Design Choices and Theoretical Issues for Relative Feature Importance, a Metric for Nonparametric Discriminatory Power

Hilary J. Holz[1,2] and Murray H. Loew[1]

[1]Department of Electrical and Computer Engineering, The George Washington University, Washington, DC, 20002
[2](Corresponding Author): Phone: (301) 983-8652, Fax: (301) 983-4899
hholz@seas.gwu.edu

Abstract. We have developed *relative feature importance* (RFI), a metric for the classifier-independent ranking of features. Previously, we have shown the metric to rank accurately features for a wide variety of artificial and natural problems, for both two-class and multi-class problems. In this paper, we present the design of the metric, including both theoretical considerations and statistical analysis of the possible components.

Keywords: discriminatory power, feature selection, feature extraction, feature analysis, non-parametric, classifier-independent, relative feature importance, multi-class

1 Classifier-Independent Feature Analysis

In all feature analysis problems some initial set of *candidate features* must be identified. The candidate features are the result of some external analysis or search process. Feature analysis techniques analyze the usefulness of the candidate features. They cannot guarantee that there does not exist an as yet undiscovered feature which may be more useful. They also cannot guarantee that classification error overall could not be reduced using features not in the candidate feature set.

Since classifier-independent feature analysis is based on the structure of the data, features can be analyzed only on the basis of a *learning sample.* The learning sample is a set of correctly classified objects that are represented by feature values for the features in the candidate feature set. Since classifier-independent feature analysis is driven by the learning sample, a high degree of confidence in the learning sample is important.

The learning sample is taken as baseline truth, therefore the classes represented in a learning sample are necessarily *collectively exhaustive*. The problem of missing

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 696–705, 2000.

classes (sometimes called new class discovery [1]) is a separate problem from the feature analysis problem. In new class discovery the *features* are assumed to represent the objects accurately, and are used to explore the structure of the *logical* space. In feature analysis the *classes* are assumed to be known accurately, and are used to explore the structure of the *feature* space. Thus, the classes can be assumed to be collectively exhaustive without significant loss of generality. In contrast, a significant loss in generality does result from assuming that the classes are *mutually exclusive*. In medical diagnosis, for example, assuming in the general case that a patient can have at most one pathology is unrealistic.

The goal of classifier-independent feature analysis for classification is to measure the usefulness of the features in the candidate feature set. Nonetheless, classification performance on the learning sample cannot be used in and of itself as a basis for analyzing the features for several reasons. First, it has been shown that, in the general case, features that optimize classification performance for one classifier may not perform at all well in another classifier [2]. More fundamentally, though, classifier-independent feature analysis tries to measure the *potential* for discrimination between classes of the features in the candidate feature set, which potential may not be realizable in practice.

Once classification performance has been eliminated as a measure of usefulness, what remains is the separability between the classes. Separability is not subject to the theoretical constraints of classification performance. When expressed as Bayes error, the separation between class-conditional joint feature distributions places a lower bound on classification error that is classifier-independent. Unfortunately, Bayesian error is not calculable for many problems. Nonetheless, separation between class-conditional joint feature distributions gives rise to the potential for classification. Issues of calculation aside, classifier-independent feature analysis uses separability between classes as the basis for the usefulness of a feature.

A theoretical constraint placed on feature analysis is that feature rankings are subset dependent. Even under the assumption of feature independence, feature rankings can change as a function of adding and removing features [3]. Nonetheless, ranking the features is a critical component of feature analysis: in medical diagnosis, when test results are ambiguous, the physician needs guidance as to their relative value for discrimination. Therefore, ranking is given within a subset, with the critical ranking being that within the optimal subset. The optimal subset of the candidate feature set is defined as the smallest subset with the maximum potential for separability between classes.

2 Measuring Separation: Discriminant Analysis

Discriminant analysis can be used to extract features that maximize the ratio of the separation between classes to the spread within classes, as measured by the between-class and within-class scatter matrices. Within-class scatter is a measure of the scatter of a class relative to its own mean. Between-class scatter is a measure of the distance from each class to the mean(s) of the other classes. Within-class and between-class scatter can be defined parametrically or non-parametrically. Parametric scatter matrices use the learning sample to estimate the distributions of the features through estimation of parameters for an assumed distributional structure. Non-parametric scatter

matrices use the learning sample to perform local density estimation around individual samples, and then measures scatter using the local density estimates.

The parametric versions of the within-class and between-class scatter matrices estimate the means of the classes based on the entire learning sample. The parametric versions assume that a distribution can be characterized by its mean and covariance. Let P_i be the *a priori* probability of class ω_i, Σ_i be the covariance matrix and M_i be the mean of class ω_i, N be the total number of samples, and L be the number of classes present in the learning sample. Parametric within-class scatter is defined as the averaged covariance. The *a priori* probability is estimated from the learning sample as N_i/N, where N_i is the number of samples from ω_i. Σ_i is estimated by $\hat{\Sigma}_i$, the sample covariance matrix. Parametric between-class scatter is the scatter of the expected means around the mixture means.The components of the between-class scatter matrix are estimated using the learning sample in the same manner as the within-class scatter matrix.

RFI uses non-parametric versions of the scatter matrices based on versions proposed by Fukunaga and Mantock [4]. They based their non-parametric scatter estimates on local density estimates using the *k*-nearest neighbors (*kNN*) technique. They defined the ω_i *local mean* for a given class ω_i and a given sample $\bar{x}_j$ as

$$\mathcal{M}_i^k(\bar{x}_j) = \frac{1}{k} \sum_{q=1}^{k} \bar{x}_{qNN}^{(i)} \tag{1}$$

where $\bar{x}_{qNN}^{(i)}$ is the *qth*-nearest-neighbor in ω_i. Because Fukunaga and Mantock experimented only with two-class problems, they could use the ω_i-local mean for calculating both within- and between-class scatter.

While use of the local mean introduces the parameter *k*, its behavior is well studied. With infinite sample size, the accuracy of the local density estimation improves as *k* increases. With finite sample size, *k* is subject to the problem of oversampling, otherwise known as Hughes phenomenon [5]. A value of *k* which is too large for the sample size performs local density estimation on non-local samples! A value of *k* which is too small for the sample size reduces the accuracy of the local density estimation. In practice, *k* is generally set to a small fraction of the number of samples [6].

To generalize Fukunaga and Mantock's approach to more than two classes, the *local out-of-class mixture mean* for each sample $\bar{x}_j^{(i)}$ is defined as

$$\mathcal{M}_{r \neq i}^k\left(\bar{x}_j^{(i)}\right) = \frac{1}{k} \sum_{q=1}^{k} \bar{x}_{qNN}^{(r \neq i)} \tag{2}$$

where $\bar{x}_{qNN}^{(r \neq i)}$ is the *qth*-nearest-neighbor outside of ω_i. The local mixture mean differs from the parametric mixture mean in that it excludes data from a sample's own class.

Non-parametric within-class scatter is defined as the averaged scatter, where scatter is around the local means:

$$S_w = \sum_{i=1}^{L} P_i \sum_{j=1}^{N_i} (\bar{x}_j^{(i)} - \mathcal{M}_i^k(\bar{x}_j^{(i)}))(\bar{x}_j^{(i)} - \mathcal{M}_i^k(\bar{x}_j^{(i)}))^T \quad (3)$$

When $k = N_i$, the local mean reduces to the parametric mean, and therefore the non-parametric within-class scatter matrix reduces to the parametric version. Non-parametric between-class scatter is measured as the scatter around the out-of-class mixture means:

$$S_b = \sum_{i=1}^{L} P_i \sum_{j=1}^{N_i} (\bar{x}_j^{(i)} - \mathcal{M}_{r \neq i}^k(\bar{x}_j^{(i)}))(\bar{x}_j^{(i)} - \mathcal{M}_{r \neq i}^k(\bar{x}_j^{(i)}))^T \quad (4)$$

The between-class non-parametric scatter matrix does not reduce to its parametric form as does the within-class, because the out-of-class mixture means necessarily exclude same class samples, but the relationship is close when $k = N_i$.

The use of the k-nearest-neighbor local density estimates introduces the need to choose a distance metric for determining the distance between a sample and its neighbors. Many distance measures have been proposed for use with kNN error estimation [7]. Two commonly used metrics are the Euclidean distance and the Mahalanobis distance [6]. Fukunaga and Mantock used Euclidean distance in their original work. Mahalanobis distance should also be considered (especially using Fukunaga and Mantock's original algorithm), since it incorporates information concerning the relative variance of the features.

A further refinement introduced by Fukunaga and Mantock was the use of a weighting factor, w_j, to de-emphasize samples which lie far away from the classification boundary. RFI uses the natural multi-class extension of Fukunaga and Mantock's weighting factor as given in [8]. Using the weighting factor, the contribution of each $\bar{x}_j$ to scatter is inversely proportional to its distance from the nearest classification boundary.

Thus, the final forms for non-parametric within-class and between-class scatter are (estimating components as necessary using the learning sample):

$$S_w = \sum_{i=1}^{L} \frac{1}{N} \sum_{j=1}^{N_i} w_j(\bar{x}_j^{(i)} - \mathcal{M}_i^k(\bar{x}_j^{(i)}))(\bar{x}_j^{(i)} - \mathcal{M}_i^k(\bar{x}_j^{(i)}))^T \quad (5)$$

and

$$S_b = \sum_{i=1}^{L} \frac{1}{N} \sum_{j=1}^{N_i} w_j(\bar{x}_j^{(i)} - \mathcal{M}_{r \neq i}^k(\bar{x}_j^{(i)}))(\bar{x}_j^{(i)} - \mathcal{M}_{r \neq i}^k(\bar{x}_j^{(i)}))^T \quad (6)$$

3 Theoretical Implications

The optimal extracted features are found by eigensystem decomposition of the ratio of the between-class to within-class scatter matrices. Specifically, the optimality criterion

used is the trace:

$$J = tr(S_w^{-1} S_b) \tag{7}$$

Thus, for both the parametric and the non-parametric forms, the eigenvectors form the linear transform which maximizes J, the ratio of the between-class to within-class scatter. The eigenvalues measure the amount of separation induced in the extracted space. The extracted features are optimal in the sense that they maximize separation between the class-conditional joint feature distributions in the rotated space.

Using the non-parametric scatter matrices, feature extraction is based on local density estimation. Thus the results are a compromise between information provided in the various clusters or regions belonging to a class [8].

While it is not possible to define the class of problems for which the non-parametric scatter matrices accurately capture the discriminatory power of the features, it is nevertheless possible to characterize those problems which are pathological. A problem under consideration can then be compared to the pathological problems in an attempt to determine the suitability of RFI for the problem.

One class of problems which are pathological for RFI, regardless of the use of parametric or non-parametric scatter, are problems which violate the assumption that proximity in feature space can be used to determine class membership. These problems are problems which k-nearest neighbor classifiers cannot solve. Figure 1 (a) illustrates one

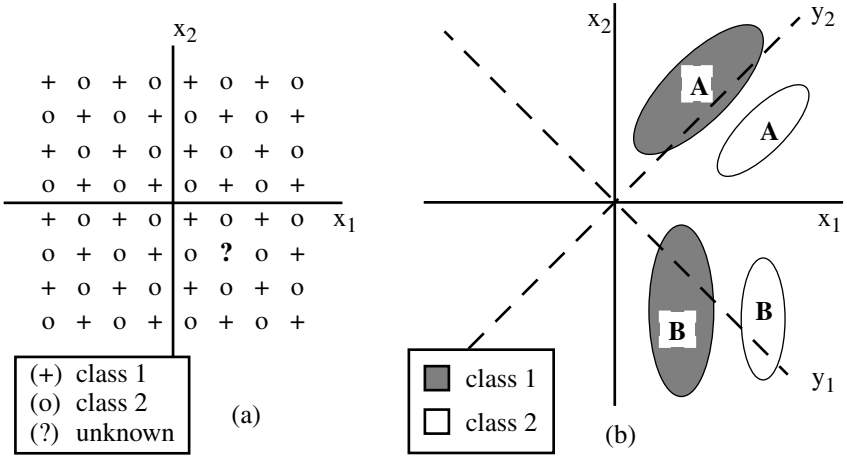

Figure 1: Two pathological problems for RFI.

such classical problem, the checkerboard. Any problem which is a pathological problem for k-nearest neighbor density estimation is a pathological problem for RFI.

A second class of pathological problems derives from the combining of information from multiple clusters or regions in the non-parametric scatter matrices. By constructing a problem wherein the transformations necessary to optimize the ratio of between-class to within-class scatter conflict, the non-parametric scatter matrices' ability to

combine local information can be exploited as a weakness. Figure 1 (b) illustrates such a problem. Note, however, that the parametric scatter matrices would do no better with the problems in Figure 1.

Because non-parametric discriminatory power measures the potential of the features for inducing separability between classes, it is desirable that measures of non-parametric discriminatory power be invariant with regard to rotation, scaling, and shift of features. Rotational and shift invariance eliminate the impact of irrelevant details of the measurement method for the features. Scale invariance eliminates the need for normalization of the features while preserving the critical information of the ratio of between-class to within-class scatter.

RFI is a function of the eigenvalues and eigenvectors of the parametric and non-parametric scatter matrices. While the non-parametric scatter matrices are not as well understood as the parametric scatter matrices, the non-parametric forms are still symmetric, as can be seen by observation of equations 5 and 6. Therefore, functions of eigenvectors and eigenvalues retain the same properties for both parametric and non-parametric scatter matrices.

Rotational invariance results from the extraction technique; since the optimal features are extracted from the original features, rotation in the original feature space has no impact. Scale invariance results from the use of the ratio of between-class to within-class scatter; since both within-class and between-class scatter are equally affected by scaling a feature, the ratio removes the effects of scaling. Shift invariance results from the use of scatter around the means, therefore the technique is self-centering.

All three forms of invariance reduce to the issue of preserving class separability, which is invariant under any nonsingular transformation (including rotation, scaling, and shift) [9]. Those transformations affect separability in the individual features (*i.e.,* in the marginal feature distributions), but not between the classes themselves. Thus, so long as none of the extracted features is discarded, RFI is invariant.

4 Finding the Optimal Subset

The optimal subset of features is the smallest subset with the maximum potential for separability between classes. RFI extracts a set of *optimal* features from a set of *original* features, without the use of classifier-specific assumptions. The optimal subset of the original features can be found by maximizing the separation induced between the class-conditional joint feature distributions across all possible subsets of the original features, as measured using the optimal extracted features. The optimal subset of features is thus the smallest subset of original features which produces the maximum separation, measured in the rotated space.

Given the presence of redundant features, more than one subset of the same size may produce the same amount of separation. When two or more smallest subsets produce the same amount of separation, and that separation is the maximum separation found, then more than one optimal subset exists. The presence of more than one optimal subset is not a problem; in both assisted and automatic classification, it offers more options in the design of the classification system.

The criteria commonly used in parametric discriminant analysis to find the optimal subset of features are not appropriate for the non-parametric case. Criteria such as the

trace of the ratio of the between-class to within-class scatter matrices are based on the same simplifying assumptions as the parametric scatter matrices. The trace, when calculated on parametric scatter matrices, is monotonic as a function of subset size, reflecting the theoretical assumption that Bayes error also decreases monotonically as a function of subset size.

Under conditions of limited sample size, the monotonicity assumption does not hold even for well-behaved data sets with unimodal Gaussian distributions, if the true distributions are not known and must be estimated. As the number of features increases for a fixed sample size, so does the error in the estimation. A second concern is the cost of including each feature, in computer time, in complexity, and sometimes in human suffering, as can be the case in medical diagnosis. In practice, whether for automatic classification or assisted classification, more features is not always better.

A non-parametric approach is to select the optimal subset based on the k-nearest-neighbor error *in the extracted space*. Because *kNN* error is based solely on proximity in feature space, it does not introduce any new classifier-specific assumptions. Moreover, because *kNN* error is asymptotically at most twice the Bayes error, calculating *kNN* error in the *extracted* space estimates the theoretical lower limit on the *potential* classification error [10]. Using *kNN* introduces a new parameter (k, the number of nearest neighbors used to calculate e_{kNN}). Fortunately, as discussed in Section 2, the behavior of k is well understood.

Finding the optimal subset requires exhaustive search, since any non-exhaustive technique can do arbitrarily poorly in the general case [11]. The assumption of monotonicity, necessary for branch-and-bound algorithms to guarantee performance, is extremely restrictive, and rarely justified in real problems [12]. Whenever possible, exhaustive search should be done. For the purposes of evaluating different configurations of RFI, or for comparing estimators for non-parametric discriminatory power, exhaustive search is required. When applying RFI directly to real problems which are too large to execute exhaustive search, sub-optimal techniques must be used.

Since the criterion, J, used by RFI to estimate the inherent Bayes error in each feature subset is a random variable, it is necessary to determine statistically whether the difference between separation in the subsets is due to the variance in the learning sample or the effects of the subsets. An analysis-of-variance (ANOVA) is performed on the results from multiple data sets. Each subset is considered a different treatment for the purpose of the ANOVA.

Calculating J for all possible subsets for each data set reduces the noise in the experiment. Each data set is a *block* in a block ANOVA. Calculation of J for each subset on a particular data set constitutes the experimental units within that block. The use of blocking reduces the noise in the data by reducing the number of data sets for the same number of experimental units. Since RFI does not carry over any information from one treatment to the next, the concept of *order* in applying the treatments is meaningless, and can be considered to be random. Thus, the model used by RFI is randomized block ANOVA.

A sensitivity analysis was performed to measure the impact of the algorithmic variations of Sections 2 and 3 on the ability of RFI to find the optimal subset. The problem chosen, (see Table 1) has multiple clusters, mixed distributions, a noise feature, and

Table 1. Sensitivity Analysis Problem

Feature	Bayes error (B,C)	Class 1		Class 2		Rank
		Cluster A	Cluster B	Cluster C	Cluster D	
1	37.5%	U[-3.0, -2.0]	U[-1.0, 0.0]	U[-0.75, 0.25]	U[4.0, 5.0]	1
2	10%	U[-3.0, -2.0]	U[-0.5, 0.5]	U[0.3, 1.3]	U[4.0, 5.0]	3
3	25%	U[-3.0, -2.0]	U[-1.5, -0.5]	U[-1.0, 0.0]	U[4.0, 5.0]	2
4	noise	N (0,1)	N (0,1)	N (0,1)	N (0,1)	0

three different ranks of non-noise features. Despite its complexity, the sensitivity analysis problem can still be solved using *600* samples per cluster, or *2400* samples in all.

A full factorial design was used, with two levels per factor. The coding chart for the experiments is given in Table 2. Four design points found the correct subset: non-parametric within-class scatter with euclidean distance, using either parametric or non-parametric between-class scatter and either setting for k.

Table 2. Coding chart for Sensitivity Analysis Part 1

Factor	-	+
Within-class scatter	Parametric	Non-parametric
Between-class scatter	Parametric	Non-parametric
Distance measure	Mahalanobis	Euclidean
k value	1	5

5 Ranking the Features

RFI ranks features based on the contribution of the original features to the separation in the rotated space. The contribution of the original features to the separation in the extracted space can be estimated using the eigenvectors and eigenvalues of the optimal transformation. The magnitudes of the eigenvalues measure the amount that each original feature contributes to each extracted feature. The normalized eigenvalues estimate the amount of separability contributed by each extracted feature to separation in the extracted space. Thus the normalized eigenvalues can be used to estimate separability in the rotated space, and the eigenvectors can be used to estimate the amount which each original features contributes to that separability.

The contribution of the original features to the separation in the extracted space can be estimated without tuning parameters by using the Weighted Absolute Weight Size (*WAWS*) of [14]. WAWS uses the normalized eigenvalues to measure the contributions of the original features to the extracted features by the proportion of separation the extracted features contribute to separation in the extracted space.

Features with statistically distinct WAWS values are given different ranks. To determine whether WAWS values are distinct, a second randomized block ANOVA is performed, and intervals constructed around the differences between treatment means using the multiple comparisons formula. Each feature is thus a treatment, and each data set (optimal subset only), a block.

Features with intervals around the differences from all other features are given dis-

tinct ranks. Groups of features in which some features have distinct WAWS values, but others do not, are given a single rank. Features not in the optimal subset have rank zero. Features (or groups of features) with distinct ranks are ranked based on their treatment means, with the largest distinct treatment mean being assigned the highest rank. Higher ranks indicate greater discriminatory power.

The sensitivity analysis was performed a second time to measure the impact of the design alternatives on the ability of RFI to correctly rank the features, given the optimal subset. Three design points ranked the features correctly (see Table 3). Non-para-

Table 3. Design points which correctly rank the features, given the optimal subset.

Within-class scatter	Between-class scatter	Distance measure	k value	Ranking Method
-	+	-	+	+
+	+	+	-	+
+	+	+	+	+

metric between-class scatter is clearly shown to be necessary. The setting for k is, again, shown to not be critical, as would be expected. Using parametric within-class scatter with Mahalanobis distance also ranks the features correctly, given the optimal subset. Thus, the use of Mahalanobis distance compensates to some degree for the information lost by the parametric scatter matrix

6 Complete Algorithm

In practice, RFI first finds the optimal subset, and then ranks the features within that subset. A final sensitivity analysis was performed, using the complete algorithm. Two configurations of RFI correctly solved the problem. The only factor that was not critical was the number of nearest neighbors. Within-class scatter had to be calculated non-parametrically using euclidean distance to find the correct subset. Between-class scatter had to be calculated non-parametrically to rank the features correctly. To correctly rank the features overall (assigning zero to features outside the optimal subset), both within-class and between-class scatter had to be calculated non-parametrically and, euclidean distances had to be used. Note that the insensitivity to k may have been due to the use of uniformly distributed signal values. Earlier research with Gaussian signal features has demonstrated greater sensitivity to k [15].

7 Conclusions and Future Research

A number of choices were considered and resolved in the design of RFI. RFI must use non-parametric scatter matrices for both within-class and between-class scatter, based on the results of the sensitivity analysis. RFI selects the optimal subset of the candidate features based on their potential for inducing class separability, thus, RFI uses *kNN* error in the *rotated* space to find the optimal subset. The choice of *kNN* error was made because it asymptotically approaches twice the Bayes error with increasing sample size. In addition, *kNN* error introduces no new assumptions, being based on the assumption that proximity in feature space can be used to determine class membership. RFI uses randomized block ANOVA to determine whether one subset (or set of sub-

sets) has statistically better class separability than the other subsets.

The design of RFI presented here has been shown to correctly rank features for a variety of two-class and multi-class artificial and natural data problems [8,14,16]. Planned enhancements of RFI include incorporation of cost information and categorical features in the *kNN* density estimation. In addition, the computational cost of the algorithm might be reduced through the application of such techniques as adaptive or edited kNN.

References

1 I. Chang and M.H. Loew, "Pattern Recognition with New Class Discovery," *Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Computer Vision,* pp. 438-443., 1991.

2 M. Ben-Bassat, "f-Entropies, Probability of Error, and Feature Selection," *Information and Control,* vol. 39, pp. 227-242, 1978.

3 T.M. Cover, "The Best Two Independent Measurements are Not the Two Best," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 4, no. 1, pp. 116-117, Jan.1974.

4 K. Fukunaga and J.M. Mantock, "Nonparametric Discriminant Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 5, no. 6, pp. 671-678, Nov. 1983.

5 G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory,* vol. IT-14, pp. 55-63, Jan. 1968.

6 R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis,* New York: Wiley & Sons, 1973.

7 R.D. Short and K. Fukunaga, "The Optimal Distance Measure for Nearest Neighbor Classification," *IEEE Transactions on Information Theory,* vol. IT-27, no. 5, pp. 622-627, Sept. 1981.

8 H.J. Holz and M.H. Loew, "Multi-class classifier-independent feature analysis," *Pattern Recognition Letters,* vol. 18, no. 11-13, pp. 1219-1224, Nov. 1997.

9 K. Fukunaga, *Introduction to Statistical Pattern Recognition: 2nd Edition,* Academic Press, Inc. 1990.

10 T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory,* vol. IT-13, pp. 21-27, Jan. 1967.

11 J.M. Van Campenhout, "The Arbitrary Relation Between Probability of Error and Measurement Subset," *Journal of the American Statistical Association,* vol. 75, no. 369, pp. 104-109, March 1980.

12 P.M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers,* vol. 26, no. 9, pp. 917-922, Sept. 1977.

13 N.L. Johnson and F.C. Leone, *Statistics and Experimental Design in Engineering and Physical Sciences,* pp. 614-661, New York: Wiley, 1977.

14 H.J. Holz and M.H. Loew, "Relative Feature Importance: A Classifier-Independent Approach to Feature Selection," *Pattern Recognition in Practice IV,* E.S. Gelsema and L.N. Kanal, eds, pp. 473-487, Elsevier Science B.V., 1994.

15 _____, "Non-Parametric Discriminatory Power," *Proceedings of the 1994 IEEE-IMS Workshop on Information Theory and Statistics,* pp. 65, Alexandria, VA, 1994.

16 _____, "Validation of Relative Feature Importance Using a Natural Data Set," *15th International Conference on Pattern Recognition,* Barcelona, Spain, September 2000.