Gaussian Mixture Model Classifiers

Bertrand Scherrer

February 5, 2007

This summary attempts to give a quick presentation of one of the most common classifiers today. Some key concepts are introduced in the first part. Using one particular piece of work, the basic principle of GMM classification will be inestigated. Finally, some additional points of interests not included here are mentionned.

1 Principle

1.1 Classification Model

Before presenting in more details the Gaussian Mixture Model (GMM) classification process, it is worthwhile to consider what "classification" actually means. According to [3], a "classification model" is made of three main parts :

- a **transducer** : in the case of music this would typically be the A/D conversion chain of the sound.
- a **feature extractor** : it extracts significant features from the information coming from the transducer (e.g. the spectral centroid of frames of signal). These features should be chosen in such a way that clear groups or classes of data can be identified.
- a **classifier** : its role is to assign the input data represented by their features to a number of different categories (e.g. different types of instruments).

1.2 GMM an unsupervised classifier

To describe the GMM classifier more accurately, we should not that it belongs to the "unsupervised" classifiers category [3], [8]. This means that the training samples of a classifier are not labelled to show their category membership [3]. More precisely, what makes GMM unsupervised is that during the training of the classifier, we try to estimate the underlying probablity density functions (pdf's) of the observations.

1.3 The building block of GMM, the multivariate Gaussian pdf

In the GMM classifier, the conditional-pdf of the observation vector with respect to the different classes is modelled as a linear combination of multivariate Gaussian pdf's. Each of them has the following general form :

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^2} e^{\left[\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]} \tag{1}$$

where :

- $\mathbf{x}$ is a d-component feature vector.
- $\boldsymbol{\mu}$ is the d-component vector containing the mean of each feature.

- Σ is the d-by-d covariance matrix, and $|\Sigma|$ is its determinant. It characterises the dispersion of the data on the d-dimensions fo the feature vector. The diagonal element σ_{ii} is the variance of x_i, and the non diagonal elements are the covariances between features. Often, the assumption is made that the features are independent [1]. Thus, Σ is diagonal and $p(\mathbf{x})$ can actually be written as the product of the univariate probability densities for the elements of $\mathbf{x}$:

It is important to note that each multivariate Gaussian pdf is completely defined if we know $\boldsymbol{\theta} = [\boldsymbol{\mu}, \Sigma]$.

1.4 The main assumptions

Extracting information from an unlabelled set of data can only be possible if we make certain assumptions [3]. The assumptions made to build a GMM are the following:

- The samples come from a known number c of classes.
- The a priori probabilities $P(\omega_j)$ for each class ω_j are known (they are often taken to be all equal to $\frac{1}{c}$ [7]).
- The forms of the class-conditional probability densities $p(\mathbf{x}|\omega_j, \boldsymbol{\theta_j})$ are known for all classes, $j = 1 \dots c$ (we assume that they are a sum of K multivariate gaussian probability density functions).
- The unknowns are the values of the c parameter vectors $\boldsymbol{\theta_{j=1\dots c}}$ (for each class, the respective weights of the N gaussian pdf's, as well as their mean vector and covariance matrix).

2 Example of application of the GMM classifier to music

GMM classifiers have been used in many fields from image pattern recognition [10] to text-independent speaker recognition [9]. Of course, it has also been used in the field of MIR.

In the following, we are going to investigate in detail one example of classification applied to music: [7]. This work was chosen for the clarity with which it presented the basic principle of the classification process. There are many other MIR papers using GMM. See [6, 4, 5].

In [7], the goal is to differentiate between eight different instruments[1] from very short (0.2s) monophonic musical extracts. The principle is as follows :

We consider a set $\mathbf{X}$ of m observations of d features (cepstral, mel-cepstral and LPC coefficients) : $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2} \dots \mathbf{x_m}]$.

Assuming that the observations are independent and identically distributed, the likelihood that the entire set of observations has been produced by a violin (the class C_0 for example) is :

$$p\left(\mathbf{X}|C_0\right) = \prod_{t=1}^{m} p(\mathbf{x_t}|C_0) \tag{2}$$

We assume that $p(\mathbf{x_t}|C_0)$ is a mixture of K multivariate gaussians :

$$p(\mathbf{x_t}|C_0) = \sum_{l=1}^{K} P(l|C_0).p(\mathbf{x_t}|l, C_0) \tag{3}$$

where $p(\mathbf{x_t}|l, C_0) = N(\boldsymbol{\mu}_{l,0}, \Sigma_{l,0})$ is the probability of $\mathbf{x_t}$ being produced by the gaussian of index l in the instrument class 0. On the other hand, $P(l|C_0)$ is the prior probability of having a gaussian l for the instrument class 0. It is a weight that changes with the class of instrument.

[1] bagpipe, clarinet, flute, harpsichord, organ, piano, trombone and violin

2.1 GMM Training

To achieve the classification of short segments of sound using feature vectors in musical instrument classes the GMM needs training. At this stage, one tries to estimate for all the classes of instrument the parameters of the GMM: $P(l|C_i)$, $\boldsymbol{\mu}_{l,i}$ and $\boldsymbol{\Sigma}_{l,i}$ with $l = 1 \dots K$.

One ideal way to do this is to use the maximum likelihood estimation (a.k.a. MLE). The MLE theoretically consists in finding the set of values $\boldsymbol{\theta}_{il} = [P(l|C_i), \boldsymbol{\mu}_{l,i}\boldsymbol{\Sigma}_{l,i}]$, for $l = 1 \dots K$, maximizing the likelihood of observing $\mathbf{x}$ as being produced by the instrument i. Nevertheless, in the case where all the parameters are unknown, "the maximum likelihood yields useless singular solutions" [3]. Thus there is a need for an alternate method.

In the surveyed literature, the Expectation Maximisation algorithm (a.k.a EM) [2] is the most often used solution to that problem. I will not go into the details of this algorithm. It is enough to say that it requires a fair amount of algebra [8] to derive a closed-form expression of the parameters of the GMM that correspond to a local extremum of the likelihood $p(\mathbf{X}|\ GMM\ parameters)$.

Once this is done we make a first guess on the values of the $\boldsymbol{\theta}_{il}$'s and start iterations from there. This algorithm is assured to converge to a local optimum [8].

Let us note that the training set provided to the GMM has to be well thought out in order for the model to be general enough (and avoid the common problem of overfitting the training data).

2.2 Classification test

Assuming all is well and that we have managed to train the GMM, we can proceed to the classification test : a feature vector $\mathbf{x_t}$ is said to belong to an instrument class if it maximizes $p(C_i|\mathbf{x_t}) = p(\mathbf{x_t}|C_i).p(C_i)$[2]. In the case where we assume that all the classes can occur with the same probability, we are actually concerned by maximizing $p(\mathbf{x_t}|C_i)$ for every possible class of instruments.

To go back to the specific study undertaken in [7], mel cepstral feature vectors (16 elements in a vector) were extracted from very short monophonic audio recordings and the GMM was of order 2. The system achieved an overall error rate of 37%. This is thus not a marvelous result but this might also be due to the very short length of the musical extracts.

3 Conclusion

Hopefully, this very quick introduction provides the reader with basic pointers for the comprehension and use of GMM classifiers in MIR. A lot of very interesting topics have not been included here but would be worth investigating more in depth. For example, the optimal choice of features to allow classification seems to be a very interesting subject. Also, the choice of alternate estimation techniques for the estimation of the model's parameters could also be very valuable. Finally, it would be worth finding studies of the influence of the order of the GMM on the quality of the classification.

References

[1] Brown, J. C., Houix, O., and McAdams, S. [2001]. Feature dependence in the automatic identification of musical woodwind instruments. *Journal of the Acoustical Society of America*, vol. 109: pp. 1064–1072.

[2] Dempster, P., Laird, N. M., and Rubin, D. B. [1977]. Maximum likelihood from incomplete data using the em algorithm. *Journal of the Royal Society of Statistics*, vol. 39(1): pp. 1–38.

[3] Duda, R. O. and Hart, P. E. [1973]. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc.

[2] this is the Bayes Rule

[4] Eggink, J. and Brown, G. J. [2003]. A missing feature approach to instrument identification. In *Audio Speech and Sound Processing, 2003, Proceedings of the International Conference of*.

[5] Heittola, T. and Klapuri, A. [2002]. Locating segments with drums in music signals. In *Proceeding of the 3rd International Conference on Music Information Retrieval*. pp. 271–272.

[6] Marolt, M. [2004]. Gaussian mixture models for extraction of melodic lines from audio recordings. In *Proceedings of the 2004 International Conference on Music Information Retrieval*.

[7] Marques, J. and Moreno, P. J. [1999]. A study of musical instrument classification using gaussian mixture models and support vector machines. Tech. rep., COMPAQ, Cambridge Research Laboratory.

[8] Moore, A. W. [2004]. Clustering with gaussian mixtures. URL http://www.autonlab.org/tutorials/gmm.html. Tutorial Slides.

[9] Reynolds, D. and Rose, R. [1995]. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, vol. 3(1): pp. 72–83.

[10] Sanderson, C. and Paliwal, K. [2002]. Likelihood normalization for face authentication in variable recording conditions. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. pp. I–301–I–304vol.1.