# Promoting Diversity in Gaussian Mixture Ensembles: An Application to Signature Verification

Jonas Richiardi, Andrzej Drygajlo, and Laetitia Todesco

Institute of Electrical Engineering
Swiss Federal Institute of Technology Lausanne
Switzerland
{jonas.richiardi,andrzej.drygajlo}@epfl.ch
http://scgwww.epfl.ch/

**Abstract.** Classifiers based on Gaussian mixture models are good performers in many pattern recognition tasks. Unlike decision trees, they can be described as stable classifier: a small change in the sampling of the training set will produce not a large change in the parameters of the trained classifier. Given that ensembling techniques often rely on instability of the base classifiers to produce diverse ensembles, thereby reaching better performance than individual classifiers, how can we form ensembles of Gaussian mixture models? This paper proposes methods to optimise coverage in ensembles of Gaussian mixture classifiers by promoting diversity amongst these stable base classifiers. We show that changes in the signal processing chain and modelling parameters can lead to significant complementarity between classifiers, even if trained on the same source signal. We illustrate the approach by applying it to a signature verification problem, and show that very good results are obtained, as verified in the large-scale international evaluation campaign BMEC 2007.

## 1 Introduction

Successul ensembling methods such as bagging [3] and boosting [5] rely on the fact that the ensemble member classifiers are *unstable*, that is, a small change in the sampling of the training set will produce a large change in the trained classifier. Unstable classifiers include decision trees and neural networks [3], while others such as naïve Bayes are considered stable [8]. In reality, there is a continuum of stability, in the sense that the *amount* of output change incurred by different classifiers with respect to changes in the training set is not simply binary ("stable" or "unstable") [2].

Training several unstable classifiers with different sampling of the training set is one way to produce an ensemble that is diverse. The hope is that the training procedure produces classifiers whose output is complementary: they yield erroneous outputs for different samples. By combining these classifiers, the total variance can be reduced, typically leading to reductions in expected error rates.

B. Schouten et al. (Eds.): BIOID 2008, LNCS 5372, pp. 140–149, 2008.

In many applications dealing with real-life signals, a classifier that systematically yields good results is the Gaussian mixture model (see e.g. [13]). Example applications are speaker verification [16] or face recognition [21]. Leaving out effects of critically small training sample sizes with respect to the model complexity, Gaussian mixture models can be considered as stable classifiers. Given that multiple-classifier systems can outperform single-classifier systems on a large number of tasks and datasets [12], it would seem beneficial to build ensembles of Gaussian mixture classifiers. However, as pointed out above, diversity is an important factor for successful ensembling. How, then, can we increase diversity in ensembles of stable classifiers?

Recent work has shown that adding components to stable classifiers before ensembling could improve results over standard techniques such as bagging for these classifiers classifiers. For example, in the Random Oracle technique applied to naïve Bayes classifiers [19], the training set is split at random between the two classifiers, and at test time the base classifier is also selected at random. Another technique based on a hybrid of naïve Bayes and decision trees, called Levelled Naïve Bayesian Trees [22], is to grow a decision tree whose leaves are naïve Bayes classifiers. The hope there is that the naïve Bayes classifiers will inherit the instability of the decision tree growing procedure, and make them more amenable to boosting.

In this paper, to optimise the coverage of the ensemble, we propose instead to act at different levels of the pattern recognition processing chain of individual classifiers in order to increase diversity in ensembles of Gaussian mixture classifiers, and note that this does not prevent the application of other destabilising techniques. We should also note that, while it seems "diversity" is a desirable property of classifier ensembles in order to reduce error rate, there is no consensus on how to measure it and how it relates to ensemble performance [11], although theoretical work in this area is progressing [14].

The rest of this paper is laid out as follows: In Section 2 we present in more details techniques that can be used to increase diversity in ensembles of stable classifier. In Section 3, we show the detailed application of these principles to a multiple-classifier signature verification system based on Gaussian mixture models. In Section 4 we provide experimental results on a signature verification database, and Section 5 concludes the article.

## 2 Increasing Diversity in Ensembles of Stable Classifiers

A pattern recognition systems consists of a front-end responsible for extracting features, a training procedure to learn the parameters of the classifier, and a testing algorithm to obtain soft or hard output from the classifier. We will now examine these levels in more details and how they can be modified to influence the output of a classifier, which in turn can promote diversity in an ensemble. In the application field of biometrics, some of these techniques fall under the general heading of "multibiometrics" [20].

### 2.1 Changes to the Front-End

The front-end to pattern recognition systems uses a signal processing chain that starts with real-world analogue signals. A schematic view is shown on Figure 1.

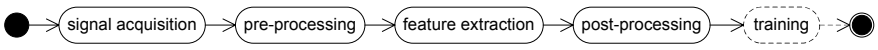



**Fig. 1.** Front-end for pattern recognition

Changes in any of the processing steps will affect all other steps further downstream, and lead to various amounts of classifier diversity. Even within the same modality (say, infrared images), changing the sensor at the **signal acquisition** stage can lead to significant differences between classifiers. In this regard, multimodal pattern recognition can be seen as a way to obtain diverse ensembles.

The **pre-processing** performed on the data can have a large influence on the feature extraction process. Filtering, denoising, imputing missing data and other linear and non-linear transformations of the digitised signal can lead to significant differences further down the processing chain.

The representation of the signal as vectors of features typically involves a non-linear transformation of the pre-processed signal. For example, the use of Fourier transforms and related transforms such as the DCT at the **feature extraction** stage change signal representation and may permit the extraction of features that lead to classifiers complementary to those trained on other signal representations. This technique is used in many applications such as language recognition, where different parameterisations of speech are often combined [15], or fingerprint recognition, where minutiae can be combined with skin pores [10]. Even within the same signal representation, it is possible to use random feature subspace methods [7] to purposefully obtain diverse classifiers.

Finally, the **post-processing** stage, which typically consists of some form of statistical normalisation of the feature vectors (mean removal being typical in speech applications [6]), can also introduce important changes to the trained parameters of the classifier by applying linear or non-linear transformations to the original feature space.

### 2.2 Changes in the Sampling of the Training Set

By our definition of stability, varying the sampling of the training set, a common strategy for achieving diversity in ensembles, will not be effective for increasing diversity in ensemble of stable classifiers (although see [19] for a more sophisticated approach). Thus, we propose to concentrate efforts on other parts of the pattern recognition system.

### 2.3 Change in Model Complexity

Classifiers implemented as statistical models (Gaussian mixture models, generative Bayesian networks) form a family in which the number of parameters has a

great influence on classification results. For example, modifying the covariance matrix structure (say, from diagonal to full) can substantially alter the output of the classifier. Likewise, by modifying the number of hidden variables in a Bayesian network corresponding to the number of components in a mixture of Gaussians, and thereby changing the number of parameters in the model, it is possible to decorrelate stable models that are trained from feature vectors where everything else in the front-end (acquisition, pre-processing, feature extraction, post-processing, samplig of the training set) is identical.

### 2.4 Change in Scoring Procedure

The same model can be used to compute a score in different ways. Depending on the model type, this is a way to promote diversity. In this regard, the recent technique presented in [23], whereby a hidden Markov model is used to produce likelihood output and a Viterbi-related output which are then combined, can be seen as a way to exploit complementarity in classifier output. However, for GMMs, it is likely that gains obtained from combining all-components scoring with top-components-scoring[1] would be small.

## 3 Application: A Gaussian Mixture Ensemble for Signature Verification

In this section, we present an application of the techniques exposed in Section 2 to the problem of signature verification, where the Gaussian mixture model is one of the best-performing classifiers [17]. The goal is to train a diverse set of signature verification classifiers, so that they can be effectively combined. The Gaussian mixture ensemble we present consists of $L$ =6 different Gaussian mixture model classifiers. In fact, since biometric verification problem can be cast as a series of 2-class problems, each of the $U$ users is modelled by one of the $U$ Gaussian mixture ensembles.

We do not use a measure of diversity based on the label (hard, binary) outputs of the classifiers [11], but rather the normalised mutual information between the scores (soft, continuous) outputs of the classifiers. We assume that having lower mutual information between pairs of classifiers is equivalent to having a higher diversity in the ensemble[2]. We use the following definition for normalised mutual information:

$$\bar{I}(Sc_1; Sc_2) \triangleq \frac{I(Sc_1; Sc_2)}{\sqrt{H(Sc_1)H(Sc_2)}}, \tag{1}$$

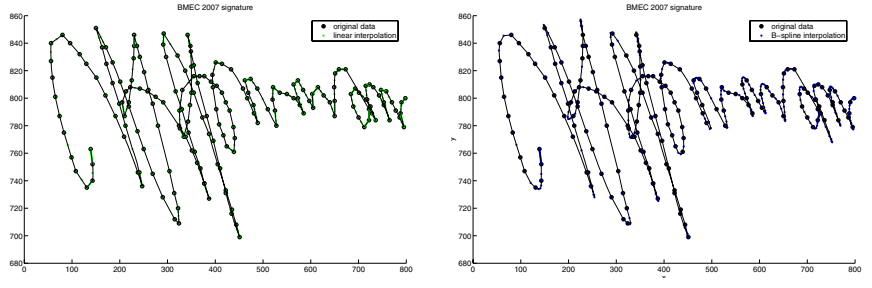
where $I(Sc_1; Sc_2)$ is the mutual information between the scores output of two classifiers, and $H(Sc_l)$ is the entropy of the scores output of the $l$th classifier.

[1] This is a common technique in speaker recognition [1], where high model orders and large datasets warrant the summing of *some* of the Gaussian components in the likelihood computation

[2] Using conditional mutual information would allow us to take into account the effect of already having included certain classifiers in the ensemble.

### 3.1 Preprocessing

On some low-power signature acquisition platforms such as personal digital assistants, data acquisition may produce missing values intermittently. Missing data is also a frequent occurence in slow and fast strokes. In this case, an effective approach is to interpolate the missing data. By using different interoplation algorithms, or none at all, it is possible to introduce variability in the signal which will be reflected further down the processing chain. Figure 2 shows the result of two different interpolation methods on the same data. Looking at a single classifier, it is not obvious which interpolation method is the best in terms of accuracy.




(a) Linear interpolation on pen-down points only

(b) B-spline interpolation

**Fig. 2.** Signature preprocessing for recovery of missing data on BMEC 2007

A second pre-processing technique that could lead to diversity is rotation normalisation. Indeed, in some situations, such as handheld device-based acquisition, it is likely that the orientation of the signature with respect to the horizontal axis of the acquisition surface can be very variable. We estimate the principal axis of the signature by eigendecomposition: The eigenvector associated with the largest eigenvalue indicates the axis of greatest variance. Again, from looking at the accuracy of a single classifier it is not obvious whether this really is of help, but it can be used to force diversity in an ensemble.

The preprocessing used by the local and global classifiers in our ensemble is detailed in Table 1.

### 3.2 Feature Extraction

In the parametric paradigm, local, segmental, or global parameters are computed from the pre-processed signals and used as features.

*Local features* are extracted at the same rate as the incoming signal: that is, each input sample data vector corresponds to a local feature vector.

*Segmental features* are extracted once the signature has been cut into segments. A segment typically consists of a sequence of points for which some definition of coherence holds.

*Global features* summarise some property of the complete observed signature; for instance the total signing time, pen-up to pen-down ratio, etc.

Changing the signal representation and combining the resulting classifiers is a common technique in pattern recognition, and has been applied also to signature verification [4]. Our Gaussian mixture ensemble consists of 5 classifiers trained on local features, and 1 classifier trained on global features (see Table 1).

**Table 1.** Details of classifier in the ensemble

| Name | $GL_1$ | $GL_2$ | $GL_3$ | $GL_4$ | $GL_5$ | $GG$ |
|---|---|---|---|---|---|---|
| Interpolation | LI | B-S | LI | LI | B-S | LI |
| Rotation | y | n | y | n | y | n |
| feature set | 1 | 1 | 1 | 2 | 3 | 4 |
| user model order | 24 | 36 | | | | 2 |
| world model order | 4 | | | | | |

### 3.3 Modelling

Diversity can be enforced in ensembles of Gaussian mixture models by changing the number of parameters used for the constituant classifiers, for instance by changing the type of covariance matrix (diagonal, full, spherical...), or by choosing a different number of Gaussian components in the mixture. A further way of increasing diversity is by using a MAP adapation scheme instead of direct training.

### 3.4 Diversity in the Ensemble

The 5 GMM classifiers based on local features, denoted $GL_{1...5}$, and the GMM classifier based on global features, denoted $GG$, use the specific combinations of preprocessing, feature extraction, and model orders shown in Table 1. In the table, LI referes to linear interpolation, while B-S refers to B-spline interpolation. *Rotation* indicates whether rotation normalisation is performed or not. The feature sets are as follows: feature set 1 comprises $\{x_t, y_t, \Delta, \Delta\Delta\}$, where $x_t$ and $y_t$ are the sampled horizontal and vertical position of the pen. The $\Delta$ and $\Delta\Delta$ features are numerically approximated first, respectively second derivatives of the base features. Feature set 2 is $\{x_t, y_t, \theta_t, v_t, \Delta, \Delta\Delta\}$, where $\theta_t$ is the writing angle and $v_t$ is the instantaneous writing speed. Feature set 3 is $\{x_t, y_t, z_t, \Delta, \Delta\Delta\}$, where $z_t$ is a binary variable representing pressure. Feature set 4 comprises 11 global features, described in [18]. Lastly, different number of components are used in the mixture, denoted *user model order*.

In terms of classifier output, these changes result in a diverse ensemble of GMMs, with complementarity clearly showing on Figure 3. As could be expected, the different parameterisation of the signal (local or global) result in the largest diversity, but it can also be observed that changing the model order or the preprocessing can also lead to important changes in classifier output. To put the

results in perspective, the normalised mutual information between a vector $\mathbf{x}$ consisting of 1000 samples drawn at random from a uniform distribution between 0 and 1 and the vector-valued $sin(\mathbf{x})$, a near-linear relationship, is 0.75. The normalised mutual information between two vectors of dimension 1000 randomly drawn from a uniform distribution between 0 and 1 is 0.02. Thus, it can be seen that significant reductions in dependence between classifiers can be achieved by applying the approach proposed here: for example classifiers $GL_1$ and $GL_3$ have a normalised mutual information of 0.41, while the only difference between them is the model order (and the random initialisation of the EM algorithm).

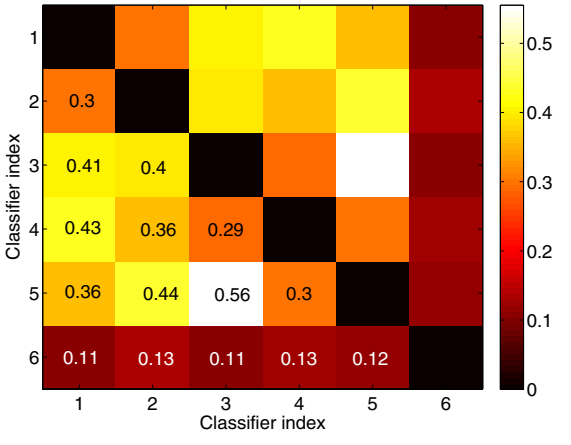



**Fig. 3.** Mutual information between classifiers in the ensemble. Note that the diagonal (equivalent to the entropy of each classifier) has been set to 0 for enhanced contrast.

## 4 Verification Experiments and Results

### 4.1 Database

The BMEC2007 development database contains 50 users, each with two sessions, and is part of the larger BioSecure DS3 dataset. For each user, the first session contains 5 genuine signatures and 10 skilled forgeries[3]. The second session contains 15 genuine signatures and 10 skilled forgeries. Signatures are acquired on a low-power mobile platform (Ipaq PDA). This means that some data is missing, and interpolation approaches outlined in Section 3.1 have to be applied. Furthermore, the orientation of the signatures is haphazard. The acquisition platform only captures binary pressure (on/off) and x,y signals. No pen orientation information is available. The low quality of the data explains why error rates are in general high on this database compared to other signature databases.

[3] These forgeries fall between levels 6 and 8 in [9, Table 3], as the forger has no visual contact with the victim, but is allowed to see several times the dynamics of signing.

### 4.2 Protocol

For each user, We train their set of classifiers ($GL_{1...5}$ and $GG$) on the 5 genuine signatures of the first session. We then run these classifiers on the remaining held-out test data. Thus, for each user we obtain 15 genuine and 20 skilled forgery scores, resulting in a total of 750 genuine signature scores and 1000 skilled forgery scores.

The ensemble classifier (in the present case a simple mean rule, but similar results are obtained using logistic regression) is then trained and tested with this score data using 5-fold cross-validation.

### 4.3 Results

Glancing at Figure 4, it appears that the local classifiers in the ensemble offer approximately the same performance, while the global classifier trails behind. By ensembling local classifiers via the mean rule, it is already possible to substantially lower the error rate, indicating that our coverage optimisation approach based on changes in preprocessing, feature subsets, and modelling complexity is effective. Further adding a global classifier, itself with different features and modelling complexity, yields improved performance. This could be expected given that global information is complementary with local information, and that time information (signature length) is incorporated in the global feature set. While not reported here, we have performed experiments on other signature databases with similar results. It is interesting to note that, while classifiers $GL_3$ and $GL_4$ have virtually identical performances, their mutual information is low (0.3); this is to be accounted for mainly by the rotation normalisation and the inclusion of tangent angles in one feature set. None of them stands out in isolation, but they

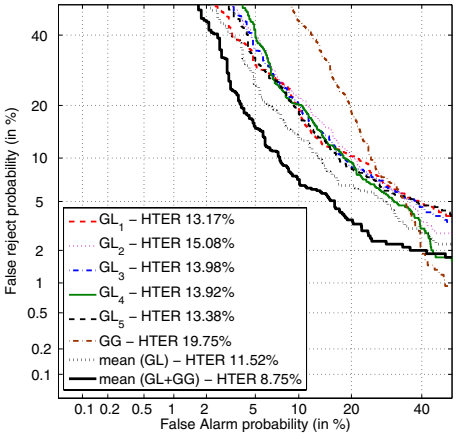



**Fig. 4.** Verification results (skilled forgeries) for base classifiers ($GG_{1...5}$ and $GG$) and Gaussian mixture ensemble with only local classifiers (mean $GG$) and local and global classifiers (mean $GL + GG$)

can be usefully combined because of their diversity. It is certainly possible to reduce the complexity of this ensemble by removing a few local classifiers, while still preserving an adequate accuracy.

This ensemble performed well in the BMEC 2007 competition, comprising a database 430 users, and has taken first place for random forgeries (about 4.0% EER), second place for skilled forgeries (about 13.6% EER), and first place for synthetic forgeries (about 10.7% EER).

## 5 Conclusions

In biometric verification applications, Gaussian mixture models are generally top performers. Other classifiers commonly used in pattern recognition, such as decision trees or random forests, are not often used as base classifiers. We have shown that despite their being categorised as stable, Gaussian mixture models can serve as base classifiers in ensembles if coverage is optimised adequately. To this end, the signal processing chain and other components of the pattern recognition pipeline has to be modified to introduce variability. While the resulting classifiers have roughly the same accuracy, they are complementary and can be usefully combined in an ensemble.

## References

1. Bonastre, J.-F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP, Philadelphia, USA, March 2005, pp. 737–740 (2005)
2. Bousquetand, O., Elisseeff, A.: Stability and generalization. J. of Machine Learning Research 2, 499–526 (2002)
3. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
4. Fiérrez-Aguilar, J., Nanni, L., Lopez-Peñalba, J., Ortega-Garcia, J., Maltoni, D.: An on-line signature verification system based on fusion of local and global information. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 523–532. Springer, Heidelberg (2005)
5. Freund, Y.: Boosting a weak learning algorithm by majority. Information and Computation 121(2), 256–285 (1995)
6. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans.on Acoustics, Speech, and Signal Processing 29(2), 254–272 (1981)
7. Ho, T.K.: The random space method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
8. Hoare, Z.: Landscapes of naïve Bayes classifiers. Pattern Analysis and Applications 11(1), 59–72 (2007)
9. ISO/IEC JTC 1/SC 37 Biometrics. TR 19795-3, biometric performance testing and reporting, part 3: Modality specific testing. Technical report, International Standards Organization (2007)
10. Kryszczuk, K., Morier, P., Drygajlo, A.: Study of the distinctiveness of level 2 and level 3 features in fragmentary fingerprint comparison. In: International Conference on Computer Vision, Biometric Authentication Workshop, Prague, Czech Republic (May 2004)

11. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51(2), 181–207 (2003)
12. Kuncheva, L.I.: Combining Pattern Classifiers. Wiley and sons, Chichester (2004)
13. McLachlan, G.J., Basford, K.E.: Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York (1987)
14. Meynet, J., Thiran, J.-P.: Information theoretic combination of classifiers with application to adaboost. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 171–179. Springer, Heidelberg (2007)
15. Müller, C., Biel, J.-I.: The ICSI 2007 language recognition system. In: Proc. Odyssey 2008: The Speaker and Language Recognition Workshop, Stellenbosch, South Africa (January 2008)
16. Reynolds, D.A.: Speaker identification and verification using gaussian mixture speaker models. Speech Communication 17, 91–108 (1995)
17. Richiardi, J., Drygajlo, A.: Gaussian mixture models for on-line signature verification. In: Proc. ACM SIGMM Multimedia, Workshop on Biometrics methods and applications (WBMA), Berkeley, USA, pp. 115–122 (November 2003)
18. Richiardi, J., Ketabdar, H., Drygajlo, A.: Local and global feature selection for on-line signature verification. In: Proc. IAPR 8th International Conference on Document Analysis and Recognition (ICDAR 2005), Seoul, Korea, vol. 2, pp. 625–629 (August-September 2005)
19. Rodrìguez, J., Kuncheva, L.: Naïve bayes ensembles with a random oracle. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 450–458. Springer, Heidelberg (2007)
20. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, Heidelberg (2006)
21. Sanderson, C.: Automatic Person Verification Using Speech and Face Information. PhD thesis, Griffith University, Queensland, Australia (2002)
22. Ting, K., Zheng, Z.: Improving the performance of boosting for naive bayesian classification. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS, vol. 1574, pp. 296–305. Springer, Heidelberg (1999)
23. Van Bao, L., Garcia-Salicetti, S., Dorizzi, B.: On Using the Viterbi Path Along With HMM Likelihood Information for Online Signature Verification. IEEE Trans. on Systems, Man, and Cybernetics, Part B 37(5), 1237–1247 (2007)