ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Gaussian Mixture Models for multiclass problems with performance constraints

Nisrine Jrad, Edith Grall-Maës and Pierre Beauseroy

Université de Technologie de Troyes ICD (FRE CNRS 2848), LM2S
12 rue Marie Curie, BP 2060, 10010 Troyes cedex - France

Abstract. This paper proposes a method using labelled data to learn a decision rule for multiclass problems with class-selective rejection and performance constraints. The method is based on class-conditional density estimations obtained by using the Gaussian Mixture Models (GMM). The rule is thus determined by plugging these estimations in the statistical hypothesis framework and solving an optimization problem. Two simulations are then carried out to corroborate the efficiency of the proposed method. Experimental results show that it compares well with a non-parametric solution using Parzen estimator.

1 Introduction

Classification of an unknown pattern into one of a finite number of known classes is a well-known problem in many fields of science and engineering. Generally, a classification system is designed to optimize a given loss function, for instance the error rate. For some cases, the loss function should be more general. First, some applications, like face identification, may favor withholding decision than taking a wrong one. Consequently, the introduction of rejection options should be considered. In other words, a pattern can be rejected from one, some or all classes in order to ensure a higher reliability. In this class-selective rejection scheme, the loss function penalizes differently the wrong decisions and partially correct ones. Second, these applications may require to satisfy some constraints. Hence, the loss function should also take into account these constraints. A general formulation for this problem was proposed in [1, 2] and the optimal decision rule was given in the framework of decision theory. To learn a classifier for such problems, one approach consists of using labelled data and determining the optimal decision rule with the estimated conditional probability densities instead of the theoretical ones.

In the past decades, mixture models were applied as an expressive class of models for density estimation. In particular, Gaussian Mixture Models (GMM) were used as effective models with high identification accuracy. The most popular algorithm to learn mixture models is the Expectation-Maximization (EM) algorithm [3]. Starting from a random configuration, many learning GMM methods estimate parameters in order to optimize a "goodness of fit" criterion such as the likelihood [4, 5, 6, 7]. The greedy EM algorithm [6] is used in this manuscript due to its insensitivity to the initialization step, its growing nature useful when the number of mixing components is unknown and its ability to find the optimal likelihood maximum. The class-conditional probabilities, estimated using GMM are plugged into the hypothesis tests to determine a supervised decision rule. Moreover, GMM results are compared with a non-parametric solution using Parzen windows estimator [8] in order to asses the efficiency of both methods.

This paper is outlined as follows. Section 2 describes the multiclass decision problems with performance constraints. A brief description of greedy EM is given in section 3. Section 4 presents the proposed training and quality assessment

algorithms. GMM based method is compared to the Parzen windows estimators based method using two simulated examples in section 5. Finally, a conclusion and perspectives are given in section 6.

2 Multiclass decision problem

Assuming that a classification problem is characterized by N classes $w_1 ... \ w_N$ and that any observation $x \in \Re^d$ belongs to one class, a decision rule consists of a partition Z of $\Re^d$ in I sets Z_i composed of elements x assigned to the decision option ψ_i. In the class-selective rejection scheme, options are defined by an admissible class or a subset of classes (i.e. $x \in \psi_i = \{1; 3\}$ means that x is assigned to w_1 and w_3 with ambiguity). The probability that elements of w_j are assigned to ψ_i is:

$$P(D_i/w_j) = \int_{Z_i} P(x/w_j)dx.$$

The problem consists of finding the decision rule Z^* that minimizes a given loss $\overline{c}(Z)$ and respects K given constraints respectively defined by:

$$\overline{c}(Z) = \sum_{i=1}^{I}\sum_{j=1}^{N} c_{i,j}P_jP(D_i/w_j) \quad \text{and} \quad e^{(k)} = \sum_{i=1}^{I}\sum_{j=1}^{N} \alpha_{i,j}^{(k)}P_jP(D_i/w_j) \leq \gamma^{(k)}$$

γ^k are thresholds, $c_{i,j}$ and $\alpha_{i,j}^{(k)} \in \Re$ are the costs of deciding that an element x belongs to the set ψ_i when it belongs to the class w_j, in the expressions of the loss and the constraints respectively. P_j are the a priori probabilities. The optimal decision rule Z^* is the solution of the following problem:

$$\min_{Z} \ \overline{c}(Z) \quad \text{subject to } e^{(k)}(Z) \leq \gamma^{(k)} \ \forall k = 1, ..., K.$$

This problem may be defined as an optimization problem with constraints. Its solution can be obtained by using the Lagrangian $L(Z, \boldsymbol{\mu}) = \overline{c}(Z) + \sum_{k=1}^{K} \mu_k(e^{(k)} - \gamma^{(k)})$ where $\boldsymbol{\mu}$ is the vector of the Lagrange multipliers $\mu_k \geq 0, k = 1, ..., K$ associated to the constraints. The determination of the solution in the statistical decision theory was expounded in [2]. It consists of finding the saddle point $(Z^*, \boldsymbol{\mu^*})$ of $L(Z, \boldsymbol{\mu})$ by solving:

$$\max_{\mu \in \Re^{K+}} \{\min_{Z} L(Z, \boldsymbol{\mu})\} \tag{1}$$

The optimal decision rule is defined by the partitions $Z_i^*(\boldsymbol{\mu^*})$ for $i = 1, ..., I$ where $Z_i(\boldsymbol{\mu})$ is given by the set $\{x/\lambda_i(x, \boldsymbol{\mu}) < \lambda_l(x, \boldsymbol{\mu}), l = 1, ..., I, l \neq i\}$ and $\lambda_i(x, \boldsymbol{\mu}) = \sum_{j=1}^{N} P_jP(x/w_j)\left(c_{i,j} + \boldsymbol{\mu}^T\alpha_{i,j}\right)$.

3 Gaussian Mixture Models

The GMM are investigated to design a supervised decision rule by using the estimates of the conditional probability functions. The GMM are given by $\hat{P}_T(x/w) = \sum_{t=1}^{T} \pi_t\phi(x/w; \theta_t)$ where the t-th gaussian component of a given class w, $\phi(x/w; \theta_t)$, is the d-dimensional gaussian density parameterized by $\theta_t = \{m_t, S_t\}$:

$$\phi(x/w; \theta_t) = (2\pi)^{\frac{-d}{2}} \mid S \mid^{\frac{-1}{2}} \exp[-0.5(x - m_t)^T S_t^{-1}(x - m_t)]$$

T is the number of components for the modelization of class w, π_t is the mixing

weight satisfying $\sum_{t=1}^{T} \pi_t = 1$ and $\pi_t \geq 0$, m_t is the mean and S_t is the covariance matrix of the t-th component. Given a set $\{x_1, ..., x_n\}$ drawn from w, the task is to estimate the parameters π_t, m_t, S_t and the number T of components that maximize the log-likelihood $\mathcal{L}_T = \sum_{p=1}^{n} \log \hat{P}_T(x_p/w)$. The log-likelihood maximization can be carried out by the greedy EM algorithm based on the theoretical results of [9]. In this latter, Li and Barron show that the difference in Kullback-Leibler divergence achievable by T-component mixtures and the Kullback-Leibler distance achievable by any (possibly non-finite) mixture from the same family of components tends to zero with the rate c/T with c a constant dependant from the component family. Furthermore, this bound is reachable by employing the greedy procedure. Therefore, the maximum likelihood of the mixture can be determined by adding iteratively a new component to the mixture. In this work, the greedy EM [6, 7] algorithm for learning GMM is used since it is able to find the global likelihood maximum and to estimate the unknown number of the mixture components. This algorithm can be summarized as follows. Starting from a 1-component mixture ($T = 1$), the optimal parameters are obtained by an EM procedure until convergence($\mid \mathcal{L}^{iteration} - \mathcal{L}^{iteration-1} \mid \leq \varepsilon$). Then, a search for a new component $\phi(x/w; \theta^*)$ location and a corresponding weight a^* is performed in order to maximize the new log-likelihood:

$$\mathcal{L}_{T+1} = \sum_{p=1}^{n} \log \hat{P}_{T_+1}(x_p/w) = \sum_{p=1}^{n} \log[(1-a)\hat{P}_T(x_p/w) + a\phi(x_p/w; \theta)] \qquad (2)$$

with $\hat{P}_T$ remaining fixed. It is obvious that the crucial step of this algorithm is the search of a new component location. It can be shown that $\mathcal{L}_{T+1}$ is concave as function of a but can have multiple maxima as function of θ. Hence, a global search is required. One way pointed in [7] proposes to use all the points as initial candidates of the sought component. Every point is the mean of a corresponding candidate with the same covariance matrix $\sigma^2 I$, where σ is set according to [10]. For each candidate component, a is set to the mixing weight maximizing the second order Taylor approximation of $\mathcal{L}_{T+1}$ around $a = 0.5$. The candidate yielding to the highest log-likelihood when added to $\hat{P}_T$ in (2) is selected and updated using partial EM until convergence. The new component is added to $\hat{P}_T$ and the research is repeated until reaching the maximum likelihood on a validation set. An improved version of this global search [6] is used in this work. At each iteration, it selects the best component from a set of candidates whose size increases linearly with T yielding to better and faster performances.

4 Supervised learning and quality assessment of the rule

In the statistical decision theory framework, the determination of a multiclass rule that satisfies performance constraints consists of finding the optimal Z^* and the optimal Lagrange multipliers $\boldsymbol{\mu^*}$ by solving the optimization problem invoked in (1) [1, 2]. However, in the supervised learning framework, P_j and $P(D_i/w_j)$ are unknown. One strategy to learn a supervised classifier is to estimate these probability density functions and determine the corresponding optimal supervised rule $\widehat{Z^*}$ and estimated Lagrange multipliers $\widehat{\boldsymbol{\mu^*}}$. In this paper, we study the repercussion due to the estimation of $P(x/w_j)$ and we consider that P_j is known. Two estimators, Parzen windows method [8], with a smoothness parameter h, and GMM fitting are used. The probability estimates depend on the labelled set and on the density estimators parameters, h of Parzen and T of

the GMM. These parameters are determined by maximizing the log-likelihood of a validation set using 10-Cross Validation. Supervised rules obtained by using these estimations are then compared.

To assess the quality of the supervised rules obtained using the two density estimators, the Lagrangian of the supervised rules $\widehat{Z^*}$ should be estimated on a test set as a comparison criterion. The validity and the relevance of this criterion were experimentally shown in [11]. Since the aim of this work is to study the GMM and Parzen estimators ability learning a decision rule, it is important to compare different rules $\widehat{Z^*}$ with comparable criterions computed with a unique estimator of $P(x/w_j)$ and the same value of $\widehat{\boldsymbol{\mu}^*}$. In this paper, criterions will be computed on an infinite test set (theoretical densities) with theoretical Lagrange multipliers (those associated with the theoretical rule) in order to get the theoretical performance of the rule. The learning-testing procedures of the GMM and Parzen windows estimator algorithm are as follow:

1. For each class w_j, estimating the GMM or the Parzen windows distributions $\hat{P}(x/w_j)$ using a training set and a validation set.
2. Learning the decision rule by solving the optimization problem (1) with the estimated $\hat{P}(x/w_j)$, namely, finding the optimal supervised rule $\widehat{Z^*}$ and the optimal $\widehat{\boldsymbol{\mu}^*}$.
3. Quality assessment of the rule: computing the Lagrangian $\widehat{L}(\widehat{Z^*}, \boldsymbol{\mu}^*)$ of the supervised rule $\widehat{Z^*}$ on an infinite test set using theoretical $\boldsymbol{\mu}^*$.

5 Simulation results and discussions

To evaluate the performances of the supervised learning approaches, two 2-D problems with three equiprobable classes and performance constraints were considered. For both problems, GMM and Parzen estimators were used. Synthetic data and experimental results are presented and compared below.

5.1 Toy problems

The first problem is defined by three classes, each one is a 3-gaussian component distribution with unbalanced weights, leading to a trimodal distribution. The aims of this experiment is first to study the case where the distributions correspond to the hypothesis of GMM and second to asses the Parzen estimator ability fitting multimodal distributions. The second problem is given by three bivariate gamma distributions in order to study the case where the hypothesis of GMM is not fulfilled by data distributions. The corresponding theoretical densities were represented using isodensity curves in figures 1.a and 2.a.

For both problems, the possible decision options are given by: $\psi_1 = \{1\}$, $\psi_2 = \{2\}$, $\psi_3 = \{3\}$, $\psi_4 = \{1, 2\}$, $\psi_5 = \{1; 3\}$, $\psi_6 = \{2; 3\}$ and $\psi_7 = \{1; 2; 3\}$. The constraints are defined by $P_E \leq 0.05$ and $P_I \leq 0.1$ for the first problem and $P_E \leq 0.1$ and $P_I \leq 0.15$ for the second one. P_E is the probability of error and P_I is the probability of indistinctness, namely, the probability to assign a pattern x of w_j to a subset of classes ψ_i that contains more than one class of which one is w_j. The Lagrangian is defined by $L(Z, \boldsymbol{\mu}) = P_E + 0.5P_I + P(D_7) + \sum_{k=1}^{K} \mu_k(e^{(k)} - \gamma^{(k)})$ where $P(D_7)$ corresponds to no decision. Theoretical decision rules are given by the partitions reported in figures 1.a and 2.a.

For both experiments, the learning-testing algorithm described above was carried out on three groups of 40, 20 and 10 learning sets of 50, 100 and 200 observations per class respectively.

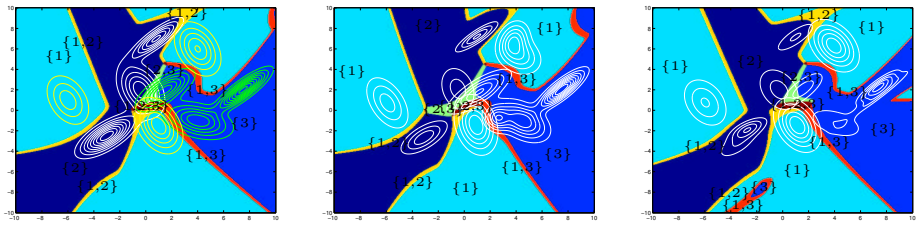

Fig. 1: 3-gaussian component problem: density probabilities and the corresponding partition. From left to right: Theoretical case (1.a) and an example of estimators in the case of Parzen (1.b) and GMM (1.c)

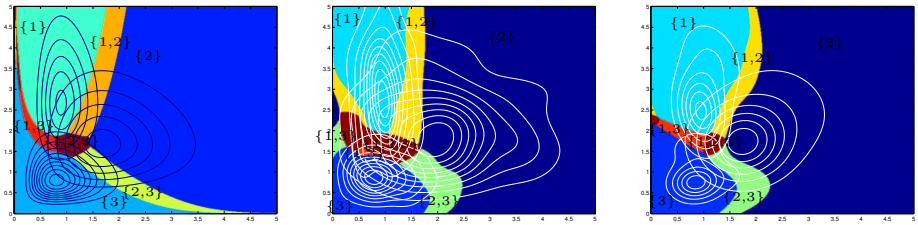

Fig. 2: Bivariate gamma problem: density probabilities and the corresponding partition. From left to right: Theoretical case (2.a) and an example of estimators in the case of Parzen (2.b) and GMM (2.c)

5.2 Experimental results and discussions

For both problems, P_E, P_I, $\overline{c}(Z^*)$ and $L(Z^*, \boldsymbol{\mu}^*)$ were computed for the optimal rule. Besides, their estimated values $\widehat{P_E}$, $\widehat{P_I}$, $\widehat{\overline{c}}(\widehat{Z^*})$ and $\widehat{L}(\widehat{Z^*}, \boldsymbol{\mu}^*)$ were computed on supervised partitions using an infinite test set and the theoretical $\boldsymbol{\mu}^*$. The mean and the standard deviations are reported in tables 1 and 2. An example of optimal rules built with GMM and Parzen densities are shown in figures 1.b and 1.c (for the 3-gaussian component distribution problem) and 2.b and 2.c (for the gamma distribution problem). They are obtained using the density estimated on a set with 200 observations per class. Results show that decision rules built with GMM and Parzen estimates are relevant. Their accuracy increases as long as the learning set size increases. Furthermore, GMM can be considered as a good family of non-symmetrical density estimators. They achieve results superior to Parzen estimators in term of losses, especially when the learning set size decreases. These results can be explained by several reasons: (i) Parzen estimators converge asymptotically to the real densities. (ii) Parzen density estimates are sums of as many local windows as the size of learning set, while the GMM estimates are compact functions parameterized according to a global search over all the learning set. Thus, the local nature of the Parzen estimator can lead to overfitting. Moreover, for the first problem, the errors, as the function of the number of observations of the GMM, decrease faster than those of Parzen. It is a logic result since the distributions correspond to the GMM hypothesis and GMM are more accurate fitting a multimodal distribution.

6 Conclusion

A decision rule learning method for multiclass problems with class-selective rejection and performance constraints is proposed. First, the optimal class-conditional density estimations are selected by maximizing the likelihood. Sec-

	GMM			Parzen			Theo.
	50 obs	100 obs	200 obs	50 obs	100 obs	200 obs	
$\widehat{P_E}$	0.084 ± 0.023	0.060 ± 0.011	0.053 ± 0.006	0.033 ± 0.015	0.034 ± 0.010	0.030 ± 0.008	0.050
$\widehat{P_I}$	0.082 ± 0.027	0.094 ± 0.016	0.098 ± 0.006	0.090 ± 0.011	0.087 ± 0.008	0.085 ± 0.006	0.100
$\widehat{\overline{c}}$	0.161 ± 0.040	0.133 ± 0.017	0.136 ± 0.018	0.248 ± 0.039	0.212 ± 0.028	0.207 ± 0.029	0.131
$\widehat{L}$	0.205 ± 0.033	0.147 ± 0.012	0.140 ± 0.011	0.224 ± 0.024	0.189 ± 0.015	0.177 ± 0.017	0.132

Table 1: Values of the theoretical and estimated $\widehat{P_E}$, $\widehat{P_I}$, $\widehat{\overline{c}}(\widehat{Z^*})$ and $\widehat{L}(\widehat{Z^*}, \boldsymbol{\mu^*})$ using GMM and Parzen estimators for the 3-gaussian component problem

	GMM			Parzen			Theo.
	50 obs	100 obs	200 obs	50 obs	100 obs	200 obs	
$\widehat{P_E}$	0.114 ± 0.021	0.110 ± 0.010	0.109 ± 0.008	0.083 ± 0.028	0.079 ± 0.018	0.086 ± 0.010	0.100
$\widehat{P_I}$	0.143 ± 0.023	0.143 ± 0.011	0.143 ± 0.012	0.147 ± 0.019	0.148 ± 0.012	0.147 ± 0.010	0.150
$\widehat{\overline{c}}$	0.235 ± 0.022	0.239 ± 0.012	0.234 ± 0.009	0.304 ± 0.044	0.290 ± 0.030	0.268 ± 0.016	0.229
$\widehat{L}$	0.251 ± 0.021	0.249 ± 0.008	0.247 ± 0.006	0.280 ± 0.025	0.260 ± 0.010	0.248 ± 0.003	0.229

Table 2: Values of the theoretical and estimated $\widehat{P_E}$, $\widehat{P_I}$, $\widehat{\overline{c}}(\widehat{Z^*})$ and $\widehat{L}(\widehat{Z^*}, \boldsymbol{\mu^*})$ using GMM and Parzen estimators for the gamma distributions problem

ond, the supervised rule, associated to the optimal estimates, is selected by optimizing the Lagrangian. GMM fitting and Parzen windows are used as density estimators. Supervised rules obtained by these two estimators are compared. Simulations on two synthetic problems were carried out. Results show that for some complex distributions like trimodal distributions, GMM fitting can be a good model yielding to an accurate decision rule. Moreover, GMM algorithm can be efficient, accurate and considerably fast-easy way to predict decision rules for a wide variety of statistical distributions. Outgoing work may tackle multiclass problems with evolutionary constraints. Varying constraints in the context of this method leads only to re-optimize $\boldsymbol{\mu}$. Thus, this flexible aspect allows to handle problems with continuous or discrete time sequences of constraints.

References

[1] E. Grall, P. Beauseroy and A. Bounsiar, Multilabel classification rule with performance constraints. In *proceedings of IEEE conference ICASSP'06*, France, 2006.

[2] E. Grall and P. Beauseroy, Optimal Decision Rule with Class-Selective Rejection and Performance Constraints, To appear in *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[3] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Roy. Statist*, 39:1-38, 1977.

[4] R. Nakano, Z. Ghahramani, G.E. Hinton and N. Ueda, SMEM Algorithm for Mixture Models, *Neural Computation*, 12:2109–2128, 2000.

[5] M.A.T Figueiredo and A.K. Jain, Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396, 2002.

[6] N. Vlassis, J.J. Verbeek and B. Kröse, Efficient greedy learning of gaussian mixture models, *Neural Computation*, 15:469–485, 2003.

[7] N. Vlassis and A. Likas, A Greedy EM Algorithm for Gaussian Mixture Learning, *Neural Processing Letters*, 15:77–87, Springer, 2002.

[8] E. Parzen, On estimation of a probability density function and mode, *Ann. Math. Statistics*, 33 :1065-1076, 1962.

[9] J. Q. Li and A.R. Barron, Mixture density estimation *Advances in Neural Information Processing Systems 12*, MIT Press, 2000.

[10] M. P. Wand, Fast computation of multivariate kernel estimators, *j-J-COMPUT-GRAPH-STAT*, 433–445,1999.

[11] N. Jrad, E. Grall and P. Beauseroy, Supervised learning rule selection for multiclass decision with performance constraints. In *ICPR 08*, USA, 2008.