

SVM Ensemble Classification of NMR Spectra Based on Different Configurations of Data Processing Techniques

Kai Lienemann, Thomas Plötz, and Gernot A. Fink
Dortmund University of Technology, Dortmund, Germany
{Kai.Lienemann,Thomas.Ploetz,Gernot.Fink}@udo.edu

Abstract

The early detection of drug-induced organ toxicities is one of the major goals in safety pharmacology. Automating this process by classification of metabolic changes based on the analysis of ^1H Nuclear Magnetic Resonance spectra improves this process. In this paper we propose an Ensemble classification system based on Support Vector Machines trained on diverse “views” on the data. These views are created by variation of preprocessing techniques and the final classification is achieved by voting on an optimized selection of all experts. Results of an experimental evaluation on a challenging data-set from industrial safety pharmacology show the effectiveness of the proposed approach w.r.t. the detection of drug-induced toxicity.

1. Introduction

One of the major challenges in drug design applications within safety pharmacology is the reliable detection of drug-induced adverse effects being toxic for particular (regions of) organs. The research field of *Metabonomics* aims, among others, at the early detection of putative toxicities of particular pharmaceuticals by analyzing changes of certain metabolites’ concentrations. In this context ^1H Nuclear Magnetic Resonance (NMR) spectroscopy of biofluids collected from treated organisms has been proven very effective.

The analysis of NMR-spectra can be interpreted as a classical pattern recognition task. High-dimensional data (cf. figure 1) is examined for characteristic changes correlated to different class assignments (toxic or control). Typically only a few hundred spectra are available each containing several thousand measurement points (“ n small, p large”). Complicating the situation NMR-data often contains certain variance not related to putative toxicities but to changes in sample conditions – NMR-spectra are typically rather noisy.

In order to deal with this noisy data in an automatic classification approach certain preprocessing steps are

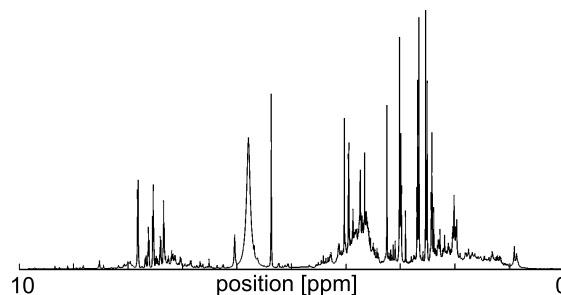


Figure 1. Exemplary ^1H -NMR-spectrum

necessary aiming at normalized spectra. Various criteria relevant for optimizing the NMR-spectra have been proposed (bucketing, scaling, etc.) but, unfortunately, no commonly accepted standard procedure could be established. The basic dilemma is, that one has to decide for some preprocessing. However, favoring one method over the other, might substantially bias the results of the subsequent classification procedure.

Apparently, it is impossible to decide exclusively for one particular preprocessing step without losing information relevant for the classification. Thus, in this paper we tackle the aforementioned dilemma by the application of an Ensemble of strong classifiers. It consists of Support Vector Machines (SVMs) [12] trained on different “views” of the data each obtained by the application of preprocessing variants to the spectra. By means of automatic model selection the most relevant *experts* are integrated into the Ensemble, which outperforms all single classifier approaches. We integrated this Ensemble approach into an NMR-classification framework developed in our previous work (cf. [8, 9]) and successfully evaluated it on a challenging data-set consisting of real NMR data from industrial safety pharmacology.

In the next section we will briefly review related work. Section 3 describes our Ensemble based classification system for NMR spectra. Following this, the results of the experimental evaluation are presented. The paper concludes with a discussion.

2. Related work

Concentrations of metabolites in NMR spectra, interpreted as (high-dimensional) vectorial data, are, basically, determined by the analysis of peaks and their positions, respectively. The detection of specific changes in “peak patterns” within the data is the basis for their classification w.r.t. toxicity of particular pharmaceuticals. Automatic classification approaches are usually applied in order to support the drug discovery process.

In order to focus on those peak changes induced by some actually toxic substance, especially the spectral variance related to changes in sample conditions – so-called peak shifts – needs to be compensated. As a standard technique the so-called bucketing [13] is usually applied, integrating small, equally spaced spectral parts (buckets) of a defined width to a single value by summation. Additionally, often a normalization of the spectra is performed, e.g. applying *Standard normal variate* (SNV) scaling [1] for the compensation of baseline distortions. Every spectrum is normalized to zero mean and unit variance.

By means of techniques from multivariate data analysis (e.g. PCA, PLS [11]) metabolic changes in normalized NMR spectra are identified. For the automatic classification of NMR spectra, usually, statistical models are applied. Prominent examples are SIMCA [6], or PLS discriminant analysis [6]. Within the *Consortium for Metabonomic Toxicity* (COMET) project [4], a mixture density model like classification approach, namely *CLOUDS (Classification of Unknowns by Density Superposition)* [3], has been used for toxicity prediction.

3. SVM-Ensemble classification based on variation of data processing techniques

Reconsidering related work it becomes clear, that all current NMR analysis approaches rely on fixed configurations for spectral preprocessing plus optional feature extraction. In fact in most publications the focus is not on the particular preprocessing step. Instead, often preprocessing like scaling or (standard) feature extraction is applied without further optimization. According to our practical experiences the initial treatment of NMR data is, however, a very crucial step w.r.t. the effectiveness of the subsequent classification. The exclusive decision for one particular fixed configuration (e.g. manually determined fixed bucket-widths or scaling w.r.t. some constant defined “by experience”) is often optimal for subsets of the sample data only while being unfavorable for others. When relying exclusively on the fixed configuration of preprocessing techniques the overall classification process is erroneously biased.

This bias generally influences the analysis process in a way that it tends to miss certain substantial fea-

tures of subsets of the underlying data which could have been covered if additionally respecting alternative techniques or just different configurations. Thus, we propose to overcome this problem by taking different preprocessing configurations into account in an Ensemble approach and by automatically learning the optimal combination of these preprocessing techniques. Similar to our previous work (cf. [8, 9]) the Ensemble consists of multiple SVMs as base classifiers each trained on different “views” of the raw NMR data. These views originate from the application of (combinations of) certain preprocessing techniques.

3.1. Different views of NMR data

Preprocessing techniques for NMR data can, generally, be subdivided into two main categories: normalization and feature extraction (see also section 2). For improved toxicity classification we generate different views on NMR spectra by combining different configurations of preprocessing techniques of either type. Note that putative redundancies in these views will be removed automatically by the expert selection procedure applied for Ensemble creation (cf. section 3.2).

Initial dimensionality reduction and compensation of peak-shift effects is achieved by bucketing. Increasing the bucket-width implies more effective compensation of (erroneous) peak-shifts but, unfortunately, also reduces the spectral resolution which corresponds to loss of information. Thus, it is rather crucial to find a reasonable compromise satisfying both criteria. The optimal bucket-width can only be determined for the data analyzed, which, unfortunately, does not generalize to unknown data as desired. Consequently, there is no generally accepted standard bucket-width defined.

In our work we perform a normalization of variables using SNV scaling which results in a reduction of disturbing baseline distortions. It is applied to every spectrum prior to putative feature extraction. For feature extraction we apply PLS transformation which results in dimensionality reduction of the (normalized) NMR spectra. The parametrization of the feature extraction process again represents a compromise which is not possible to find on sample data thereby reasonably generalizing to unknown data. Increasing the number of components used for the PLS model results in a more detailed description of the data but, at the same time, increases the overall dimensionality.

For the proposed approach the aforementioned methods are varied in their parametrization in order to achieve different views on the NMR data. Bucketing is applied as default method and the bucket-width is varied. Subsequently, SNV and PLS transformation are (optionally each) applied, thereby varying the target dimensionality of PLS in a predefined range.

3.2. An SVM Ensemble for NMR classification

The parametrization of preprocessing methods corresponds to a non-trivial optimization task with multiple objectives. We propose to determine the combination of these techniques and their appropriate configurations which is optimal w.r.t. classification accuracy using an Ensemble approach. By means of the combination of specialized base classifiers the overall classification performance is increased since the ‘‘opinions’’ of multiple experts are integrated (cf. e.g. [7]). We used SVMs as base classifiers trained separately for every view of the sample data.

Technically the process of Ensemble creation can be summarized as follows (figure 2). Different configurations of spectral preprocessing and feature extraction techniques (cf. section 4 for details) are used, each creating a qualitatively differing view on the data. These different data sets X_i are used for training of C-SVMs [12] SVM_i . The optimal parameter setting for every SVM_i is determined by cross-validation using a logarithmic grid search (cf. [8]).

The final classification is achieved by majority voting on a subset of all experts’ predictions due to the putative presence of redundancy and experts of minor performance in the Ensemble. In order to overcome the combinatorial problem of full evaluation of all possible expert selections, a genetic algorithm is applied. The initial population of possible selections is initialized randomly and new populations are created following the standard recombination and mutation rules (cf. [5]). Individuals are selected for every new generation according to the classification performance of the particular expert selection they represent. The iterative process is stopped after a defined time-limit and the best selection used for majority voting

4. Experimental evaluation

In order to evaluate the effectiveness of the proposed approach we conducted various practical experiments aiming at toxicity prediction w.r.t. proximal tubule (kidney). The basis for this was a truly challenging real-world data set consisting of NMR spectra of 47 different pharmaceuticals as currently analyzed in industrial safety pharmacology. It contains 896 samples (637 control, 259 toxic) each initially interpreted as 130 000-dimensional data vectors. Ground truth labeling of this data was given by experts’ predictions of toxicities based on literature investigations and histological judgment (further details are given in [9]).

Preserving the ratios of toxic and control samples for all substances analyzed the data-set was split into disjoint sub-sets for five-fold cross-validation. In every

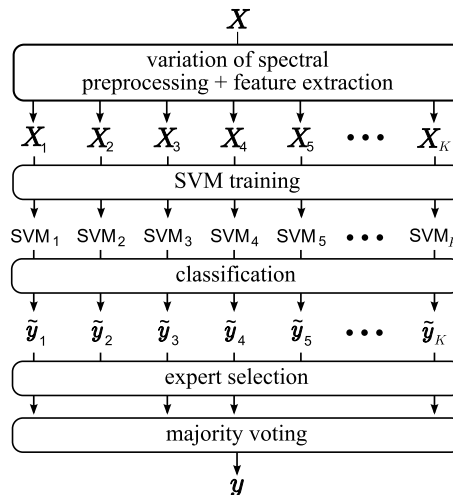


Figure 2. Ensemble classification system

of the five configurations possible 3/5 were selected for training, 1/5 for cross validation and 1/5 for test. The final classification rates were averaged over the results achieved on the five test sets (the particularly remaining fifths). The evaluation criterion for all training and optimization procedures is the *Matthews Correlation Coefficient* [10] – *MC* (normalized to $[-1 \dots 1]$), indicating the correlation between real and predicted class labels:

$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

(TP: number of true positive predictions, FP: false positives, TN: true negatives, FN: false negatives)

Different views on the data were created by varying bucket-widths between 0.01, 0.02 and 0.04 ppm, plus optional SNV scaling. The target dimensionality of the PLS transformation was varied from five to fifty in steps of five. In total this resulted in 66 different configurations of preprocessing techniques. Linear C-SVMs based on the libSVM [2] implementation were trained by cross-validation. The best selection of experts out of 50 runs of the genetic algorithm with population size of 100, single-crossover rate of 0.6 and mutation rate of 0.05 was used for the final Ensemble classification.

According to the actual pharmacological application the final classification result is not intended to be on the level of each spectrum. Instead, toxicity classification of applied pharmaceuticals at a certain dose is desired. Therefore, predictions of all samples corresponding to a single drug are combined by majority voting to a final substance-dose classification. False predictions caused by biological noise due to different individual responses to the applied pharmaceutical are, thereby, compensated by averaging over different experimental animals, collection time-points and genders.

Table 1. Evaluation results for sample / substance-dose classification

| Measure | single SVM | SVM Ensemble |
|----------------------|--------------------|----------------------|
| cross-validation set | | |
| Accuracy [%] | 76.9 / 82.7 | 80.4 / 88.5 |
| Specificity [%] | 85.2 / 82.4 | 92.2 / 91.2 |
| Sensitivity [%] | 56.4 / 83.3 | 51.4 / 83.3 |
| MC | 0.426 / 0.637 | 0.489 / 0.745 |
| test set | | |
| Accuracy [%] | 68.2 / 73.1 | 73.4 / 80.8 |
| Specificity [%] | 77.9 / 73.5 | 86.8 / 79.4 |
| Sensitivity [%] | 44.4 / 72.2 | 40.5 / 83.3 |
| MC | 0.223 / 0.441 | 0.304 / 0.604 |

In table 1 the results are given as pairs of sample-wise / substance-dose wise classification rates. Results achieved using a single SVM serve as reference for those achieved when using the proposed Ensemble approach (based on 13 experts). It can be seen that the classification accuracy increases significantly (level of significance = 2.6 %) when applying the SVM Ensemble. In all cases the specificity increases substantially, which in fact is the most relevant measure for practical applications in pharmacology. The sensitivity is slightly decreased which is, however, uncritical.

The combination of samples' classification to a final prediction (substance-dose classification using 15 optimized experts – second figure each) leads to a general improvement in classification performance.

5. Discussion and conclusions

In this paper we presented a new approach for the classification of NMR spectra utilizing an SVM Ensemble. Different configurations of spectral preprocessing and feature extraction techniques were used to create different views on the data. SVMs are trained for every view and integrated into an Ensemble. The final classification decision is achieved by majority voting among a selection of SVMs, determined by a genetic algorithm. The proposed approach significantly outperforms single SVM-based classification for sample- and substance-dose wise evaluation.

As a practical outcome of this paper for pharmacological applications the SVM Ensemble also provides confidence values for classification. For this, the percentage of votes for some toxicity class in case of substance-dose classification is utilized. Contrary to hard decisions of standard approaches, additional information w.r.t. classification is provided for further interpretation. This benefit has already been used successfully for the generation of dose-dependent responses.

Automatic classification of NMR spectra increases efficiency of drug design in safety pharmacology by fast and subject-independent classification without need for expert knowledge and manual analysis of results from clinical chemistry. Therefore, the proposed approach reasonably supports the development of new drugs by classification of experimental pharmaceuticals regarding different organ toxicities.

6. Acknowledgments

Parts of this work have been funded by Boehringer Ingelheim Pharma GmbH & Co. KG., Genomics group. We would like to thank the General Pharmacology Group of Boehringer for providing the sample set.

References

- [1] R. Barnes et al. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5):772–777, 1989.
- [2] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001.
- [3] T. Ebbels et al. Toxicity classification from metabonomic data using a density superposition approach: CLOUDS. *Anal. Chim. Acta*, 490:109–122, 2003.
- [4] T. Ebbels et al. Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The consortium on metabonomic toxicology screening approach. *Journal of Proteome Research*, 6(11):4407–4422, 2007.
- [5] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., 1989.
- [6] E. Holmes et al. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chemical Research in Toxicology*, 13(6):471–478, 2000.
- [7] L. I. Kuncheva. *Combining Pattern Classifiers – Methods and Algorithms*. Wiley Interscience, 2004.
- [8] K. Lienemann et al. On the application of SVM-Ensembles based on adapted random subspace sampling for automatic classification of NMR data. In *Multiple Classifier Systems*, LNCS, pages 42–51, 2007.
- [9] K. Lienemann et al. NMR-based urine analysis in rats: Prediction of proximal tubule kidney toxicity and phospholipidosis. *Journal of Pharmacological & Toxicological Methods*, doi:10.1016/j.vascn.2008.06.002, 2008.
- [10] B. W. Matthews. Comparison of the predicted and observed secondary structure of the T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451, 1975.
- [11] L. I. Nord et al. Multivariate analysis of ¹H NMR spectra for saponins from quillaja saponaria molina. *Anal. Chim. Acta*, 446:197–207, 2001.
- [12] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [13] M. Spraul et al. Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *Journal of Pharmaceutical & Biomedical Analysis*, 12:1215–1225, 1994.