

Utilizing Non-Uniform Cost Learning for Active Control of Inter-class Confusion

Dwi Sianto Mansjur, Qiang Fu, Biing Hwang Juang
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, USA 30332
{dwi,qfu,juang}@ece.gatech.edu

Abstract

In this paper, we demonstrate the use of learning with non-uniform error-cost as a novel technique to design a multiclass cost-sensitive classifier. We investigate two important aspects of the design. First, we show that the learning is effective enough for active control of the multiclass confusion matrix using the cost-matrix. Second, we study the cases when the classifiers have mild model mismatch problems, and conclude that our design still have better performance compared to the conventional cost-sensitive classifier design.

1 Introduction

A straightforward way to convert an error-rate based classifier to a cost-sensitive classifier is to simply apply additional expected-cost criteria. Because this criteria only need to be applied during classification, it uses the same error-based classifier even when the cost changes [3, 1]. The underlying assumption of this criteria is that the error-rate based classifier provides well-tuned *a posteriori* probability for each class. In practice, the well-tuned *a posteriori* probability cannot be obtained easily due to the lack of the knowledge of the true distribution function. Furthermore, there are usually limited numbers of observations available to estimate the parameters of the distribution function chosen to represent the classifier design.

A non-uniform error-cost criteria has been proposed in our earlier work as a novel learning algorithm to obtain a cost-sensitive classifier for an arbitrary cost matrix [2]. The main characteristic of the learning algorithm is that it offers a framework to combine the cost-sensitive decision rule and the classifier performance (i.e. the minimum cost) into a novel cost function so that the system performance can be evaluated and op-

timized. In the earlier development, we conclude that the non-uniform error-cost learning is effective in minimizing design cost according to system requirements. This paper investigates further two important aspects of the non-uniform error-cost. First, it studies a typical behavior of the cost learning in term of the multiclass confusion matrix. Second, it demonstrates that even if the classifier models have mild model mismatch with respect to the true class distribution, the algorithm still provides a smaller cost than the conventional cost-sensitive classifier.

This paper is organized as follows: Section 2 reviews the foundation of non-uniform error-criteria learning. Section 3 explains the development of the non-uniform error-cost criteria learning. The application of the learning criteria to obtain cost-sensitive Gaussian Mixture Model (GMM) is presented in Section 4. Experiments on the artificial datasets are provided in the Section 5.

2 Optimal Decision for Non-uniform Error Cost

Here, we introduce the foundation of the non-uniform error-cost criteria: i.e., the Bayes decision theory. Consider a classification task into M classes (e.g., the task of recognizing handwritten digit with $M=10$). An unknown pattern denoted by X is observed and classified into one of the M classes. Thus, the classifier is a function C that maps X into a class identity C_i , where $i \in I_M = \{i, i = 1, \dots, M\}$. We denote this function as a decision function $C(X)$. Associated with each decision is a cost, which can be expressed as an entry ϵ_{ij} in an $M \times M$ matrix, where $i, j \in I_M$. The entry ϵ_{ij} indicates the cost of classifying a pattern from the j^{th} class as one of the i^{th} class. Suppose at our disposal, we have the knowledge of the *a posteriori* probabilities $P(C_i|X), \forall i \in I_M$. According to Bayes' decision theory, given X , the conditional cost of making a decision $C(X) = C_i$ can be defined as

$R(C_i|X) = \sum_{j=1}^M \epsilon_{ij} P(C_j|X)$ and the system performance in term of the *expected loss* is defined as

$$\mathcal{L} = E\{R(C(X)|X)\} = \int R(C(X)|X)p(X)dX. \quad (1)$$

Error-based classifier uses 0-1 error cost (0 for correct and 1 for incorrect decision), but in non-uniform/cost-sensitive learning the error cost is zero for correct decision and positive-cost for incorrect decision. We can institute the decision rule with the non-uniform error cost function as

$$C(X) = \arg \min_i R(C_i|X) \quad (2)$$

Traditional decision rule with 0-1 error cost is defined as $C(X) = \arg \min_i R(C_i|X) = \arg \max_i P(C_i|X)$, and it is the well-known *maximum a posteriori* (MAP) decision rule.

3 Non-Uniform Error-Cost Criteria

The non-uniform error-cost criteria is a novel objective function, which encapsulates both the minimum cost decision rule and system performance. Thus, throughout the development of the criteria, attention has to be focus on the *decision rule* and the *system performance*. Let $g_i(X; \Lambda) \geq 0$ be a discriminant fuction for the i^{th} class, $i = 1, 2, \dots, M$ where Λ is the parameter set that defines the functions. The classification decision is reached according to:

$$C(X) = \arg \max_i g_i(X; \Lambda). \quad (3)$$

That is the classifier choose the class that leads to the largest value among all discriminants evaluated on X . If the true *a posteriori* probability is available, then a monotonically decreasing function of the conditional risk would be appropriate. For example,

$$g_i(X; \Lambda) = \exp\{-R(C_i|X)\} = \exp\{-\sum_{j \in I_M} \epsilon_{ij} P(C_j|X)\} \quad (4)$$

To accumulate the error cost of each sample into the objective function, the expected system loss needs to be expressed in term of the empirical loss with the decision rule embedded in it. For clarity, let $i_x = C(X)$ be the identity index as decided by the classifier and j_x be the true identity of X and a set of training samples $\Omega = \{X^{(n)}\}_{n=1}^N$. The cost incurred by a single sample is defined as:

$$l_{i_x}(X; \Lambda) = \epsilon_{i_x j_x} \quad (5)$$

Therefore, if the empirical system loss is defined over the realized sample-based costs, an alternative non-uniform cost can be defined as follows:

$$L = \frac{1}{N} \sum_{X \in \Omega} \epsilon_{i_x j_x} \rightarrow \int \epsilon_{i_x j_x} p(X) dX \quad (6)$$

Suppose each class is prescribed as discriminant function $g_j(X; \Lambda) \forall j$, then the decision rule is defined as follows:

$$C(X) = i_X = \arg \max_k g_k(X; \Lambda) \quad (7)$$

The empirical system loss of (6) based on the set of training samples Ω becomes:

$$L = \frac{1}{N} \sum_{X \in \Omega} \left(\sum_{i \in I_M} \sum_{j \in I_M} \epsilon_{ij} \mathbf{1}\{j_X = j\} \mathbf{1}\{i = \arg \max_k g_k(X; \Lambda)\} \right) \quad (8)$$

The definition above uses two indicator functions. The indicator function $\mathbf{1}$ is to indicate membership of an element in a set, i.e. it assumes the value of 1 if the argument is true and zero otherwise. The first function indicates observation X belong to class C_j , i.e. $\mathbf{1}\{j_x = j\} = \mathbf{1}\{X \in C_j\}$. The second function $\mathbf{1}\{i = \arg \max_k g_k(X; \Lambda)\}$ denotes the decision rule (7).

The remaining challenge is to turn the objective function L in (8) into a smooth function suitable for optimization. Consider $L = \sum_{j \in I_M} L_j$ where each L_j is defined as follows:

$$L_j = \frac{1}{N} \sum_{x \in \Omega} \sum_{i \in I_M} \epsilon_{ij} \mathbf{1}[X \in C_j] \mathbf{1}\{i = \arg \max_k g_k(X; \Lambda)\} \quad (9)$$

That is L_j is the empirical error cost collected over samples in Ω with $j_X = j$. This approximation needs to be made to the summands. This can be accomplished by:

$$\sum_{i \in I_M} \epsilon_{ij} \mathbf{1}\{i = \arg \max_k g_k(X; \Lambda)\} \approx \sum_{i \in I_M} \epsilon_{ij} \frac{g_i(X; \Lambda)}{G(X; \Lambda)} \quad (10)$$

where $G(X; \Lambda) = [\sum_{i \in I_M} g_i(X; \Lambda)]^{1/\eta}$. Note that as the design parameter $\eta \rightarrow \infty$:

$$\frac{g_i(X; \Lambda)}{G(X; \Lambda)} \approx \begin{cases} 1, & \text{if } G(X; \Lambda) = \max_k g_k(X; \Lambda) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

Finally, the smoothed empirical error-cost is as follows:

$$L \approx \frac{1}{N} \sum_{X \in \Omega} \sum_{j \in I_M} \left(\sum_{i \in I_M} \epsilon_{ij} \frac{g_i(X; \Lambda)}{G(X; \Lambda)} \right) \mathbf{1}[X \in C_j] \quad (12)$$

4 Application in Mixture Model Learning

In this section, we set the discriminant function g_j for class C_i in term of a Gaussian mixture model (GMM) with diagonal covariance matrix as defined as follows:

$$g_i = g_i(X; \Lambda) = \exp \left(- \sum_{j=1}^K \epsilon_{ij} b_j P(C_j) \right) \quad (13)$$

$\begin{bmatrix} 0 & 1 & 1 & \boxed{1} \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$ (a) cost matrix	$\begin{bmatrix} 107.1 & 14.3 & 0.7 & \boxed{8.4} \\ 9.5 & 104.7 & 11.5 & 0.2 \\ 0.7 & 8.7 & 98.7 & 7.9 \\ 10.6 & 0.3 & 17.1 & 111.4 \end{bmatrix}$ (b) true model	$\begin{bmatrix} 103.5 & 16.7 & 1.1 & \boxed{9.3} \\ 11.3 & 97.8 & 11.0 & 0.3 \\ 1.8 & 13.0 & 100.7 & 17.9 \\ 11.5 & 0.4 & 15.2 & 100.4 \end{bmatrix}$ (c) baseline	$\begin{bmatrix} 104.7 & 15.2 & 1.1 & \boxed{9.5} \\ 11.1 & 98.2 & 9.8 & 0.4 \\ 1.2 & 14.2 & 102.1 & 17.1 \\ 11.0 & 0.4 & 15.0 & 101.0 \end{bmatrix}$ (d) cost-optimized
--	--	---	--

Table 1. Confusion matrices of true (b), baseline (c) and cost-optimized (d) models based on cost matrix (a) for cost-insensitive classification.

$\begin{bmatrix} 0 & 1 & 1 & \boxed{2} \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$ (a) cost matrix	$\begin{bmatrix} 100.5 & 14.2 & 0.1 & \boxed{3.6} \\ 9.5 & 104.7 & 11.5 & 0.3 \\ 0.8 & 8.6 & 98.8 & 7.9 \\ 17.1 & 0.4 & 17.4 & 116.1 \end{bmatrix}$ (b) true model	$\begin{bmatrix} 97.3 & 16.4 & 0.5 & \boxed{5.3} \\ 11.3 & 97.8 & 11.0 & 0.3 \\ 2.0 & 13.0 & 100.8 & 18.2 \\ 17.4 & 0.7 & 15.6 & 104.1 \end{bmatrix}$ (c) baseline	$\begin{bmatrix} 98.6 & 15.3 & 0.6 & \boxed{4.7} \\ 10.8 & 98.0 & 9.9 & 0.3 \\ 1.8 & 13.9 & 102.3 & 18.4 \\ 16.7 & 0.6 & 15.0 & 104.4 \end{bmatrix}$ (d) cost-optimized
--	--	--	---

Table 2. Confusion matrices of true (b), baseline (c) and cost-optimized (d) models based on cost matrix (a) for cost-sensitive classification.

where $b_j = p(X|C_i) = \sum_{k=1}^K c_k \mathcal{N}(X; \mu_k, \sigma_k^2)$ is the likelihood function and $P(C_j)$ is the prior probability of class j . Thus, the parameter sets $\Lambda = \{\mu_k, \sigma_k^2, c_k\}$.

The notations used are as follows: j is the index of the class identity, k denotes the mixture number and l indicates the dimension starting from 1 to D . For clarity purpose, $P(C_i)$, $g_i(X; \Lambda)$ and $G(X; \Lambda)$ are written into P_i , g_i and G respectively. First, to help the convergence of the learning process, parameter transformation is applied to the mean μ_{jkl} to obtain $\tilde{\mu}_{jkl}$, where, $\tilde{\mu}_{jkl} = \frac{\mu_{jkl}}{\sigma_{jkl}}$. The learning process for the mean $\tilde{\mu}_{jkl}$ vector is as follows:

$$\tilde{\mu}_{jkl}(t+1) = \tilde{\mu}_{jkl}(t) - \epsilon^t \frac{\partial L_i}{\partial \tilde{\mu}_{jkl}} \quad (14)$$

where ϵ^t is the learning step during t iteration, and the gradient is defined as $\frac{\partial L_i}{\partial \tilde{\mu}_{jkl}} = \frac{\partial L_i}{\partial g_i} \frac{\partial g_i}{\partial b_j} \frac{\partial b_j}{\partial \mu_{jkl}} \frac{\partial \mu_{jkl}}{\partial \tilde{\mu}_{jkl}}$. The gradient can be computed from (15) and (16)

$$\frac{\partial L_n}{\partial b_m} = \frac{P_m}{\sum_{r=1}^M b_r P_r} \left\{ \left(\frac{-\sum_{k=1}^M \epsilon_{kj} (\epsilon_{km} + \log(g_k)) g_k}{G} \right) + \left(\frac{L_n \sum_{k=1}^M (\epsilon_{km} + \log(g_k)) g_k^\eta}{G^\eta} \right) \right\} \quad (15)$$

$$\frac{\partial b_j}{\partial \tilde{\mu}_{jkl}} = c_{jk} (2\pi)^{-D/2} \exp \left\{ -\frac{1}{2} \sum_{l=1}^D \left(\frac{x_l - \mu_{jkl}}{\sigma_{jkl}} \right)^2 \right\} \left(\frac{x_l - \mu_{jkl}}{\sigma_{jkl}} \right) \left(\prod_l \sigma_{jkl} \right)^{-1} \quad (16)$$

The new μ_{jkl} is obtained from $\mu_{jkl} = \tilde{\mu}_{jkl} \sigma_{jkl}$ afterwards. Learning for the variance and weight vectors follow the same procedure.

5 Experiment

We make use of the multiclass *confusion matrix* to evaluate our system. The rows of the confusion matrix

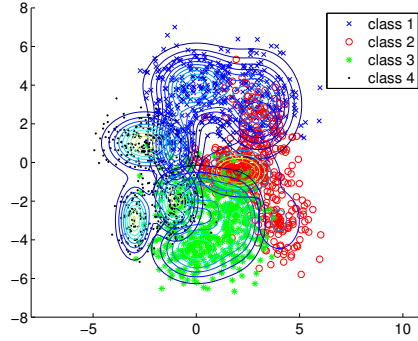


Figure 1. Scatter plot of the true model

represent the system decisions, and the columns represent the true observation labels. Ideally, the values of off-diagonal entries are zeros which means none of the samples have been misclassified. This row and column convention is the same as that used in the *cost matrix*.

We have used multi-class multi-dimensional random observations in our experiments. We generated 4 classes of 2-dimensional data with 128 samples for each class and a total of 512 samples for training and testing sets. Each class has the same number of mixture but with different class prior. The scatter plot for the classes is shown in Figure 1. All of those mixtures have different covariance matrices and different mixture weight. This is what we referred to as the *true models*. The *conventional models* or the *baseline models* are based on maximum likelihood density estimations. Then additional expected error criteria are applied to the estimation to obtain conventional cost-sensitive classifier.

We started the experiments with the standard 0-1 cost

column 1 row 4 entry	3 mixtures			4 mixtures		
	cost 2	cost 4	cost 8	cost 2	cost 4	cost 8
true models	3.60(1.83)	1.03(0.89)	0.30(0.53)	3.60(1.83)	1.03(0.89)	0.30(0.53)
baseline models	5.30(3.36)	2.80(2.44)	1.73(1.95)	6.20(3.31)	3.13(2.43)	1.60(1.75)
cost-optimized models	4.70(3.04)	2.37(2.27)	1.37(1.87)	5.63(2.90)	2.53(1.68)	1.27(1.36)

Table 3. The mean (standard deviation) of row 1 and column 4 of the multiclass confusion matrix across true models, baseline models and cost-optimized models.

matrices and Table 1 list the cost matrix and the confusion matrices of the non cost-sensitive (cost-insensitive) classification. Then, we arbitrary choose to minimize the misclassification of true class 4 as class 1 (the value on row 1, column 4 of the cost matrix). Specifically, the value of 2, 4 and 8 were used as the cost of misclassification. For each cost value, we collect the statistics on the confusion matrix based on 30 trials. Table 2 shows the performance result when the misclassification cost is equal to 2. The cost matrix is shown in Table 2(a) and the confusion matrix results of the true, baseline and cost-optimized models are shown in Table 2(b), 2(c) and 2(d) respectively. Notice that the cost-optimized models are closer to the ground truth (the true models) than the baseline models (i.e., $\text{true}[1,4] < \text{cost-optimized}[1,4] < \text{baseline}[1,4]$). The main reason is that the proposed technique uses the cost matrix to optimize the classifier models. However, the baseline models are only based on the conventional density estimation and the classifier models are not optimized for the cost matrix.

The entry in row 1 column 4 is the only difference between the cost matrices used to obtained the result in Table 1 and Table 2. Comparing these two tables, we can observed the impact of the cost matrices on the confusion matrices. The misclassifications from class 4 into class 1 decrease if we compare between the corresponding models in those two tables. The reason is that all models use the same additional expected error criteria in their decision rule. It is also important to notice that the row 1 column 4 entry of the baseline models is lower than the cost-optimized models in Table 1. However, the baseline models has higher error rate compared to cost-optimized models because the minimization of error-cost with 0-1 cost matrix is equivalent to the minimization of the error-rate.

We expect that as the cost value increases, the misclassification between two classes would decrease. Therefore, we experimented further with the cost values of 4 and 8. We also expect that either the wrong or the correct number of mixture were used, the cost-optimized models still perform better than the baseline models. We compared model match with 3 mixtures and model mismatch with 4 mixtures. Table 3 shows

the results of the misclassification (class 4 as class 1) over 30 trials. The results show the average number of misclassification decrease for all techniques as the cost value increases from 2 to 8 since fewer samples are categorized into class 1. As expected, the cost-optimized models in general perform better than the baseline models even with the wrong number of mixture.

It is important to notice that we chose the entry of row 1 and column 4 arbitrary as a proof of concept that non-uniform cost learning is capable of active control of a confusion matrix using the cost matrix. Indeed, the non-uniform error learning is general enough to be used with any valid cost matrix.

6 Conclusions

Converting an error-rate based classifier to a cost-sensitive classifier involves more steps than simply the application of expected cost criteria during the classification process because the well-tuned *a posteriori* probability is hard to obtain. A novel machine learning technique based on non-uniform cost-criteria has been introduced to obtain a better cost-sensitive classifier in term of expected error-cost. Moreover, the resulting cost-sensitive classifier can have active control of inter-class confusion. Our experimental results have confirmed this for both model match and a mild model mismatch cases.

References

- [1] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 973–978, 2001.
- [2] Q. Fu, D. S. Mansjur, and B. H. Juang. Non-uniform error criteria for automatic pattern and speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 2008.
- [3] S. Viaene and G. Dedene. Cost-sensitive learning and decision making revisited. *European Journal of Operational Research*, 127(1):212–220, October 2005.