Gaussian mixture pdf in one-class classification: computing and utilizing confidence values

J. Ilonen, P. Paalanen, J.-K. Kamarainen, H. Kälviäinen

Department of Information Technology
Lappeenranta University of Technology
P.O.Box 20, FI-53851 Lappeenranta, Finland

Abstract

In this study a confidence measure for probability density functions (pdfs) is presented. The measure can be used in one-class classification to select a pdf threshold for class inclusion. In addition, confidence information can be used to verify correctness of a decision in a multi-class case where for example the Bayesian decision rule reveals which class is the most probable. Additionally, using confidence values – which represent in which quantile of the probability mass a pdf value resides ($[0, 1]$) – is often straightforward compared to using arbitrarily scaled pdf values. As the main contributions, use of confidence information in classification is described and a method for confidence estimation is presented.

1 Introduction

One-class classification, also called as novelty detection, outlier detection, or data description ([7]), can be used to detect uncharacteristic observations. One-class classification is necessary when samples can be obtained only from a single known class, for example, normal operation mode in motor condition monitoring where all failure modes are not known. One-class classification is also useful when the background class contains enormous variations making its estimation unfeasible, for example, background class in object detection: the background class should contain everything except the object to be detected.

The most straightforward method for obtaining an one-class classifier is to estimate the probability density of the data and to set a density value threshold. Gaussian mixture models (GMM) have been widely used in classification and general density estimation tasks, and they are also suitable for one-class classification. The expectation-maximization (EM) is a general method for estimating mixture model parameters, and the EM algorithm is proved to converge to the global maximum likelihood estimate if the overlap between Gaussians in the model is sufficiently small and there is a sufficient amount of data [5]. The proposed confidence measure is based on a density quantile defined for Gaussian mixture model probability density function.

The main contribution of this study is a method for computing confidence value which denotes in which quantile of the probability mass a pdf value resides: the confidence value is always between 0 and 1. The confidence is formulated to measure reliability of a class assignment – the lower the confidence, the more probable a classification mistake is. Confidence can be used in one-class classification for selecting a pdf threshold for class inclusion. In addition, confidence information can be used to verify the correctness of a classification decision in multi-class classification where, for example, the Bayesian decision rule is used to select the most probable class. Moreover, working with confidence values is often easier than working directly with arbitrarily scaled pdf values. The proposed confidence measure is defined for GMMs, but can be extended to any pdf fulfilling the required conditions. Two algorithms, one utilizing only training data and the second also generated data, are proposed. The latter approach overcomes the problem of limited training data producing coarsely quantized confidence values.

2 Gaussian mixture pdf

Finite mixture models and their typical parameter estimation methods can approximate a wide variety of pdfs and are thus attractive solutions for cases where single function forms, such as a single normal distribution, fail. Generally the basic distribution function can be of any type, but the multivariate normal distribution, the Gaussian distribution, is undoubtedly one of the most well-known and useful distributions in statistics, playing a predominant role in many

areas of applications [8].

A non-singular multivariate normal distribution of a D dimensional random variable $\boldsymbol{X} \mapsto \boldsymbol{x}$ can be defined as

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{1}$$

where $\boldsymbol{\mu}$ is the mean vector and Σ the covariance matrix of the normally distributed random variable $\boldsymbol{X}$. The multivariate Gaussian pdfs belong to the class of elliptically contoured distributions where the equiprobability surfaces of the Gaussian are $\boldsymbol{\mu}$-centered hyperellipsoids [8].

The Gaussian distribution in Eq. 1 can be used to describe a pdf of real valued random vector ($\boldsymbol{x} \in \mathbb{R}^D$). However, a similar form can be derived for complex random vectors ($\boldsymbol{x} \in \mathbb{C}^D$) as (e.g. [2])

$$\mathcal{N}^{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\pi^D|\Sigma|} \exp\left[-(\boldsymbol{x}-\boldsymbol{\mu})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{2}$$

where * denotes the adjoint matrix.

For a multimodal random variable, where values are generated by several randomly occurring independent sources instead of a single source, a finite mixture model can be used to approximate the true pdf. If the Gaussian form is sufficient for single sources then a Gaussian mixture model (GMM) can be used in the approximation. It should be noted that the underlying distributions do not necessarily need to be Gaussians but GMMs can also approximate many other types of distributions.

The GMM probability density function can be defined as a weighted sum of Gaussians

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c) \tag{3}$$

where α_c is the weight of cth component. The weight can be interpreted as *a priori* probability that a value of the random variable is generated by the cth source, and thus, $0 \leq \alpha_c \leq 1$ and $\sum_{c=1}^{C} \alpha_c = 1$. Now, a Gaussian mixture model probability density function is completely defined by a parameter list [1]

$$\boldsymbol{\theta} = \{\alpha_1, \boldsymbol{\mu}_1, \Sigma_1, \ldots, \alpha_C, \boldsymbol{\mu}_C, \Sigma_C\} \ . \tag{4}$$

3 Classification using confidence

In our case confidence is used to estimate the reliability of a classification result where a class label is assigned to an unknown observation. If the confidence is low it is more probable that a wrong decision has been made. Intuitively a value of class conditional pdf at an observation corresponds to decision confidence for favor of the corresponding class: the higher the pdf value is, the more class instances appear similar to the observation. However, using pdf values directly can be difficult since they are arbitrarily scaled. Confidence values are always in the range $[0, 1]$.

The most straightforward use of confidence is to find a pdf value threshold for a class [6]. The threshold can be used to decide whether an observation is sufficiently similar to the class in question. The threshold can be selected based on the training data, for example, by selecting a pdf threshold for which half of the training data yields higher pdf values (median). Another possibility is to select the threshold using confidence: finding a threshold which includes a certain proportion of the total probability mass. It should be noted that the pdf type is not limited to a single Gaussian distribution but mixture models with an arbitrary number of components can be used. The selection method can be easily generalized for other types of pdfs.

3.1 Interpretation of confidence

Definition 1 *Confidence value $\kappa \in [0, 1]$ and a confidence region $\mathcal{R} \subseteq \Omega$ for a probability density function $0 \leq p(\boldsymbol{x}) < \infty$, $\forall \boldsymbol{x} \in \Omega$. κ is a confidence value related to a non-unique confidence region $\mathcal{R}$ such that*

$$\int_{\Omega \setminus \mathcal{R}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \kappa \ . \tag{5}$$

The confidence in Definition 1 is easily interpretable via the confidence region $\mathcal{R}$. It is a region which covers a proportion $1 - \kappa$ of the probability mass of $p(\boldsymbol{x})$ because for probability distributions $\int_\Omega p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. It is clear that $\kappa = 1$ for $\mathcal{R} = \emptyset$ and $\kappa = 0$ for $\mathcal{R} = \Omega$. It should be noted that the confidence value has no use until the region $\mathcal{R}$ is defined as the minimal volume region which satisfies Definition 1. The minimal volume region is called the highest density region (HDR) [4]. The HDR can be non-unique (e.g., the uniform distribution).

A confidence value corresponds to a proportion of a probability mass that retains in the area $\mathcal{R}_k$ for the class ω_k. In a classification task where certain confidence for decision making is required the confidence value itself is not used but the confidence region $\mathcal{R}_k$ is important since a sample vector $\boldsymbol{x}$ is allowed to enter the class ω_k only if $\boldsymbol{x} \in \mathcal{R}_k$. If a sample is not within the confidence region of any of the classes it must be classified to a background class. The background class is a special class and samples assigned to the class need special attention; for example, in a two-class problem where data is available only from one class the background class may represent another class with an unknown distribution.

To find the confidence region a reverse approach can be used to find a pdf value τ which is at the border of the confidence region. It is assumed that the gradient of the pdf is never zero in the neighborhood of any point where the pdf value is nonzero. τ must be equal everywhere in the

border, otherwise the region cannot be the minimal volume region [4]. τ can be computed by rank-order statistics using the density quantile $F(\tau)$ (e.g., [4]) and by generating data according to the pdf.

3.2 Estimation algorithms

An analytical solution to the GMM confidence region cannot be solved and therefore estimation must be used. Estimation can be based on the GMM training data directly, or it can be based on randomly generated data derived from the estimated pdf.

A pdf value threshold for $p(\boldsymbol{x})$ can be selected with the help of training data. First, a cumulative pdf value histogram H for the data $\boldsymbol{x}_{1..N}$ is created (Algorithm 1). Second, the threshold can be found using the cumulative histogram H and the required confidence value $c = 1 - F(\tau)$ using Algorithm 2.

Algorithm 1 *Create cumulative confidence histogram H for pdf $p(\boldsymbol{x})$ with sample vectors $\boldsymbol{x}_{1..N}$ (training data)*

for $k = 1..N$ ***do***
  Calculate pdf value for $\boldsymbol{x}_k$, $H_k = p(\boldsymbol{x}_k)$
end for
Sort H in ascending order, H=sort(H)
Return H.

Algorithm 2 *Select pdf threshold value τ for the confidence value c using the confidence histogram $H_{1..N}$*

Select histogram position, m=round($c * N$)
Return $\tau = H_m$.

Confidence value for a new sample $\boldsymbol{x}$ can be calculated using Algorithm 3.

Algorithm 3 *Return confidence value c for a sample vector $\boldsymbol{x}$ using confidence histogram $H_{1..N}$ of the pdf $p(\boldsymbol{x})$*

Calculate pdf value for the sample vector $\boldsymbol{x}$, $p_x = p(\boldsymbol{x})$
Select position of the closest pdf value to p_x in H, $m = argmin_i |H_i - p_x|$
Return $c = m/N$.

In Algorithms 2 and 3 interpolation can be used instead of simply selecting the nearest value.

In the case of Gaussian mixture models, it may be beneficial to use randomly generated data. An algorithm for generating random data for any GMM is presented in Algorithm 4. An algorithm for generating data based on a single multi-variate Gaussian distribution has been presented in [8], and the algorithm has been extended here to GMM pdfs with multiple components.

Algorithm 4 *Generate N random samples, X, for a D-dimensional GMM of C components with weights $\alpha_{1..C}$, mean vectors $\boldsymbol{\mu}_{1..C}$ and covariance matrices $\Sigma_{1..C}$*

k=1
for $c = 1..C$ ***do***
  $T = chol(\Sigma_c)$ {*Cholesky decomposition* }
  {*Number of generated samples depends on the weight of the component,* α_c }
  for $1..round(\alpha_c N)$ ***do***
    $Z = randn(1 \times D)$ {*Generate D independent normally distributed ($\mu = 0$, $\sigma = 1$) random variables*}
    $X_k = ZT + \boldsymbol{\mu}_c$
    k=k+1
  end for
end for

4 Experiments

4.1 Data generation

The performance of the confidence and threshold computation methods with Algorithms 1 to 4 depends only on the amount of data – assuming that the data and the estimated GMM represent the same underlying distribution. If that holds, the only inaccuracy in the confidence values is caused by sparse sampling, i.e., limited amount of data. If the distributions deviate slightly from each other, which is typically caused by the GMM parameter estimation, the confidence values may be biased. If there is a large discrepancy between the distributions the confidence values may become completely useless, for example, all become binarized to either 0 or 1.

The number of required random samples increases with respect to the data dimensionality. The effect is demonstrated in Fig. 1. To avoid the issue of distribution mismatch, the GMM pdf was generated semi-randomly and data was derived from the generated GMMs. A pdf threshold for a D-dimensional Gaussian with confidence $c = 0.5$ was searched. First, random data was generated with Alg. 4 and then a pdf threshold was searched with Algs. 1 and 2. For each value of D the number of needed samples was evaluated repeatedly; each evaluation consisted of creating a semi-random covariance matrix and finding a number of samples at which standard deviation of the found pdf threshold value was varying at most 1% from the mean value. The number of needed samples increased linearly with the data dimensionality (Fig. 1). The linear dependency is as expected based on the data generating Algorithm 4: a D-dimensional sample is generated using D random numbers, despite the fact that the size of the covariance matrix increases quadratically.

4.2 Image feature detection

In the second example we demonstrate the use of confidence information in image processing – detection of image features [3]. In face detection methods based on detection of smaller facial parts, image features, features are extracted from every spatial location and for all locations one of the

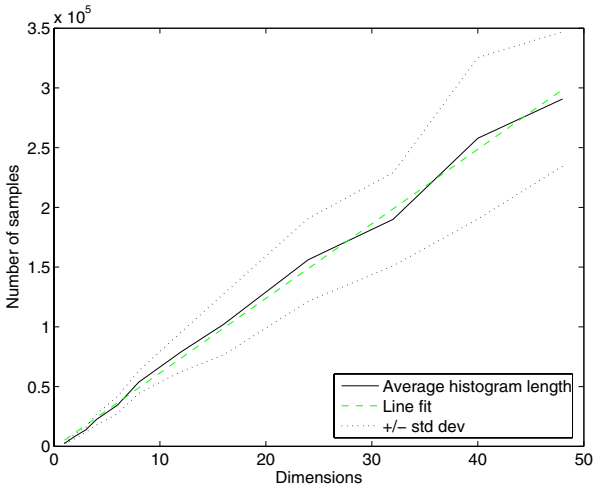

Figure 1. Required number of generated samples for a pdf threshold estimate ($c = 0.5$).

feature classes is assigned (e.g., [3]). The final detection is based on the inspection of spatial configuration of the image features. In Fig. 2(a) is shown a facial image and 10 marked image features [3]. In the training phase Gabor features were extracted from training set images and class conditional GMM pdfs were estimated for the each feature class. In Fig. 2(b) is shown a pdf surface for the image feature number 7 corresponding to the left (in the image) nostril. In Figs. 2(c) and 2(d) are shown only the confidence regions corresponding to 0.5 and 0.95 confidence values. It is clear that the correct image feature location, the left nostril, was already included in the 0.95 confidence region, and thus, image feature search space for the next processing level was reduced dramatically [3]. Using the confidence information image feature based detection and recognition methods can be sped up significantly.

5 Conclusions

The main contribution of this study was a method for computing confidence values for GMM pdfs by utilizing an approximation approach. A low confidence corresponds to the high probability of a wrong classification decision, and conversely, a high confidence that the classification decision was likely to be correct. The proposed measure is based on the pdf density quantile. The results are useful in reducing search space in the image feature based object detection and recognition.

Acknowledgments

This work has been supported by Academy of Finland (project# 204708).

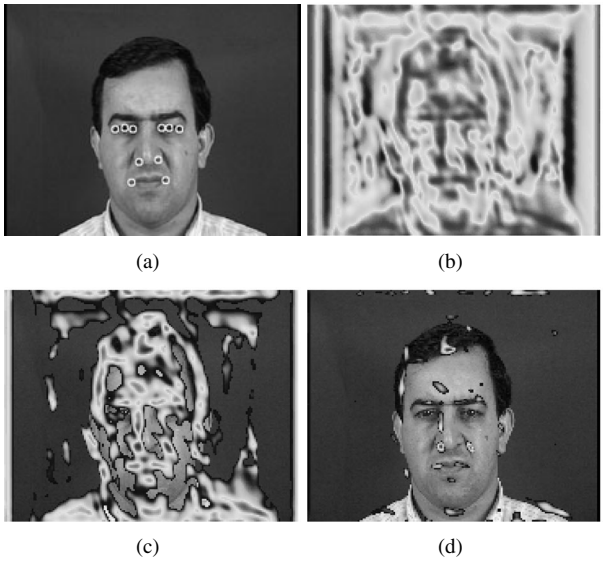

Figure 2. Example of using density quantile for defining confidence regions : (a) face image and 10 marked image feature classes; (b) pdf value surface for the left (in the image) nostril class; (c) confidence threshold 0.5 ($F(\tau) = 0.5$); (d) confidence threshold 0.95 ($F(\tau) = 0.05$).

References

[1] B. Everitt and D. Hand. *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics. Chapman and Hall, 1981.

[2] N. Goodman. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, 34(1):152–177, March 1963.

[3] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, and H. Kälviäinen. Affine-invariant face detection and localization using GMM-based feature detector and enhanced appearance model. In *Proc. of the 6th Int. Conf. on Automatic Face and Gesture Recognition*, pages 67–72, Seoul, Korea, 2004.

[4] R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, May 1996.

[5] J. Ma and S. Fu. On the correct convergence of the EM algorithm for Gaussian mixtures. *Pattern Recognition*, 38:2602–2611, 2005.

[6] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th international confenrence on artificial neural networks*, pages 442–447, 1995.

[7] D. Tax and R. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

[8] Y. Tong. *The Multivariate Normal Distribution*. Springer Series in Statistics. Springer-Verlag, 1990.