Confidence Estimation in Classification Decision: A Method for Detecting Unseen Patterns

Pandu R Devarakota[†‡] and Bruno Mirbach[†]

[†]*IEE S. A., ZAE Weiergewan,*
11, rue Edmond Reuter,
L-5326 Contern, Luxembourg
E-mail: {pdv,bmi}@iee.lu

Björn Ottersten[‡]

[‡]*Signal Processing Group*
School of Electrical Engineering
Royal Institute of Technology (KTH)
SE–100 44 Stockholm, Sweden

The classification task for a real world application shall include a confidence estimation to handle unseen patterns i.e., patterns which were not considered during the learning stage of a classifier. This is important especially for safety critical applications where the goal is to assign these situations as "unknown" before they can lead to a false classification. Several methods were proposed in the past which were based on choosing a threshold on the estimated class membership probability. In this paper we extend the use of Gaussian mixture model (*GMM*)to estimate the uncertainty of the estimated class membership probability in terms of confidence interval around the estimated class membership probability. This uncertainty measure takes into account the number of training patterns available in the local neighborhood of a test pattern. Accordingly, the lower bound of the confidence interval or the number of training samples around a test pattern, can be used to detect the unseen patterns. Experimental results on a real-world application are discussed.

Keywords: Pattern classification, confidence based classifier, density estimation, confidence intervals

1. Introduction

A statistical classifier estimates for each pattern a class membership probability. Correspondingly, the pattern is assigned to a class which has a maximum class membership probability. In this respect, challenges are to reduce the misclassification rate and to reject unseen patterns by assigning them with a low confidence (reliability). To overcome the above challenges, a classifier with a *reject* option is one approach which is often exercised in the literature.[2] If the conditional probability exceeds a required minimum threshold, then the respective decision is rejected. As a consequence of this method, a test pattern close to a decision boarder implicitly defined by the classifier is prone to be rejected, while a test pattern far away from the boarder will be assigned to a class. This approach, however needs an optimum Bayesian decision rule.[2] It is, however, known that it is very difficult to determine the Bayesian error in real problems[1] as it needs a complete knowledge of the distribution of data. As a consequence of the above fact, a statistical classifier in practice overly estimates the class membership and thus always has an uncertainty related to each estimate. In order to use the classifier with a reject option the above uncertainty should be subtracted upon the classification.

To our best knowledge there exists yet no a method which is able to define statistically the uncertainty in the estimated class membership probability which can allow further to define a confidence interval for the same. In this paper we extend the use of Gaussian mixture model to estimate the uncertainty of the estimated class membership probability. The basic idea of the proposed approach is that it takes into account the number of training patterns falling in the local neighborhood of a test pattern. The motivation behind this approach is: If there is a small number of training patterns falling in the local neighborhood of a test pattern, the uncertainty in the estimate will be large and if there is a large number of training patterns around a test pattern, the uncertainty will be small.

The paper is organized as follows: The proposed method for estimating the number of training patterns in the neighborhood of a test pattern is presented in Section 2 and its simplification using Gaussian mixture models is described in Section 3. The computation of confidence intervals which takes into account the uncertainty is also introduced in the

same section. Results are presented in Section 4, and the concluding remarks are given in Section 5.

2. The density estimation of a test pattern

Let D be a training set $\{(x_i, y_i)\}_{i=1}^{N}$ with each sample $x_i \in \mathcal{R}^d$ and the corresponding output label $y_i \in 1, \ldots, N_c$, where N_c is the number of classes.
Then, the density of training patterns at a data point x is given by

$$\rho(x) = \sum_{i=1}^{N} \delta(x - x_i) \tag{1}$$

where δ denotes the Dirac delta function.
Accordingly the number of training patterns falling in the region R centered at x can be calculated as

$$N_R(x) = \int_{x' \in R(x)} \rho(x')dx' \tag{2}$$

The solution to Eq. 2 may result a discrete value and in practice, a continuous value of density is desirable. This can be achieved by considering a window function φ which is centered at x and covers a region R. Equation. 2 may be rewritten now as

$$N_R(x) = \int \rho(x')\varphi(x - x')dx' \tag{3}$$

The window function φ has a flexibility that the region R can be tuned by including width parameter r into it.

Consider a special case where a radial basis function of width r is selected as a window function. The number of training patterns falling in a sphere of radius r which is centered at x is now calculated as

$$N_r(x) = \int \rho(x') \frac{1}{\sqrt{(2\pi r^2)^d}} \exp\left(-\frac{(x' - x)^2}{2r^2}\right)dx' \tag{4}$$

Furthermore, Eq. 4 can be generalized by considering an hyperellipsoid centered at x as a window function (nothing but a Mahalanobis distance):

$$N_r(x) = \int \rho(x') \frac{1}{\sqrt{(2\pi r^2)^d \det C}} * \exp\left(-(x' - x)^T C^{-1}(x' - x)/2r^2\right)dx' \tag{5}$$

This approach should be considered if the components of training data show some correlation. In that case C should be chosen to be the covariance matrix of the training set D.
The formation in Eq. 5 is straightforward to compute the number of training patterns in the hyperellipsoid centered at the test pattern x. However it needs the storage of the entire training data in the memory which is computationally very demanding and hence not useful for a real-world application. In this work, we consider a Gaussian mixture model approximation to estimate $\rho(x)$ which is reasonable and sufficient in practice, allow us to develop explicit expressions to compute $N_r(x)$ and requires only a small number of estimated parameters as input.

3. The Gaussian mixture model

Assume that the density function of the training data is expressible as the mixture of Gaussian functions in the following way:

$$\rho(x) = \sum_{k=1}^{K} \frac{N_k}{\sqrt{(2\pi)^d \det S_k}} * \exp[-(x - \mu_k)^T S_k^{-1}(x - \mu_k)/2] \tag{6}$$

where K represents the number of Gaussian functions in the mixture, μ_k the center of the k-th Gaussian function, S_k a matrix describing the widths of the k-th Gaussian function, d the dimension of the training data and the N_k represents normalization factors fulfilling

$$N = \sum_{k}^{K} N_k \tag{7}$$

where N is the total number of training patterns. The parameters μ_k, S_k and N_k can be computed offline using the Expectation-Maximization algorithm.[3]
With $\rho(x)$ as defined in Eq. 6, the integration in Eq. 5 can easily solved by the convolution of two Gaussian functions and the final expression may be written as

$$N_r(x) = \sum_{k}^{K} N_k (\det(T_k S_k^{-1}))^{1/2} * \exp[-\frac{1}{2}(x - \mu_k)^T \cdots S_k^{-1}(1 - T_k S_k^{-1})(x - \mu_k)] \tag{8}$$

where T_k^{-1} is given as follows,

$$T_k^{-1} = \frac{C^{-1}}{r^2} + S_k^{-1} \tag{9}$$

Finally, the Eqn. 8 for computing $N_r(x)$ can be brought into the form given below:

$$N_r(x) = \sum_k N_k^{'} \exp[(x-\mu_k)^T {S_k^{'}}^{-1}(x-\mu_k)/2] \tag{10}$$

where $N_k^{'} = N_k(\det(T_k S_k^{-1}))^{1/2}$ and $S_k^{'} = (1 - T_k S_k^{-1})^{-1} S_k$.
The parameters $N_k^{'}$ and $S_k^{'}$ can be computed off-line and the number of training samples around a test sample can now easily be calculated using the explicit form of the Eq. 10.

Let $p_e(x)$ be the estimated class membership probability and $N_r(x)$ be the number of training patterns in the neighborhood of the test pattern x. Then, one can calculate the confidence interval for p_e using the standard Wilson interval for a Binomial distribution[a],[4]

$$\begin{aligned} p_\pm &= \frac{p_e + \frac{\lambda^2}{2N_r}}{1+\frac{\lambda^2}{N_r}} \pm \Delta(p_e) \\ \Delta(p_e) &= \frac{\sqrt{\lambda^2/N_r}\sqrt{p_e(1-p_e)+\lambda^2/(4N_r)}}{1+\frac{\lambda^2}{N_r}} \end{aligned} \tag{11}$$

where $p_\pm$ is the upper, respectively the lower bound of the confidence interval (CI), and λ is determined by the confidence level that has been chosen. $\Delta(p_e)$ can be regarded as uncertainty in the estimated class membership probability p_e. A reject of pattern as "unseen" can then be easily established by applying a threshold to the lower bound of the confidence interval. Alternatively one could also establish a reject based on the width Δ of the confidence interval or based on the number of training samples $N_r(x)$ itself. Here, the lower bound of the confidence interval is chosen.

4. Experimental Results

The aim of experiments is not to improve the classification rate but to see how the classifier reacts to the unknown patterns which were not included in the training data. The effectiveness of the proposed method to estimate the confidence is tested on a real-world application[5] and the results are then compared with a rejection threshold on a class membership probability alone which is a *state-of-art* in the literature.[2] In Ref. 5, an occupant classification system was evaluated where the goal is to detect the occupancy of passenger seat by an optical system and then classify it into one of the following four classes 1. Empty seat 2. Rearward facing infant seat (RFIS) 3. Forward facing child seat (FFCS) 4. Adult (P) (see Fig. 1). A low-resolution range sensor based on the time-of-flight principle is employed for capturing the image sequences. For more details about the system see Ref. 5. The challenge is to cope with the large variation in the scene. If designed properly, in addition to the classification task, such a system should also be able to overcome the unexpected situations that may occur in real life which were not defined in the training data. The goal of confidence measure is there to reject those situations by assigning them as "unseen".

The data set used for training and testing the classifier is shown in Table 1. The number $n \times m$ represents the number of sequences n and the number of frames m in each sequence respectively. In order to take the variation of occupant scenes into account, different occupants with varying hand postures, leg postures, and torso gestures were recorded. Two-third of the data used to train the classifier and one-third is used for testing. Refer to the Ref. 5 for details about the Camera used in this application, pre-processing, and feature extraction. As an unseen data, we recorded few sequences of data where the passenger seat is occupied with an object, for instance, boxes, rucksack which were not in the part of training data. One example for unseen data where the passenger seat is occupied with a box can be seen in Fig. 1. The unseen data information is also listed in Table 1. For the classification task, the classifier based on a polynomial regression is considered. The advantage of the Polynomial classifier is that it makes no assumptions on the statistical distribution of the data and leads, at least when using the least

Table 1. Dataset used for evaluating the confidence measure

Class	No.of Sequences
Empty	18 × 50
RFIS	236 × 20
FFCS	30 × 50
P	45 × 200
Unseen data	4 × 50

[a]This problem of estimating the confidence interval is well established for a so-called Bernoulli process, which follows statistically a Binomial distribution.

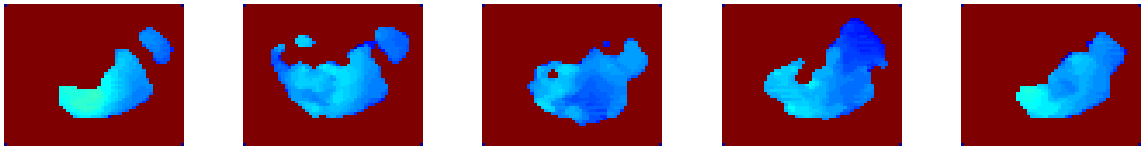

Fig. 1. Range images of possible occupancy in a vehicle (a) Empty (b) RFIS (c) FFCS (d)Adult respectively. Blue represents the closest point, and red represents the furthest point to the camera. One possibility of unseen class is shown in (e) where the passenger seat is occupied with a box. Note that the shown images are preprocessed images.

mean-square error optimization criterion, to a closed solution of the optimization problem. The more details about the Polynomial classifier can be found in Ref. 6. But the output of a polynomial classifier does not corresponds to a probabilistic output thus the estimate of class membership probability is not readily available. A method to transform the SVM classifier output into a probabilistic output using a sigmoid approximation was proposed.[7] The same method is adapted here to produce probabilistic outputs for the Polynomial classifier.

The evaluation of confidence is as follows: On the basis of the performance on the training set patterns, we establish a reject criterion once based on the estimated class membership probability and once based on the lower bound of the CI. In both cases a rejection threshold is chosen such that less than 5% of the training data are rejected. After that the reject criterion is applied to the test data and to the unseen data based on the chosen threshold value. A radius of $r = 1$ and $K = 10$ are chosen in Eq. 9 for all experiments, and a 95% confidence level is chosen to calculate the λ value in Eq. 11.

In Fig. 2, the percentage of patterns rejected based on the class membership probability for the train data, test data and unseen data is plotted at different thresholds. The vertical dotted line shows the corresponding threshold value to reject 5% of the training data. It can be seen that at the respective threshold, 5.5% of the test data and only 3% of the unseen data is rejected. Figure 3 shows the percentage of patterns rejected based on the lower bound of CI for the train data, test data and unseen data is plotted at different thresholds. At the respective threshold, the lower bound of the CI is able to reject 86.5% of the "unseen" data at the expense of only 4% rejection in the test data. Thus taking the uncertainty of the class membership clearly shows the ability of detecting the "unseen" patterns.

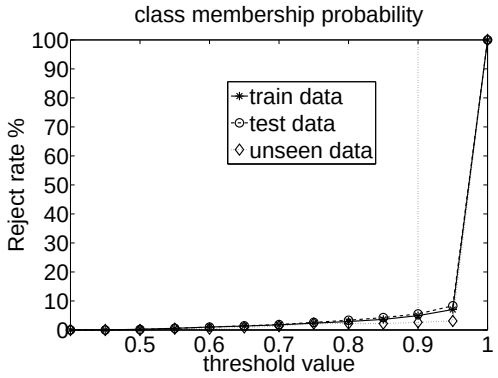

Fig. 2. % of patterns rejected based on the class membership probability

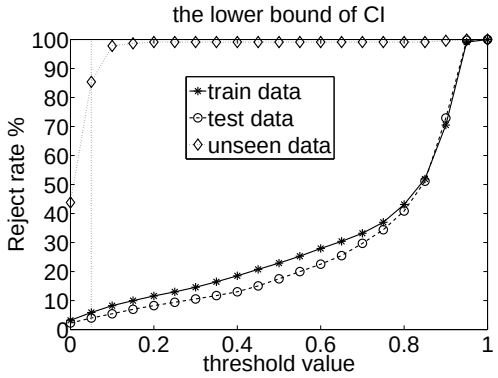

Fig. 3. % of patterns rejected based on the lower bound of confidence interval.

5. Conclusion

In this paper, we have formulated a method for estimating the uncertainty (reliability) of the class membership probability estimated by a statistical classi-

fier. The basic idea of the approach is, it takes into account the the number of training patterns in the neighborhood of a test pattern. The density of the training data around a test pattern is defined and then an explicit expression is derived for calculating the number of training patterns in a neighborhood of a test pattern. The use of Gaussian mixture model is extended for this purpose. It is shown that this method needs a small number of estimated parameters and thus this method does not need the training patterns to be stored in the memory. Furthermore, the uncertainty is represented in terms of confidence interval for the estimated class membership probability. For this a standard Wilson interval for a Binomial distribution is considered. Though we have shown a particular application in a pattern recognition where the presented method is useful, it can be applicable to several fields of applications where the "novel pattern" detection is important.

6. Acknowledgements

This project is funded by IEE S. A. and Luxembourg International Advanced Studies in Information Technology (LIASIT), Luxembourg.

References

1. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, 1991).
2. C. Chow, *IEEE Trans. on Information Theory* **It-16 1**, 41 (1970).
3. M. A. Figueiredo and A. K. Jain, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24**, 1(March 2002).
4. L. D. Brown, T. T. Cai and A. DasGupta, *Statistical Science* **16**, 101(May 2001).
5. P. R. Devarakota, B. Mirbach, M. Castillo-Franco and B. Ottersten, *To appear in IEEE Trans. on Vehicular Technology* (2006).
6. J. Schürmann, *Pattern Classification: Statistical and Neural Network based Approach* (John Wiley and Sons, Inc., New York, 1990).
7. J. Platt, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods* (In: A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (eds.): *Advances in Large Margin Classifiers, Cambridge, MA, 2000).*