Simplifying the model-based classifiers for multi-modal problems in classification of spectra.

Pavel Paclík, David M.J.Tax, Serguei Verzakov, Robert P.W.Duin

Information and Communication Theory Group
Faculty of Elec.Eng, Math. and Comp.Science
Delft University of Technology, The Netherlands
{P.Paclik, D.M.J.Tax, S.Verzakov, R.P.W.Duin}@ewi.tudelft.nl

Abstract

Industrial applications of spectral imaging introduce novel types of problems where generally applied assumption of class unimodality does not hold anymore. Object sorting problems, for example, become multi-modal as soon as the terminal high-level classes are defined as ad-hoc collections of material types. This paper discusses design of model-based multi-modal classifiers in problems, where the class modes are apriori known during training.

We consider derivation of data representation and classifier design as connected issues. Several designs simplifying the full Gaussian mixture-based model using diverse representation building strategies and regularization are discussed. Apart of model-based algorithms employing mode descriptors such as Gaussian mixtures or multi-modal SIMCA, also an alternative method based on inter-mode discriminants is considered.

A set of experiments is conducted on an artificial dataset modeling some aspects of multi-modal spectral classification problems and two real-world datasets from industrial object-sorting application. The behaviour of different model-based methods is studied using learning curves. The main conclusion of this paper is that incorporation of supervised information significantly improves classification performance, reduces the complexity of the final system and speeds-up its execution.

1. Introduction

Industrial applications employing spectroscopy or spectral imaging often deal with inherently multi-modal pattern recognition problems. An example is an object sorting into high-level classes defined as ad-hoc collections of materials. The motivation of this research is derivation of high-accuracy classifiers for multi-modal problems capable of effective execution and thereby applicable in on-line processing.

In this study, we consider construction of data representation and model building as connected issues. We aim at understanding of the interplay between both stages and its impact on model-complexity, classifier performance, and execution speed. Although a model-based classifier may be built directly on the original spectra, such model is very complex due to high dimensionality. Model simplification may yield significant improvements of classification performance given the same amount of training examples.

The paper discusses several approaches to simplification of model-based classifiers, such as regularization, derivation of a simpler representation for the complete problem or for each mode separately and supervised feature extraction. An alternative model-based classification strategy, proposed by us in [7], is also studied reducing the model complexity even further by building a multi-modal classifier from inter-mode discriminants, rather than by modeling the entire mode domain. Our motivation here is to understand if this decomposition-based multi-modal discriminant (DMMD) may be beneficial compared to model-based descriptors.

In order to compare behaviours of different model-simplification strategies, the learning curves are employed estimating the generalization performance of a model-based classifier as a function of the training set size. Two types of problems are considered. The first is an artificial dataset constructed in order to simulate some aspects of spectral classification problems such as large variability in directions uninformative for classification, large number of uninformative features or strong feature correlation. Furthermore, two datasets originating from the real-world sorting application illustrate multi-modal problems of different levels of difficulty. The computational complexity in execution is eventually discussed in relation to the classifier performance.

2. Multi-modal pattern recognition problem

Let us define a multi-modal pattern recognition problem as an allocation of observations into high-level classes, defined as collections of lower-level concepts. Formally, a D-dimensional feature vector $\vec{x} \in R^D$ as assigned into one class from the set $\Omega = \{\omega : \omega = 1, 2, ..., C\}$ of C pre-defined high-level classes. Each class $\omega \in \Omega$ is defined as a collection of concepts $\{m_j^\omega : j = 1, \ldots, M_\omega\}$, where M_ω denotes the number of concepts in the high-level class ω. (In other words, each class $\omega \in \Omega$ is divided into M_ω subclasses denoted by $m_j^\omega, j = 1, \ldots, M_\omega$, $M = \sum_\omega M_\omega$). Each observation $\vec{x}$ belongs to a single concept m_j^ω and thereby also to a single high-level class ω.

We assume that both the information on the low-level concepts and on high-level classes is available during training. In a typical sorting problem, the low-level concepts often represent types of material and the high-level classes the sorting categories. The discussed approaches are valid also in situations where such prior knowledge is missing. In such cases, the mapping of observations into modes needs to be extracted by cluster analysis [7].

In the following, we refer to the concepts, constituting the high-level classes as *modes*. Note that our definition of mode differs from the statistical viewpoint where mode usually represents a unimodal peak of the probability density function. Modes of a sorting system such as material types may, however, exhibit internal statistical multi-modality.

3. Model-based classification strategies for multi-modal problems

In this analysis, we consider a full Gaussian mixture model derived on the raw spectral data as a base method. The mixture-model may be generally viewed as a *decomposition-based descriptor*. The decomposition is performed by defining the class modes which are then described by a suitable model. The Section 4 describes the Gaussian mixture model and several approaches for its simplification.

In [7], we proposed an alternative strategy for the classification of multi-modal data, called the *decomposition-based discriminant*. Here the decomposition yields a set classification sub-problems discriminating between individual modes. The eventual multi-modal classifier is built as a combination of discriminants, derived on these sub-problems. This strategy is described further in Section 5.

4. Gaussian mixture model

A finite mixture model is a probabilistic model of the form:

$$p(\vec{x}|\Theta) = \sum_{k=1}^{K} w_k p(\vec{x}|\theta_k), \quad \vec{x} \in R^D \tag{1}$$

where K denotes the number of mixture components, w_k are the mixing weights ($w_k \geq 0, \sum_{k=1}^{K} w_k = 1$), and $p(\vec{x}|\theta_k)$ represent component densities specified by a parameter vector θ_k. Θ denotes the set of parameters $\{w_1, ..., w_K, \theta_1, ..., \theta_K\}$. In the following, we consider a Gaussian model with a mean vector and a covariance matrix parameters $\theta_k = \{\mu_k, \Sigma_k\}$.

In general, the membership of observations to the mixture components is unknown and is estimated from the training set using the Expectation-Maximization (EM) algorithm. The result of training is the estimated soft allocation of the training examples into components together with estimates of the per-component parameters θ_k.

However, in case of the multi-modal problem defined in Section 2, the mixture model may take advantage of the known mode membership of the training examples. Considering the known low-level concepts m_k in the multi-modal problem as components of a mixture model, we simplify the mixture training to estimation of the components parameters θ_k only from the observations apriori-known to originate from the low-level concept m_k. Similarly, the mixing weights are estimated by apparent mode-priors.

We consider the following strategies simplifying the full Gaussian mixture model.

Regularization Regularization is a widely used strategy for model simplification stabilizing the model parameters and hence restricting the group of admissible classifiers. We consider the regularization of the covariance estimates by adding a small value to its diagonal elements (variances). Due to large data dependency, the regularization parameter needs to be tuned for the particular situation.

Unsupervised dimensionality reduction. Instead of stabilizing a model in a high-dimensional representation, a simpler model may be built in a lower-dimensional space. Because spectra represent high-dimensional measurements with low intrinsic dimensionality, dimensionality reduction is a basic tool for simplifying the model-based classifiers [1]. Principal Component Analysis (PCA) represents a classical unsupervised dimensionality reduction technique. The data is projected to the linear subspace preserving a pre-defined fraction of the overall variance. The approach may, however, yield entirely uninformative data representation is situations where the dataset contains directions uninformative for the sake of class separation yet exhibiting large variance.

Mode-specific dimensionality reduction In order to leverage the dimensionality reduction approach in a supervised problem, this may be performed individually for each class. This approach, originating from the chemometric community, is known as SIMCA [10]. For each class separately, a PCA subspace-model is derived. New examples are classified on the basis of a distance measure combining the Mahalanobis distance of the projected observation to the model and the Euclidean distance of the observation to the model subspace.

In order to apply SIMCA classifier to the multi-modal data, we built separate subspace models on individual modes. Due to the use of the Mahalanobis in-model distance the SIMCA classifier trained on modes closely resembles the Gaussian mixture model.

Supervised feature extraction Naturally, the mixture model may be built in a feature space, derived by a supervised feature extraction. Although the number of spectra-specific feature extractors have been proposed [6, 5, 8], most of the approaches leverage linear projections. We have illustrated in [7], that building a linear discriminant on the complete spectra often results in better classification performance than building classifiers on sets of linear features, derived on separate groups of wavelengths. Based on this observation, we adopt here the linear discriminant analysis (LDA) as a prototypical feature extraction technique [3].

The LDA maps the input data into $C-1$ dimensional linear subspace, where C represents the number of classes. Because this would produce only 1D representation in a typical two-class sorting problem, we leverage the apriori-known modes as classes in the LDA mapping. The Gaussian mixture model is built in the projected feature space.

5. Decomposition-based multi-modal discriminant

The decomposition-based multi-modal discriminant (DMMD) was motivated by the following two observations. Firstly, we realized that if the final objective is the separation of the high-level classes, the full statistical description of the individual modes is unnecessary. Modeling the full domain of the data, the descriptors also requires more evidence than the discriminants [9]. Therefore, we proposed to build a multi-modal classifier on the basis of the inter-mode discriminants.

Our second observation was, that multi-modal classifiers are currently designed analogously to any simple classifier system i.e. in two sequential stages separating the data representation building from the design of a single classifier. Because both steps optimize the full multi-modal problem at once, the resulting algorithms must inevitably become complex. Our proposal was to limit the complexity of representations and classifiers by problem decomposition. We decompose the complicated problem into a set of simpler sub-problems, tackle these independently (by building the inter-mode discriminants) and combine their decisions into the eventual high-level class assignment.

The training and execution of the DMMD algorithm is described by Algorithm 1.

Algorithm 1: Decomposition-based multi-modal discriminant.

Training:

1. *sub-problem creation*. A sub-problem is a two-class dataset where the classes correspond to pairs of modes originating from different high-level classes. There is $(M^2 - \sum_\omega (M_\omega)^2)/2$ sub-problems.
2. *derivation of sub-problem classifiers*. For each sub-problem derive the specific data representation (feature extraction, dissimilarities,...) and construct the sub-problem classifier.
3. *collecting the sub-problem classifier outputs*. Execute the trained sub-problem extractors and classifiers on the complete training set and collect their outputs. Apply the normalization mapping on the outputs (optional). Construct the second-stage dataset from the collected outputs. Each training example retains its high-level label.
4. *combiner training*. Train the second-stage classifier on the second-stage dataset.

Execution:

1. *execution of sub-problem classifiers*. Execute the stored representation-building procedures and trained classifiers on the input data.
2. *output normalization*. If applicable, normalize the classifier outputs appropriately.
3. *collecting the sub-problem classifier outputs*. Collect the outputs of the sub-problem classifiers in the identical order as during training.
4. *combiner application*. Apply the trained combiner to the collected outputs and assign the observations into the high-level classes.

Note although the normalization of sub-problem classifier outputs is not necessary for the sake of building a trainable combiner it is a beneficial strategy for incorporation of the non-linearity into otherwise fully linear system.

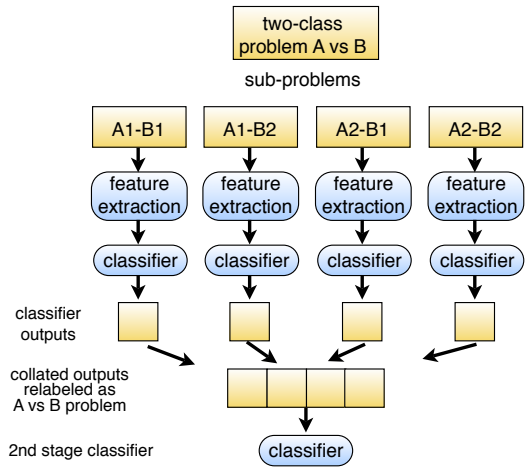

Figure 1: A schematic view of DMMD training in a two-class, two-mode situation. A and B denote classes, A1 denotes the first mode of class A.

6. Experiments

6.1. Algorithms

The following algorithms are considered in the study:

MOGC *Gaussian mixture model on full spectra*. The mean and the full covariance matrix are estimated for each mode.

reg-MOGC *Regularized Gaussian mixture model on full spectra.* The regularization is performed by adding a λ-fraction of the mean of variance-elements of the covariance matrix to its diagonal. For a given problem the fraction λ is optimized using the validation set approach. Available training dataset is split into two parts (80/20 ratio). The large part is used for training of mixture models regularized with the five logarithmically-scaled λ-fractions. The mean classification error rate on the smaller (validation) set is employed as the selection criterion to choose the best λ-fraction.

PCA-MOGC *Gaussian mixture model in the PCA-reduced space.* 99.99% of total variance is preserved.

mode-SIMCA Three versions were considered:

- preserving 99.99% of mode variance, using both in-model and out-of-model distances
- preserving 99% of mode variance, using both in-model and out-of-model distances
- preserving 99% of mode variance, using only in-model distance

LDA-MOGC *Gaussian mixture model in feature space derived by supervised extraction.* Training data, labeled by the available mode-labels are used to construct an LDA mapping to the $M - 1$ dimensional subspace, where a mixture model is built.

DMMD *Decomposition-based multi-modal discriminant.* Without prior feature extraction, FLD is trained at the sub-problem level. The outputs of the sub-problem classifiers are normalized using sigmoid mapping [2]. The sigmoid parameter is optimized on the sub-problem training set. FLD is also employed as the eventual combiner.

For the sake of comparison, the Fisher linear discriminant (FLD), trained on high-level class labels is also provided for comparison.

6.2. Artificial dataset

In order to investigate behaviour of the above-mentioned model-based algorithms, we constructed an artificial problem reflecting issues generally present in real-world spectral datasets:

Multiple modes within classes This may be a direct consequence of the ad-hoc definition of the high-level classes as collections of material types. We consider a configuration of modes which is not linearly separable.

Highly correlated features. High correlation of features is typically present for neighboring bands in spectral data.

Large overall data variance unrelated to the class separation. This may be observed in situations where spectral measurements of objects vary due to object inhomogeneity or biological diversity. Another possible culprit may

be the acquisition system where the three-dimensional macroscopic objects are imaged by means of surface reflectance. The variability due to reflectance changes may outweigh the differences between materials to be classified.

Large number of uninformative features. Spectral measurements typical comprise tens or hundreds of readings in narrow spectral bands. As the most of this information may be unrelated to the classification problem, the large number of dimensions may be considered noisy.

The artificial dataset comprises two classes, each with two modes. The modes are modeled by Gaussian distributions. Only two first features bear information capable of class discrimination. Remaining features are modeled by zero-centered Gaussian with unit variance for all modes.

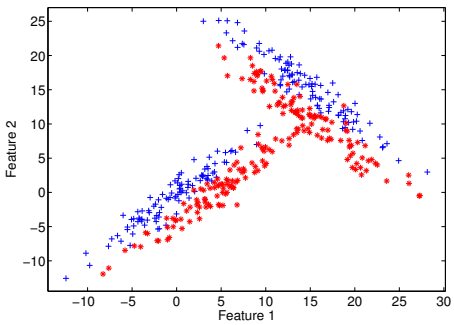

Figure 2: Scatter plot of the two informative features in the artificial dataset.

Let us investigate two situations, differing in the number of uninformative features present in the dataset. The dataset *A* contains 10 dimensions (two informative and eight uninformative) and the dataset *B* 100 dimensions. For each dataset, the design and the test sets were independently generated, each with 1000 data samples per mode.

For each of the investigated algorithms, the learning curve was estimated varying the training set size. The training set of a given size was randomly drawn from the design set (random draw was performed per mode). The algorithm, trained on this training set was executed on the independent test set. The procedure was repeated 20 times averaging the results. The results are presented in Figures 3 and 4 as mean error rates per class and the standard deviations of error rates.

Discussion on Figure 3:

- Poor performance of the FLD on both problems illustrates that non-linear classifier is needed.
- On the 10D dataset *A*, the LDA-MOGC approach yields worse performance than remaining non-linear classifiers suggesting that linear feature extraction over complete problem is not sufficient.
- The discriminant-based DMMD yields significantly better performance than mixture models for very small sample sizes (<50). Being a simpler model, it also exhibits asymptotically higher error than more complex mixture-based approaches.

Discussion on Figure 4:

- The 100D dataset *B* appears to be a significantly harder problem than the 10D dataset *A*.
- The performance of a MOGC classifier almost entirely coincides with the performance of PCA-MOGC method.

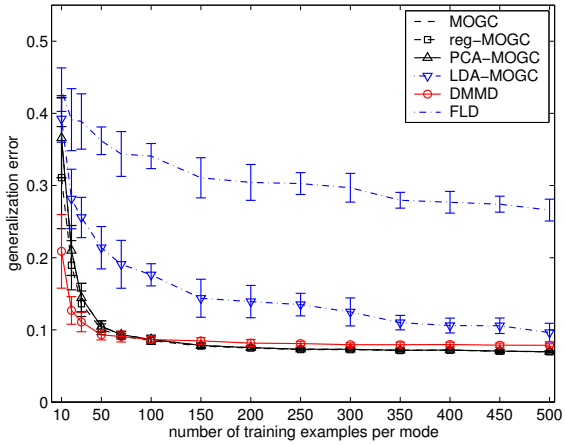

Figure 3: Learning curves for the 10D artificial dataset *A*.

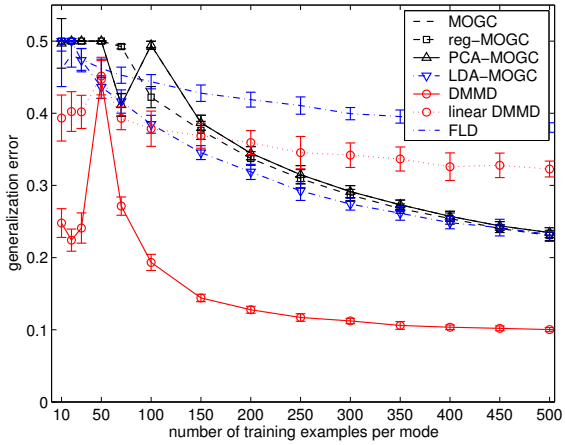

Figure 4: Learning curves for the 100D artificial dataset *B*.

- The extraction of the features using a supervised method (LDA-MOGC) appears to be more beneficial than unsupervised approach (PCA-MOGC) or the use of full spectra (reg-MOGC), the difference however vanishes with growing number of training examples.
- The proposed DMMD method is significantly better than all other approaches. We have included a learning curve of the DMMD algorithm without non-linear mapping of sub-problem classifier outputs i.e. fully linear system. Note that even the fully linear DMMD classifier improves over the simple FLD. We assume the reason lays in its use of the prior knowledge on modes and in combination of multiple linear discriminants.

  The distinct peak at 50 examples per mode is caused by the sub-problem FLDs, trained using 100 examples in 100D feature space. This, so called *peaking phenomenon* is caused by numerically problematic estimation of covariance matrices from datasets where sample size nears or equals the dimensionality [4]. Note that classifiers are built also for problems with less examples than dimensions. Here the pseudo inverse in employed.

6.3. Industrial object sorting dataset

The second investigated dataset originates from the real-world object-sorting problem based on spectral imaging [7]. Hyper-

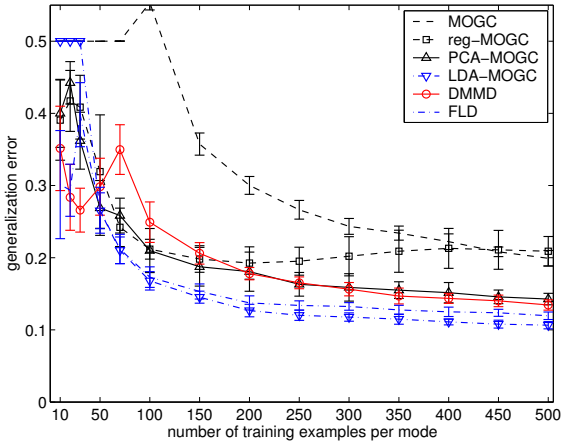

Figure 5: Learning curves for the *two-mode* dataset.

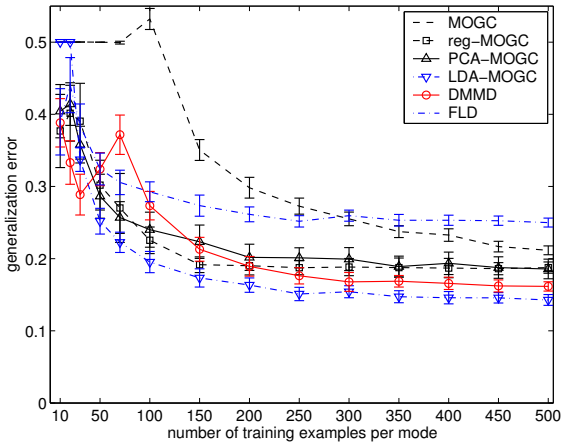

Figure 6: Learning curves for the *three-mode* dataset.

spectral images of objects on the conveyor belt were acquired by an imaging spectrograph. The spectra, consisting of 128 narrow spectral bands, were normalized using the black and white reference images. The objects originate from two high-level classes to be distinguished. Each object is entirely composed of a single material and the material type is known in training.

The original object-sorting problem comprises four levels, namely individual pixels/spectra, objects, material types and high-level classes. For the sake of this study, we omit the object level and simplify the problem into classification of spectra with known material types and high-level classes.

Two real-world datasets were constructed based on this problem. The first one comprises two modes per class, and the second three modes per class. We refer to these datasets as *two-mode* and *three-mode* dataset, respectively. Similarly to the artificial dataset, the respective design sets and independent test sets were constructed. In both cases the designs sets contain 1500 examples per mode, and the test sets 800 examples per mode.

The spectra in the test sets are drawn from entirely different sets of original objects than spectra in the design sets. This is necessary in order to avoid performance estimation bias caused by the presence of neighboring and thereby almost identical spectra in both algorithm training and evaluation.

Following the identical 20-fold cross-validation methodology as explained in Section 6.2, the learning curves were estimated. The results are presented in Figures 5 and 6.

Discussion on Figure 5:

- The overall excellent performance of FLD on the *two-mode* and even better result of the LDA-MOGC algorithm suggests the problem is slightly non-linearly separable. This is a practical illustration of the fact, that multi-modality of the dataset does not necessarily translates into (heavy) non-linearity of the classification problem. We can also observe the peaking phenomenon at 30 examples per mode (i.e. 120 examples in a four-mode problem) due to close match with the spectra dimensionality (128).
- The performance of the MOGC trained on full spectra may be significantly improved by regularization. Regularization is especially beneficial for small sample sizes where it reaches performance the full mixture model attains only for over three-times more training examples.
- The trends of the full mixture model (MOGC) and the of mixture model built in PCA-reduced space (PCA-MOGC) suggest MOGC will need *very* large training sets to attain similar performance.
- The DMMD method exhibits a peaking phenomenon around 70 examples per mode caused by the the sub-problem FLD classifiers (140 examples in 128D space).

Discussion on Figure 6:

- The performance of FLD compared to other, non-linear discriminants suggests that the problem with three modes per class ceased to be linearly separable.
- The peak at the DMMD curve keeps position because it is related to the sub-problem size, not the number of modes.
- Interestingly, the linear feature extraction (LDA-MOGC) results in a low-dimensional representation where mixture model significantly outperforms even the DMMD classifier over the whole range of training set sizes.
- For both *two-mode* and *three-mode* datasets, the DMMD method again outperforms the mixture-based classifiers for very small sample sizes. This suggests the benefits of the use of discriminants instead of descriptors.

6.4. Note on the mode-SIMCA performance

The results of mode-SIMCA algorithms were not included into the plots for the sake of brevity. On the artificial dataset *A*, the mode-SIMCA classifiers provides analogous performance to other mixture-based methods. Also on the dataset *B*, the mode-SIMCA algorithm preserving 99.99% of variance per mode copies the trend of other mixture-based approaches. This is understandable as no dimensionality reduction was effectively performed. However, the algorithm preserving only 99% of variance per mode ($\approx$ 96D models) exhibit significant error increase even above the FLD performance. Interestingly, the mode-SIMCA classifier built in the same subspace, but considering only the in-model distance and neglecting the out-of-model component brings the performance back to the trend of other mixture models. We conclude that in very noisy situations, the mode-SIMCA algorithm needs careful tuning because even few out-of-model directions may lead to its failure. The mode-SIMCA error on the real datasets does not fall under 25% error for the simpler *two-mode* dataset and under 32% on the *three-mode* dataset.

6.5. Computational complexity in execution

Table 1 illustrates relation between classification performance of the discussed algorithms and their computational complexity in execution. The theoretical number of operations, required for execution of each algorithm on a single spectrum was estimated by assuming that all quantities that can be precomputed were precomputed. The trained algorithms for the *three-mode* problem, 500 examples per class were used in Table 1. The dimensionalities of feature spaces used for building the models are also provided. Note the broad range of computational complexities of mixture models build in different representations. The best-performing method is also the fastest one (LDA-MOGC), building the Gaussian models in 5D space.

algorithm	*approx.dim*	*error (std) [%]*	*ops./spec.*
MOGC	128	21.16 (0.63)	198 156
reg-MOGC	128	18.76 (0.76)	198 924
PCA-MOGC	≈ 26	18.57 (1.34)	15 996
mode-SIMCA			
99.99%	≈ 26	32.44 (1.21)	42 318
99%	≈ 4	32.21 (0.63)	7 344
99%, in-model	≈ 4	37.71 (1.21)	5 704
LDA-MOGC	5	14.28 (0.75)	1 396
DMMD	128	16.15 (0.63)	4 772
FLD	128	25.00 (0.63)	514

Table 1: Computational complexity versus mean generalization error for the three-mode dataset. Complexity is given in number of operations per spectrum. Dimensionality where the model is built is provided (approx.dim.)

7. Conclusions

In this paper, we investigated behaviour of several model-based algorithms in multi-modal problems where samples are apriori known to originate from sub-classes of the high-level classes. This study aimed at building the understanding of the effects of various approaches on simplification of the model-based classifiers.

An artificial dataset was constructed simulating several conditions found in multi-modal spectral classification problems such as large data variance in uninformative directions, non-linear separability, and large number of uninformative features. Studying the learning curves revealed that mixture models and SIMCA are severely affected by the presence of large number of noisy features trying to model uninformative directions in the data. The discriminant-based DMMD algorithm is significantly more robust in such situations.

Two real-world datasets were also studied. Both datasets illustrated less severe conditions where regularized mixture or dimensionality reduction significantly improved mixture applicability to smaller sample size problems. The best result is attained by the mixture model, trained in LDA-derived space using prior knowledge on modes.

For very small sample sizes or in presence of severe noise, the DMMD method may provide simpler and thereby potentially better solutions than mixture models. However, in situations where small and informative representation may be derived, mixtures appear to yield better performance. Using supervised information appears to be overall better strategy than the use on unsupervised dimensionality reduction. The estimated computational complexity in execution illustrates that the use of supervised feature extraction and prior information also significantly increases execution speed.

Some aspects of the artificial and real datasets differ resulting in the deviations between the observed behaviours of the DMMD and LDA-MOGC classifiers. We hypothesize that this may be related to informativeness of features in artificial and real datasets. While the artificial dataset contained only two informative features, the discriminatory information in real spectra is probably distributed over groups of bands. That might explain poor performance of the LDA-MOGC method on the dataset *B* compared to the DMMD classifier.

Acknowledgments The authors would like to thank Carmen Lai for fruitful discussions and support. This research is/was supported by the Technology Foundation STW, applied science division of NWO and the technology program of the Ministry of Economic Affairs.

8. References

[1] C.I. Chang. *Hyperpectral Imaging: Techniques for Spectral Detection and Classification*. Kluwer Academic, 2003.

[2] Robert P W Duin and David M J Tax. Classifier conditional posterior probabilities. In A Amin, D Dori, P Pudil, and H Freeman, editors, *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 611–619. Springer, Berlin, Augost 1998. Proc. Joint IAPR Int. Workshops SSPR'98 and SPR'98, Sydney, Australia.

[3] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[4] A.K. Jain and B. Chandrasekaran. *Handbook of Statistics 2*, chapter Dimensionality and Sample Size Consideration in Patter Recognition Practice, pages 835–855. North-Holland, Amsterdam, 1987.

[5] S Kumar, J Gosh, and M M Crawford. Best-bases feature extraction algorithm for classification of hyperspectral data. *IEEE Trans.on Geoscience and Remote Sensing*, 39(7):1368–1379, 2001.

[6] David Landgrebe. *Signal Theory methods in Multispectral Remote Sensing*. John Wiley & Sons, 2003.

[7] P. Paclík and R. P. W. Duin. Designing multi-modal classifiers of spectra: a study on industrial sorting application. In *accepted for the 2nd Spectral Imaging workshop, Villach, Austria*, 2005.

[8] M. Skurichina, P. Paclik, R.P.W. Duin, D.C.G. de Veld, H.J.C.M. Sterenborg, M.J.H. Witjes, and J.L.N. Roodenburg. Selection/extraction of spectral regions for autofluorescence spectra measured in the oral cavity. In *Proc.of SSSPR 2004*, volume vol. 3138 LNCS, Springer, Berlin, pages 1096–1104, 2004.

[9] D.M.J. Tax. *One-class classification*. PhD thesis, Delft University of Technology, June 2001, `http://www-ict.ewi.tudelft.nl/~davidt/`

[10] S. Wold and M. Sjostrom. *SIMCA: a method for analysing chemical data in terms of similarity and analogy*. ACS Symposium Series 52, 1977.