Designing multi-modal classifiers of spectra: a study on industrial sorting application

P.Paclík and R.P.W. Duin

Information and Communication Theory Group
Delft University of Technology, The Netherlands

Abstract:
Spectral imaging is frequently used in various industrial sorting applications. Spectral images of objects, moving on the conveyor belt are processed by a pattern recognition system and classified into one of pre-specified high-level classes. Because these terminal decisions are often defined in terms of material types (e.g. of lower-level classes), a sorting application generally poses a multi-modal pattern recognition problem with multiple levels (spectra, objects, materials, and high-level decisions). It remains an open question how to design high-performance and fast classifiers applicable in such situations.

In this paper, we propose a strategy based on a decomposition of the multi-modal problem into a set of two-class sub-problems formed by pairs of clusters originating from different classes. Based on individual spectra, a specific feature representation and classifier is derived for each sub-problem. The outcomes of the sub-problem classifiers are fused using a trainable combiner. The final assignment on the object level is carried on by majority voting of the per-spectrum decisions. On a set of experiments with an industrial sorting problem, we discuss the proposed methodology comparing it with several state-of-the art techniques.

1 Introduction

Spectral imaging has become a frequently used technique in numerous industrial applications such as quality control, material inspection or sorting. Spectral images of objects, moving on a conveyor belt, are acquired by the imaging spectrometer and processed by pattern recognition algorithms assigning each object to one of pre-defined categories. Although the basic building blocks used in sorting pattern recognition systems are individual spectra or pixels, the objective is to sort objects i.e. connected groups of spectra. Moreover, as the eventual system decisions are often defined in terms of multiple material (sub)types, a sorting application represents a pattern recognition system operating on multiple levels, namely spectra, objects, materials and the high-level decisions. This multi-level nature of sorting systems translates into an inherent multi-modality of corresponding spectral datasets.

The problem of multi-modal data distributions has been recognized by several researchers. Hoffbeck and Landgrebe [5] proposed to cluster the multi-modal data and describe individual modes using uni-modal Gaussian components. Classification is then performed using Gaussian

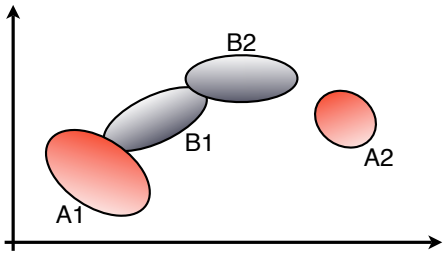


Figure 1: Schematic picture of a two-class multi-modal classification problem with classes A and B. Each class contains two clusters.

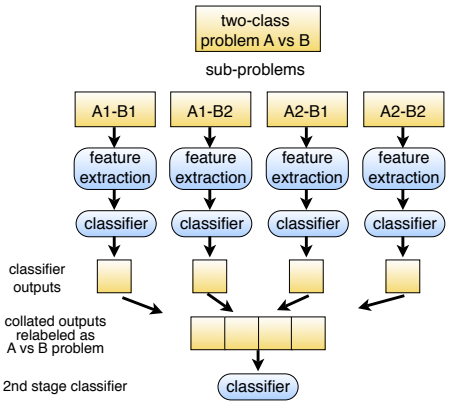


Figure 2: Training of the proposed decomposition-based algorithm on the dataset 1

mixture model. Berge and Solberg [2] discuss the benefits of penalized likelihood for the sake of robust estimation of a multi-modal mixture model using the EM algorithm. Bachmann et.al. [1] use the Projection Pursuit technique to produce lower-dimensional data representations preserving multi-modality instead of Principal Component Analysis (PCA) which is incapable of detecting multi-modal or non-Gaussian situations. An alternative approach dealing with multi-modal situations is based on non-parametric techniques such as the Parzen classifier or the nearest neighbor rule. Especially the dissimilarity-based techniques have been shown to perform well in classification of spectra [7, 9].

Currently used approaches deal with multi-modality in spectral datasets in one step i.e. derive a single, more complex, classifier attempting to solve the complete problem at once. We propose an alternative solution decomposing a multi-modal problem into a collection of simpler sub-problems that can be tackled independently. In this paper, we describe a preliminary study on properties and behavior of this decomposition-based algorithm and provide first experimental results on a real-world dataset.

2 Decomposition-based algorithm

Our starting point is a classification problem where some or all of the classes exhibit internal multi-modality. The modes of high-level classes may be defined apriori as different materials, types of defects or known specimen varieties. In some situations, prior information on structure of high-level classes is not available. Then, the class modes may be identified using cluster analysis. A schematic example of the multi-modal classification problem is given in Figure 1. Each of the two classes A and B contains two clusters.

Our proposal is to decompose a complex multi-modal classification problem into a set of simpler sub-problems and tackle these independently. Because the eventual goal is separation of high-level classes, we define the two-class sub-problems by considering all pairs of clusters

from different classes. For each of the sub-problems, we construct a specific data representation and a classifier. As only data examples from the two clusters are considered in this step, not the complete dataset, the sub-problem classifiers may focus on the context specific clues. Following the example in Figure 1, the sub-problem consisting of cluster A1 against cluster B2 will require a simpler data representation and classifier than the sub-problem consisting of clusters A1 and B1.

We propose to leverage the local expertise of sub-problem classifiers by combining their outputs. We assume that the sub-problem classifiers estimate confidences that a given observation originates from each of the sub-problem clusters. We adopt a trained combining rule learning the pattern in the sub-problem classifier responses from the available training data (see Figure 2). The advantage of this approach over the fixed combining rules is that arbitrary classifier architectures may be mixed. In order to construct a trainable combiner, a second-stage training dataset comprising sub-problem classifier outputs must be build. This is done by executing the trained sub-problem classifiers on the entire available training dataset[1)] The second-stage classifier is trained on the outputs of all sub-problem classifiers, collated into one set and labeled by the high-level class labels.

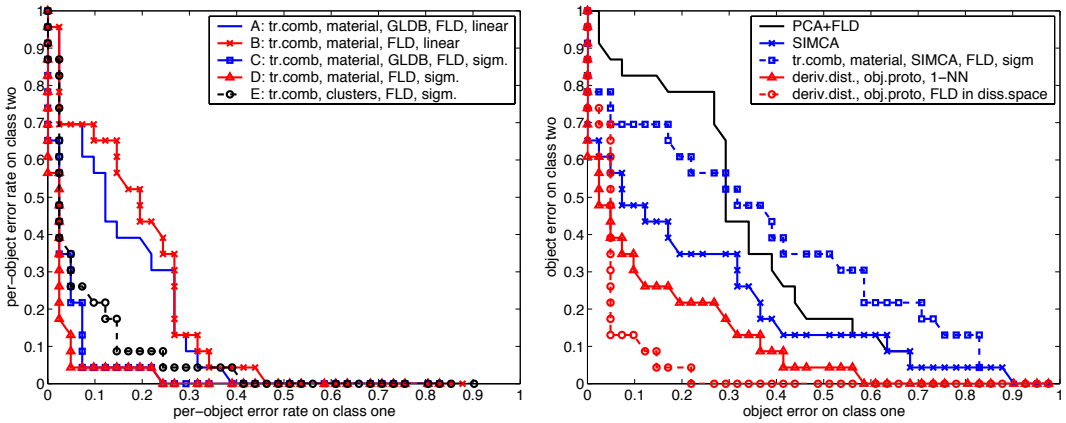
The proposed strategy shares some similarities with the multi-class Bayesian Pairwise Classifier (BPC) proposed by Kumar et.al. in [6]. BPC classifier derives for *each pair of classes* a separate set of features and a classifier. The combination is performed by the majority voting on the decisions of elementary classifiers. The main difference is that the approach, proposed by us, tackles multi-modal, multi-class problems while BPC was proposed as a general multi-class classifier.

3 Experimental setup

In order to assess the performance of the proposed decomposition-based algorithm and to compare it with other techniques, we conducted a set of experiments on an industrial object sorting problem. Hyperspectral images of material pieces (objects) on the conveyor belt were acquired by an imaging spectrograph N17 from Specim Ltd. using a SU-128 InGaAs camera. The spectra were normalized using black and white reference images. The dataset contains in total 108 373 spectra measured on 290 objects. The objects are grouped into two high-level classes, one comprised of three and the other of six material types. The spectral measurements consist of 128 wavelengths. The goal of our experiments is to understand differences between various pattern recognition strategies for object sorting. We assume a perfect object detector and focus solely on the evaluation of object classification algorithms. The common strategy, used in all experiments, was to train a classifier of individual spectra using either terminal

[1)] This strategy is recommended only for datasets with large number of examples. In small sample-size problems, re-using the training set for both stages will yield a biased combiner. In such situations, the second-stage training set may be constructed using a different strategy, such as stacked generalized [10].

high-level classes or material labels. When executed on new data, all object pixels are first labeled by a given spectral classifier. The decision on the object as a whole is performed by majority voting. Note that this approach is based on our prior knowledge that each object is composed of a single material. A more complex post-processing scheme for situations where an object may contain multiple material classes is discussed by Leitner et.al. in [7].



(a) methods based on the proposed problem decomposition using sub-problem classifiers and followed by a trainable combiner.

(b) comparison with unimodal and dissimilarity-based algorithms

Figure 3: Per-object ROC curves expressing errors on both classes for different algorithms on the test set.

For the sake of algorithm development, a design set of 133 randomly selected objects (48 721 spectra) was constructed. During development, the algorithms were evaluated on the design set in a cross-validation fashion. The experimental results, presented in this paper, were obtained by training a selected subset of algorithms on the full design set and evaluating them using an independent set of 64 objects (25 474 spectra) unseen during the design stage.

Existing studies on hyperspectral object sorting systems discuss in detail per-spectrum error rates averaged over classes [7, 12]. Although this gives an indication on the separability of given classes by the spectral sensor, it does not provide any information on the eventual performance of the overall sorting system. In order to estimate the sorting performance, the *error over objects* needs to be computed. Therefore, we present the experimental results in a form of *per-object* error ROC curves relating the error rates on both classes (see Figure 3). Due to its independence on prior probabilities, the ROC curve provides for a two-class situation more informative performance characteristics than the average error rate. The following section gives a more detailed description of the used algorithms and comments on the experimental results.

4 Discussion

Figure 3(b) illustrates that the methods assuming uni-modality such as PCA, followed by Fisher Linear Discriminant (FLD) or SIMCA [13] provide only poor per-object classification performances. The dissimilarity-based methods exhibit more robustness. We have exploited the prior knowledge on the existence of objects in the dataset and used mean objects spectra as prototypes. A significant improvement over the performance of the nearest neighbor rule (triangular markers) is reached when the correlations between dissimilarities to prototypes are exploited (circular markers). This is done by training the FLD classifier in the dissimilarity space [8, 9]. Other experiments, omitted here for the sake of brevity, show that the derivative-based dissimilarity measure [8, 9] performs better than the Spectral Angle Mapper and that comparable performances may be also reached using large sets of randomly selected prototypes. This, however, leads to excessive computational requirements as execution complexity grows linearly with the number of prototypes.

The Figure 3(a) comprises five variants of the proposed decomposition-based approach, denoted A to E. In all cases, the FLD was used for both the base (sub-problem) classifiers and the trained combiner. We have investigated three different design questions:

1. *Is it beneficial to extract specific features for each sub-problem or to use the original wavelengths directly as features?* We have employed the top-down Generalized Local Discriminant Bases (GLDB) feature extractor, proposed by Kumar et.al. [6]. The GLDB algorithm selects groups of adjacent wavelengths and deriving a linear extractor for each group.

2. *What are the merits of normalizing the outputs of base classifiers by a non-linear sigmoid?* The sigmoid transformation is used in order to obtain comparable confidence-like outputs [4]. It introduces a non-linearity into the second-stage training set and thereby yields a non-linear object classifier. The sigmoidal parameter is trained in a maximum-likelihood manner.

3. *What is the difference between a decomposition leveraging prior information on material classes and a clustering-based approach?* The high-level classes are composed of three and six material types respectively. This means, that 18 two-class sub-problems were considered and the combiner was trained on a 36D second-stage dataset. A stability-based clustering [11] was employed by method E using the mode-seeking algorithm [3] operating on the derivative dissimilarity. Interestingly, also the automatically selected solution consisted of three and six clusters.

An interesting outcome of our experiments is that the non-linearity, introduced by the sigmoidal mapping, significantly improves the algorithm performance (compare the method A versus C, and B versus D). The non-linear methods C and D even outperform the FLD classifier trained in dissimilarity space, Figure 3(b), which is also non-linear due to the used

derivative-based distance. Note that the methods C and D somewhat resemble a neural network architecture. The proposed algorithm differs from neural network structures in the problem decomposition step, and deterministic two-stage training process which is less prone to over-training.

While the extraction of specific features for each sub-problem is beneficial in case of fully-linear algorithms A and B, it is slightly decreasing the performance in case the sigmoidal mapping is used. The decomposition based on prior knowledge on materials is clearly beneficial to the unsupervised, clustering-based approach. Nonetheless, the clustering-based algorithm still yields significantly better performance than classical tools or dissimilarity-based methods employing the nearest neighbor rule.

Training the SIMCA classifier on apriori known materials and combining the per-material models using the same trainable combiner approach (Figure 3(b), dashed line, square markers) yields worse results than when SIMCA is applied to the multi-modal high-level classes directly (solid line with cross markers). The main difference between the per-material SIMCA algorithm and the proposed method lays in the nature of the used sub-problems. As a *class descriptor*, SIMCA models each of the materials by a PCA-like method. On the other hand, the proposed decomposition-strategy uses between-cluster *discriminants*. We hypothesize that it may be the combination of discriminants rather than of descriptors what yields high-accuracy multi-modal classifiers.

5 Conclusions

In this paper, we proposed to discriminate highly multi-modal classes using a two-stage algorithm decomposing the problem into several simpler sub-problems, building discriminants for pairs of clusters from different classes, and combining their outputs by a trainable combiner. The preliminary results on a real-world object sorting application suggest this approach delivers performance comparable to advanced dissimilarity-based classifiers for a fraction of execution cost.

Acknowledgments We would like to express our gratitude to CTR Austria for providing the dataset. 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.

References

[1] C.M. Bachmann, T.F. Donato, G.M. Lamela, W.J. Rhea, M.H. Bettenhausen, R.A. Fusina, K.R. Du Bois, J.H. Porter, and B.R. Truitt. Automatic classification of land cover on smith island, va, using hymap imagery. *IEEE Trans.on Geoscience and Remote Sensing*, 40(10):2313–2330, October 2002.

[2] A. Berge and A.S. Solberg. Robust classification of hyperspectral data. In *Proceedings of IEEE International Geoscience and Remote Sensing Symposium, IGARSS'04, Anchorage*, September 2004.

[3] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[4] R.P.W. Duin, P. Juszczak, D. de Ridder, P. Paclík, E. Pekalska, and D. M. J. Tax. PR-Tools 4.0, a Matlab toolbox for pattern recognition. Technical report, ICT Group, TU Delft, The Netherlands, January 2004. `http://www.prtools.org`.

[5] J.P. Hoffbeck and D. Landgrebe. Classification of high-dimensional multispectral data. Technical Report TR-EE 95-14, Purdue University, West Lafayette, 1995.

[6] 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.

[7] R. Leitner, H. Mairer, and A. Kercek. Real-time classification of polymers with nir spectral imaging and blob analysis. *Real-Time Imaging*, 9(4):245–251, 2003.

[8] P. Paclík and R. P. W. Duin. Classifying spectral data using relational representation. In R. Leitner, editor, *Spectral Imaging (Proc. Int.Workshop on Spectral Imaging, Austrian Computer Society, Vienna*, pages 31–34, April 2003.

[9] P. Paclík and R.P.W. Duin. Dissimilarity-based classification of spectra: computational issues. *Real-Time Imaging*, 9(4):237–244, 2003.

[10] P. Paclík, T. C. W. Landgrebe, R. P. W. Duin, and D. M. J. Tax. On deriving the second-stage training set for trainable combiners. In *Lecture Notes in Computer Science, vol. 3541*, pages 136–146. Springer, Berlin, 2005.

[11] Pavel Paclík. *Building Road Sign Classifiers.* PhD thesis, Czech Technical University Prague, 2004. `http://www-ict.ewi.tudelft.nl/~pavel/`.

[12] P. Tatzer, M. Wolf, and T. Panner. Industrial application for inline material sorting using hyperspectral imaging in the nir range. *Real-Time Imaging*, 11:99–107, 2005.

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