

Unsupervised feature selection using a neuro-fuzzy approach

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

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

Abstract

A neuro-fuzzy methodology is described which involves connectionist minimization of a fuzzy feature evaluation index with unsupervised training. The concept of a flexible membership function incorporating weighed distance is introduced in the evaluation index to make the modeling of clusters more appropriate. A set of optimal weighing coefficients in terms of networks parameters representing individual feature importance is obtained through connectionist minimization. Besides, the investigation includes the development of another algorithm for ranking of different feature subsets using the aforesaid fuzzy evaluation index without neural networks. Results demonstrating the effectiveness of the algorithms for various real life data are provided.

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 $\mathbf{x}(x_1, x_2, \dots, x_n)$ in an n -dimensional measurement space (\mathbb{R}^n) is transformed into a point $\mathbf{x}'(x'_1, x'_2, \dots, x'_n)$ in an n' -dimensional ($n' < n$) feature space ($\mathbb{R}^{n'}$). The problem of feature selection deals with choosing some of the x_i s from the measurement space to constitute the feature space. On the other hand, the problem of feature extraction deals with generating new x'_j s (constituting the feature space) based on some x_i s in the measurement space. The main objective of these processes is to retain the optimum salient characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable al-

gorithms can be devised for efficient categorization.

Fuzzy set theory enables one to deal with uncertainties in different tasks of a pattern recognition system, arising from deficiency (e.g., vagueness, incompleteness, etc.) in information, in an efficient manner. Artificial Neural Networks (ANNs), having the capability of fault tolerance, adaptivity and generalization, and scope for massive parallelism, are widely used in dealing with learning and optimization tasks. Fuzzy set theoretic approaches for feature selection are mainly based on measures of entropy and index of fuzziness (Pal and Chakraborty, 1986; Pal, 1992), fuzzy c-means and fuzzy ISODATA algorithms (Bezdek and Castelaz, 1977). Some of the recent attempts made for feature selection in the framework of ANN are mainly based on multilayer feedforward networks and self-organizing networks (Priddy et al., 1993; Steppe and Bauer, Jr., 1996; De et al., 1997; Pregoner et al., 1996). Note that, depending on whether the class information of the samples is

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

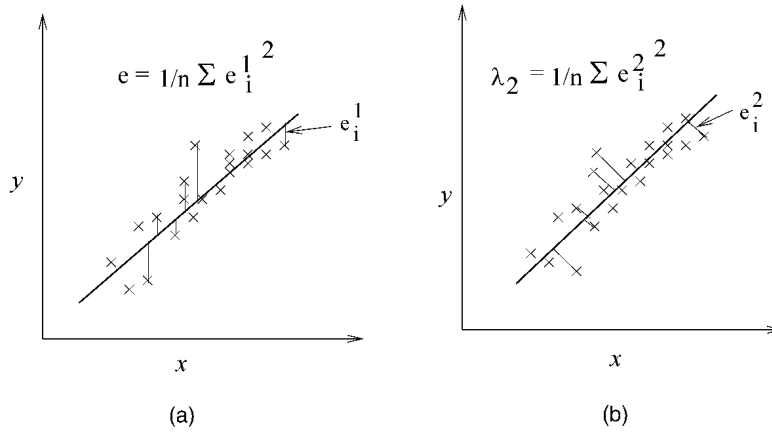


Fig. 1. Nature of errors in linear regression: (a) Least-square fit (e) and (b) least-square projection fit (λ_2).

is the error in predicting y from the linear model $y = a + bx$. a and b are the regression coefficients obtained by minimizing the mean square error

$$e(x, y)^2 = \frac{1}{n} \sum (e(x, y)_i)^2,$$

$e(x, y)_i = y_i - a - bx_i$. The coefficients are given by $a = \bar{y}$ and $b = \frac{\text{cov}(x, y)}{\text{var}(x)}$ and the mean square error $e(x, y)$ is given by $e(x, y) = \text{var}(y)(1 - \rho(x, y)^2)$. If y and x are linearly related $e(x, y) = 0$, and if x and y are completely uncorrelated $e(x, y) = \text{var}(y)$. The measure e^2 is also known as the *residual variance*. It is the amount of variance of y unexplained by the linear model. Some properties of the e are:

1. $0 \leq e(x, y) \leq \text{var}(y)$.
2. $e(x, y) = 0$ if and only if x and y are linearly related.
3. $e(x, y) \neq e(y, x)$ (unsymmetric).
4. If $u = x/c$ and $v = y/d$ for some constant a, b, c, d , then $e(x, y) = d^2 e(u, v)$, i.e., the measure e is *sensitive to scaling* of the variables. It is also clear that e is *invariant to translation* of the variables.
5. The measure e is *sensitive to rotation* of the scatter diagram in $x - y$ plane.

Note that the measure e is not symmetric (property 3). Moreover, it is sensitive to rotation (property 5).

Now, we suggest a measure of linear dependency which has many desirable properties for feature selection not present in the above two measures.

Maximal Information Compression Index (λ_2). Let Σ be the covariance matrix of random variables x and y . Define, *maximal information compression index* as $\lambda_2(x, y) =$ smallest eigenvalue of Σ , i.e.,

$$2\lambda_2(x, y) = (\text{var}(x) + \text{var}(y)) - \sqrt{(\text{var}(x) + \text{var}(y))^2 - 4\text{var}(x)\text{var}(y)(1 - \rho(x, y)^2)}.$$

The value of λ_2 is zero when the features are linearly dependent and increases as the amount of dependency decreases. It may be noted that the measure λ_2 is nothing but the eigenvalue for the direction normal to the principle component direction of feature pair (x, y) . It is shown in [2]

that maximum information compression is achieved if a multivariate (in this case bivariate) data is projected along its principal component direction. The corresponding loss of information in reconstruction of the pattern (in terms of second order statistics) is equal to the eigenvalue along the direction normal to the principal component. Hence, λ_2 is the amount of reconstruction error committed if the data is projected to a reduced (in this case reduced from two to one) dimension in the best possible way. Therefore, it is a measure of the *minimum amount of information loss* or the *maximum amount of information compression*, possible.

The significance of λ_2 can also be explained geometrically in terms of linear regression. It can be easily shown [22] that the value of λ_2 is equal to the sum of the squares of the perpendicular distances of the points (x, y) to the best fit line $y = \hat{a} + \hat{b}x$, obtained by minimizing the sum of the squared perpendicular distances. The coefficients of such a best fit line are given by $\hat{a} = \bar{x} \cot \theta + \bar{y}$ and $\hat{b} = -\cot \theta$, where

$$\theta = 2 \tan^{-1} \left(\frac{2\text{cov}(x, y)}{\text{var}(x)^2 - \text{var}(y)^2} \right).$$

The nature of errors and the best fit lines for least-square regression and principal component analysis are illustrated in Fig. 1. λ_2 has the following properties:

1. $0 \leq \lambda_2(x, y) \leq 0.5(\text{var}(x) + \text{var}(y))$.
2. $\lambda_2(x, y) = 0$ if and only if x and y are linearly related.
3. $\lambda_2(x, y) = \lambda_2(y, x)$ (symmetric).
4. If $u = \frac{x}{c}$ and $v = \frac{y}{d}$ for some constant a, b, c, d , then $\lambda_2(x, y) \neq \lambda_2(u, v)$, i.e., the measure is *sensitive to scaling* of the variables. Since the expression of λ_2 does not contain mean, but only the variance and covariance terms, it is *invariant to translation* of the data set.
5. λ_2 is *invariant to rotation* of the variables about the origin (this can be easily verified from the geometric interpretation of λ_2 considering the property that the perpendicular distance of a point to a line does not change with rotation of the axes).

The measure λ_2 possesses several desirable properties like symmetry (property 3), sensitivity to scaling (property 4), and invariance to rotation (property 5). It is a property of

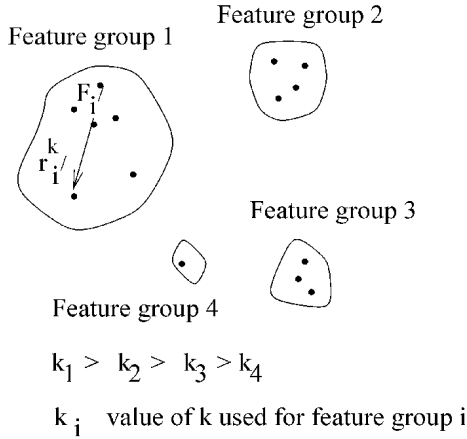


Fig. 2. Feature clusters.

the variable pair (x, y) reflecting the amount of error committed if maximal information compression is performed by reducing the variable pair to a single variable. Hence, it may be suitably used in redundancy reduction.

3 FEATURE SELECTION METHOD

The task of feature selection involves two steps, namely, partitioning the original feature set into a number of homogeneous subsets (clusters) and selecting a representative feature from each such cluster. Partitioning of the features is done based on the k -NN principle using one of the feature similarity measures described in Section 2. In doing so, we first compute the k nearest features of each feature. Among them the feature having the most compact subset (as determined by its distance to the farthest neighbor) is selected, and its k neighboring features are discarded. The process is repeated for the remaining features until all of them are either selected or discarded.

While determining the k nearest-neighbors of features, we assign a constant error threshold (ϵ) which is set equal to the distance of the k th nearest-neighbor of the feature selected in the first iteration. In subsequent iterations, we check the λ_2 value, corresponding to the subset of a feature, whether it is greater than ϵ or not. If yes, then we decrease the value of k . Therefore, k may be varying over iterations. The concept of clustering features into homogeneous groups of varying sizes is illustrated in Fig. 2. The algorithm may be stated as follows:

Algorithm:

Let the original number of features be D , and the original feature set be $O = \{F_i, i = 1, \dots, D\}$. Represent the dissimilarity between features F_i and F_j by $S(F_i, F_j)$. Higher the value of S is, the more dissimilar are the features. The measures of linear dependency (e.g., ρ, e, λ_2) described in Section 2 may be used in computing S . Let r_i^k represent the dissimilarity between feature F_i and its k th nearest-neighbor feature in R . Then

Step 1: Choose an initial value of $k \leq D - 1$. Initialize the reduced feature subset R to the original feature set O ,

i.e., $R \leftarrow O$.

Step 2: For each feature $F_i \in R$, compute r_i^k .

Step 3: Find feature $F_{i'}$ for which $r_{i'}^k$ is minimum.

Retain this feature in R and discard k nearest features of $F_{i'}$. (Note: $F_{i'}$ denotes the feature for which removing k nearest-neighbors will cause minimum error among all the features in R). Let $\epsilon = r_{i'}^k$.

Step 4: If $k > \text{cardinality}(R) - 1$: $k = \text{cardinality}(R) - 1$.

Step 5: If $k = 1$: **Go to Step 8.**

Step 6: **While** $r_{i'}^k > \epsilon$ **do:**

(a) $k = k - 1$.

$r_{i'}^k = \inf_{F_i \in R} r_i^k$.

("k" is decremented by 1, until the

" k th nearest-neighbor" of at least one of the features in R is less than ϵ -dissimilar with the feature)

(b) **If** $k = 1$: **Go to Step 8.**

(if no feature in R has less than ϵ -dissimilar "nearest-neighbor" select all the remaining features in R)

End While

Step 7: **Go to Step 2.**

Step 8: Return feature set R as the reduced feature set.

3.1 Remarks

Computational Complexity. The algorithm has low-computational complexity with respect to both number of features and number of samples of the original data. With respect to the dimension (D), the method has complexity $\mathcal{O}(D^2)$. Among the existing search based schemes only sequential forward and backward search have complexity $\mathcal{O}(D^2)$, though each evaluation is more time consuming. Other algorithms like plus- l -take- r , sequential floating search and branch and bound algorithm [2] have complexity higher than quadratic. Most probabilistic search algorithms also require more than quadratic number of evaluations.

The second factor which contributes to the speedup achieved by the proposed algorithm is the low-computational complexity of evaluating the linear dependency measures of feature similarity. If the data set contains l samples, evaluation of the similarity measure for a feature pair is of complexity $\mathcal{O}(l)$. Thus, the feature selection scheme has overall complexity $\mathcal{O}(D^2l)$. Almost all other supervised and unsupervised feature evaluation indices (e.g., entropy, class separability, K -NN classification accuracy) have at least $\mathcal{O}(l^2)$ complexity of computation. Moreover, evaluation of the linear dependency measures involves computation using one-dimensional variables only, while the other measures often involve distance computations at higher dimensions. All these factors contribute to the large speedup achieved by the proposed algorithm compared to other feature selection schemes.

Notion of Scale in Feature Selection and Choice of k . In our algorithm k controls the size of the reduced set. Since k determines the error threshold (ϵ), the representation of the data at different degrees of details is controlled by its choice. This characteristic is useful in data mining where *multiscale* representation of the data is often necessary. Note that the said property may not always be possessed by

TABLE 1
Comparison of Feature Selection Algorithms for Large-Dimensional Data Sets

Dataset	Method	Evaluation Criteria						CPU Time (sec)	
		E	FFEI	S	KNNA (%)		BayesA (%)		
					Mean	SD	Mean		SD
Isolet d=310 D=617 k = 305	SFS	0.52	0.41	1.09	95.02	0.89	92.03	0.52	14.01 × 10 ⁴
	SWC	0.71	0.55	2.70	72.01	0.71	68.01	0.44	431
	Relief-F	0.70	0.52	2.24	95.81	0.81	95.52	0.47	5.03 × 10 ³
	Proposed	0.50	0.40	1.07	96.00	0.78	95.01	0.52	440
Mult. Feat. d=325 D=649 k = 322	SFS	0.67	0.47	0.45	77.01	0.24	75.02	0.14	5.00 × 10 ⁴
	SWC	0.79	0.55	0.59	52.00	0.19	50.05	0.10	401
	Relief-F	0.71	0.50	0.52	78.37	0.22	75.25	0.11	1.10 × 10 ³
	Proposed	0.68	0.48	0.45	78.34	0.22	75.28	0.10	451
Arrhythmia d=100 D=195 k = 95	SFS	0.74	0.44	0.25	52.02	0.55	50.21	0.43	1511
	SWC	0.82	0.59	0.41	40.01	0.52	38.45	0.38	70
	Relief-F	0.78	0.55	0.27	56.04	0.54	54.55	0.40	404
	Proposed	0.72	0.40	0.17	58.93	0.54	56.00	0.41	74

E: Entropy, *FFEI*: Fuzzy Feature Evaluation Index, *S*: Class Separability, *KNNA*: *k*-NN classification accuracy, *BayesA*: naive Bayes classification accuracy, and *SD*: standard deviation. *SFS*: Sequential Forward Search and *SWC*: Stepwise Clustering. *d*: number of selected features, *D*: number of original features and *k*: parameter used by the proposed method.

other algorithms where the input is usually the desired size of the reduced feature set. The reason is that changing the size of the reduced set may not necessarily result in any change in the levels of details. In contrast, for the proposed algorithm, *k* acts as a scale parameter which controls the degree of details in a more direct manner.

Nonmetric Nature of Similarity Measure. The similarity measures used in the proposed algorithm need not be a metric. Unlike conventional agglomerative clustering algorithms, it does not utilize metric property of the similarity measures. Also, unlike the stepwise clustering method [21] used previously for feature selection, our clustering algorithm is partitional and nonhierarchical in nature.

4 FEATURE EVALUATION INDICES

Now, let us describe some indices that have been considered for evaluating the effectiveness of the selected feature subsets. The first three indices, namely class separability, *K*-NN classification accuracy, and naive Bayes classification accuracy, do need class information of the samples while the remaining three namely, entropy, fuzzy feature evaluation index, and representation entropy, do not. Before we discuss them, we mention, for convenience, the following notations: Let *l* be the number of sample points in the data set, *c* be the number of classes present in the data set, *D* be the number of features in the original feature set *O*, *d* be the number of features in the reduced feature set *R*, Ω_O be the original feature space with

dimension *D*, and Ω_R be the transformed feature space with dimension *d*.

1. *Class Separability* [2]. Class separability *S* of a data set is defined as $S = \text{trace}(S_b^{-1}S_w)$. S_w is the within class scatter matrix and S_b is the between class scatter matrix, defined as:

$$\begin{aligned}
 S_w &= \sum_{j=1}^c \pi_j E\{(X - \mu_j)(X - \mu_j)^T | \omega_j\} = \sum_{j=1}^c \pi_j \Sigma_j \\
 S_b &= \sum_{j=1}^c (\mu_j - M_o)(\mu_j - M_o)^T \\
 M_o &= E\{X\} = \sum_{j=1}^c \pi_j \mu_j,
 \end{aligned} \tag{1}$$

where π_j is the a priori probability that a pattern belongs to class ω_j , *X* is the feature vector, μ_j is the sample mean vector of class ω_j , M_o is the sample mean vector for the entire data points, Σ_j is the sample covariance matrix of class ω_j , and $E\{\cdot\}$ is the expectation operator. A lower value of the separability criteria *S* ensures that the classes are well separated by their scatter means.

2. *K-NN Classification Accuracy.* Here, we have used the *K*-NN rule for evaluating the effectiveness of the reduced set for classification. Cross-validation is performed in the following manner—we randomly select 10 percent of the data as training set and

TABLE 2
Comparison of Feature Selection Algorithms for Medium-Dimensional Data Sets

Dataset	Method	Evaluation Criteria						CPU Time (sec)	
		E	FFEI	S	KNNA (%)		BayesA (%)		
					Mean	SD	Mean		SD
Spambase d=29 D=57 k = 27	BB	0.50	0.30	0.28	90.01	0.71	88.17	0.55	1579
	SFFS	0.50	0.30	0.28	90.01	0.72	88.17	0.55	1109
	SFS	0.52	0.34	0.29	87.03	0.68	86.20	0.54	121.36
	SWC	0.59	0.37	0.41	82.04	0.68	79.10	0.55	11.02
	Relief-F	0.59	0.36	0.34	87.04	0.70	86.01	0.52	70.80
	Proposed	0.50	0.30	0.28	90.01	0.71	88.19	0.52	13.36
Waveform d=20 D=40 k = 17	BB	0.67	0.47	0.29	78.02	0.47	62.27	0.41	1019
	SFFS	0.68	0.48	0.31	77.55	0.45	62.22	0.41	627
	SFS	0.69	0.49	0.37	74.37	0.44	59.01	0.42	71.53
	SWC	0.72	0.55	0.41	62.03	0.40	47.50	0.40	8.01
	Relief-F	0.73	0.54	0.38	74.88	0.41	62.88	0.40	50.22
	Proposed	0.68	0.48	0.30	75.20	0.43	63.01	0.40	8.28
Ionosphere d=16 D=32 k = 11	BB	0.65	0.44	0.07	75.96	0.35	65.10	0.28	150.11
	SFFS	0.65	0.44	0.08	74.73	0.37	65.08	0.31	50.36
	SFS	0.65	0.44	0.10	69.94	0.32	62.00	0.27	10.70
	SWC	0.66	0.47	0.22	62.03	0.32	59.02	0.25	1.04
	Relief-F	0.62	0.47	0.15	72.90	0.34	64.55	0.27	8.20
	Proposed	0.64	0.43	0.10	78.77	0.35	65.92	0.28	1.07

BB: Branch and Bound, SFFS: Sequential Floating Forward Search

classify the remaining 90 percent points. Ten such independent runs are performed and the average classification accuracy on test set is used. The value of K , chosen for the K -NN rule, is the square root of the number of data points in the training set.

3. *Naive Bayes Classification Accuracy.* A Bayes maximum likelihood classifier [2], assuming normal distribution of classes, is also used for evaluating the classification performance. Mean and covariance of the classes are estimated from a randomly selected 10 percent training sample, and the remaining 90 percent of the points are used as test set. Ten such independent runs are performed and the average classification accuracy on test set is provided.
4. *Entropy* [10]. Let the distance between two data points p, q be

$$D_{pq} = \left[\sum_{j=1}^M \left(\frac{x_{p,j} - x_{q,j}}{\max_j - \min_j} \right)^2 \right]^{1/2},$$

where $x_{p,j}$ denotes feature value for p along j th direction, and \max_j, \min_j are the maximum and minimum values computed over all the samples along

j th axis, M is the number of features. Similarity between p, q is given by $\text{sim}(p, q) = e^{-\alpha D_{pq}}$, where α is a positive constant. A possible value of α is $\frac{-\ln 0.5}{\bar{D}}$. \bar{D} is the average distance between data points computed over the entire data set. Entropy is defined as:

$$E = - \sum_{p=1}^l \sum_{q=1}^l (\text{sim}(p, q) \times \log \text{sim}(p, q) + (1 - \text{sim}(p, q)) \times \log (1 - \text{sim}(p, q))). \quad (2)$$

If the data is uniformly distributed in the feature space, entropy is maximum. When the data has well-formed clusters uncertainty is low and so is entropy.

5. *Fuzzy Feature Evaluation Index* [13]. Fuzzy feature evaluation index (FFEI) is defined as:

$$FFEI = \frac{2}{l(l-1)} \sum_p \sum_{q \neq p} \frac{1}{2} \left[\mu_{pq}^R (1 - \mu_{pq}^O) + \mu_{pq}^O (1 - \mu_{pq}^R) \right], \quad (3)$$

where μ_{pq}^O and μ_{pq}^R are the degrees that both patterns p and q belong to the same cluster in the feature

TABLE 3
Comparison of Feature Selection Algorithms for Low-Dimensional Data Sets

Dataset	Method	Evaluation Criteria						CPU Time (sec)	
		E	FFEI	S	KNNA (%)		BayesA (%)		
					Mean	SD	Mean		SD
Forest d=5 D=10 k = 5	BB	0.65	0.40	0.90	64.03	0.41	63.55	0.40	4.01×10^4
	SFFS	0.64	0.39	0.81	67.75	0.43	66.22	0.41	3.02×10^4
	SFS	0.64	0.41	0.98	62.03	0.41	61.09	0.40	7.00×10^3
	SWC	0.68	0.45	1.00	54.70	0.37	53.25	0.35	50.03
	Relief-F	0.65	0.40	0.90	64.03	0.41	63.55	0.40	2.80×10^4
	Proposed	0.65	0.40	0.90	64.03	0.41	63.55	0.40	55.50
Cancer d=4 D=9 k = 5	BB	0.59	0.36	1.84	94.90	0.17	94.45	0.14	3.39
	SFFS	0.59	0.36	1.84	94.90	0.17	94.45	0.14	6.82
	SFS	0.61	0.37	2.68	92.20	0.17	91.05	0.15	1.16
	SWC	0.60	0.37	2.69	90.01	0.19	89.11	0.17	0.10
	Relief-F	0.59	0.36	1.84	94.90	0.17	94.25	0.17	0.91
	Proposed	0.56	0.34	1.70	95.56	0.17	94.88	0.17	0.10
Iris d=2 D=4 k = 2	BB	0.55	0.34	22.0	96.80	0.14	97.33	0.10	0.56
	SFFS	0.55	0.34	22.0	96.80	0.14	97.33	0.10	0.71
	SFS	0.57	0.35	27.0	92.55	0.17	93.10	0.14	0.25
	SWC	0.60	0.37	29.2	92.19	0.19	93.02	0.17	0.01
	Relief-F	0.55	0.34	22.0	96.80	0.14	97.33	0.10	0.14
	Proposed	0.55	0.34	22.0	96.80	0.14	97.33	0.10	0.01

spaces Ω_O and Ω_R , respectively. Membership function μ_{pq} may be defined as

$$\mu_{pq} = 1 - \frac{d_{pq}}{D_{max}} \quad \text{if } d_{pq} \leq D_{max}$$

$$= 0, \quad \text{otherwise.}$$

d_{pq} is the distance between patterns p and q , and D_{max} is the maximum separation between patterns in the respective feature spaces.

The value of FFEI decreases as the intercluster/intracluster distances increase/decrease. Hence, the lower the value of FFEI, the more crisp is the cluster structure. Note that the first two indices, class separability and K-NN accuracy, measure the effectiveness of the feature subsets for classification, while the indices entropy and fuzzy feature evaluation index evaluate the clustering performance of the feature subsets. Let us now describe a quantitative index which measures the amount of redundancy present in the reduced subset.

6. *Representation Entropy* [2]. Let the eigenvalues of the $d \times d$ covariance matrix of a feature set of size d be $\lambda_j, j = 1, \dots, d$. Let

$$\tilde{\lambda}_j = \frac{\lambda_j}{\sum_{j=1}^d \lambda_j}.$$

$\tilde{\lambda}_j$ has similar properties like probability, namely, $0 \leq \tilde{\lambda}_j \leq 1$ and $\sum_{j=1}^d \tilde{\lambda}_j = 1$. Hence, an entropy function can be defined as

$$H_R = - \sum_{j=1}^d \tilde{\lambda}_j \log \tilde{\lambda}_j. \quad (4)$$

The function H_R attains a minimum value (zero) when all the eigenvalues except one are zero or, in other words, when all the information is present along a single coordinate direction. If all the eigenvalues are equal, i.e., information is equally distributed among all the features, H_R is maximum and so is the uncertainty involved in feature reduction. The above measure is known as *representation entropy*. It is a property of the data set as represented by a particular set of features, and is a measure of the amount of information compression possible by dimensionality reduction. This is equivalent to the amount of redundancy present in that particular representation of the data set. Since the

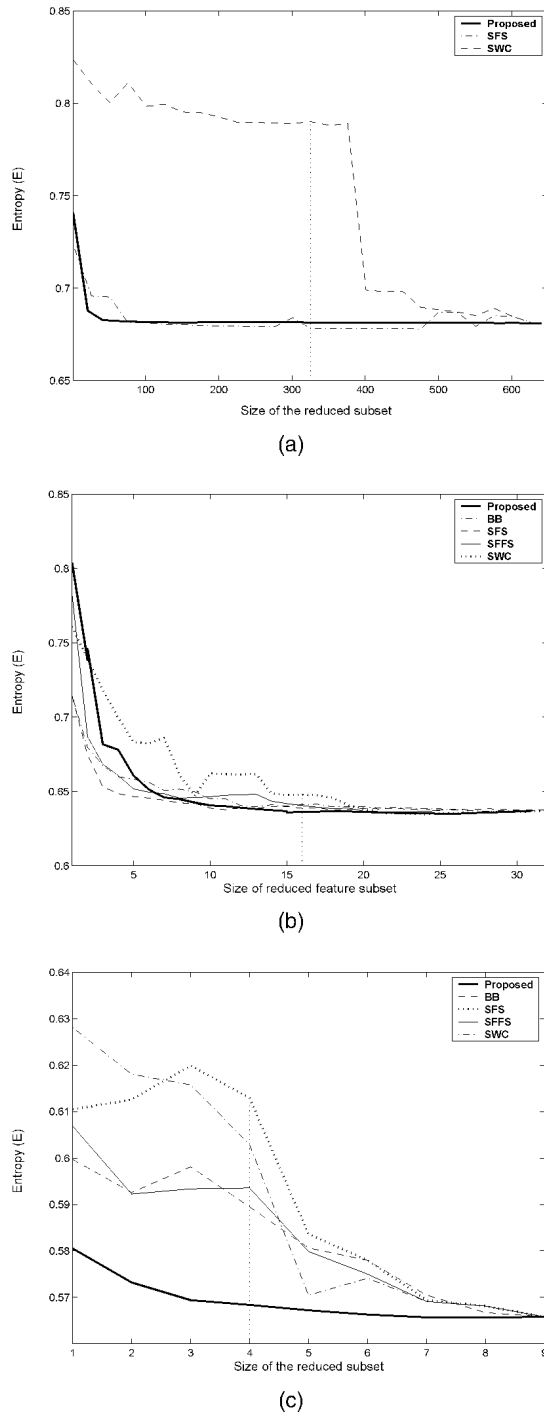


Fig. 3. Variation in classification accuracy with size of the reduced subset for—(a) Multiple features, (b) ionosphere, and (c) cancer data sets. The vertical dotted line marks the point for which results are reported in Tables 1, 2, and 3.

proposed algorithm involves partitioning of the original feature set into a number of homogeneous (highly compressible) clusters, it is expected that representation entropy of the individual clusters are as low as possible, while that of the final reduced set of features has low redundancy, i.e., a high value of representation entropy.

It may be noted that among all the d -dimensional subspaces of an original D -dimensional data set, the one corresponding to the Karhunen-Loeve coordinates [2] (for the first d eigenvalues) has the highest representation entropy, i.e., is least redundant. However, for large-dimensional data sets K-L transform directions are difficult to compute. Also, K-L transform results, in general, transformed variables and not exact subsets of the original features.

5 EXPERIMENTAL RESULTS AND COMPARISONS

Organization of the experimental results is as follows: First, the characteristics of the nine data sets used are discussed briefly. Then, performance of the proposed algorithm in terms of the feature evaluation indices, discussed in Section 4, is compared with five other feature selection schemes. Next, we have studied the redundancy reduction aspect of the algorithm quantitatively along with comparisons. Effect of varying the parameter k , used in feature clustering, is also studied.

Three categories of real-life public domain data sets are used: low-dimensional ($D \leq 10$), medium-dimensional ($10 < D \leq 100$), and high-dimensional ($D > 100$), containing both large and relatively smaller number of points. They are available from the UCI Machine Learning Repository [23]. Their characteristics are described below.

1. *Isolet*. The data consists of several spectral coefficients of utterances of English alphabets by 150 subjects. There are 617 features all real in the range $[0, 1]$, 7,797 instances, and 26 classes.
2. *Multiple Features*. This data set consists of features of handwritten numerals ("0"- "9") extracted from a collection of Dutch utility maps. There are total 2,000 patterns, 649 features, and 10 classes.
3. *Arrhythmia*. It contains 452 samples, each having 279 attributes. Among the attributes 195 are real valued and are used for our experiments. The attributes represent parameters of ECG measurements and the task is to classify a patient into one of the 16 classes of cardiac Arrhythmia.
4. *Spambase*. The task is to classify an email into spam or nonspam category. There are 4,601 instances, 57 continuous valued attributes denoting word frequencies, and 2 classes.
5. *Waveform*. This consists of 5,000 instances having 40 attributes each. The attributes are continuous valued, and some of them are noise. The task is to classify an instance into one of the three categories of waves.
6. *Ionosphere*. The data represents autocorrelation functions of radar measurements. The task is to classify them into two classes denoting passage or obstruction in ionosphere. There are 351 instances and 34 attributes, all continuous.
7. *Forest Cover Type*. This is a GIS data set representing the forest cover type of a region. There are 54 attributes out of which we select 10 numeric valued attributes. There are 581,012 instances and eight classes.

TABLE 4

Comparison of Feature Selection Algorithms for Large Data Sets when Search Algorithms use FFEI as the Selection Criterion

Dataset	Method	Evaluation Criteria						CPU Time (sec)	
		FFEI	E	S	KNNA (%)		BayesA (%)		
					Mean	SD	Mean		SD
Isolet d=310, D=617	SFS	0.40	0.54	0.98	95.81	0.82	92.19	0.72	28.01×10^4
	Proposed	0.40	0.50	1.07	96.00	0.78	95.01	0.52	440
Mult. Feat. d=325, D=649	SFS	0.44	0.67	0.44	77.71	0.44	75.81	0.17	9.20×10^4
	Proposed	0.48	0.68	0.45	78.34	0.22	75.28	0.10	451
Arrhythmia d=100, D=195	SFS	0.40	0.77	0.21	53.22	0.59	52.25	0.44	2008
	Proposed	0.40	0.72	0.17	58.93	0.54	56.00	0.41	74
Forest d=5, D=10	BB	0.40	0.65	0.90	64.03	0.41	63.55	0.40	9.21×10^4
	SFFS	0.40	0.66	0.83	67.01	0.45	66.00	0.44	7.52×10^4
	SFS	0.43	0.66	1.01	61.41	0.44	60.01	0.41	17.19×10^3
	Proposed	0.40	0.65	0.90	64.03	0.41	63.55	0.40	55.50

8. *Wisconsin Cancer*. The popular Wisconsin breast cancer data set contains nine features, 684 instances, and two classes.
9. *Iris*. The data set contains 150 instances, four features, and three classes of Iris flowers.

5.1 Comparison: Classification and Clustering Performance

Four indices, namely, entropy (2), fuzzy feature evaluation index (3), class separability (1), and K-NN and naive Bayes classification accuracy are considered to demonstrate the efficacy of the proposed methodology and for comparing it with other methods. Four unsupervised feature selection schemes considered for comparison are:

1. Branch and Bound Algorithm (BB) [2]. A search method in which all possible subsets are implicitly inspected without exhaustive search. If the feature selection criterion is monotonic BB returns the optimal subset.
2. Sequential Forward Search (SFS) [2]. A suboptimal search procedure where one feature at a time is added to the current feature set. At each stage, the feature to be included in the feature set is selected from among the remaining available features so that the new enlarged feature set yields a maximum value of the criterion function used.
3. Sequential Floating Forward Search (SFFS) [3]. A near optimal search procedure with lower computational cost than BB. It performs sequential forward search with provision for backtracking.
4. Stepwise Clustering (using correlation coefficient) (SWC) [21]. A nonsearch based scheme which obtains a reduced subset by discarding correlated features.

In our experiments, we have mainly used entropy (2) as the feature selection criterion with the first three search algorithms.

Comparisons in terms of five indices was made for different sizes of the reduced feature subsets. Tables 1, 2, and 3 provide such a comparative result corresponding to high, medium, and low-dimensional data sets when the size of the reduced feature subset is taken to be about half of the original size as an example. Comparison for other sizes of the reduced feature set is provided in Fig. 3 considering one data set from each of the three categories, namely, multiple features (high), ionosphere (medium), and cancer (low). The CPU time required by each of the algorithms on a Sun UltraSparc 350 MHz workstation are also reported in Tables 1, 2, and 3. Since the branch and bound (BB) and the sequential floating forward search (SFFS) algorithms require infeasibly high computation time for the large data sets, we could not provide the figures for them in Table 1. For the classification accuracies (using K-NN and Bayes), both mean and standard deviations (SD) computed for ten independent runs are presented.

Compared to the search-based algorithms (BB, SFFS, and SFS), the performance of the proposed scheme is comparable or slightly superior, while the computational time requirement is much less for the proposed scheme. On the other hand, compared to the similarity based SWC method the performance of the proposed algorithm is much superior, keeping the time requirement comparable. It is further noted that the superiority in terms of computational time increases as the dimensionality and sample size increase. For example, in the case of low-dimensional data sets the speedup factor of the proposed scheme compared to BB and SFFS algorithms is about 30-50, for Forest data

TABLE 5
Representation Entropy H_R^s of Subsets Selected using Some Algorithms

Dataset	BB	SFFS	SFS	SWC	Relief-F	Proposed
Isolet	-	-	2.91	2.87	2.89	3.50
Mult. Ftrs.	-	-	2.02	1.90	1.92	3.41
Arrhythmia	-	-	2.11	2.05	2.02	3.77
Spambase	2.02	1.90	1.70	1.44	1.72	2.71
Waveform	1.04	1.02	0.98	0.81	0.92	1.21
Ionosphere	1.71	1.71	1.70	0.91	1.52	1.81
Forest	0.91	0.82	0.82	0.77	0.91	0.91
Cancer	0.71	0.71	0.55	0.55	0.59	0.82
Iris	0.47	0.47	0.41	0.31	0.47	0.47

which is low-dimensional but has large sample size the factor is about 100, for medium-dimensional data sets, BB and SFFS are about 100 times slower and SFS about ten times slower, while for the high-dimensional data sets SFS is about 100 times slower, and BB and SFFS could not be compared as they require infeasibly high run time.

It may be noted that the aforesaid unsupervised feature selection algorithms (namely, BB, SFFS, SFS) usually consider “entropy” as the selection criterion. Keeping this in mind detailed results are provided in Tables 1, 2, and 3. However, we have also run the experiments using another unsupervised measure, namely, fuzzy feature evaluation index (FFEI) (3). Table 4 shows, as an illustration, the results only for the four large data sets (Isolet, Multiple Features,

Arrhythmia, and Forest Covertype). These results corroborate the findings obtained using entropy.

In a part of the experiments, we compared the performance with a supervised method Relief-F, which is widely used. We have used 50 percent of the samples as design set for the Relief-F algorithm. Results are presented in Tables 1, 2, and 3. The Relief-F algorithm provides classification performance comparable to the proposed scheme in spite of using class label information. Moreover, it has much higher time requirement, specially for data sets with large number of samples, e.g., the Forest data. Its performance in terms of the unsupervised indices is also poor.

Statistical significance of the classification performance of the proposed method compared to those of the other algorithms is tested. Means and SD values of the accuracies,

TABLE 6
Redundancy Reduction using Different Feature Similarity Measures

Dataset	Similarity Measure: λ_2		Similarity Measure: e		Similarity Measure: ρ	
	H_R^g	H_R^s	H_R^g	H_R^s	H_R^g	H_R^s
Isolet	0.001	3.50	0.007	3.01	0.003	3.41
Mult. Ftrs.	0.002	3.41	0.008	2.95	0.007	3.01
Arrhythmia	0.007	3.77	0.017	2.80	0.010	3.41
Spambase	0.04	2.71	0.07	2.01	0.05	2.53
Waveform	0.10	1.21	0.14	1.04	0.11	1.08
Ionosphere	0.05	1.81	0.07	1.54	0.07	1.54
Forest	0.10	0.91	0.17	0.82	0.11	0.91
Cancer	0.19	0.82	0.22	0.71	0.19	0.82
Iris	0.17	0.47	0.22	0.31	0.17	0.47

H_R^g : Average representation entropy of feature groups, H_R^s : representation entropy of selected subset, λ_2 : maximal information compression index, e : least-square regression error, and ρ : correlation coefficients.

computed over 10 independent runs, are used for this purpose. A generalized version of paired t -test suitable for both unequal means and variances is used. The above problem is the classical Behrens-Fisher problem in hypothesis testing, a suitable test statistic is described and tabled in [24] and [25], respectively.¹ It is observed that the proposed method has significantly better performance compared to the SWC algorithm for all the data sets, and the SFS algorithm for most of the data sets. For the other algorithms, namely Relief-F, BB, and SFFS, the performance is comparable, i.e., the difference of the mean values of the classification scores is statistically insignificant.

5.2 Redundancy Reduction: Quantitative Study

As mentioned before, the proposed algorithm involves partitioning the original feature set into certain number of homogeneous groups and then replacing each group by a single feature, thereby resulting in the reduced feature set. Representation entropy (H_R), defined in Section 4, is used to measure the redundancy in both the homogeneous clusters and the final selected feature subset. H_R when computed over the individual clusters should be as low as possible (indicating high redundancy among the features belonging to a single cluster), while giving as high value as possible for the selected subset (indicating minimum redundancy). Let us denote the average value of H_R computed over the homogeneous groups by H_R^g and the value of H_R for the final selected subset by H_R^s .

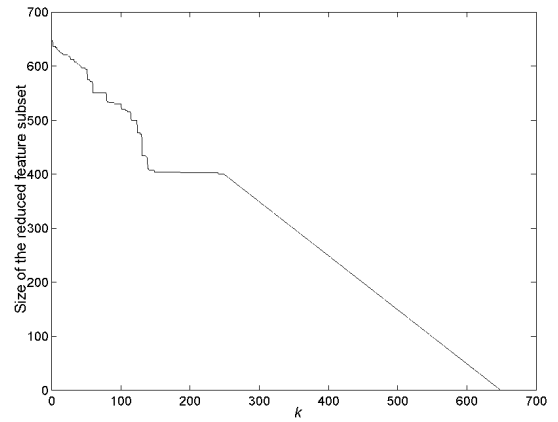
Table 5 shows the comparative results of the proposed method with other feature selection algorithms in terms of H_R^s . It is seen that the subset obtained by the proposed scheme is least redundant having the highest H_R^s values.

To demonstrate the superiority of the *maximal information compression index* λ_2 , compared to the other two feature similarity measures (ρ and e) used previously, we provide Table 6, where we have compared both H_R^s and H_R^g values obtained using each of the similarity measures, in our clustering algorithm. It is seen from Table 6 that, λ_2 has superior information compression capability compared to the other two measures as indicated by the lowest and highest values of H_R^g and H_R^s , respectively.

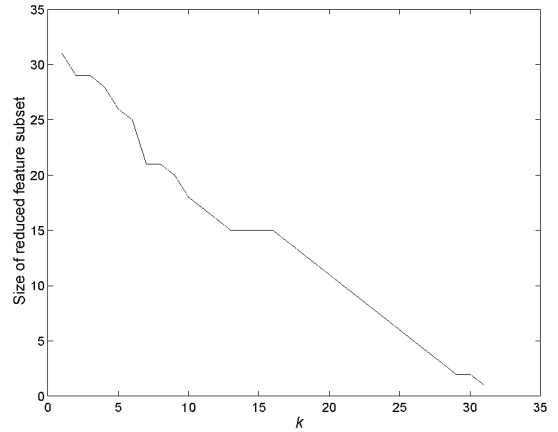
5.3 Effect of Parameter k

In our algorithm, the size of the reduced feature subset and hence, the scale of details of data representation is controlled by the parameter k . Fig. 4 illustrates such an effect for three data sets—multiple features, ionosphere, and cancer, considering one data from each of the high, medium, and low categories. As expected, the size of the reduced subset decreases overall with increase in k . However, for medium and particularly large-dimensional data (Fig. 4a), it is observed that for certain ranges of k at the

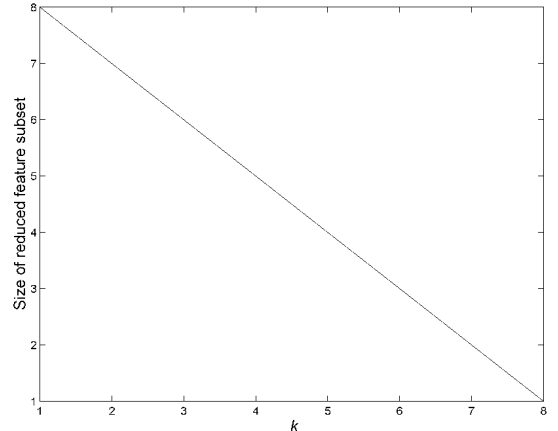
1. The test statistic is of the form $v = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\lambda_1 s_1^2 + \lambda_2 s_2^2}}$, where \bar{x}_1, \bar{x}_2 are the means, s_1, s_2 the standard deviations, and $\lambda_1 = 1/n_1, \lambda_2 = 1/n_2, n_1, n_2$ are the number of observations.



(a)



(b)



(c)

Fig. 4. Variation in size of the reduced subset with parameter k for— (a) multiple features, (b) ionosphere, and (c) cancer data.

lower side, there is no change in the size of the reduced subset, i.e., no reduction in dimension occurs. Another interesting fact observed in all the data sets considered is that, for all values of k in the case of small dimensional data sets, and for high values of k in the case of medium and large-dimensional data sets, the size of the selected subset varies linearly with k . Further, it is seen in those cases,

$d + k \approx D$, where d is the size of the reduced subset and D is the size of the original feature set.

6 CONCLUSION AND DISCUSSIONS

An algorithm for unsupervised feature selection using feature similarity measures is described. The novelty of the scheme, as compared to other conventional feature selection algorithms, is the absence of search process which contributes to the high-computational time requirement of those feature selection algorithms. Our algorithm is based on pairwise feature similarity measures, which are fast to compute. It is found to require several orders less CPU time compared to other schemes. Unlike other approaches, which are based on optimizing either classification or clustering performance explicitly, here we determine a set of maximally independent features by discarding the redundant ones. This enhances the applicability of the resulting features to compression and other tasks like forecasting, summarization, and association mining in addition to classification/clustering. Another characteristics of the proposed algorithm is its capability for multiscale representation of data sets. The scale parameter k used for feature clustering efficiently parametrizes the tradeoff between representation accuracy and feature subset size. All these make it suitable for a wide variety of data mining tasks involving large (in terms of both dimension and size) data sets.

Besides formulating the novel clustering algorithm, we have defined a feature similarity measure called *maximal information compression index*. One may note that the definition of the said parameter is not new, it is its use in feature subset selection framework which is novel. The superiority of this measure for feature selection is established experimentally. It is also demonstrated through extensive experiments that *representation entropy* can be used as an index for quantifying both redundancy reduction and information loss in a feature selection method.

In the present article, we have measured the information loss in terms of second order statistics. The similarity measure used for the feature selection algorithm is selected/defined accordingly. One may modify these measures suitably in case even higher order statistics are used. In this regard, modifications of correlation indices [22] which measure higher order polynomial dependency between variables may be considered. Also, the similarity measure is valid only for numeric features; its extension to accommodate other kinds of variables (e.g., symbolic, categorical, hybrid) may also be investigated.

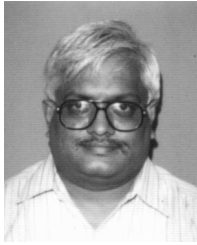
REFERENCES

- [1] U. Fayyad and R. Uthurusamy, "Data Mining and Knowledge Discovery in Databases," *Comm. ACM*, vol. 39, no. 11, pp. 24-27, Nov. 1996.
- [2] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs: Prentice Hall, 1982.
- [3] P. Pudil, J. Novovicová, and J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters*, vol. 15, pp. 1119-1125, 1994.
- [4] D.W. Aha and R.L. Bankert, "A Comparative Evaluation of Sequential Feature Selection Algorithms," *Artificial Intelligence and Statistics V*, D. Fisher and J.-H. Lenz, eds., New York: Springer Verlag, 1996.
- [5] *Genetic Algorithms for Pattern Recognition*, S.K. Pal and, P.P. Wang, eds. Boca Raton: CRC Press, 1996.
- [6] M. Kudo and J. Sklansky, "Comparison of Algorithms that Selects Features for Pattern Classifiers," *Pattern Recognition*, vol. 33, pp. 25-41, 2000.
- [7] D. Skalak, "Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms," *Proc. 11th Int'l. Machine Learning Conf.*, pp. 293-301, 1994.
- [8] A.W. Moore and M.S. Lee, "Efficient Algorithms for Minimizing Cross Validation Error," *Proc. 11th Int'l. Conf. Machine Learning*, 1994.
- [9] H. Liu and R. Setiono, "Some Issues in Scalable Feature Selection," *Expert Systems with Applications*, vol. 15, pp. 333-339, 1998.
- [10] M. Dash and H. Liu, "Unsupervised Feature Selection," *Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining*, pp. 110-121, 2000.
- [11] J. Dy and C. Brodley, "Feature Subset Selection and Order Identification for Unsupervised Learning," *Proc. 17th Int'l. Conf. Machine Learning*, 2000.
- [12] S. Basu, C.A. Micchelli, and P. Olsen, "Maximum Entropy and Maximum Likelihood Criteria for Feature Selection from Multivariate Data," *Proc. IEEE Int'l. Symp. Circuits and Systems*, pp. III-267-270, 2000.
- [13] S.K. Pal, R.K. De, and J. Basak, "Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach," *IEEE Trans. Neural Network*, vol. 11, pp. 366-376, 2000.
- [14] M.A. Hall, "Correlation Based Feature Selection for Discrete and Numeric Class Machine Learning," *Proc. 17th Int'l. Conf. Machine Learning*, 2000.
- [15] R.P. Heydorn, "Redundancy in Feature Extraction," *IEEE Trans. Computers*, pp. 1051-1054, 1971.
- [16] S.K. Das, "Feature Selection with a Linear Dependence Measure," *IEEE Trans. Computers*, pp. 1106-1109, 1971.
- [17] G.T. Toussaint and T.R. Vilmansen, "Comments on Feature Selection with a Linear Dependence Measure," *IEEE Trans. Computers*, p. 408, 1972.
- [18] K. Kira and L. Rendell, "A Practical Approach to Feature Selection," *Proc. Ninth Int'l. Workshop Machine Learning*, pp. 249-256, 1992.
- [19] I. Kononenko, "Estimating Attributes: Analysis and Extension of Relief," *Proc. Seventh European Machine Learning Conf.*, pp. 171-182, 1994.
- [20] D. Koller and M. Sahami, "Towards Optimal Feature Selection," *Proc. 13th Int'l. Conf. Machine Learning*, pp. 284-292, 1996.
- [21] B. King, "Step-Wise Clustering Procedures," *J. Am. Statistical Assoc.*, pp. 86-101, 1967.
- [22] C.R. Rao, *Linear Statistical Inference and Its Applications*. John Wiley, 1973.
- [23] C.L. Blake and C.J. Merz, *UCI Repository of Machine Learning Databases*, Univ. of California, Irvine, Dept. of Information and Computer Sciences, <http://www.ics.uci.edu/mllearn/MLRepository.html>, 1998.
- [24] E.L. Lehmann, *Testing of Statistical Hypotheses*. New York: John Wiley, 1976.
- [25] A. Aspin, "Tables for Use in Comparisons Whose Accuracy Involves Two Variances," *Biometrika*, vol. 36, pp. 245-271, 1949.



Pabiitra Mitra received the BTech degree in electrical engineering from the Indian Institute of Technology, Kharagpur, in 1996. He worked as a scientist with the Institute for Robotics and Intelligent Systems, India. Currently, he is a senior research fellow at the Indian Statistical Institute, Calcutta. His research interests are in the area of data mining and knowledge discovery, pattern recognition, learning theory, and soft computing. He is a student member of the

IEEE.



C.A. Murthy received the BStat (Hons), MStat, and PhD degrees from the Indian Statistical Institute (ISI). He visited Michigan State University, East Lansing, in 1991-92, and the Pennsylvania State University, University Park, in 1996-97. He is a professor in the Machine Intelligence Unit of ISI. His fields of interest include pattern recognition, image processing, machine learning, neural networks, fractals, genetic algorithms, wavelets, and data mining. He received the best paper award in 1996 in

computer science from the Institute of Engineers, India. He is a fellow of the National Academy of Engineering, India.



Sankar K. Pal (M '81-SM '84-F '93) (04230363) received the MTech and PhD degrees in radio physics and electronics in 1974 and 1979, respectively, from the University of Calcutta. In 1982, he received another PhD degree in electrical engineering along with DIC from Imperial College, University of London. He is a professor and distinguished scientist at the Indian Statistical Institute, Calcutta. He is also the founding head of Machine Intelligence Unit.

He worked at the University of California, Berkeley, and the University of Maryland, College Park, during 1986-87 as a Fulbright Postdoctoral Visiting Fellow, at the NASA Johnson Space Center, Houston, Texas during 1990-92, and, in 1994, as a guest investigator under the NRC-NASA Senior Research Associateship program; and at the Hong Kong Polytechnic University, Hong Kong, in 1999 as a visiting professor. He served as a distinguished visitor of the IEEE Computer Society (USA) for the Asia-Pacific Region during 1997-99. He is a fellow of the IEEE, Third World Academy of Sciences, Italy, and all the four National Academies for Science/Engineering in India. His research interests include pattern recognition, image processing, data mining, soft computing, neural nets, genetic algorithms, and fuzzy systems. He is a coauthor/coeditor of eight books including *Fuzzy Mathematical Approach to Pattern Recognition*, John Wiley (Halsted), New York, 1986, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*, John Wiley, New York 1999, and has published more than 300 research publications. He has received the 1990 S.S. Bhatnagar Prize (which is the most coveted award for a scientist in India), 1993 Jawaharlal Nehru Fellowship, 1993 Vikram Sarabhai Research Award, 1993 NASA Tech Brief Award, 1994 IEEE Transactions on Neural Networks Outstanding Paper Award, 1995 NASA Patent Application Award, 1997 IETE-Ram Lal Wadhwa Gold Medal, 1998 Om Bhasin Foundation Award, 1999 G.D. Birla Award for Scientific Research, the 2000 Khwarizmi International Award (1st winner) from the Islamic Republic of Iran, and the 2001 Syed Husain Zaheer Medal from Indian National Science Academy. Professor Pal has been an associate editor for *IEEE Transactions on Neural Networks* (1994-98), *Pattern Recognition Letters*, *International Journal Pattern Recognition and Artificial Intelligence*, *Neurocomputing*, *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, and *Fundamenta Informaticae*. He is a member of the Executive Advisory Editorial Board, *IEEE Transactions Fuzzy Systems*, *International Journal on Image and Graphics*, and *International Journal of Approximate Reasoning*, and a guest editor of many journals including *IEEE Computer*.

▷ For more information on this or any computing topic, please visit our Digital Library at <http://computer.org/publications/dilib>.