Proceedings of the 1998 IEEE
ISIC/CIRA/ISAS Joint Conference
Gaithersburg, MD • September 14-17, 1998

CLASS-SPECIFIC FEATURE SETS IN CLASSIFICATION

Dr. Paul M. Baggenstoss

Naval Undersea Warfare Center
1176 Howell St
Newport RI, 02841
401-841-7505 x 38240 (TEL)
401-841-7453 (FAX)
p.m.baggenstoss@ieee.org (EMAIL)

ABSTRACT

The classical Bayesian approach to classification requires knowledge of the probability density function (PDF) of the data or sufficient statistic under all class hypotheses. Because it is difficult or impossible to obtain a single low-dimensional sufficient statistic, it is often necessary to utilize a sub-optimal yet still relatively high-dimensional feature set. The performance of such an approach is severely limited by the ability to estimate the PDF on a high-dimensiopnal space from training data. A new theorem shows that by introducing a special "noise-only" signal class (H0), it is possible to re-formulate the classical approach based upon M sufficient statistics, one corresponding to each signal class. Furthermore, the optimal classifier requires knowledge of only the PDF's of the sufficient statistics under the corresponding signal class and under noise-only condition. We present simulation results of a 9-class synthetic problem showing dramatic improvements over the traditional high - dimensional approach.

1. INTRODUCTION

In M-ary classification, one is often given the original data set $\mathbf{x}$, which is usually reduced to a set of statistics $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$, which we represent by $\{\mathbf{x}_i\}_{i=1}^M$. The optimal Baysian classifier is given by

$$\arg\max_j p(\{\mathbf{x}_i\}_{i=1}^M | H_j) p(H_j) \tag{1}$$

The problem with this often-used formulation is the following. Very often some of the features are chosen to be descriptive of a particular class. For example, if H_j was a narrowband data model, then it would stand to reason that one of the feature sets, say $\mathbf{x}_j$, ought to be based on a fourier analysis of the data $\mathbf{x}$. The data under hypothesis H_j may be based on a statistical model with a fairly small number of parameters which are often closely or loosely associated with a corresponding small set of features. It is then common practice to snatch defeat from the jaws of victory by lumping all these features together into a high-dimensional superset $\{\mathbf{x}_i\}_{i=1}^M$.

The complexity of the high-dimensional space rapidly exceeds our ability to estimate the distribution. The exponential increase in complexity of systems has been termed the *curse of dimensionality* by Richard Bellman [1]. In complex problems, there may be as many as a hundred separate measurements involved. This dimensionality is entirely unmanageable. It is recognized by a number of researchers that attempting to estimate PDF's nonparametrically above 5 dimensions is difficult and above 20 dimensions is futile [2].

We are motivated to find a classifier formulation based on using small sets of features separately, not lumped together. The following theorems shows one way to arrive there.

2. CLASS-SPECIFIC FORMULATION

The ideas of sufficient statistics [3] are not entirely new in classification [4],[5],[3]. However, up until now there has not been a general method for building an optimal classifier based on a non-homogenious set of features, that is feature sets selected separately for each data class, aside from the traditional method of treating the entire set together. This theorem fills this gap:

Theorem 1 *Let there be M hypotheses $H_1, \ldots, H_M$. Under each hypothesis, say H_j, let the data $\mathbf{x}$ be completely parameterized by a parameter set θ_j. Furthermore, for each class j, let there be a sufficient statistic for $\boldsymbol{\theta}_j$, $\mathbf{x}_j = T_j(\mathbf{x})$. Let the span of θ_j include a null case θ_j^0 which corresponds to signal not present. Because each hypothesis contains this equivalent case, we have*

$$\begin{aligned} p(\{\mathbf{x}_i\}_{i=1}^M | H_1, \boldsymbol{\theta}_1^0) &= p(\{\mathbf{x}_i\}_{i=1}^M | H_2, \boldsymbol{\theta}_2^0) \\ &= \cdots \\ &= p(\{\mathbf{x}_i\}_{i=1}^M | H_M, \boldsymbol{\theta}_M^0) \end{aligned} \tag{2}$$

Then, the optimum Bayes classifier reduces to

$$\arg\max_j \frac{p(\mathbf{x}_j | H_j)}{p(\mathbf{x}_j | H_j, \boldsymbol{\theta}_j = \boldsymbol{\theta}_j^0)} p(H_j) \tag{3}$$

Note that this formulation uses only low-dimensional distributions. The proof is provided in the appendix.

This result shows that if $\{\mathbf{x}_i\}$ are sufficient for the parameterizations of corresponding class, then the optimum Bayes classifier reduces to a classifier based only on the low-dimensional distributions. This is very important in the context of high-dimensional problems.

Note that class H_0 needs to be accessible from all classes through the parameter set. The only natural class to use would be the noise-only class. This has a distinct advantage because the likelihood ratios in (3) can be thresholded in order to reject all the classes except H_0, a convenience when classifying weak-signal data.

This work supported by the Office of Naval Research

0-7803-4423-5/98 $10.00 © 1998 IEEE

3. PRACTICAL CONSIDERATIONS

To utilize (3), it is necessary to obtain estimates of $p(\mathbf{z}_j|H_k)$ for both $k = 0$ and $k = j$. For $k = j$, it is clear that exemplars of $\mathbf{z}_j$ from a training data set may be used to train a density estimate, for example using Gaussian Mixtures via the EM algorithm. Likewise, for $k = 0$, a large number of exemplars may be created under the noise-only assumption by simulation. However, a numerical problem arises for feature vectors which differ greatly from the noise-only hypothesis (i.e. high-SNR). Then, the denominator density $p(\mathbf{z}_j|H_0)$ will be outside its useful range in which it can approximate the density. We are left with several choices:

1. Obtain theoretical densities under H_0 by deriving them analytically. This is aided by the fact that the number of features is (hopefully) small and that H_0 is straight-forward (i.e. iid Gaussian noise).
2. Use large sample approximations based on central limit theorem, etc.
3. If analytic expressions are not available, it is often the case that analytic expressions for the characteristic function is available. then numerical solutions are possible.
4. Asymptotic analysis of the tail behavior may be possible if exact expressions are not available for $p(\mathbf{z}_j|H_0)$.
5. Approximations to $p(\mathbf{z}_j|H_0)$ are possible by perturbation analysis of the feature extraction algorithm $\mathbf{z}_j = T_j(\mathbf{x})$. This will identify the Jacobian of the transformation and allow numerical evaluation of the density of $\mathbf{z}_j$. If $T(\mathbf{x})$ is not 1:1, problems arise, but they are not insurmountable.

We now present an example in which both choices 1 and 3 are used.

4. 9-CLASS EXAMPLE

In this example, we consider 9 data classes denoted $H_1, \ldots, H_9$.

- Class H_0: Noise only
- Class H_1: Long Sinewave
- Class H_2: Medium Sinewave
- Class H_3: Short Sinewave
- Class H_4: Long Gaussian Signal
- Class H_5: Short Gaussian Signal
- Class H_6: Short Impulse Signal
- Class H_7: Long Impulse Signal
- Class H_8: Long Laplacian Distributed Noise
- Class H_9: Short Laplacian Distributed Noise

Examples of these signals are provided in Figure 1. For each case, the model includes an amplitude parameter, which is unknown, and possibly additional unknown nuisance parameters such as phase. For each model, we derive:

1. An exact or approximate sufficient statistic or maximal invariant for the binary hypothesis testing problem involving H_j vs. H_0, denoted $\mathbf{z}_j(\mathbf{x})$.
2. The distribution of $\mathbf{z}_j(\mathbf{x})$ under H_0. This is used to implement the denominator of (3),

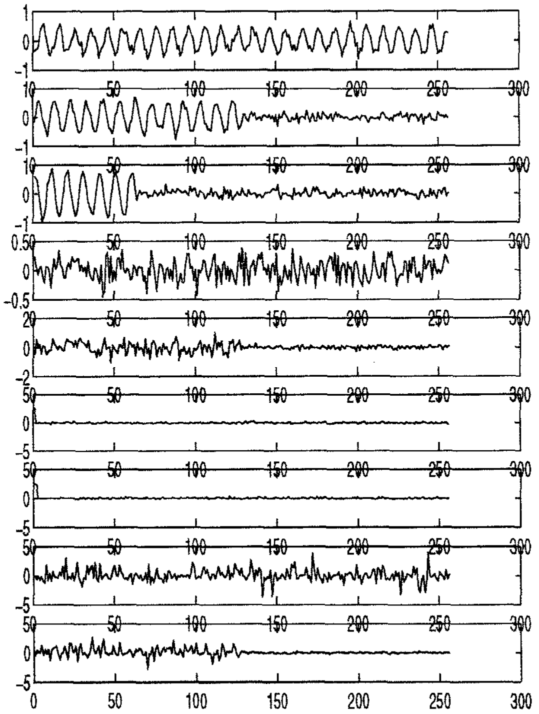

Figure 1: Examples of the nine signal types. Signal-to-Noise (SNR) has been increased for clarity. Actual SNR varies.

For brevity, these derivations could not be included. The statistics and their associated distributions under H_0 are tabulated below in Tables 1,2. Note that for signals H_8 and H_9, the given statistics are not sufficient and the distributions under H_0 are approximations based on the central limit theorem.

In many situations, the sufficient statistics for a model are still of relatively high dimension. In such cases it is convenient to apply the *principle of invariance*. A rich theory exists on the subject [6],[7]. The basic idea is that a natural symmetry exists in many problems that can be represented by a set of transformations. We want our statistic to have the desired symmetry (invariance to the transformations) while at the same time exhibiting the maximum information. The *maximal invariant* statistic, which is derived from the sufficient statistic and is of lower dimension, extracts all the information in the data that is invariant to the transformations. In most cases, the maximal invariant is intuitive and/or is related to the likelihood function maximized over the unknown parameters.

Example: the signal is an unknown constant level C in additive Gaussian noise of unknown variance σ^2.

$$p(\mathbf{x}|C) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x_i - C)^2}$$

We would like to detect the presence of a non-zero constant. Since we have no prior knowledge about either C or σ^2, we expect (and demand!) that our statistical test be invariant to data scaling which preserves SNR. It certainly would not be a good algorithm if it was not invariant to scaling. The sufficient statistics are the sample mean and variance. The maximal invariant in this case is the square of the sample mean divided by the sample variance, i.e. an SNR estimate.

Whenever the natural symmetry is associated with a nuisance parameter, the generalized likelihood ratio test (GLRT) is closely tied to the maximal invariant. In fact, in the example above, the GLRT will depend on the data only through the maximal invariant. In what follows, we will use maximal invariants in place of sufficient statistics when necessary.

$$\mathbf{z}_1 = \log\left\{\left[\sum_{i=1}^{N} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N} x_i \sin(\omega i)\right]^2\right\}$$

$$\mathbf{z}_2 = \log\left\{\left[\sum_{i=1}^{N/2} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/2} x_i \sin(\omega i)\right]^2\right\}$$

$$\mathbf{z}_3 = \log\left\{\left[\sum_{i=1}^{N/4} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/4} x_i \sin(\omega i)\right]^2\right\}$$

$$\mathbf{z}_4 = \sum_{i=1}^{N} x_i^2$$

$$\mathbf{z}_5 = \sum_{i=1}^{N/2} x_i^2$$

$$\mathbf{z}_6(\mathbf{x}) = \log(x_1^2)$$

$$\mathbf{z}_7(\mathbf{x}) = \log(x_1^2 + x_2^2)$$

$$\mathbf{z}_8 = \begin{bmatrix} \sum_{i=1}^{N} |x_i| \\ \sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

$$\mathbf{z}_9 = \begin{bmatrix} \sum_{i=1}^{N/2} |x_i| \\ \sum_{i=1}^{N/2} x_i^2 \end{bmatrix}$$

Table 1: Class-Specific Statistics

5. SIMULATION RESULTS

The following experiment was performed. A total of 8192 samples from each of classes H_1 through H_9 were created. Each sample consisted of a statistically independent realization of a $N = 256$ time series generated under the corresponding hypothesis. For each hypothesis, the values of pertinent model parameters were selected at random using the a-priori distributions which are provided in the appendix. For each time series produced, the the statistics (features) $\mathbf{z}_1, \ldots, \mathbf{z}_9$ were computed.

As a check on the determination of theoretical PDF under H_0, data was also generated for pure Gaussian noise. Histograms of the H_0 distributions overlaid on the theoretical curves are provided in Figure 2. Notice that $\mathbf{z}_8$ and $\mathbf{z}_9$ are two-dimensional and a planar plot is needed.

The feature data was used in holdout trials to determine probability of correct classification (P_{cc}) as a function of the number of training samples (NTRAIN). For each value of NTRAIN, four independent holdout trials were performed, using all the data not used in training for determining P_{cc}. The results of the experiment

$$p(\mathbf{z}_1|H_0) = \left(\tfrac{e^{\mathbf{z}_1}}{N\sigma^2}\right) \exp\left\{-\tfrac{e^{\mathbf{z}_1}}{N\sigma^2}\right\}$$

$$p(\mathbf{z}_2|H_0) = \left(\tfrac{2e^{\mathbf{z}_2}}{N\sigma^2}\right) \exp\left\{-\tfrac{2e^{\mathbf{z}_2}}{N\sigma^2}\right\}$$

$$p(\mathbf{z}_3|H_0) = \left(\tfrac{4e^{\mathbf{z}_3}}{N\sigma^2}\right) \exp\left\{-\tfrac{4e^{\mathbf{z}_3}}{N\sigma^2}\right\}$$

$$p(\mathbf{z}_4|H_0) = \tfrac{1}{\sigma^2}\Gamma^{-1}(\tfrac{N}{2})\, 2^{-\frac{N}{2}} (\tfrac{\mathbf{z}_4}{\sigma^2})^{\frac{N}{2}-1} \exp\left\{-\tfrac{\mathbf{z}_4}{2\sigma^2}\right\}$$

$$p(\mathbf{z}_5|H_0) = \tfrac{1}{\sigma^2}\Gamma^{-1}(\tfrac{N}{4})\, 2^{-\frac{N}{4}} (\tfrac{\mathbf{z}_5}{\sigma^2})^{\frac{N}{2}-1} \exp\left\{-\tfrac{\mathbf{z}_5}{2\sigma^2}\right\}$$

$$p(\mathbf{z}_6|H_0) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\tfrac{e^{\mathbf{z}_6}}{2\sigma^2}\right\} e^{\mathbf{z}_6/2}$$

$$p(\mathbf{z}_7|H_0) = \left(4\pi\sigma^2\right)^{-1/2} \exp\left\{-\tfrac{e^{\mathbf{z}_7}}{4\sigma^2}\right\} e^{\mathbf{z}_7/2}$$

$p(\mathbf{z}_8|H_0)$ Gaussian for $N \to \infty$:

$$E\left(\mathbf{z}_8|H_0\right) = N \begin{bmatrix} \sqrt{\frac{2}{\pi}} \\ 1 \end{bmatrix}$$

$$\text{cov}\left(\mathbf{z}_8|H_0\right) = N \begin{bmatrix} 1-\frac{2}{\pi} & \sqrt{\frac{2}{\pi}} \\ \sqrt{\frac{2}{\pi}} & 2 \end{bmatrix}$$

$p(\mathbf{z}_9|H_0)$ Gaussian for $N \to \infty$:

$$E\left(\mathbf{z}_9|H_0\right) = \tfrac{N}{2} \begin{bmatrix} \sqrt{\frac{2}{\pi}} \\ 1 \end{bmatrix}$$

$$\text{cov}\left(\mathbf{z}_9|H_0\right) = \tfrac{N}{2} \begin{bmatrix} 1-\frac{2}{\pi} & \sqrt{\frac{2}{\pi}} \\ \sqrt{\frac{2}{\pi}} & 2 \end{bmatrix}$$

Table 2: Distributions of Class-Specific Statistics

are provided in Figure 3 for three classifiers:

1. K-nearest neighbor classifier with $K = 3$.
2. Full-dimensional (FD) classifier implementing equation (1).
3. Class-specific (CS) classifier implementing equation (3).

Both the full-dimensional and class-specific classifier are implemented using heteroscedastic Gaussian Mixture approximations to the various PDF's [8],[9].

Two claims of this paper are supported by the graph. First that the lower dimensional formulation performs better with fewer training samples. Second, that both formulations are equivalent (given sufficient data). The latter claim is supported by the asymptotic convergence to similar performance levels. Of course, the approximations used for classes H_8, H_9 could account for some sub-optimal behavior of the class-specific formulation. Due to practical limitations, the FD performance could not be evaluated at higher that 8192 training samples. Further evidence is obtained from the confusion matrices of the FD and CS classifiers for 8192 and 128 training samples. There was insufficient space to include these tables, however they were nearly identical.

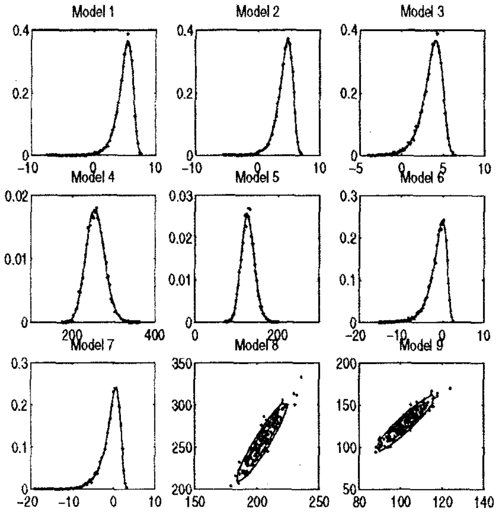

Figure 2: Histograms of statistics under H_0 with theoretical distributions.

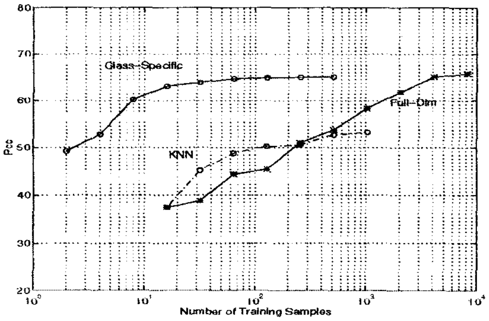

Figure 3: Percent correct vs. number of training samples for three classifiers. Each data point is the average of 4 independent trials.

6. CONCLUSIONS

The benefit of the class-specific formulation of the optimum Bayesian classifier is clearly demonstrated in a synthetic 9-class problem. More that 2 orders of magnitude more training data is required by the traditional approach. We have also seen that vast improvements are possible even if approximate sufficiency and approximate noise-only distributions are used.

7. REFERENCES

[1] R. E. Bellman, *Adaptive Control Processes*. Priceton, New Jersey, USA: Princeton Univ. Press, 1961.

[2] D. W. Scott, *Multivariate Density Estimation*. Wiley, 1992.

[3] E. H. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.

[4] E. C. Real, "Feature extraction and sufficient statistics in detection and classification," in *Proc. ICASSP 1996*, pp. 3049–3052, 1996.

[5] Q. Li and D. W. Tufts, "Principal feature classification," *IEEE Trans Neural Networks*, vol. 8, no. 1, pp. 155–160, 1997.

[6] E. H. Lehmann, *Testing Statistical Hypotheses*. New York: Wiley, 1959.

[7] L. Scharf and D. W. Lytle, "Signal detection in Gaussian noise of unknown level: An invariance application," *IEEE Trans. Info. Theory*, vol. IT-17, pp. 404–411, November 1971.

[8] R. L. Streit, "A neural network for optimum Neyman-Pearson classification," *IEEE Trans. Neural Nets*, vol. 5, no. 5, pp. 764–783, 1994.

[9] L. I. Perlovsky, "A model-based neural network for transient signal processing," *Neural Networks*, vol. 7, no. 3, pp. 565–572, 1994.

8. APPENDIX

Proof of (3):
We may write

$$\begin{aligned} p(\{\mathbf{x}_i\}_{i=1}^M|H_j) &= \int p(\{\mathbf{x}_i\}_{i=1}^M|H_j,\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|H_j)d\boldsymbol{\theta}_j \\ &= \int p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j,\boldsymbol{\theta}_j) \\ &\qquad p(\mathbf{x}_j|H_j,\boldsymbol{\theta}_j)\;p(\boldsymbol{\theta}_j|H_j)\;d\boldsymbol{\theta}_j \\ &\qquad \text{which, because of sufficiency:} \\ &= p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j,\boldsymbol{\theta}_j)\int p(\mathbf{x}_j|H_j,\boldsymbol{\theta}_j) \\ &\qquad p(\boldsymbol{\theta}_j|H_j)d\boldsymbol{\theta}_j \\ &= p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j,\boldsymbol{\theta}_j)\;p(\mathbf{x}_j|H_j) \end{aligned}$$

where $p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j,\boldsymbol{\theta}_j)$ is independent of $\boldsymbol{\theta}_j$, however we retain $\boldsymbol{\theta}_j$ in the arguments for reasons that will become clear. Now, $p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j,\boldsymbol{\theta}_j)$ may be expanded:

$$p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j,\boldsymbol{\theta}_j) = \frac{p(\{\mathbf{x}_i\}_{i=1}^M|H_j,\boldsymbol{\theta}_j)}{p(\mathbf{x}_j|H_j,\boldsymbol{\theta}_j)}$$

We note that since the quotient is independent of $\boldsymbol{\theta}_j$, we might as well use $\boldsymbol{\theta}_j^0$. Thus, dropping the dependence on $\boldsymbol{\theta}_j$,

$$p(\{\mathbf{x}_i\}_{i=1,i\neq j}^M|\mathbf{x}_j,H_j) = \frac{p(\{\mathbf{x}_i\}_{i=1}^M|H_j,\boldsymbol{\theta}_j^0)}{p(\mathbf{x}_j|H_j,\boldsymbol{\theta}_j^0)}$$

Now, $p(\{\mathbf{x}_i\}_{i=1}^M|H_j,\boldsymbol{\theta}_j^0)$ is independent of j as a result of (2), and we write it $p(\{\mathbf{x}_i\}_{i=1}^M|H_0)$, even though H_0 is not actually another class. Thus,

$$p(\{\mathbf{x}_i\}_{i=1}^M|H_j) = \frac{p(\mathbf{x}_j|H_j)}{p(\mathbf{x}_j|H_j,\boldsymbol{\theta}_j^0)}p(\{\mathbf{x}_i\}_{i=1}^M|H_0)$$

Now, plugging into (1), and dividing out $p(\{\mathbf{x}_i\}_{i=1}^M|H_0)$, which does not depend on j, we get

$$\begin{aligned} &\arg\max_j p(\{\mathbf{x}_i\}_{i=1}^M|H_j)p(H_j) \\ &= \arg\max_j \frac{p(\mathbf{x}_j|H_j)}{p(\mathbf{x}_j|H_j,\boldsymbol{\theta}_j=\boldsymbol{\theta}_j^0)}p(H_j) \end{aligned} \tag{4}$$