The sugar dataset - A multimodal hyperspectral dataset for classification and research

Friedrich Melchert[1,2], Andrea Matros[3], Michael Biehl[1], and Udo Seiffert[2]

[1] *University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, Groningen, The Netherlands*

[2] *Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany*

[3] *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany*

This file is provides further information on the Sugar dataset and is provided together with the dataset files. Within this document you will find information on the used sensor systems, the chemical compounds investigated as well as the formating of the dataset files.

1 Introduction

We present a multi modal hyperspectral dataset (available online at [5]) that cannot only be used to evaluate and compare classification performance, but also enables research on new topics.

In the development of algorithms for hyperspectral data classification several benchmark dataset became common, e.g. the Tecator dataset [7] and the Wine dataset [6], to name two examples. These datasets are mainly used as benchmark problems for different algorithms and classification systems as in [1, 4], although they just compile a set of labeled spectra. In opposite to these well established datasets the *Sugar* dataset offers multiple sets of spectra for each of the available classes. Using a variety of different sensors and hyperspectral cameras the spectral information within the dataset is given over different wavelength ranges. An overview over the sensors and their corresponding wavelength ranges is given in Table 1.

Table 1: Key properties of the different sensors used to record the sugar dataset.

Sensor name	Manufacturer	wavelength range [nm]	sampling points
EOS 70D	Canon	RGB	3
Fieldspec	ASD	350 - 2500	2151
VNIR-1600	NEO	400 - 1000	160
VNIR-1800	NEO	400 - 1000	186
Nuance EX	Nuance	520 - 880	37
SWIR-320m-e	NEO	1000 - 2500	256
SWIR-384	NEO	1000 - 2500	288

2 Chemical compounds

As a training set we selected nine sugar and sugar related compounds with common optical appearance, which are not to be distinguishable by conventional optical imaging. We included were three monomeric sugars (D-galactose, D-glucose, and D-fructose), two sugar alcohols (D-sorbitol and D-mannitol), as well as four sugar esters (S170, S770, S1570, and P1570). Monomeric sugars containing six carbon atoms are also referred to as hexoses, having the chemical formula $C_6H_{12}O_6$. Hexoses occur in many stereoisomers and are classified into aldohexoses, having an aldehyde at position 1 (e.g. D-galactose and D-glucose), and ketohexoses having a ketone at position 2 (e.g. D-fructose). Sugar alcohols are typically derived from sugars by a reduction reaction, changing the aldehyde group to a hydroxyl group. We selected two hexose-derived sugar alcohols with the molecular formula $C_6H_{14}O_6$. Sugar esters, also called sucrose fatty acid esters, are non-ionic surfactants consisting of sucrose as hydrophilic group and fatty acid as lipophilic group. Sugar esters can vary in the nature of the attached fatty acid, such as palmitate (P1570) or stearate (S170, S770, S1570) as well as in the number of attached fatty acids (called mono-, di-, tri-, tetraester). In our case, compounds with variation of both parameters have been chosen: S-170 (sucrose stearate, ratio 1% monoester, 99% di-, tri-, and polyester), S-770 (sucrose stearate, ratio 40% monoester, 60% di-, tri-, and polyester), S-1570 (sucrose stearate, ratio 70% monoester, 30% di-, tri-, and polyester), and P-1570 (sucrose palmitate, ratio 70% monoester, 30% di-, tri-, and polyester). According to the high variation in stereochemistry and composition we expected a high degree of diversity in our data set. All compounds appear as white powder, whereas D-fructose looked more crystalline. Spectral profiles were acquired using a variety of different sensors and hyperspectral cameras (Table 1). Given the nine different compounds it is possible to define five classification problems. The mapping of the compounds to the different classification problems is given in Table 2.

Table 2: Definition of the different classification problems. The numbers in the table represent the class index of the substance with regard to the classification problem. Empty cells indicate, that this substance is not used within the concrete classification problem (row). Borders are used to illustrate the pooling of multiple substances to a single class.

	number of classes	Sugar ester S170	Sugar ester S770	Sugar ester S1570	Sugar ester P1570	D-Mannitol	D-Sorbitol	D-Glucose	D-Galactose	D-Fructose
problem 0	9	1	2	3	4	5	6	7	8	9
problem 1	3	1				2		3		
problem 2	4	1	2	3	4					
problem 3	2					1	2			
problem 4	3							1	2	3

3 Research Questions

Besides the use as a benchmark dataset, the unique structure of the *Sugar* dataset offers the opportunity to encourage research on further topics. Three of the possible research questions are briefly discussed in the following

Dimensionality reduction The data within this dataset is compiled from high dimensional feature vectors. Various machine learning algorithms suffer from the presence of high dimensional inputs, which imply a high number of adaptive parameters, leading to convergence problems, overfitting effects and suboptimal results [2].

Taking into account the functional characteristics of spectral data, the high number of input dimensions is not justifiable. For functional data, such as the hyperspectral data in this dataset, a high correlation of neighbored features is expected. Thus the dataset can serve as a basis for the development and evaluation of dimension reduction algorithms. The varying number of input dimensions ($37 - 2151$, cf. Table 1) within the dataset facilitates a solid benchmarking of the performance and scaling of novel approaches.

Sensor invariant classification models In the design of classification models the formatting of the input data plays a major role. In most cases a change of the input data format is simply not possible. Furthermore trained classification models often implicitly incorporate sensor specific properties during optimization. So the change of measurement equipment can lead to a loss in classification performance. Since the training of classification models is usually time consuming the generation of a new classifier after a hardware change is costly.

For the composition of this dataset the spectral information of certain substances were recorded using multiple different sensors. Given the overlapping wavelength ranges (again cf. Table 1) the dataset contains data, that represents the same spectral information recorded with different sensors. This data can be used for the development and validation of algorithms, which are capable of handling different input formats such as variable sized feature vectors and slight shifts in the positions of the spectral sampling points, as well as sensor specific patterns and fragments within the data.

High dimensional data exploration For the generation of industrial classification systems based on spectral information the selection of a suitable sensor system is one of the key issues. In most of the cases the wavelengths which are relevant for the classification task are not known in advance, so the selection of a sensor system usually follows an educated guess or is guided by financial issues.

Having wide and limited wavelength range sensor data within the presented dataset, the question of proper sensors selection may be tackled in a more systematic way. Using the provided data it is possible to develop relevance learning schemes, which quantify the importance of wavelengths. Relevances emerging from the classification of wide spectral bandwidth data (which may have a low number of samples) can be used for a proper sensor selection, on which classification models may be tuned afterwards.

Apart from the sensor selection the dataset provides also opportunities for the challenging visualization of high dimensional data, which is a key issue for data exploration [3].

These questions may serve as a starting point for further research. Nevertheless this list is not complete (neither it is meant to be). The presented Sugar dataset is unique in terms of its structure and extent, and hopefully serves as a basis for future improvements in the classification of hyperspectral data, as well as the outlined research topics.

4 File formats

All the data belonging to the sugar dataset is provided as plain text files following the comma-separated values format. The filenames are compiled according to the sensor/camera used when recording the data, starting with a prefix "`sugar_`" followed by a string that identifies the sensor, e.g. "`neoSWIR320`" and terminated by the file extension "`.csv`". So a full filename could be e.g. "`sugar_neoSWIR320.csv`". The complete sugar dataset is composed from two files per sensor used, so twelve data files in total. For each sensor there is a training data file, and a validation data file. In addition to the filename of the training dataset (as described above), the filename for the validation data is altered by adding "`_val`" before the filename extension, e.g. "`sugar_neoSWIR320_val.csv`". The structure of the validation set files follows the same specifications as the training set files. Validation and training dataset differ in the day of recording. The validation data was recorded on a different day, with separate specimens of the different sugar substances, but using the same senors and experiment setup. Therefore the provided validation dataset especially encounters the question of the reproducibility of achieved classification results.

Table 3: Format of the data files provided in the sugar dataset.

		column index				
		1	2	3	4	5
row index	1	99	λ_1	λ_2	$\dots$	λ_d
	2	label(x_1)	x_{1,λ_1}	x_{1,λ_2}	$\dots$	x_{1,λ_d}
	3	label(x_2)	x_{2,λ_1}	x_{2,λ_2}	$\dots$	x_{2,λ_d}
	1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	4	label(x_n)	x_{n,λ_1}	x_{d,λ_2}	$\dots$	x_{n,λ_d}

The content of each data file is organized as described in the following. We assume the number of samples to be n and the number of different values per sample to be d. Therefore d also represents the number of different wavelengths for the given sensor. The dataset file represents an $((n + 1) \times (d + 1))$ sized table. The very first entry in the table represents the value 99. It is fixed and may serve as an option to check, if the file has been read correctly. The following d entries of the first line represent the different wavelengths, at which the data values were sampled[1]. For subsequent lines the value of the first column represents the label of the sample as given above *problem 0*. The other columns represent the feature values of the spectra, according to the wavelengths given in the first row. The structure of the dataset file is illustrated in table 3.

5 Additional Files

In order to simplify the usage of the dataset there are different MATLAB$^{\text{TM}}$ scripts provided together with the dataset, that enables an easy handling of the data in MATLAB$^{\text{TM}}$. The files provided are:

loadDataset.m This function provides a simple routine for loading the dataset files into the local workspace.

reformulateDataset.m Given a dataset in its plain format this function can be used to derive different classification problems specified in Table 2.

labels.txt This is not a MATLAB$^{\text{TM}}$ script file, but a plain text file containing the labels of the different classes according to problem 0 as specified in Table 2. Each line in the file contains the class index (e.g. 1) followed by a white space and the complete label of this class(e.g. Sugar Ester S170). This file only contains correct labels for classification problem 0 and should be translated to the other problems on demand.

[1] For the data recorded with the RGB camera Canon EOS 70D the concept of wavelengths does not apply. To keep the format of all data files consistent, arbitrary (but plausible) wavelengths were chosen for the three channels, namely 660nm for the red, 540nm for the green and 470nm for the blue channel.

References

[1] Marika Kästner, Barbara Hammer, Michael Biehl, and Thomas Villmann. Functional relevance learning in generalized learning vector quantization. *Neurocomputing*, 90:85–95, 2012.

[2] Yann LeCun, J. S. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In David Touretzky, editor, *Advances in Neural Information Processing Systems (NIPS 1989)*, volume 2, Denver, CO, 1990. Morgan Kaufman.

[3] Allen R Martin and Matthew O Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the 6th Conference on Visualization'95*, page 271. IEEE Computer Society, 1995.

[4] Belen Martin-Barragan, Rosa Lillo, and Juan Romo. Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232(1):146–155, 2014.

[5] Friedrich Melchert, Andrea Matros, Micheal Biehl, and Udo Seiffert. The sugar dataset. University of Groningen, `http://dx.doi.org/10.4121/uuid:fad486b6-0dfb-4d23-8025-b24407a08698`, 2016. Dataset.

[6] Marc Meurens. Wine mean infrared spectra dataset. University of Louvain, `http://mlg.info.ucl.ac.be/index.php?page=DataBases`. Dataset.

[7] Hans Henrik Thodberg. Tecator meat sample dataset. StatLib Datasets Archive, `http://lib.stat.cmu.edu/datasets/tecator`, 1995. Dataset.