AN E-M ALGORITHM FOR JOINT MODEL ESTIMATION

Dr. Paul M. Baggenstoss and T. E. Luginbuhl

Naval Undersea Warfare Center
Newport RI, 02841
p.m.baggenstoss@ieee.org (EMAIL)

ABSTRACT

In the unlabeled data problem, data contains signals from various sources whose identities are not known *apriori*, yet the parameters of the individual sources must be estimated. To do this optimally, it is necessary to optimize the data PDF, which may be modeled as a mixture density, jointly over the parameters of all the signal models. This can present a problem of enormous complexity if the number of signal classes is large. This paper describes a algorithm for jointly estimating the parameters of the various signal types, each with different parameterizations and associated sufficient statistics. In doing so, it maximizes the likelihood function of all the parameters jointly, but does so without incurring the full dimensionality of the problem. It allows lower-dimensional sufficient statistics to be utilized for each signal model, yet still achieves joint optimality. It uses an extension of the class-specific decomposition of the Bayes minimum error probability classifier.

1. INTRODUCTION

In many real-world problems, there are a variety of co-existing signal types imbedded in noise and the exact nature or classification of signals as they arrive at the sensor are unknown. This is sometimes known as the *unlabeled data* problem. However, it is often necessary to obtain statistical characterizations, or probability density functions (PDF's), of the individual signal types. We can envision two general types of problems where this is necessary. In the first type of problem (Type I), the varous signal types are considered to be a *background* that is distinguished from some *desired* signal type. In active sonar, for example, there are a large variety of reflecting boundaries in the ocean that must be distinguished from the reflection from a ship or submarine. In such problems, the sub-classification of the exact background signal type is not important but the PDF of the background is necessary in order to obtain optimal classification performance against the desired signal. In the second type of problem (Type II), the PDF's of the individual sub-classes of background *are* important, but it is not practical to manually classify the data for the purposes of estimating the individual PDF's, i.e. the training data is *unlabeled*.

We present below an algorithm that solves the global problem of estimating the parameters of the PDF's of the individual models from the unlabeled data. It does so in the global maximum likelihood sense, but without the high-dimensional search. Suppose K samples of signal-plus-noise are observed from the mixture PDF

$$p(\mathbf{X};\lambda) = \sum_{m=1}^{M} p(\mathbf{X}|H_m;\lambda_m)p(H_m) \qquad (1)$$

where $\mathbf{X} = \{\mathbf{X}_k\}_{k=1}^{K}$, $\mathbf{X}_k \in \mathcal{R}^N$ and H_m is a distinct signal class (or PDF) of data parameterized by $\lambda_m \in \Lambda_m$. The collection of all parameters is denoted λ where

$$\lambda = \left\{p(H_m); \{\lambda_m\}_{m=1}^{M}\right\} \; ; \lambda \in \Lambda.$$

The joint maximum likelihood estimate is the λ which maximizes

$$\max_{\lambda\in\Lambda} \prod_{k=1}^{K} p(\mathbf{X}_k;\lambda), \qquad (2)$$

There are two aspects of this problem that threaten to make it impossible to solve in practice. First, the varous terms in the mixture PDF interact, so it involves a search for a single maximum point in the high-dimensional space Λ. This is in contrast to the labeled data problem which consists a set of simpler problems

$$\max_{\lambda_m\in\Lambda_m} \prod_{k=1}^{K} p(\mathbf{X}_k^m;\lambda_m), \qquad (3)$$

for each m, where $\mathbf{X}_k^m$ are data samples known to be from class m. Second, (3) may be aided by reduction of $\mathbf{X}$ to a sufficient statistic $\mathbf{Z}_m = T_m(\mathbf{X})$ apropriate for each signal class. This cannot be done in (2) unless the sufficient statistics are the same for all m. Now, we solve both of these problems.

2. ASSUMPTIONS AND MATHEMATICAL RESULTS

The following two assumptions are needed for the main results:

Assumption 1 *Let H_0 be a noise-only class written $p(\mathbf{X}|H_0;\lambda_0)$. Within Λ_m, for each m, there exists the same "noise-only" condition λ_m^0, i.e.,*

$$\lim_{\lambda_m\to\lambda_m^0} p(\mathbf{X}|H_m;\lambda_m) = p(\mathbf{X}|H_0;\lambda_0)$$

Assumption 2 *Suppose for each Class PDF, $p(\mathbf{X}|H_m)$, there exists a sufficient statistic for the parameter $\lambda_m \in \Lambda_m$. Denote this sufficient statistic by $\mathbf{Z}_m \triangleq T_m(\mathbf{X})$.*

Applying the above two assumptions, we have

$$\frac{p(\mathbf{X}|H_m;\lambda_m)}{p(\mathbf{X}|H_0;\lambda_0)} = \frac{p(\mathbf{Z}_m|H_m;\lambda_m)}{p(\mathbf{Z}_m|H_0;\lambda_0)}$$

by sufficiency. Therefore,

$$\frac{p(\mathbf{X};\lambda)}{p(\mathbf{X}|H_0;\lambda_0)} = \sum_{m=1}^{M} \frac{p(\mathbf{Z}_m|H_m;\lambda_m)}{p(\mathbf{Z}_m|H_0;\lambda_0)} p(H_m) \qquad (4)$$

0-7803-5041-3/99 $10.00 © 1999 IEEE

Define the likelihood function to be

$$L(\mathbf{X};\lambda) = \prod_{k=1}^{K} \frac{p(\mathbf{X}_k;\lambda)}{p(\mathbf{X}_k|H_0;\lambda_0)} = \prod_{k=1}^{K}\sum_{m=1}^{M} \frac{p(\mathbf{Z}_{m,k}|H_m;\lambda_m)}{p(\mathbf{Z}_{m,k}|H_0;\lambda_0)} p(H_m) \quad (5)$$

where $\mathbf{Z}_{m,k} \triangleq T_m(\mathbf{X}_k)$, $\lambda = \left\{p(H_m); \{\lambda_m\}_{m=1}^{M}\right\}$, and λ_0 is presumed to be known.

The objective is to estimate λ using the E.M. algorithm. It is shown in another section that the E-step consists of maximizing

$$Q(\lambda;\lambda') = \sum_{k=1}^{K}\sum_{m=1}^{M} [\log p(H_m) + \log p(\mathbf{Z}_{m,k}|H_m;\lambda_m) - \log p(\mathbf{Z}_{m,k}|H_0;\lambda_0)]\, \gamma_{km}(\mathbf{Z}_{m,k},\lambda') \quad (6)$$

over λ, where

$$\gamma_{km}(\mathbf{Z}_{m,k},\lambda) = \frac{\frac{p(\mathbf{Z}_{m,k}|H_m;\lambda_m)}{p(\mathbf{Z}_{m,k}|H_0;\lambda_0)} p(H_m)}{\sum_{l=1}^{M} \frac{p(\mathbf{Z}_{l,k}|H_l;\lambda_l)}{p(\mathbf{Z}_{l,k}|H_0;\lambda_0)} p(H_l)} \quad (7)$$

From hereafter, we drom the notational dependence on $\mathbf{Z}_{m,k}, \lambda$, simplifying the notation to γ_{km}. The M-step produces estimates of $p(H_m)$:

$$p(H_m) = \frac{1}{K}\sum_{k=1}^{K} \gamma_{km} \quad (8)$$

2.1. Discussion

Notice that in (6), the maximization of $Q(\lambda;\lambda')$ with respect to λ requires the functions

$$Q_m(\lambda_m;\lambda') = \sum_{k=1}^{K} \log p(\mathbf{Z}_{m,k}|H_m;\lambda_m)\, \gamma_{km} \quad (9)$$

to be *independently* maximized over λ_m, for each m. This is in contrast to (5), which contains mixed terms and requires joint maximization. Equation (9) is, in effect, is a probabilistic weighting of each data sample, a minor modification of individual maximum likelihood estimators represented by (3). If an existing algorithm exists for maximization of $\sum_{k=1}^{K} \log p(\mathbf{Z}_{m,k}|H_m;\lambda_m)$, then this algorithm may be used with a minor modification. One way to do it, albeit impractical, would be to scale γ_{km} by a large constant C, round to an integer $n_{km} = \lfloor C\ \gamma_{km} \rfloor$, then form a larger data set composed of a replication of each data sample $\mathbf{Z}_{m,k}$ by the corresponding integer:

$$\mathbf{Z}' \triangleq \{\{\mathbf{Z}_{m,1} \cdots \mathbf{Z}_{m,1}\}, \ldots, \{\mathbf{Z}_{m,K} \cdots \mathbf{Z}_{m,K}\}\},$$

then maximize

$$\sum_{k=1}^{K'} \log p(\mathbf{Z}'_k|H_m;\lambda_m)$$

where $K' = \sum_{k=1}^{K} n_{km}$. A more practical, yet suboptimal method would be to threshold γ_{km} and include only those data samples in $\mathbf{X}$ that exceed the threshold. Of course, the best approach would be to integrate the weighting directly in the algorithm. If an EM-algorithm is used within the ML estimator for λ_m, it is straightforward. Below, we give an example in which $p(\mathbf{Z}_{m,k}|H_m;\lambda_m)$ are Gaussian mixtures. The result is an EM algorithm operating simultaneously on M different feature spaces.

3. AN EM ALGORITHM FOR A NON-HOMOGENIOUS GAUSSIAN MIXTURE

We now show how to integrate the data weighting γ_{km} into the update equations for the EM algorithm for the individual class PDF's using an example where $p(\mathbf{Z}_m|H_m)$ are Gaussian Mixtures. Consider a Gaussian muxture for $\mathbf{Z}_m \in \mathcal{R}^{N_m}$ under class m

$$p(\mathbf{Z}_m|H_m) = \sum_{i=1}^{L_m} \alpha_{mi}\, \mathcal{N}_{mi}(\mathbf{Z}_m) \quad (10)$$

where

$$\mathcal{N}_{mi}(\mathbf{Z}_m) = (2\pi)^{-N_m/2}\, |\mathbf{\Sigma}_{mi}|^{-1/2} \exp\left\{-\tfrac{1}{2} (\mathbf{Z}_m - \boldsymbol{\mu}_{mi})'\ \mathbf{\Sigma}_{mi}^{-1}\ (\mathbf{Z}_m - \boldsymbol{\mu}_{mi})\right\}$$

By substituting (10) into (4), we have an expression for $p(\mathbf{X}|\lambda)$ that is a kind of mixture of mixtures, with each sub-mixture a function of a different sufficient statistic. We call this a *non-homogeneous* Gaussian mixture. The update procedure for the parameters

$$\lambda = \{p(H_m),\ \alpha_{mi},\ \boldsymbol{\mu}_{mi},\ \mathbf{\Sigma}_{mi}\}$$

are provided in Table 3.

It may be noted that the above update equations are identical to the usual Gaussian Mixture EM updates except for the inclusion of γ_{km} in the first step. Notice also that because of the way $w_{i,k}$ are normalized, the data weights γ_{km} may be arbitrarliy scaled without changing the algorithm.

Because of possible numerical issues, it may be necessary to add a constant to the diagonal elements of $\mathbf{\Sigma}_{mi}$ at each iteration. These constants may be regarded as prior knowledge in the form of independent measurement error variancess, and should be chosen carefully. This method has been employed with much success.

4. SIMULATION RESULTS (7-CLASS EXAMPLE)

The example problem to be discussed here is a subset of the 9-class synthetic problem discused in a previous paper [1]. We consider the following 7 data classes denoted $H_1, \ldots, H_7$.

- 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

The sufficient statistics are tabulated in Table 2 and their distributions under H_0 are listed in Table 3.

Futher details of the signal types and their distributions under H_0 may be found elsewhere [2].

- Repeat for $m = 1, \ldots, M$:
  1. Compute data weights

$$w_{i,k} = \frac{\alpha_{mi} \ \mathcal{N}_{mi}(\mathbf{Z}_{m,k}) \ \gamma_{km}}{\sum_{i=1}^{L_m} \alpha_{mi} \ \mathcal{N}_{mi}(\mathbf{Z}_{m,k})},$$

  for $i = 1, \ldots, L_m$.
  2. Let

$$a_i = \sum_{k=1}^{K} w_{i,k},$$

  for $i = 1, \ldots, L_m$.
  3. Update the means

$$\mu_{mi} = \frac{1}{a_i} \sum_{k=1}^{K} w_{i,k} \ \mathbf{Z}_{m,k},$$

  for $i = 1, \ldots, L_m$.
  4. Update the covariances

$$\Sigma_{mi} = \frac{1}{a_i} \sum_{k=1}^{K} w_{i,k} \ (\mathbf{Z}_{m,k} - \mu_{mi}) \ (\mathbf{Z}_{m,k} - \mu_{mi})',$$

  for $i = 1, \ldots, L_m$.
  5. Update mode weights

$$\alpha_{mi} = \frac{a_i}{\sum_{k=1}^{K} \gamma_{km}}$$

  for $i = 1, \ldots, L_m$.

  End
- Use (7) to update γ_{km} for all k, m.
- Use (8) to update $P(H_m)$ for all m.

Table 1: Update Equations for Non-Homogeneous Mixture

4.1. Data Set

To simulate a data set from a mixture of the seven data classes, an equal share of 1024 samples from each data class were created. Each input data sample was a time-series of 256 data points. The samples were then joined together into a single data set. The true class index of each sample was not used by the algorithm, but was remembered for use later in validation. For each input data sample, the sufficient statistics were computed for all class indexes.

4.2. Algorithm Initialization

Initial values of μ_{mi} were set equal to randomly chosen input data samples. Initial values of Σ_{mi} were set equal to the sample covariance of the entire data set. Initial values of α_{mi} were all equal, as were the initial values of $P(H_m)$. The number of Gaussian mixture components per data class was 10.

$\mathbf{Z}_1 = \left[\sum_{i=1}^{N} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N} x_i \sin(\omega i)\right]^2$
$\mathbf{Z}_2 = \left[\sum_{i=1}^{N/2} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/2} x_i \sin(\omega i)\right]^2$
$\mathbf{Z}_3 = \left[\sum_{i=1}^{N/4} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/4} x_i \sin(\omega i)\right]^2$
$\mathbf{Z}_4 = \sum_{i=1}^{N} x_i^2$
$\mathbf{Z}_5 = \sum_{i=1}^{N/2} x_i^2$
$\mathbf{Z}_6 = \log(x_1^2)$
$\mathbf{Z}_7 = \log(x_1^2 + x_2^2)$

Table 2: Class-Specific Statistics

4.3. Algorithm Performance

Algorithm performance may be measured by monitoring the likelihood function (5). Notice also that γ_{km} in (9) acts as a probablistic data weighting for each sample. It is in effect an estimate of the probability that data sample k is from class m. If the algorithm is working properly and $p(Z_m|H_m)$ are converging to the true PDF's, γ_{km} should act as data classifiers. Thus, for a given sample k, maximizing over m will produce a guess as to the class index of the sample. But this will not work in general. Specifically, if two data classes have the same or equivalent sufficient statistics, the algorithm has no way to make the separation between the classes except perhaps as different Gaussian Mixture components within a fixed m. This shortcoming of the algorithm is expected since it is designed only to estimate the PDF of the overall non-homogenious mixture (Type I problems). The separation of the sub-classes is irrelevant to its operation. However, if all the sufficient statistics are different, it has a chance of accomplishing this goal (Type II problems). In this example, we monitor the algorithm performance as a Type II problem by determining the probability of correct classification (P_{cc}). P_{cc} was determined by determining what percentage of the data was classifed correctly (i.e. when $\arg\max_m \gamma_{km}$ was equal to the true class index).

The algorithm was allowed to iterate 380 times. At each iteration, the total likelihood as well as the probability of correct classification (P_{cc}) were determined. These quantities are plotted in Figure 1. Notice that the likelihood was monotonic increasing as expected.

The algorithm was re-run using labeled data, i.e. only data from class m was used to train $p(\mathbf{Z}_m|H_m)$. The result is plotted in Figure 2. As would be expected, P_{cc} is higher, but the likelihood is lower (not discernable on the graph). Using labeled data does not nececarily maximize (5).

$p(\mathbf{Z}_1\|H_0) = \left(\frac{e^{\mathbf{Z}_1}}{N\sigma^2}\right) \exp\left\{-\frac{e^{\mathbf{Z}_1}}{N\sigma^2}\right\}$
$p(\mathbf{Z}_2\|H_0) = \left(\frac{2e^{\mathbf{Z}_2}}{N\sigma^2}\right) \exp\left\{-\frac{2e^{\mathbf{Z}_2}}{N\sigma^2}\right\}$
$p(\mathbf{Z}_3\|H_0) = \left(\frac{4e^{\mathbf{Z}_3}}{N\sigma^2}\right) \exp\left\{-\frac{4e^{\mathbf{Z}_3}}{N\sigma^2}\right\}$
$p(\mathbf{Z}_4\|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(\frac{N}{2})\, 2^{-\frac{N}{2}} (\frac{\mathbf{Z}_4}{\sigma^2})^{\frac{N}{2}-1} \exp\left\{-\frac{\mathbf{Z}_4}{2\sigma^2}\right\}$
$p(\mathbf{Z}_5\|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(\frac{N}{4})\, 2^{-\frac{N}{4}} (\frac{\mathbf{Z}_5}{\sigma^2})^{\frac{N}{2}-1} \exp\left\{-\frac{\mathbf{Z}_5}{2\sigma^2}\right\}$
$p(\mathbf{Z}_6\|H_0) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{e^{\mathbf{Z}_6}}{2\sigma^2}\right\} e^{\mathbf{Z}_6/2}$
$p(\mathbf{Z}_7\|H_0) = \left(4\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{e^{\mathbf{Z}_7}}{4\sigma^2}\right\} e^{\mathbf{Z}_7/2}$

Table 3: Distributions of Class-Specific Statistics

The PDF estimate of $p(\mathbf{Z}_4|H_4)$ from unlabeled data is plotted in Figure 3. Superimposed on the graph are the histograms of $\mathbf{Z}_4$ for all data classes and for just class 4. The fact that the PDF estimate matches the histogram for class 4 illustrates the fact that PDF estimates may be obtained from unlabeled data.

5. CONCLUSIONS

An E-M algorithm has been derived for the case when the input data is a mixture density of several data classes, with each data class dependent on a different set of parameters, By taking advantage of different sufficient statistics for each data class, it is possible to jointly estimate the parameters efficiently.

6. REFERENCES

[1] P. M. Baggenstoss, "Class-specific features in classification.," *IEEE Trans SP , in review (to be published 1998 or 1999)*, 1997.

[2] P. M. Baggenstoss, "Class-specific features in classification.," in *IASTED International Conference on Signal and Image Processing*, 1998.

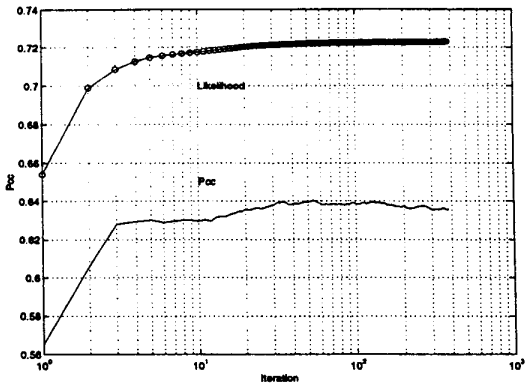

Figure 1: Algorithm Convergence Properties. Scaled log likelihood values superimposed on a plot of P_{cc}.

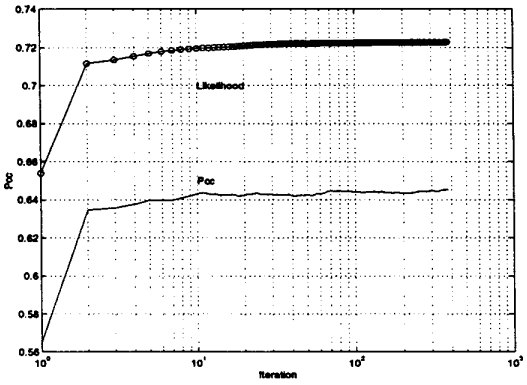

Figure 2: Repeat of Figure 1 with labeled data used to train $p(\mathbf{Z}_m|H_m)$.

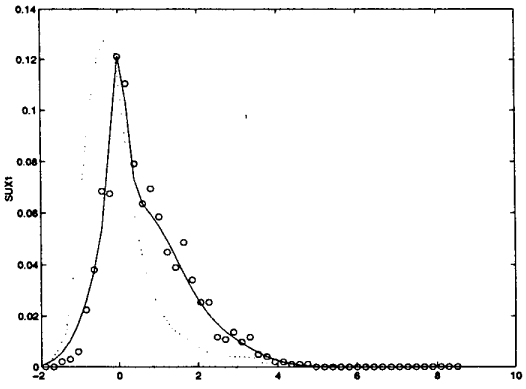

Figure 3: PDF estimation results for feature $\mathbf{Z}_4$ using unlabeled data. Graph includes histogram of data from all 7 classes (dotted), histogram of data from class 4 (circles), PDF estimate for $p(\mathbf{Z}_4|H_4)$ (solid). Data is plotted on a normalized axis. The designation "SUX1" is the name used to identify feature $\mathbf{Z}_4$.