In **Proc. Int. Workshop on Multiple Classifier Systems (MCS2007)**,
LNCS 4472, pages 42-51, Heidelberg, Germany, 2007. Springer-Verlag.

On the Application of SVM-Ensembles based on Adapted Random Subspace Sampling for Automatic Classification of NMR Data

Kai Lienemann, Thomas Plötz, and Gernot A. Fink

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

Abstract. We present an approach for the automatic classification of Nuclear Magnetic Resonance Spectroscopy data of biofluids with respect to drug induced organ toxicities. Classification is realized by an Ensemble of Support Vector Machines, trained on different subspaces according to a modified version of Random Subspace Sampling. Features most likely leading to an improved classification accuracy are favored by the determination of subspaces, resulting in an improved classification accuracy of base classifiers within the Ensemble. An experimental evaluation based on a challenging, real task from pharmacology proves the increased classification accuracy of the proposed Ensemble creation approach compared to single SVM classification and classical Random Subspace Sampling.

1 Introduction

The reliable detection of drug induced adverse effects which might be considered toxic for particular organs or regions of organs is a major prerequisite for effective drug design in pharmacology. Within the research field of *Metabonomics* putative toxicities of particular pharmaceuticals are usually indicated by the change of concentrations of metabolites. For both qualitative and quantitative measurements of such changes the so-called ^{1}H *Nuclear Magnetic Resonance (NMR) Spectroscopy* of biofluids extracted from the treated organism has been proven very effective [1]. The process of NMR spectroscopy results in (high-dimensional) spectral data (cf. figure 1) where both positions and intensities of particular peaks convey the information about particular metabolites.

The process of spectra generation including the treatment of experimental animals is a very time and cost intensive task which usually results in rather small sample sets (typically only a few hundred spectra are available each containing several thousand measurement points). In addition to this the manual analysis of these complex data-sets is very tedious and its results are often of more or less subjective type. Thus, sophisticated methods for the automatic

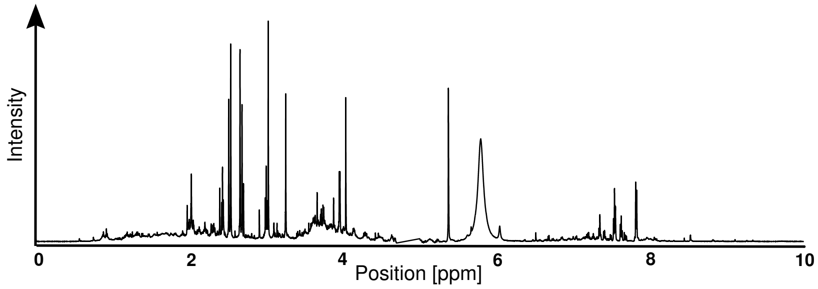

Fig. 1. Exemplary NMR spectrum consisting of approx. 130 000 measurements

classification of NMR spectra dealing with both high dimensionality of the original data and small sample sets are required. Surprisingly, so far only very few, rather straightforward techniques have already been developed for the task of automatic analysis of NMR spectra.

In previous (internal) investigations we observed that the application of *Support Vector Machines (SVMs)* [2] performs best when aiming at the automatic classification of NMR spectra with respect to certain toxicity classes. However, for the application of an automatic analysis system in productive pharmacological environments its classification rate needs to be improved.

Generally, in order to process small but complex data sets recently approaches utilizing multiple classifiers have become popular. According to this, we propose a novel approach for generating SVM Ensembles based on iterative adapted Random Subspace Sampling (RSS) exploiting small but high-dimensional sample sets of NMR spectra. Contrary to standard RSS techniques our method is based on a dimension weighting scheme. According to SVM based classification results those feature vectors' components are favoured which are most informative with respect to the overall analysis. Based on a challenging task examining real pharmacological data the effectiveness of the new approach is demonstrated.

In the following section the state-of-the-art specifically for automatic classification of NMR spectra as well as for the general application of multiple classifier systems is briefly summarized. In section 3 the proposed approach of SVM Ensembles based on adapted Random Subspace Sampling is discussed. The results of the experimental evaluation are presented in section 4. The paper concludes with a discussion of perspectives and an outlook on future work.

2 Related Work

The determination of pharmaceutical adverse effects is an important prerequisite for drug design and its automation is highly desirable. However, according to the literature only few and more or less straightforward techniques have been developed addressing the automatic classification of NMR spectra.

In order to process (very) high-dimensional raw NMR spectral data, usually a basic initial abstraction procedure is applied. Therefore, small spectral regions are aggregated and the corresponding integral value is used for further processing. By means of this *bucketing* technique [3] “feature” vectors consisting of several hundred components (instead of thousands) are created.

The most prominent related work has been pursued within the COMET (*Consortium for Metabonomic Toxicity*) project [4] aiming at a system for complete analysis of (large amounts of non-public) NMR data including their automatic classification based on *CLOUDS* (*Classification of Unknowns by Density Superposition* [5]). Using CLOUDS toxicity classes are modeled by mixture densities of Gaussians (with predefined standard deviations) centered on the training samples used. Inspecting the related literature it is, unfortunately, not clear how the system performs for small sample sets as addressed by this paper.

For general classification tasks where only small sample sets are available the application of Support Vector Machines has been proven very effective. Classification is based on linear separation of data originating from different classes. Therefore, a discriminating hyperplane is constructed utilizing a subset of training vectors as support-points and a non-linear transformation into a high-dimensional feature space allowing for linear separation. For efficient evaluation usually kernel functions are applied avoiding the actual transformation into the high-dimensional space. Since linear separability (even in the high-dimensional space) cannot be guaranteed for all sample data the hyperplane’s optimization is related to a so-called soft margin defined by *slack variables* [2].

In the last few years the application of multiple classifier systems has been proven effective for complex data sets. Therefore, different base classifiers are estimated either on modified sample sets or on alternative data representations. Both variations of the training data are derived from the original sample sets. Applying the set of classifiers to the original task results in multiple decisions which are aggregated in various ways in order to achieve a final classification. Compared to single classifiers substantial improvements in the overall classification performance of such *Classifier Ensembles* can be achieved [6].

The principle constraint for base classifiers used for Ensemble techniques is a classification accuracy of better than random – so-called *weak classifiers*. However, the Ensemble approach performs even better when *strong base classifiers* like SVMs are deployed (cf. e.g. [7]). Compared to single classifier approaches substantial improvements in classification performance can be obtained by Ensembles only when the underlying base classifiers contain substantial mutual diversity, i.e. modeling different characteristics of the training set.

As one approach for Ensemble creation utilizing a (limited) set of training data *Bagging* aggregates classifiers estimated on bootstrap replicates of all training samples [8]. Sample sets are derived (most likely) avoiding redundant or less informative samples for training and therefore possibly increasing the classification accuracy of the base classifiers. Alternatively, *Boosting* focuses on (re-)weighting of sample data for their consideration in the training procedure. During this iterative procedure the focus is concentrated on those samples which are harder to classify, i.e. causing classification errors.

Alternatively, base classifiers covering sub-spaces of the original feature space can be integrated into Ensemble approaches. Most prominently the *Random Subspace Sampling (RSS)* technique randomly selects subsets of feature components for base classifier training [9]. RSS reduces the effect of redundant or less informative dimensions and (most likely) alleviates the discrepancy between small sample-set sizes and high dimensionality.

3 SVM-Ensemble based on Adapted RSS

The analysis of our first experiments addressing the automatic classification of NMR-spectra with respect to organ toxicities empirically proved the suitability of C-SVMs [2], explicitly controlling the sum of slack-variables in soft margin classification. Thus, they were chosen as starting point for our developments. Multiple SVM base-classifiers are integrated into an Ensemble aiming for improved classification of NMR-spectra when only small training sets are available.

Even when considering the bucket representation of NMR-spectra it is very unlikely that every particular dimension of the resulting (high-dimensional) feature vectors represents a similar amount of information for the overall classification process. In order to obtain reasonably diverse but relevant sample-sets for the estimation of the abovementioned base classifiers we propose the application of (improved) Random Subspace Sampling.

According to our practical experiences standard RSS, unfortunately, does not guarantee the selection of the most relevant feature components.[1] Thus, in our modified approach the random selection process is based on an underlying probability distribution assigning weights to every feature component. Exploiting this distribution multiple sub-spaces are derived from the original 203-dimensional NMR-bucket space by RSS. By means of the resulting sample-sets SVM base classifiers are trained and integrated into an Ensemble.

Since the optimal feature components' weights are not known in advance its probability distribution is learned in an iterative adaptation procedure. For this purpose, sub-spaces are created by applying adapted RSS, and SVMs are trained accordingly. During cross-validation these base classifiers are evaluated and according to the classification accuracies the weights of the feature components are either increased or decreased, thus, propagating the most relevant components.

In addition to the overview of the new approach given above and illustrated in figure 2, in the following, details regarding SVM training, adaptation of the actual weights, and the creation of the SVM Ensemble will be described.

3.1 Automated SVM Training

The classification accuracy of C-SVMs is mainly dependent on the choice of a feasible C value and possibly on additional kernel-specific parameters. A linear

[1] Since training / evaluation of SVM based Ensembles is rather time intensive the number of base classifiers, i.e. RSS guesses, is practically limited.

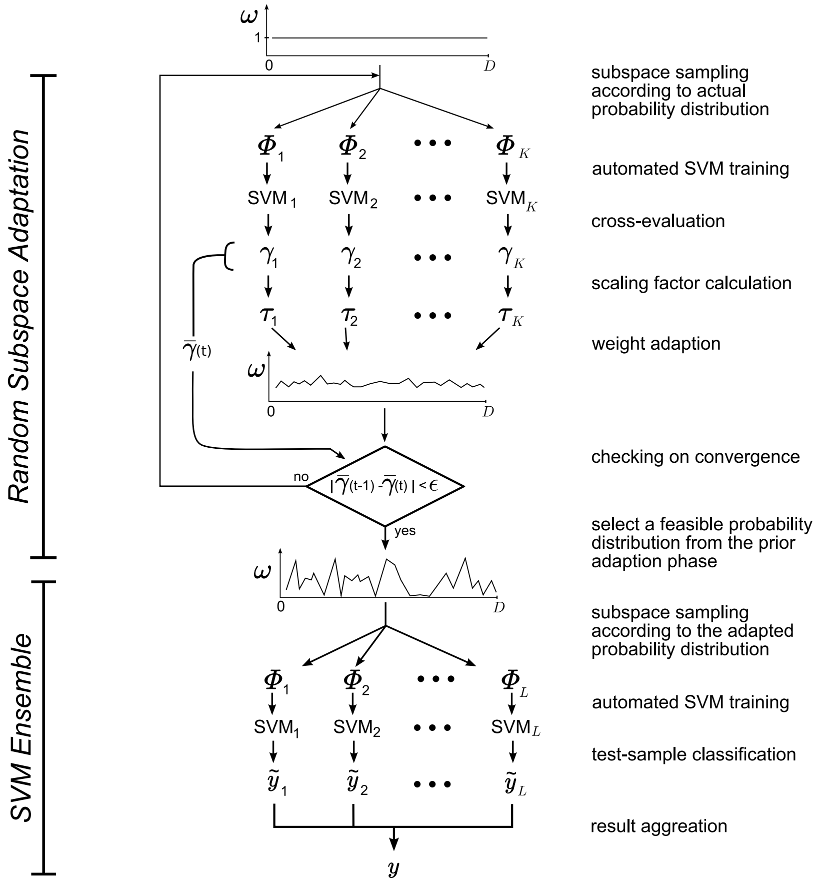
Random Subspace Adaptation
SVM Ensemble
ω
1
0
D
subspace sampling according to actual probability distribution
Φ1 Φ2 • • • ΦK
automated SVM training
SVM1 SVM2 • • • SVMK
cross-evaluation
γ1 γ2 • • • γK
scaling factor calculation
τ1 τ2 • • • τK
weight adaption
γ̄(t)
ω
0
D
checking on convergence
no
|γ̄(t-1) - γ̄(t)| < ε
yes
select a feasible probability distribution from the prior adaption phase
ω
0
D
subspace sampling according to the adapted probability distribution
Φ1 Φ2 • • • ΦL
automated SVM training
SVM1 SVM2 • • • SVML
test-sample classification
ỹ1 ỹ2 • • • ỹL
result aggreation
y

Fig. 2. Creation of SVM-Ensembles based on adapted RSS (see text for description).

kernel is not parametrized, therefore reducing the complexity and time needed in the training phase. The C-parameter is optimized by grid search and the linear kernel is used in all further investigations. Choosing too small C-values leads to a low classification accuracy and can be improved by selecting larger values up to an asymptotic behavior (cf. e.g. [10]). This process motivates an automatic grid selection. A wide and coarse logarithmic grid is defined in a first phase and the evaluation starts at a reasonable small value, stopping if convergence in classification accuracy is reached (cf. figure 3). The best classification result γ_T is determined and, based on this, the grid for the second phase is defined. Starting at the first value exceeding $\frac{\gamma_T}{2}$ up to the point of convergence, the solution space is divided into M steps equally spaced on a logarithmic scale (cf. figure 4). Increasing M results in a finer grid, but also in a longer training phase. The best parameter setting is chosen and used for the classification of test samples.

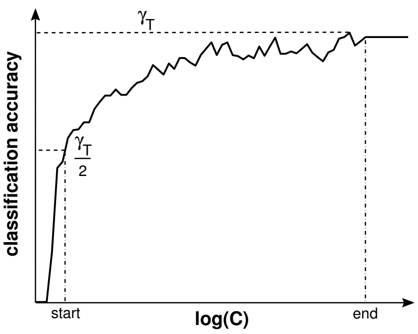

Fig. 3. Accuracy on a coarse, wide grid

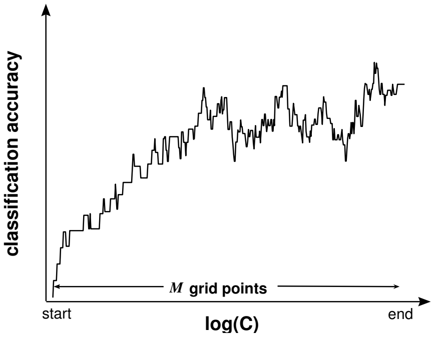

Fig. 4. Accuracy on an optimized grid

3.2 Modified Random Subspace Sampling

We propose a modification of Random Subspace Sampling (upper part of figure 2) , specifically selecting the putatively most informative features with higher probability. Increased classification accuracy of SVMs trained on the resulting subspaces can be expected due to the explicit adaptation of the probability distribution "guiding" the underlying random process of RSS towards the selection of reasonable subsets.

Given a D-dimensional data set, all weights w_i for $i = 1, \ldots, D$ are initialized to one and the selection probability p_i for each feature is calculated according to these weights.

$$p_i = \frac{w_i}{\sum_{j=1}^{D} w_j} \quad i = 1, \ldots, D$$

Based on these probabilities K (most likely) different d-dimensional ($d < D$) subspaces Φ_k for $k = 1, \ldots, K$ are determined based on the actual probability distribution and linear SVMs are trained on every subspace according to the algorithm described in section 3.1. All SVMs are sorted with respect to their classification accuracy, or an alternative evaluation measure, ranging from the best accuracy γ_B to the worst γ_W. A scaling factor τ_k is determined for each SVM, dependent on a free parameter ν and the corresponding classification accuracy γ_k.

$$\tau_k = \begin{cases} \frac{1}{\nu} + \left(\frac{2(\gamma_k - \gamma_W)}{\gamma_B - \gamma_W}\right)\left(1 - \frac{1}{\nu}\right) & \text{if } \frac{\gamma_k - \gamma_W}{\gamma_B - \gamma_W} < 0.5 \\ (2 - \nu) + \frac{(2\nu - 2)(\gamma_k - \gamma_W)}{\gamma_B - \gamma_W} & \text{otherwise} \end{cases} \qquad k = 1, \ldots, K;\ \nu > 1$$

The values of every dimension are multiplied by the scaling factor of the SVMs, it was used within, reducing the weights of dimensions possibly leading to a lower classification accuracy and vice versa. The iterative reduction of probabilities corresponding to putatively less informative dimensions leads, after several iterations, to their de-facto exclusion from RSS due to a selection probability close to zero. Consequently, the classification accuracy of trained SVMs is increased, and, simultaneously, the diversity of selected subspaces is decreased (see below).

3.3 SVM-Ensemble

The overall principle of our proposed classification system corresponds to an Ensemble of L SVMs as (strong) base classifiers (lower part of figure 2) . Different SVMs are trained based on RSS, with an adapted random selection of dimensions for subspaces, and aggregating their classifications into a final decision. All SVMs within an Ensemble are trained on diverse subspaces Φ_l for $l = 1, \ldots, L$, determined according to a probability distribution of the prior adaptation process as described in section 3.2. Unlike an increasing classification accuracy, the diversity of SVMs built within the iterative adaptation process decreases and possibly converges to one final subspace. Therefore, the final probability distribution is apparently not the optimal choice in order to build an Ensemble of diverse base classifiers and an optimal intermediate result has to be selected. In addition to the classification accuracy, several measures of diversity (cf. [11], [12]) are possible and can be applied for the selection of a feasible probability distribution for building the SVM-Ensemble.

The selected probability distribution is used for the determination of an subspace for every base SVM and these are trained according to the algorithm described in section 3.1. A final classification y is achieved by aggregating the base classifier decisions $\tilde{y}_l$ by maximum vote. An improvement in the Ensemble accuracy is expected due to the improved base classifier's accuracy and their combination in an Ensemble.

4 Experimental Evaluation

In order to demonstrate the effectiveness of the new approach for SVM-Ensembles based on adapted Random Subspace Sampling as proposed in this paper an experimental evaluation based on real NMR sample sets has been pursued. In the following the data-sets used as well as the methodology applied, and the actual results are briefly summarized.

4.1 Data-set and Methodology

The sample-set used for experimental evaluation consists of NMR-spectra analyzed in a real pharmacological task. Every spectrum originally consists of approximately 130 000 measurement points. By means of an initial bucketing step the dimensionality is reduced to 203. In summary, the data-set consists of 530 samples where every spectrum is assigned either to control (420 samples) or toxic (110), i.e. a two-class problem is considered.

For training, parameter optimization, and test the data-set was split into five disjoint sets by randomly selecting samples. Note that the actual random selection respected the imbalanced distribution of toxic and control spectra as mentioned above. By means of a five-fold cross-validation we ensured that every sample is once treated as test. In every of the five configurations possible three fifths are selected for training and one fifth for cross validation. The final classification rates are averaged over the results achieved on the five test sets (the particularly remaining fifths).

In order to avoid putative statistical artifacts all experiments related to random subspace sampling have been performed twenty times which (empirically) represents an upper limit for reasonable turn around times using current personal computers. The results reported correspond to averaging over all experiments.

Throughout the whole process of parameter training (SVM estimation, and adaptation of feature components' weights) we considered the *Matthews correlation coefficient (MC)* as optimization criterion:

$$MC = (TP{\times}TN{-}FP{\times}FN)((TP{+}FN)(TP{+}FP)(TN{+}FP)(TN{+}FN))^{-1/2}$$

with TP as number of true positive predictions, FP as number of false positives, TN as number of true negatives, and FN as number of false negatives, respectively. MC is normalized to $[-1\ldots 1]$. The larger MC the better the overall classification performance. The Matthews correlation coefficient was chosen because it is hardly sensitive to imbalanced data-sets. In addition to this, classification rates (overall (acc), and related to toxic (acc_T) and control (acc_C) samples) are reported which seems more informative for the actual Metabonomics task.

The experiments have been conducted using Matlab and the libSVM [13] interface, and our own Ensemble classification system.

4.2 Results

We first compared the SVM-Ensemble approach to single SVM classification in order to show the improvements already achieved by RSS SVM-Ensemble. In addition, the classification results achieved by our proposed method are discussed related to the single SVM and standard RSS approach.

The single SVM classifier and all further mentioned SVM base classifiers were trained according to the algorithm described in section 3.1 by cross-validation, as described in section 4.1, using $M = 300$ grid points in the second training phase. Random Subspace Sampling was performed by selecting 70 % of the original dimensionality randomly and the classification results of all base classifiers were aggregated according to the maximum vote decision rule. Under variation of L, the optimal number of base classifiers was assessed by cross-validation (cf. figure 7) and the classification results on the validation and test-set are shown in table 1. An increased MC can be achieved on the cross-validation set, but not on the test-set due to the reduced classification rate of toxic samples, which implies a better performance of the control samples.

The adaptation of prior probabilities for RSS was performed according to the algorithm described in section 3.2 using $K = 20$ SVMs in each iteration

Table 1. Classification accuracy on the cross-validation and test-set.

	cross-validation				test			
Method	**MC**	$\mathbf{acc}_{[\%]}$	$\mathbf{acc_C}_{[\%]}$	$\mathbf{acc_T}_{[\%]}$	**MC**	$\mathbf{acc}_{[\%]}$	$\mathbf{acc_C}_{[\%]}$	$\mathbf{acc_T}_{[\%]}$
single SVM	0.462	80.9	85.5	63.6	0.422	79.6	84.8	60.0
RSS(L=27)	0.537	86.1	**95.1**	51.5	0.404	82.4	**93.1**	41.4
adapted RSS (L=23)	**0.623**	**87.7**	92.6	**69.1**	**0.499**	**82.8**	87.7	**64.1**

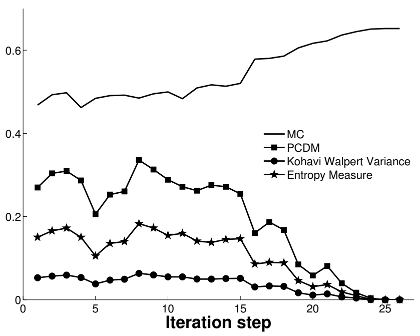

Fig. 5. MC vs. certain diversity measures

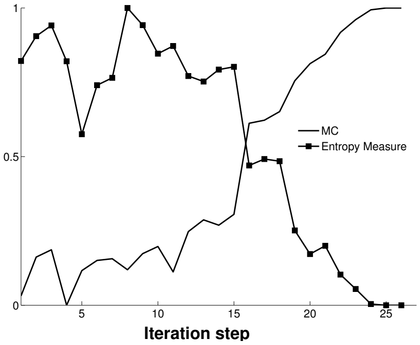

Fig. 6. MC and Entropy scaled to [0 - 1]

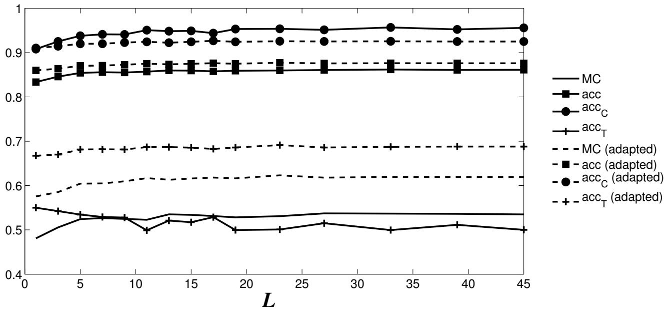

Fig. 7. Comparison of RSS and adapted RSS – Matthews correlation coefficient (MC), classification rate overall (acc) and regarding control (acc_C) and toxic (acc_T) samples.

and scaling factor $\nu = 2$. The process of increasing accuracy and decreasing diversity is shown in figure 5, using the MC for accuracy determination and the Kohavi Walpert Variance, the Entropy Measure [11] and the Percentage Correct Diversity Measure (PCDM) [12] as possible rates for the determination of diversity. According to the MC and the Entropy Measure, a probability distribution is selected by scaling both measures to the range [0 - 1] and using the distribution from the iteration closest to the intersection point as demonstrated in figure 6. The cross-validation classification results of the RSS and adapted RSS SVM-Ensemble are illustrated in figure 7 under variation of L.

Our proposed method increases nearly all evaluation results compared to single SVM classification and RSS SVM-Ensemble. The high acc_C-values of the classical RSS approach results from the low classification rate of toxic samples, thus predicting most of the test samples as control. However, with the proposed method performance can be significantly increased for toxic samples, thus also yielding an overall improvement in the MC measure.

5 Discussion

We presented a modified Random Subspace Sampling approach for the construction of SVM-Ensembles. The random selection process is based on an underlying probability distribution, assigning high probabilities to features, regarded as most informative for classification by an prior adaptation phase. Within this adaptation phase several SVMs are trained, evaluated and the weights for every feature are modified, proportional to the relative classification accuracy. The improvement of the base classifiers classification accuracy by using an adapted probability distribution for subspace sampling leads to an overall improvement in accuracy of the Ensemble.

A further improvement in classification accuracy is expected by the use of alternative SVM kernel functions like the radial basis function or sigmoid kernel. But for an experimental evaluation an *efficient* SVM training has to be developed due to the more complex process of parameter optimization.

The bucketing procedure reduces the spectral dimensionality and serves as simple feature extraction method, but decreases the resolution and correspondence between features and single peaks. Reducing the size of integrated segments within the bucketing procedure facilitates the interpretation of weights achieved in the adaptation phase. If a correspondence of most informative dimensions to peaks of single metabolites could be achieved, (possibly) new biomarkers for the detection of organ toxicities could be discovered.

Acknowledgements

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.

References

1. Bales, J., et al.: Use of high resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. Clinical Chemistry **30** (1984) 426–432
2. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, (2002)
3. Spraul, M., et al.: Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. Journal of pharmaceutical and biomedical analysis **12** (1994) 1215–1225
4. 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** (2003) 137–146
5. Ebbels, T., et al.: Toxicity classification from metabonomic data using a density superposition approach: CLOUDS. Analytica Chimica Acta **490** (2003) 109–122
6. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 226–239
7. Kim, H.C., Pang, S., Je, H.M., Kim, D., Yang Bang, S.: Constructing support vector machine ensembles. Pattern Recognition **36** (2003) 2757–2767
8. Breiman, L.: Bagging predictors. Machine Learning **24** (1996) 123–140
9. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 832–844

10. Keerthi, S.S., Lin, C.J.: Asymptotic behaviors of support vector machines with gaussian kernel. Neural Computation **15** (2003) 1667–1689
11. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning **51** (2003) 181–207
12. Banfield, R.E., et al.: A new ensemble diversity measure applied to thinning ensembles. In: Multiple Classifier Systems. Volume 2709 of LNCS. Springer (2003) 306–316
13. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.