigma^2}\right]$$

$$= \frac{\rho^2}{w_i^2} y_i^2 + \frac{\sqrt{2} w_i c_{kk'i}}{\lambda} y_i + \frac{c_{kk'i}^2 w_i^2}{2\lambda^2}.$$

Therefore, using the result of J (Eq. (A3)) we have,

$$J_{k1i} = \frac{1}{\sqrt{2\pi}\sigma} \frac{\sqrt{2}\lambda}{w_i} \frac{1}{\left(1 + \dfrac{\rho^2}{w_i^2}\right)^{1/2}} \sqrt{\pi} = \frac{\rho}{(w_i^2 + \rho^2)^{1/2}},$$

where $\alpha = (1 + \rho^2/w_i^2)$, $\beta = 0$ and $\gamma = 0$. Similarly, the expressions for $J_{kk'1i}$, $J_{kk'2i}$ and $J_{kk'3i}$ are obtained as follows.

$$J_{kk'1i} = \frac{1}{\sqrt{2\pi}\sigma}\frac{\sqrt{2}\lambda}{w_i}\frac{1}{\left(1+\dfrac{\rho^2}{w_i^2}\right)^{1/2}}\exp\left[-\frac{c_{kk'i}^2}{2\sigma^2\left(1+\dfrac{\rho^2}{w_i^2}\right)}\right]$$

$$\times\sqrt{\pi} = \frac{\rho\exp\left[-\dfrac{c_{kk'i}^2}{2\sigma^2\left(1+\dfrac{\rho^2}{w_i^2}\right)}\right]}{(\rho^2+w_i^2)^{1/2}}$$

where $\alpha = (1+\rho^2/w_i^2)$, $\beta = \sqrt{2}w_ic_{kk'i}/\lambda$ and $\gamma = c_{kk'i}^2w_i^2/2\lambda^2$.

$$J_{kk'2i} = \frac{\rho\exp\left[-\dfrac{c_{kk'i}^2\left(1+\dfrac{\rho^2}{w_i^2}\right)w_i^2}{2\sigma^2\left(2+\dfrac{\rho^2}{w_i^2}\right)}\right]}{(\rho^2+w_i^2)^{1/2}},$$

where $\alpha = (2+\rho^2/w_i^2)$, $\beta = \sqrt{2}w_ic_{kk'i}/\lambda$ and $\gamma = c_{kk'}^2w_i^2/2\lambda^2$.

$$J_{kk'3i} = \exp\left[-\frac{c_{kk'i}^2\left(1-\dfrac{w_i^2}{\rho^2}\right)w_i^2}{2\sigma^2}\right]$$

where $\alpha = \rho^2/w_i^2$, $\beta = \sqrt{2}w_ic_{kk'i}/\lambda$ and $\gamma = c_{kk'i}^2w_i^2/2\lambda^2$.

Therefore, from Eq. (A2) we have,

$$\mathscr{E}\left(\frac{\mu_k(1-\mu_k)\alpha_k}{\frac{1}{2}\sum_{k'\neq k}[\mu_k\times(1-\mu_{k'})+\mu_{k'}\times(1-\mu_k)]}\right)$$

$$= \frac{\alpha_kP_k}{M-1}\left(\left[1-\prod_i\frac{\rho}{(w_i^2+\rho^2)^{1/2}}\right]\right.$$

$$+\sum_{k'\neq k}\left[3\prod_i\frac{\rho\exp\left[-\dfrac{c_{kk'i}^2}{2\sigma^2\left(1+\dfrac{\rho^2}{w_i^2}\right)}\right]}{(\rho^2+w_i^2)^{1/2}}\right.$$

$$-2\prod_i\frac{\rho\exp\left[-\dfrac{c_{kk'i}^2\left(1+\dfrac{\rho^2}{w_i^2}\right)w_i^2}{2\sigma^2\left(2+\dfrac{\rho^2}{w_i^2}\right)}\right]}{(\rho^2+w_i^2)^{1/2}}$$

$$\left.\left.-\prod_i e^{-\frac{c_{kk'i}^2\left(1-\frac{w_i^2}{\rho^2}\right)w_i^2}{2\sigma^2}}\right]\right)$$

$$\approx\frac{\alpha_kP_k}{M-1}\left(\left[1-\prod_i\frac{\rho}{(w_i^2+\rho^2)^{1/2}}\right]\right.$$

$$+\sum_{k'\neq k}\left[3\prod_i\frac{\rho\exp\left[-\dfrac{c_{kk'i}^2}{2\sigma^2\left(1+\dfrac{\rho^2}{w_i^2}\right)}\right]}{(\rho^2+w_i^2)^{1/2}}\right.$$

$$-2\prod_i\frac{\rho\exp\left[-\dfrac{c_{kk'i}^2}{2\sigma^2\left(1+\dfrac{\rho^2}{w_i^2}\right)}\right]}{(\rho^2+w_i^2)^{1/2}}$$

$$\left.\left.-\prod_i\exp\left[-\frac{c_{kk'i}^2}{2\sigma^2\left(1+\dfrac{\rho^2}{w_i^2}\right)}\right]\right]\right) \tag{A4}$$

Now,

$$\prod_i\frac{\rho}{(w_i^2+\rho^2)^{1/2}} = \prod_i\frac{\rho}{\rho\left(1+\dfrac{w_i^2}{\rho^2}\right)^{1/2}} = \prod_i\left(1+\frac{w_i^2}{\rho^2}\right)^{-1/2}$$

$$= \prod_i\left(1-\frac{w_i^2}{2\rho^2}\right) = \left(1-\frac{\sum_i w_i^2}{2\rho^2}\right),$$

as $w_i/\rho \ll 1$. Similarly,

$$\prod_i\frac{\rho}{(2w_i^2+\rho^2)^{1/2}} = \prod_i\frac{\rho}{\rho\left(1+2\dfrac{w_i^2}{\rho^2}\right)^{1/2}}$$

$$= \prod_i\left(1+\frac{2w_i^2}{\rho^2}\right)^{-1/2} = \prod_i\left(1-\frac{w_i^2}{\rho^2}\right) = \left(1-\frac{\sum_i w_i^2}{\rho^2}\right).$$

Therefore,

$$\mathcal{E}(E) \approx \sum_{k} \frac{\alpha_k P_k}{M-1} \frac{\sum_{i} w_i^2}{2\rho^2} \times \left(1 + \sum_{k' \neq k} \exp\left[- \sum_{i} \frac{c_{kk'i}^2}{2\sigma^2\left(1 + \frac{\rho^2}{w_i^2}\right)} \right]\right). \tag{A5}$$

References

Belue, L. M., & Bauer Jr., K. W. (1995). Determining input features for multilayer perceptrons. *Neurocomputing*, *7*, 111–121.

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

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*, New York: Plenum Press.

Bhattacharjee, A. (1986). Some aspects of mango (*Mangifora indica* L) leaf growth features in varietal recognition. Master's thesis, University of Calcutta, Calcutta, India.

Davis, L. (Ed.). (1987). *Genetic algorithms and simulated annealing*. London: Pitman.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition, a statistical approach*, London: Prentice-Hall.

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 R. P. Lippmann, J. E. Moody & D. S. Touretzky, *Advances in neural information processing systems*, (pp. 578–584). Los Altos, CA: Morgan Kaufmann.

Himmelblau, D. M. (1972). *Applied nonlinear programming*, New York: McGraw-Hill.

Kowalczyk, A., & Ferra, H. L. (1994). Developing higher-order neural networks with empirically selected units. *IEEE Transactions on Neural Networks*, *5*, 698–711.

Kraaijveld, M. A., Mao, J., & Jain, A. K. (1995). A non-linear projection method based on Kohonen's topology preserving maps. *IEEE Transactions on Neural Networks*, *6*, 548–559.

Lampinen, J., & Oja, E. (1995). Distortion tolerant pattern recognition based on self-organizing feature extraction. *IEEE Transactions on Neural Networks*, *6*, 539–547.

Lowe, D., & Webb, A. R. (1991). Optimized feature extraction and Bayes decision in feed-forward classifier networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*, 355–364.

Mao, J., & Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, *6*, 296–317.

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

Pal, S. K., & Dutta Majumder, D. (1986). *Fuzzy mathematical approach to pattern recognition*, New York: Wiley (Halsted Press).

Pal, S. K., & Pramanik, P. K. (1986). Fuzzy measures in determining seed points in clustering. *Pattern Recognition Letters*, *4*, 159–164.

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

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.

Ruck, D. W., Rogers, S. K., & Kabrisky, M. (1990). Feature selection using a multilayer perceptron. *Journal of Neural Network Computing, Fall*, 40–48.

Saund, E. (1989). Dimensionality-reduction using connectionist networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*, 304–314.

Schmidt, W. A. C., & Davis, J. P. (1993). Pattern recognition properties of various feature spaces for higher order neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 795–801.

Tou, J. T., & Gonzalez, R. C. (1974). *Pattern recognition principles*, Reading, MA: Addison-Wesley.