FEATURE WEIGHTED MAHALANOBIS DISTANCE: IMPROVED ROBUSTNESS FOR GAUSSIAN CLASSIFIERS

Matthias Wölfel and Hazim Kemal Ekenel

Institut fuer Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
email: {wolfel, ekenel}@ira.uka.de, web: http://isl.ira.uka.de/~wolfel

ABSTRACT

Gaussian classifiers are strongly dependent on their underlying distance method, namely the Mahalanobis distance. Even though widely used, in the presence of noise this distance measure loses dramatically in performance, due to equal summation of the squared distances over all features. The features with large distance can mask all the other features so that the classification considers only these features, neglecting the information provided by the other features. To overcome this drawback we propose to weight the different features in the Mahalanobis distance according to their distances after the variance normalization. The idea behind this is to give less weight to noisy features and high weight to noise free features which are more reliable. Thereafter, we replace the traditional distance measure in a Gaussian classifier with the proposed. In a series of experiments we show the improved noise robustness of Gaussian classifiers by the proposed modifications in contrast to the traditional approach.

1. INTRODUCTION

The distance measure for features is of critical importance for all kind of classification methods. The well known Mahalanobis distance shows good performance for clean features, but loses performance very fast in the case of noisy features. To improve the noise robustness of the Mahalanobis distance we propose to weight the features according to their distances after the variance normalization.

The organization of the paper is as follows. In section 2 we give a brief review of estimation theory to lay the basis for the development to follow. The feature weighted Mahalanobis distance and two ways how to derive the weights are introduced in section 3. A series of experiments are performed in section 4. Finally, section 5 presents our conclusion and plans for future work.

2. ESTIMATION THEORY: A BRIEF REVIEW

In this section we briefly review the basics of estimation theory which are elementary for a good understanding of the remainder of this paper. For a detailed introduction see [1, 2]. Elementary to all estimation procedures are distance measures and a broad variety exists for different purposes [3, 4, 5, 6]. Two widely used measures are:

- The *Euclidean distance*
  (based upon Pythagorean Theorem)

$$D_i^{\text{Euclidean}} = \sqrt{(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i)} \tag{1}$$

  where μ_i represents the mean vector of class $\{I\}$ and $\mathbf{x}$ represents the sample vector to classify.

- The *Mahalanobis distance*

$$D_i^{\text{Mahalanobis}} = \sqrt{D_i} = \sqrt{(\mathbf{x}-\mu_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)} \tag{2}$$

  where $\boldsymbol{\Sigma}_i^{-1}$ represents the inverse of the covariance matrix of class $\{I\}$. The Mahalanobis distance is therefore a weighted Euclidean distance where the weighting is determined by the range of variability of the sample point; expressed by the covariance matrix.

The covariance matrix itself can be classified into three different types:

- *Spherical*: The covariance matrix is a scalar multiple of the identity matrix, $\boldsymbol{\Sigma}_j = \sigma_j^2\mathbf{I}$.
- *Diagonal*: The covariance matrix is diagonal, $\boldsymbol{\Sigma}_j = \text{diag}(\sigma_{j,1}^2, \sigma_{j,2}^2, \ldots, \sigma_{j,d}^2)$.
- *Full*: The covariance matrix is allowed to be any positive definite matrix with rank $d \times d$.

In the definition of the *Gaussian distribution*

$$\begin{aligned} p(\mathbf{x}|\omega_i) &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)\right\} \\ &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}D_i\right\}, \end{aligned} \tag{3}$$

which is later needed for the Gaussian classifier, it is apparent that the squared Mahalanobis distance is part of the equation.

Bayes' decision theory forms the basis of statistical pattern recognition. The theory is based on the assumption that the decision problem can be specified in probabilistic terms and that all of the relevant probability values are known. With the posterior probability $P(\omega_i|\mathbf{x})$ or the class-conditional probability density $p(\mathbf{x}|\omega_i)$ and the *prior* probability $P(\omega_i)$ we can write *Bayes' decision rule* as

$$i_{\text{Bayes}} = \underset{i}{\text{argmax}}\, P(\omega_i|\mathbf{x}) = \underset{i}{\text{argmax}}\, p(\mathbf{x}|\omega_i)P(\omega_i) \tag{4}$$

which maximises the classification rate.

A *Gaussian classifier* (GC) uses Bayes' decision theory where the class-conditional probability density $p(\mathbf{x}|\omega_i)$ is assumed to have a Gaussian distribution (3) for each class ω_i. Under the assumption of multivariate Gaussian densities we can write the discriminant function as:

$$d_i(\mathbf{x}) = \log p(\mathbf{x}|\omega_i)P(\omega_i) \tag{5a}$$

$$= -\frac{1}{2}(\mathbf{x}-\mu_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mu_i) + \log P(\omega_i) \tag{5b}$$

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{d}{2}\log 2\pi$$

$$= -\frac{1}{2}D_i + \log P(\omega_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{d}{2}\log 2\pi \tag{5c}$$

From (5a) to (5c) it is apparent that one part of the discriminant function is composed of the squared Mahalanobis distance. Once we have the Gaussian discriminant functions, data $\mathbf{x}$ is assigned to class i if

$$i_{\text{Bayes}} = \underset{i}{\operatorname{argmax}} \left(-\frac{1}{2} D_i + \log P(\omega_i) - \frac{1}{2} \log |\Sigma_i| \right). \quad (6)$$

Therefore, the decision process uses the squared Mahalanobis distance.

3. WEIGHTING THE DISTANCE AND CLASSIFICATION

In this section we show the drawback of the Mahalanobis distance in the presence of noise and introduce a variety of approaches to overcome this drawback.

3.1 Drawback of the Mahalanobis Distance

The drawback of the Mahalanobis distance is the equal adding up of the variance normalized squared distances of the features. In the case of noise free signals this leads to the best possible performance. But if the feature is distorted by noise, due to the squaring of the distances, a single feature can have such a high value that it covers the information provided by the other features and leads to a misclassification. Therefore, to find classification procedures which are more robust to noise we have to find a distance measure which gives less weight to the noisy features and more weight to the clean features. This can be reached by comparing the different input features to decide which feature should be given less weight or being excluded and which feature should have more weight.

The name '*weighted Mahalanobis distance*' is quite frequently used in different publications; e.g. [7, 8, 9] with quite different meanings. To not further confuse this situation the proposed features in this publication are dubbed *feature weighted Mahalanobis distance*, or simply *feature weighted distance*.

3.2 Feature Weighted Mahalanobis Distance

The Mahalanobis distance is one of the fundamental and widely used techniques as a distance measure for classification. Therefore, the Mahalanobis distance should be used as a basis of our new weighted distance metric.

The features which are distorted by noise have, in average, a higher influence on the distance measure than the less distorted features as they are further away from the feature mean of the class. Therefore, we aim to lower the influence of these features by reducing their weight. To find the features which have the strongest influence on the distance we solve the Mahalanobis distance equation for every single feature c over all input samples i and classes j and store the value in Z:

$$\forall c,i,j: \;\; Z_{i,j}[c] = (x_i[c] - \mu_j[c]) \Sigma_j[c,c]^{-1} (x_i[c] - \mu_j[c]) \quad (7)$$

The goal now is to give less weight to the features with high distance and vice versa to avoid the masking of the features with small distances.

$$\forall i,j: \;\; D_{i,j}^{\text{weighted}} = \sum_{c=1}^{N} w_{i,j}[c] \cdot Z_{i,j}[c] \quad (8)$$

under the two constraints:

$$\forall c: \;\; w[c] \geq 0, \;\; \sum_{c=1}^{N} w[c] = N \quad (9)$$

This leaves us with the question of choosing proper weights. A lot of different reasonable ways exists, e.g. giving zero weight to the feature with the highest distance and normalize the other weights. In the following sections we want to concentrate on two promising approaches.

3.3 Descent Feature Weights

In this approach the goal is to choose the weights such that all features have the same influence on the distance measure. Therefore, we have to normalize the features under consideration of their average distance. First, the vectors of the Mahalanobis distance have to be sorted over their values in ascending order $Z^{\text{sort}}[1] \leq Z^{\text{sort}}[2] \leq \cdots \leq Z^{\text{sort}}[N]$. Thereafter we have to sum over all samples and classes and normalize by the number of samples and classes

$$\forall c: \;\; Z^{\text{sum}}[c] = \frac{1}{M} \frac{1}{L} \sum_{i=1}^{M} \sum_{j=1}^{L} Z_{i,j}^{\text{sort}}[c]. \quad (10)$$

If not all samples to classify are known beforehand, the summation over all samples can be replaced with the sample to classify with nearly no loss in accuracy (on our test set it was below 0.1%*).*

To derive the weights under the two constraints (9) we first have to invert the distances and then normalize

$$\forall c: \;\; w[c] = N \cdot \frac{1/Z^{\text{sum}}[c]}{\sum_{a=1}^{N} 1/Z^{\text{sum}}[a]} \quad (11)$$

With the given weights $\forall i,j: w_{i,j}[c] = w[c]$ the descent feature weighted distance can now easily obtained by (8).

3.4 Difference Feature Weights

Another approach could give more weight to features which are similar to other features than to features which are very different. The idea here is that noisy features should be significantly different from noise free features, as long as only a small number of features are distorted by noise.

We can calculate the difference $d[c]$ for the features as

$$\forall c,i,j: \;\; d_{i,j}[c] = \sum_{a=1}^{N} \left| Z_{i,j}[c] - Z_{i,j}[a] \right|, \quad (12)$$

normalize and invert to calculate the individual weights

$$\forall c,i,j: \;\; w_{i,j}[c] = \frac{\sum_{a=1}^{N} d_{i,j}[a]}{N \cdot d_{i,j}[c]} \quad (13)$$

These weights could be directly used in (8). Additionally, we could concatenate the descent weighted approach by replacing $Z_{i,j}[c]$ of (10) with the difference weighted distances $\overline{Z}_{i,j}[c] = w_{i,j}[c] \cdot Z_{i,j}[c]$. Indeed, this concatenation was found to be better than the simple use of (8) and therefore should be referred to as the difference feature weighted distance, dropping the word descent.

Weight Number	1	2	3	4	5	6	7	8	9	10
Development Set Weights	3.277	1.730	1.153	0.946	0.846	0.764	0.637	0.310	0.196	0.142
Calculated Weights	3.335	1.886	1.296	0.966	0.750	0.592	0.465	0.354	0.243	0.112

Table 1: Comparing the weights of the a posterior derived weights on the development set and the weights calculated by (10) to (11) on the test set.

3.5 Feature Weighted Gaussian Classifier

The descent or difference feature weighted distance can now simply implemented into the GC by replacing the Mahalanobis distance in (6) by the novel distance measures. We have dubbed this new class of classifiers feature weighted GC, where the word descent or difference should be added to separate between the derivation of the weights.

4. EXPERIMENTS

For our experiments we have set up 20 Gaussians, each representing one class, in a 10 dimensional feature space. The mean and variance parameters of each Gaussian are randomly generated. For generating the mean vectors we have used a Gaussian distributed random generator with zero mean and a variance of 10. For the variance, again, a Gaussian distributed random generator with zero mean, but this time with variance of 30 was used. The priors are uniformly spread. The covariance structure is a diagonal model. With the prior definition of the Gaussian distribution we build a *test set* by generating 1000 random samples (50 for each class). All tests were repeated 50 times and their results were averaged to smooth statistical variance.

4.1 Plausibility Check of the Weights Calculation

To check if the calculation of the weights as given from (10) to (11) is reasonable we optimise the weights based on a development set, comprised of 1000 additional randomly selected samples, such as to give the best classification performance. Due to calculation time we have optimized only for a noise level of 10 over all features using the descent weighted feature approach. As we can see from Table 1 the weights using a posteriori knowledge are very similar to the ones calculated by (10) to (11). Comparing the accuracy on the test set, Table 2, we see that the traditional approach can't reach the performance of the descent feature approach which performs very similar on both sets of weights. The set of calculated weights performs even slightly better than the a posteriori derived weights. Therefore, the assumption that all features should have the same influence on the distance measure is reasonable.

Gaussian Classifier Type	Accuracy
Traditional	43.5%
Development Set Weights	49.3%
Calculated Weights	49.9%

Table 2: Accuracy on the test set for the traditional Gaussian classifier and the descent feature Gaussian classifier with a posterior derived weights on the development set and the weights calculated by (10) to (11) on the test set.

4.2 Missing Features

Here we investigate the degradation of our Gaussian classifier's performance where features are missing (set to zero). This scenario can happen, e.g., if different features are derived by different sensors and one ore more sensors are not delivering data. Comparing the different classifiers in Figure 1 we see that the novel approaches perform superior over the traditional GC. In the case of the difference feature weighted GC the performance gain is lost if more than six features are missing. For the descent feature weighted GC the gain remains.

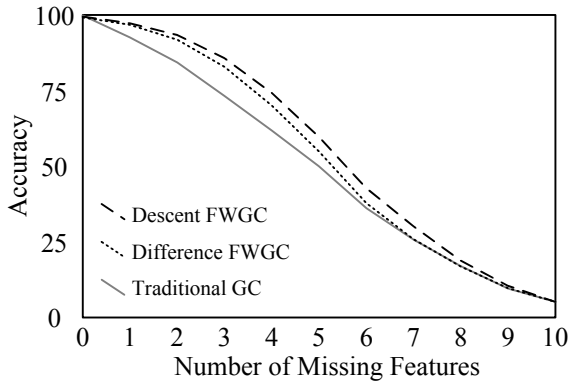

Figure 1: Accuracy of the traditional *Gaussian classifier* (GC) and the two *feature weighted Gaussian classifiers* (FWGC)s over the number of missing features.

4.3 Noisy Features

Here we investigate how additive noise is disturbing the accuracy of our Gaussian classifier on different number of features and noise level.

Figure 2, Figure 3 and Figure 4 are plotting the accuracy over one, three and five noisy features (the feature is randomly selected and therefore unknown) for the traditional GC, the descent feature weighted GC and the difference feature weighted GC. In the case of noise free features, the traditional GC performs slightly better than both novel approaches. For small noise values for one, three or five random channels the descent feature weighted GC performes best while for high noise values the difference feature weighted GC is the best choise.

The most interesting case, of course, is where all features are distorted by noise, Figure 5. Here the difference feature GC performs equally well as the traditional GC for less noise and worse for a high noise level. But, for all noise levels, the descent feature weighted GC performs best.

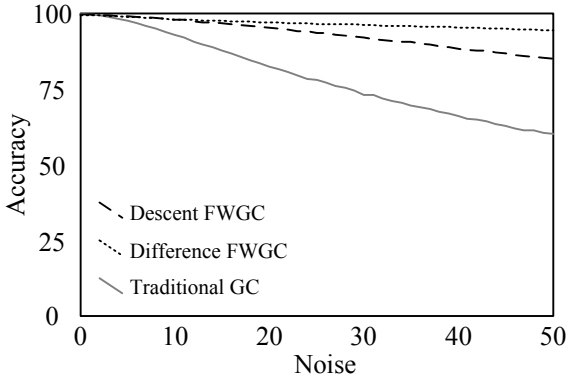

Figure 2: Accuracy of the traditional *Gaussian classifier* (GC) and the two *feature weighted Gaussian classifiers* (FWGC)s for one randomly distorted feature.

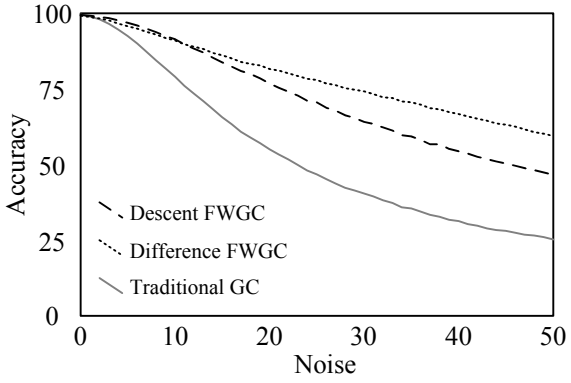

Figure 3: Accuracy of the traditional *Gaussian classifier* (GC) and the two *feature weighted Gaussian classifiers* (FWGC)s for three randomly distorted features.

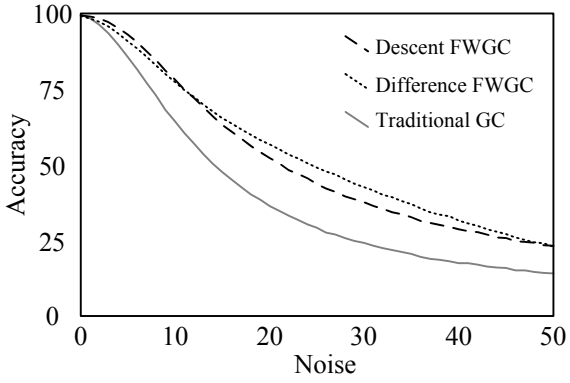

Figure 4: Accuracy of the traditional *Gaussian classifier* (GC) and the two *feature weighted Gaussian classifiers* (FWGC)s for five randomly distorted features.

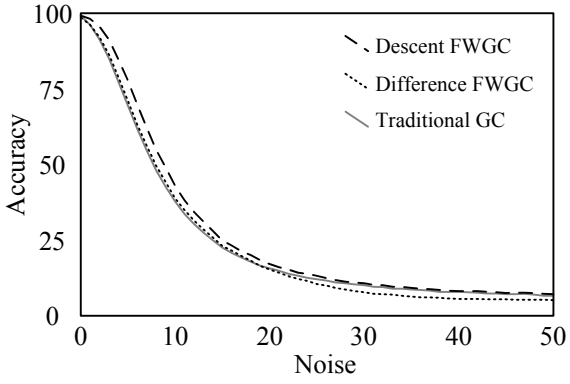

Figure 5: Accuracy of the traditional *Gaussian classifier* (GC) and the two *feature weighted Gaussian classifiers* (FWGC)s where all features are distorted.

5. CONCLUSION AND FUTURE WORK

We have shown two new approaches based on weighting of the features to improve the robustness to noise as well as missing features of the widely used Gaussian classifier. In a series of simple experiments we have shown the usefulness of the proposed approaches, where for non distorted features the new approaches perform nearly equally well and show significant gains if distorted or missing features exist. Nevertheless, so far we have worked with simulated data, thus we want to implement this approach in different applications such as speech recognition or face verification. We also plan to investigate if the gain due to the new approach is still apparent after different normalization algorithms and methods such as linear discriminant analysis.

REFERENCES

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification (2nd ed.)*. Wiley Interscience, 2003.

[2] X. Huang, A. Acero, and H.W. Hon, *Spoken language processing*. Prentice Hall, 2001.

[3] Veldhuis, R. and Klabbers, E., "On the computation of the kullbackleibler measure for spectral distances," *IEEE Transactions on speech and audio processing, VOL. 11, NO. 1*, January 2003.

[4] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

[5] F. Zheng, Z. Song, L. Li, W. Yu, F. Zheng, and W. Wu, "The distance measure for line spectrum pairs applied to speech recognition," *International Conference on Spoken Language Processing*, 1998.

[6] L. Klein, *Sensor and Data Fusion: A Tool for Information Assessment and Decision Making*. Press Monograph, 2004.

[7] K. Younis, M. Karim, R. Hardie, J. Loomis, S. Rogers, and M. DeSimio, "Cluster merging based on weighted mahalanobis distance with application in digital mammograph," *Proc. of IEEE Aerospace and Electronics Conference*, 1998.

[8] J. Peng, D.R. Heistenkamp, and H.K. Dai, "Adaptive kernel metric nearest neighbor classification," *Proc. of IEEE International Conference on Pattern Recognition*, 2002.

[9] K. Rerkrai and H. Fillbrandt, "Tracking persons under partial scene occlusion using linear regression," *8th International Student Conference on Electrical Engineering*, 2004.