Automatic Classification of NMR Spectra by Ensembles of Local Experts*

Kai Lienemann, Thomas Plötz, and Gernot A. Fink

Dortmund University of Technology, Intelligent Systems Group, Germany
{Kai.Lienemann,Thomas.Ploetz,Gernot.Fink}@udo.edu

Abstract. A new approach for the automatic detection of drug-induced organ toxicities based on Nuclear Magnetic Resonance Spectroscopy data from biofluids is presented in this paper. Spectral data from biofluids contain information on the concentration of various substances, but the combination of only a small subset of these cues is putatively useful for classification of new samples. We propose to divide the spectra into several short regions and train classifiers on them, using only a limited amount of information for class discrimination. These *local experts* are combined in an ensemble classification system and the subset of experts for the final classification is optimized automatically. Thus, only local experts for relevant spectral regions are used for the final ensemble classification. The proposed approach has been evaluated on a real data-set from industrial pharmacology, showing an improvement in classification accuracy and indicating relevant spectral regions for classification.

1 Introduction

The early detection of drug-induced adverse effects being toxic for particular organs is pursued in safety pharmacology, a crucial step in industrial drug design. The effect of applied drugs on different (regions of) organs is typically determined by measurements of various pattern of marker enzymes in urine samples in clinical chemistry followed by an interpretation by an expert. Thus, the detection of drug-induced organ toxicities is an expensive and time-consuming process requiring expert knowledge. An automation of this process is highly desirable.

The analysis of the metabolic profile from experimental animals after drug application can be automated by ^1H *Nuclear Magnetic Resonance (NMR) spectroscopy* of urine samples. Each peak in the resulting spectrum (cf. figure 1) corresponds to certain molecules' hydrogen atoms with an equal magnetic environment, which is influenced by chemical bonds and surrounding atoms. The final peak position is normalized to an added standard substance and to the measurement frequency (cf. [1]). Since the peak intensity correlates with the molecules'

* Parts of this work have been funded by a grant from Boehringer Ingelheim Pharma GmbH & Co. KG., Genomics group. The authors would like to thank the General Pharmacology Group of the company for providing the sample set.

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 790–800, 2008.

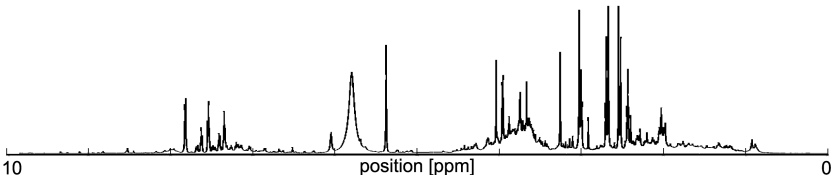

Fig. 1. Exemplary ^{1}H-NMR-spectrum from rat urine

concentrations, a change in the metabolism affects several peak intensities. For a full automation of the whole process, characteristic changes corresponding to a drug-induced organ toxicity have to be determined, facilitating the automatic classification of new samples – a classical pattern recognition task.

While in clinical chemistry the patterns or marker enzymes, with known relevance for the detection of adverse effects, are determined, NMR spectroscopy principally measures all molecules emitting signals of sufficient intensity. Since the underlying molecules from all peaks within an NMR spectrum are not known, background information on the relevance of certain molecules for the detection of adverse effects can not be introduced in the analysis and relevant peaks have to be selected in a data-driven way. This allows for the use of information from molecules not taken into account by clinical chemistry and possibly leading to a deeper understanding of metabolic changes caused by drug-induced organ-toxicities. Therefore, we propose to divide the classification of the whole spectrum into several local classifications in varying scale, incorporating a different amount of local information, and to combine these *local experts* in a multiple classifier system. Thereby, several cues from across the whole spectrum are combined for a final classification and regions not relevant are automatically excluded.

Related work aiming at an automatic classification of NMR spectra will be briefly reviewed in the next section. The proposed classification process for NMR spectra by an ensemble of local experts will be described in section 3. The results of the experimental evaluation are presented in section 4.

2 Related Work

The probably most comprehensive project within the field of Metabolomics, aiming at an automatic analysis and classification of NMR data w.r.t. drug-induced organ toxicities, was the *Consortium for Metabonomic Toxicology (COMET)* [2]. Toxicity prediction has been realized by an approach comparable to mixture density classification, namely *Classification of Unknowns by Density Superposition (CLOUDS)*. Alternative classification approaches using statistical models have also been investigated within and beyond this project (cf. [3,4,5]).

Although multiple classifier systems have been applied successfully in various fields of pattern recognition, they have been hardly used for the detection of drug-induced organ toxicities based on the analysis of NMR spectra. Basically,

several base classifiers are trained on a single sample set created by modifying samples, or different data representations. The classifiers' predictions are combined to a final ensemble classification. Thereby, the incorporation of different experts' *views* on the data into the ensemble classification can lead to an improved classification performance [6]. An ensemble approach for classification of NMR spectra presented in [7] has shown a general improvement in classification performance by using Support Vector Machines (SVMs) [8] as base classifier and a modified version of Random Subspace Sampling [9] for ensemble creation. The application of spectral preprocessing and feature extraction methods in different parametrizations is also leading to different *views* on the sample set and aggregation of SVMs trained on these views leads to an improved ensemble classification in comparison to the best single SVM (cf. [10]).

3 Ensembles of Local Experts

Our first experiments on the use of ensemble methods for classification of NMR spectra have shown, that some spectral regions are more important for classification purposes than others (cf. [7]). Incorporating weights in a Random Subspace Sampling procedure for ensemble creation has improved classification performance and apparently relevant spectral regions were indicated by the highest weights. NMR spectra were preprocessed in these experiments by a bucketing procedure [11] for the reduction of dimensionality and peak-shifts, thereby, integrating all intensity values of short spectral regions into single values by summation. However, full compensation of peak shifts can not be guaranteed by bucketing without a considerable decrease in resolution and integration of several peaks into a single bucket value, thereby combining intensity information from different molecules (cf. [12]). Therefore, the analysis should be performed in a sufficiently high resolution in order to be able to infer a single peak's intensity. Thus, it is also possible to identify metabolites showing changes in intensity subject to the degree of an investigated organ toxicity (*Biomarkers*).

We developed a new approach for the automatic classification of high-dimensional NMR spectra by the combination of several classifiers trained on short spectral regions in a multiple classifier system, as illustrated in Fig. 2. Thereby, the classification of the whole spectrum is subdivided into several classification problems using short spectral regions of interest (SROIs) in different scales, incorporating different amounts of local information. An alignment procedure is applied to each SROI for compensation of peak shifts and a classifier is trained. These classifiers serve as *local experts*, modeling class discrimination based on a limited view on the data. Local experts individually leading to a sufficient classification performance are combined in a multiple classifier system. An optimized subset of local experts is used for the final classification, combining information on intensity changes from different spectral regions of the whole spectrum. Details on the selection of SROIs, spectral alignment and combination of a subset of local experts in a multiple classifier system will be given in the following.

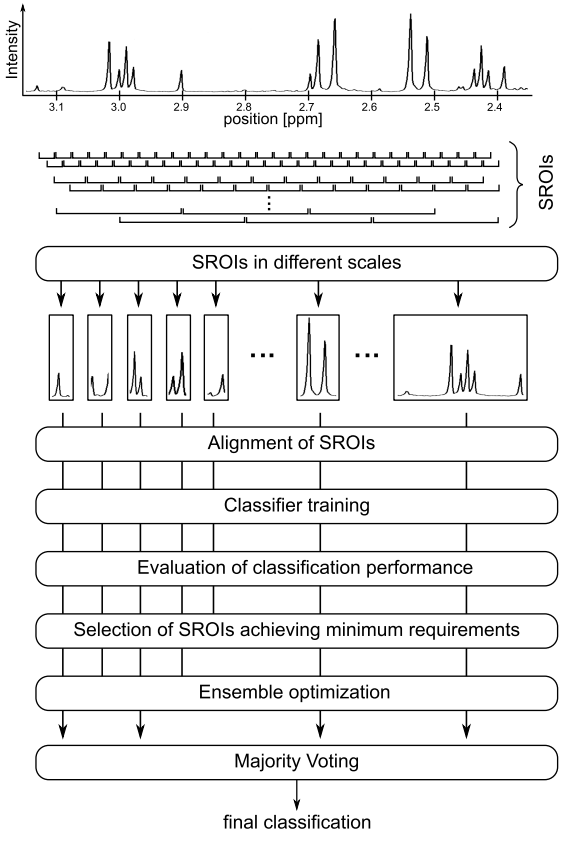

Fig. 2. Classification of NMR spectra by an ensemble of local experts

3.1 Spectral Regions of Interest

The initial selection of SROIs aims at the detection of all spectral regions putatively containing information beneficial for the further classification process. The main information within NMR spectra is principally contained in peak intensities, but the reliable detection of all peaks in a spectrum is a challenging task and their individual relevance for the current classification purposes is not known. Therefore, we propose to use a sliding window approach with overlap and varying scale, detecting SROIs among the whole spectrum including different amounts of local information within each region. Non-informative regions will be excluded from further analysis at a following stage according to their classification results. Thus the relevance of spectral regions is not defined by the pure presence of a peak, but by its local classification performance.

3.2 Alignment of SROIs

A reasonable analysis of NMR spectra in high resolution presumes an alignment procedure, normalizing the exact position of corresponding peaks among

all spectra. These are modified by peak-shifts, primarily induced by varying ion concentration and pH value of the sample to be analyzed. Shifts of different peaks predominantly occur independent of each other, thus, a global shift of the whole spectrum is not feasible for peak-shift compensation. Alignment approaches either aim at the detection of corresponding peaks among all spectra (cf. [13]) or modify the exact position of peaks for maximization of similarity among all spectra (cf. [14]). Restricting the analysis on short spectral regions simplifies the alignment process due to the very limited amount of independently varying peaks present in each region. Therefore, peak-alignment can be realized by assigning a global shift factor to the particular region of each spectrum. Thus, the shape similarity within a particular SROI across all spectra is maximized. Maximum similarity is described within the proposed alignment procedure based on the minimization of the spectrum's reconstruction error ε. This similarity measure is defined according to the first principal component $\boldsymbol{\alpha}$ of a PCA model and the mean vector $\boldsymbol{\mu}$ determined on the data set excluding the spectrum $\boldsymbol{x}$.

$$\varepsilon = \| \boldsymbol{x} - \boldsymbol{\mu} - (((\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\alpha}^T)\boldsymbol{\alpha}) \|$$

Thereby, the alignment is performed according to the shape of all spectra and is not dependent on peak-detection algorithms or a reference spectrum.

3.3 Ensemble Optimization

The initial determination of SROIs (cf. 3.1) selects numerous regions among the spectra in a non data-driven manner. In order to reduce the complexity of the subsequent ensemble optimization method, SROIs not individually leading to a sufficient classification accuracy are excluded from the further process. Subsequently, the predictions of all classifiers selected for further processing have to be aggregated to a final classification.

The most prominent and intuitive method for classifier aggregation is **majority voting**, classifying samples into the class with the maximum number of votes from all base classifiers. Furthermore, **decision templates** proposed by Kuncheva [15] are an alternative method for the combination of continuous-valued predictions of an ensemble of classifiers. Thereby, every classifiers' support for every class is organized in a decision profile matrix for every training sample. Subsequently, class-specific decision templates are obtained by averaging over all decision profile matrices of the corresponding training samples. Test samples are classified according to the class corresponding to the decision template with minimum distance to the sample's decision profile matrix.

The aforementioned combination strategies respect all classifiers' predictions within the ensemble for final classification. However, in case of redundancy or classifiers performing poorly within the ensemble, the selection of a subset of all available classifiers and the use of majority voting on their predictions can improve classification performance in the final ensemble. Especially in the proposed approach, the combination of certain local experts putatively leads to an improved classification performance due to the combination of several cues on

relevant peak-intensity changes and exclusion of non-informative regions. This combination of local experts for ensemble optimization has to be determined automatically and two different approaches will be presented in the following.

Sequential optimization is a feasible approach for the determination of an expert selection by iteratively including / excluding the best / worst expert in an ensemble until all experts are included in or excluded from the ensemble. The optimal number of classifiers is used for final classification. The best individual expert is used as starting point in *best selection* and at every iteration the next best expert is included in the ensemble. Instead of respecting only the individual experts' classification performance within every iteration, the increase of ensemble classification performance is used in *forward selection* as criterion for the selection of the expert to be added to the ensemble. This procedure is modified in *backward selection* by starting with an ensemble of all available experts and excluding at every iteration the expert, leading to the best ensemble classification performance if removed from the ensemble.

Widely used approaches for solving optimization problems are **genetic algorithms** (cf. [16]), which can easily be adapted to the problem of expert selection for ensemble classification. Each individual within a population encodes an expert selection as a string of zeros and ones, representing the inclusion or exclusion of every expert for the final ensemble classification by majority voting. The initial population is initialized randomly and each individual is selected for the next generation with a probability relative to the ensemble classification performance of its encoded ensemble selection. The genetic operations mutation and cross-over can be applied by randomly including or excluding experts in the selection and by the exchange of selection information between two individuals. Furthermore, elite selection can be applied, selecting a defined number of best individuals for the next generation without application of genetic operations. The iterative process is stopped after a defined time-limit and the best selection is used for the final ensemble classification.

4 Experimental Evaluation

The effectiveness of the proposed ensemble of local experts for automatic classification of NMR spectra has been investigated based on a real-world set of ^{1}H NMR spectra from pharmaceutical industry. Details on the data-set, evaluation methodology and results are presented in the following.

4.1 Data-Set and Methodology

The basis for an experimental evaluation of the proposed approach was a set of 896 ^{1}H NMR spectra of urine samples from rats treated with pharmaceuticals currently investigated in industrial safety pharmacology. Overall 52 pharmaceuticals were applied, whereby 34 (= 637 samples) has been labeled as non-toxic and 18 (= 259 samples) as toxic by experts' predictions w.r.t. proximal tubule (kidney) toxicity based on literature investigations and histological judgment.

All spectra were initially down-sampled by a bucketing procedure with a bucket-width 0.001 ppm in the spectral region from 0.2 ppm up to 10 ppm, excluding the water and urea peaks from 4.5 ppm up to 6.0 ppm. The selected bucket-width preserves shapes of individual peaks and provides the same discretization for each spectrum. These bucket values serve as data representation for the further process. Additionally, all spectra were scaled to a total sum of 1000 units for compensation of changes in sample concentration (further details on spectra measurements, data treatment and histological judgment are given in [10]).

Training, parameter optimization and final test were performed by means of a five-fold cross-validation and test. Therefore, all substances were grouped according to their targets and indications and divided among 5 sets structurally preserving grouping and focussing on similar ratios of non-toxic and toxic substances among all sets. Three fifth of the sets were used for training (training set), one fifth for parameter optimization (cross-validation set) and the remaining fifth for test (test set) in every possible configuration. The final rates were averaged over the results on the particular five sets. The evaluation criterion for all training and optimization procedures is the *Matthews Correlation Coefficient* [17] – *MC* (normalized to $[-1 \dots 1]$), due to its robustness to imbalanced data-sets:

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

(TP: number of true positive predictions, FP: false positives, TN: true negatives, FN: false negatives). Classification accuracy, specificity and sensitivity will be shown additionally.

The major goal in the current application is the classification of pharmaceuticals at a certain dose as non-toxic or toxic, therefore, predictions of all samples corresponding to a single drug from different collection time-points and gender have to be aggregated to a final classification. Thus, majority voting among samples corresponding to a certain drug at each collection time-point is performed and classified as toxic if the current drug is classified as toxic at one time-point. False predictions caused by biological noise due to different individual responses to the applied pharmaceutical are, thereby, compensated by averaging over different experimental animals and genders.

4.2 Results

The initial determination of SROIs was performed by sliding windows of size 0.025 ppm, 0.05 ppm, 0.1 ppm and 0.2 ppm (25, 50, 100 and 200 bucket values, respectively) and an overlap of 50 % each, resulting in an overall amount of 1 160 SROIs. These SROIs were aligned by the procedure outlined in 3.2 and the classification performance evaluated by SVMs using a radial basis function kernel function [8]. All SROIs not achieving at least a Matthews of 0.5 have been excluded from the further evaluation, resulting in 147 SROIs for the final ensemble evaluation.

Table 1. Evaluation results for different types of base classifiers

Measure	**NN**	***k*NN**	**RF**	**LSVM**	**RSVM**
cross-validation set					
Accuracy [%]	94.2	92.3	86.5	96.2	**98.1**
Specificity [%]	97.1	**100**	**100**	**100**	**100**
Sensitivity [%]	88.9	77.8	61.1	88.9	**94.4**
MC	0.872	0.834	0.712	0.916	**0.958**
test set					
Accuracy [%]	82.7	86.5	82.7	86.5	**90.4**
Specificity [%]	85.3	97.1	**100**	97.1	94.1
Sensitivity [%]	77.8	66.7	50	66.7	**83.3**
MC	0.623	0.700	0.629	0.700	**0.785**

We first compared Nearest Neighbor classifier (NN), K Nearest Neighbor classifier (*k*NN) (cf. [18]), Random Forests (RFs) [19], and SVMs with linear (LSVMs) and radial basis kernel function (RSVMs) [8] for their use as base classifier in the ensemble. The optimal number of neighbors used for *k*NN is optimized according to a fix grid from three to 31 in steps of two on each SROI, and the model parameters for LSVMs and RSVMs are determined by a grid-search algorithm. RFs are parametrized dependent on the data dimensionality v, using $\lceil log_2(v) \rceil$ decision trees in the forest and selecting $\lceil \sqrt{v} \rceil$ variables randomly at each node. The evaluation results on the cross-validation and test set, using the previously selected SROIs and ensemble optimization by forward selection are shown in table 1. RSVMs show the best classification performance on the cross-validation and test set, and are used as base classifier in the further process.

A comparison of different ensemble optimization methods' performance for RSVMs as base classifier is shown in table 2. Decision templates (DT) were used in combination with the squared euclidean distance as similarity measure (cf. [15]) and genetic algorithm optimization (GA) was performed with a population size of 100, mutation rate of 0.05, cross-over rate of 0.6 and elite selection of the 10 best individuals. Best selection (select), forward selection (fselect) and backward selection (bselect) were used as described in 3.3 and the optimal number of experts determined. Forward selection performs identical to genetic algorithm optimization on the cross-validation set, but outperforms all optimization strategies on the test set and the corresponding expert selection is used for the final ensemble classification.

The classification performance of the proposed ensemble of local experts is compared to alternative classification approaches in Table 3. In order to apply single RSVM classification on the full spectra, the data-set was preprocessed by a bucketing procedure using a bucket-width of 0.2 ppm, scaled by standard normal variate transformation [20], and features extracted by partial least squares transformation [21] using 15 model components, resulting in 15-dimensional samples.

Table 2. Evaluation results for different ensemble optimization strategies for an ensemble of RSVMs

Measure	**Majority**	**DT**	**GA**	**select**	**fselect**	**bselect**
cross-validation set						
Accuracy [%]	90.4	90.4	**98.1**	94.2	**98.1**	92.3
Specificity [%]	**100**	97.1	**100**	97.1	**100**	**100**
Sensitivity [%]	72.2	77.8	**94.4**	88.9	**94.4**	77.8
MC	0.793	0.786	**0.958**	0.872	**0.958**	0.834
test set						
Accuracy [%]	84.6	82.7	84.6	86.5	**90.4**	84.6
Specificity [%]	**97.1**	**97.1**	94.1	91.2	94.1	**97.1**
Sensitivity [%]	61.1	55.6	66.7	77.8	**83.3**	72.2
MC	0.657	0.613	0.652	0.699	**0.785**	0.743

Table 3. Comparison of single SVM classification, ensemble classification based on variation of data preprocessing, and the proposed ensemble of local experts

Measure	**single RSVM**	**preprocessing ensemble**	**local experts**
cross-validation set			
Accuracy [%]	86.5	86.5	**98.1**
Specificity [%]	88.2	88.2	**100**
Sensitivity [%]	83.3	83.3	**94.4**
MC	0.707	0.707	**0.958**
test set			
Accuracy [%]	86.5	86.5	**88.5**
Specificity [%]	85.3	85.3	**94.1**
Sensitivity [%]	**94.4**	88.9	83.3
MC	0.768	0.719	**0.785**

An ensemble based on variation of data preprocessing is used according to [10], but using RSVMs as base classifier and forward selection for ensemble optimization in order to use the same classifier in all methods compared.

The proposed ensemble of local experts leads to a general improvement in classification performance on the cross-validation set and slightly outperforms single RSVM classification on the test set. While single RSVM classification assigns hard class labels to the samples, ensemble methods provide further information on the classification by the agreement of all experts' predictions on the final classification. Furthermore, the degree of the induced organ toxicity can be derived from the percentage of votes for the toxic class, showing a dose-dependent response. The spectral regions used for final classification can be derived from the subset of local experts, indicating regions of possibly relevant peaks for classification. Thus, an improved classification is achieved and the decision can be further interpreted, which is especially important for applications in drug design.

5 Summary

We presented an ensemble approach for the detection of drug-induced adverse effects based on the classification of NMR spectra. RSVMs are trained on short spectral regions, representing local experts for the discrimination between toxic and non-toxic samples with a limited view on the data. The combination of these local experts is realized by a multiple classifier system and the subset of experts, leading to an improved final ensemble classification, is determined automatically. Thus, new samples are classified based on a combination of classifiers, focussing on a limited amount of spectral information, showing an improvement in classification performance. Furthermore, the analysis of the ensemble classification allows for interpretation of the decision, especially indicating spectral regions of relevant information for classification purposes – a very important aspect in safety pharmacology.

References

1. Freeman, R.: Magnetic resonance in chemistry and medicine. Oxford University Press, New York (2003)
2. Lindon, J.C., et al.: Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxicology and Applied Pharmacology 187(3), 137–146 (2003)
3. Holmes, E., et al.: Development of a model for classification of toxin-induced lesions using ^{1}H NMR spectroscopy of urine combined with pattern recognition. NMR in Biomedicine 11(4-5), 235–244 (1998)
4. Beckonert, O., et al.: NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. Analytica Chimica Acta 490, 3–15 (2003)
5. Fieno, T., Viswanathan, V., Tsoukalas, L.: Neural network methodology for ^{1}H NMR spectroscopy classification. In: ICIIS 1999: Proc. Int. Conf. on Information Intelligence and Systems, pp. 80–85. IEEE Computer Society, Los Alamitos (1999)
6. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
7. Lienemann, K., Plötz, T., Fink, G.A.: On the application of SVM-Ensembles based on adapted random subspace sampling for automatic classification of NMR data. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 42–51. Springer, Heidelberg (2007)
8. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
9. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
10. Lienemann, K., Plötz, T., Pestel, S.: NMR-based urine analysis in rats: Prediction of proximal tubule kidney toxicity and phospholipidosis. Journal of Pharmacological and Toxicological Methods 58(1), 41–49 (2008)
11. Spraul, M., 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)
12. Torgrip, R.J.O., et al.: New methods of data partitioning based on pars peak alignment for improvedmultivariate biomarker/biopattern detection in ^{1}H NMR spectroscopic metabolic profiling of urine. Metabolomics 2(1), 1–19 (2006)

13. Torgrip, R.J.O., Åberg, M., Karlberg, B., Jacobsson, S.P.: Peak alignment using reduced set mapping. Journal of Chemometrics 17, 573–582 (2003)
14. Skov, T., van den Berg, F., Tomasi, G., Bro, R.: Automated alignment of chromatographic data. Journal of Chemometrics 20(11-12), 484–497 (2006)
15. Kuncheva, L., Bezdek, J., Duin, R.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 34, 299–314 (2001)
16. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co. (1989)
17. Matthews, B.W.: Comparison of the predicted and observed secondary structure of the T4 phage lysozyme. Biochimica et Biophysica Acta 405, 442–451 (1975)
18. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, Chichester (2001)
19. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
20. Barnes, R.J., et al.: Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy 43(5), 772–777 (1989)
21. Nord, L.I., Kenne, L., Jacobsson, S.: Multivariate analysis of ^{1}H NMR spectra for saponins from quillaja saponaria molina. Anal. Chim. Acta 446, 197–207 (2001)