In **Proc. Int. Workshop on Multiple Classifier Systems (MCS2009)**,
LNCS 5519, pages 498-508, Heidelberg, Germany, 2009. Springer-Verlag.

Stacking for Ensembles of Local Experts in Metabonomic Applications

Kai Lienemann, Thomas Plötz, and Gernot A. Fink

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

Abstract. Recently, Ensembles of local experts have successfully been applied for the automatic detection of drug-induced organ toxicities based on spectroscopic data. For suitable Ensemble composition an expert selection optimization procedure is required that identifies the most relevant classifiers to be integrated. However, it has been observed that Ensemble optimization tends to overfit on the training data. To tackle this problem we propose to integrate a stacked classifier optimized via cross-validation that is based on the outputs of local experts. In order to achieve probabilistic outputs of Support Vector Machines used as local experts we apply a sigmoidal fitting approach. The results of an experimental evaluation on a challenging data set from safety pharmacology demonstrate the improved generalizability of the proposed approach.

1 Introduction

In the last two decades the development of new NMR (nuclear magnetic resonance) measurement techniques together with a steadily increasing spectral resolution and improved data quality, respectively, have provided the opportunity for in-depth automatic analysis of biofluids. Thereby, the ultimate goal is to detect specific changes of an organism's metabolism that is, for example, induced by drug applications in safety pharmacology. Generally, the research field of Metabonomics addresses "the quantitative measurement of the time-related multiparametric metabolic response of living systems to pathophysical stimuli or genetic modification" [1]. In addition to classical analysis methods from clinical chemistry and histopathology meanwhile also automatic classification techniques utilizing pattern recognition approaches have been applied successfully.

Recently, multiple classifier systems have been developed for the detection of drug-induced organ toxicities with applications to industrial safety pharmacology (cf. e.g. [2]). It has been shown that the use of Ensemble methods that integrate multiple classifiers each providing local views on the spectra outperforms single classifier approaches. However, when comparing the classification performance of classifier Ensembles achieved on cross-validation data with those on test-sets

it becomes clear that the systems tend to overfit. This is especially critical when only small portions of NMR spectra are available for training and optimization.

In this paper we present an enhancement of Ensembles of local experts for Metabonomic applications that explicitly focuses on improved generalizability. Our goal is to stabilize the classification performance from cross-validation to test. Therefore, a variant of stacked generalization [3] as a combination method of predictions from different models is integrated into our Ensemble system for NMR classification. Thereby, the decisions of local experts that focus on small parts of the spectra serve as probabilistic level-0 model outputs. Since for this level of classification we use Support Vector Machines that generate binary decisions pseudo-probabilities are derived by means of a sigmoidal mapping approach. Subsequently, an additional classifier – the level-1 generalizer – is applied to the vectors of pseudo-probabilities providing the final classification result. In order to find the most suitable configuration we investigated the appropriateness of certain variants of level-1 generalizers. By means of an experimental evaluation on a realistic NMR dataset from industrial safety pharmacology we demonstrate the effectiveness of the proposed approach. Using stacking for Ensembles of local experts improved generalization can be achieved for Metabonomic applications.

In the following section the general background for the automatic analysis of NMR spectra is given together with the motivation of our current work. Subsequently in section 3 the proposed Ensemble system that focuses on improved generalizability for toxicity prediction is described. The results of the experimental evaluation are presented in section 4. The paper ends with a conclusion.

2 Background & Motivation

Within an NMR spectrum of some analyzed sample the concentration of numerous molecules is represented by peak intensities. Peak positions are specific for the respective molecules as exemplary shown in figure 1. Changes in the concentration of several molecules can be detected by comparison of corresponding peak intensities between different samples. Substantial changes indicate an alteration of the organism's metabolic profile. Consequently, a major issue in Metabonomics research is the development of systems for an automatic analysis of samples for the identification of relevant peak changes.

Recently pattern recognition techniques have also been applied to Metabonomics tasks. It has been demonstrated that, generally, it is possible to detect changes in metabolisms by comparison of spectroscopic data (cf. [4–6]). Certainly the most promising approach – *Classification of Unknowns by Density Superposition* [7] – was developed within the *Consortium for Metabonomic Toxicity (COMET)* project [8]. This approach is based on spectra classification using probabilistic neural networks [9] that were trained exploiting a large database which is not publicly available. When generally analyzing related work as it has been reported in the literature it becomes clear that so far only little research has been devoted to the automatic classification of NMR data.

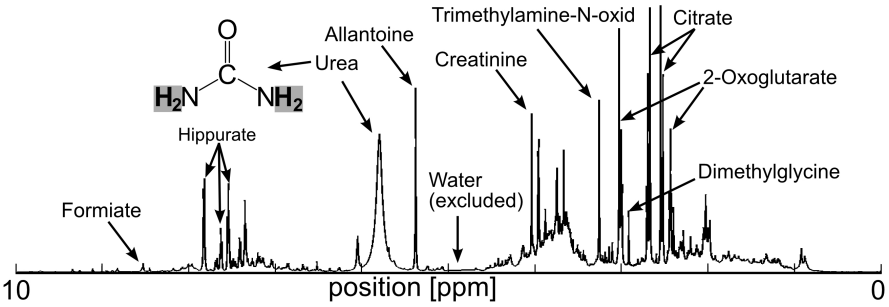

Fig. 1. Exemplary ^{1}H-NMR spectrum of an urine sample from an untreated rat. A subset of peaks and their corresponding molecules, the chemical structure of urea and the signal emitting hydrogen atoms are denoted.

In our previous work for the first time Ensemble methods were developed for the prediction of organ toxicities based on spectroscopic data [10]. We developed a multiple classifier system that introduced weighted Random Subspace Sampling (RSS) with applications to the field of Metabonomics. The weights of variables relevant for toxicity classification were iteratively optimized by an unsupervised learning approach. Subspaces were classified by Support Vector Machines (SVMs) and aggregation of the predictions was performed by majority voting. It has clearly been shown that differences in the relevance of distinct variables, i.e. parts of the NMR spectra, exist w.r.t. their significance for classification. Favoring relevant regions in RSS finally improved the Ensemble classification performance. An alternative multiple classifier approach for the analysis of NMR data has been proposed in [2] where preprocessing methods were varied for Ensemble creation. Again it has been shown that the combination of multiple classifiers outperforms single classifier approaches in Metabonomics. The idea of focusing on certain spectral regions for classification and their combination in an Ensemble system has been further investigated in [11]. In this approach an Ensemble of local experts is created by training classifiers on short spectral regions that are determined by a sliding window technique. Final aggregation of local experts' predictions is achieved by majority voting, whereas the subset of experts used for final voting is optimized in order to achieve an improved classification accuracy. Thus, the classification decision is based on specific parts of the spectra, which are supposed to reflect biologically relevant changes.

Ensemble optimization for automatic analysis of NMR data has achieved a nearly perfect classification performance on the validation sets. Unfortunately, the generalization capabilities of the system on this extremely challenging type of data is still not optimal. An experimental evaluation of several strategies for ensemble optimization has indicated a generally decreasing classification performance on unknown test data [11]. A promising approach for the combination of multiple models emphasizing the idea of cross-validation for an improvement of generalization capabilities – *generalized stacking* – was proposed by Wolpert [3],

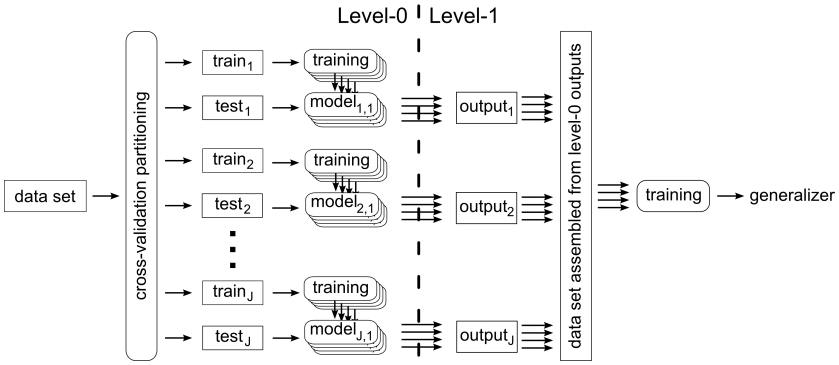

Fig. 2. Illustration of the generalized stacking approach. A data-set is subdivided into J cross-validation parts and 4 models are trained on each set, respectively. The outputs of different sets are collected in a single data set for training the final classifier.

and further discussed by Breiman [12] and Ting et al. [13] (cf figure 2). In this approach a given data set is split into J training and test sets according to the J-fold cross-validation principle. Different so-called level-0 models are estimated on the training sets and applied to predict the corresponding test sets. The predictions from all test sets are collected in a new data set and further used as input for training of a final classifier, the so-called level-1 generalizer. For a final classification level-0 models are trained on the whole data set and new samples are classified by the level-1 generalizer based on the predictions of level-0 models. Generally, the type of level-0 model and level-1 generalizer is not restricted to any specific classification algorithm. However, Ting et al. assume probabilistic outputs of level-0 models rather than class predictions for a successful application of generalized stacking [13]. By means of generalized stacking the severe problem of overfitting in Ensemble optimization methods can be avoided, which is a clear advantage contrary to optimization of expert selection for majority voting as a nontrainable combination method.

3 Stacking for Fusion of Local Spectral Information

In order to stabilize the classification results of the multiple classifier system of local experts for the analysis of NMR spectra we developed an Ensemble estimation approach that explicitly focuses on improved generalizability. Therefore, the problem of overfitting is tackled by the application of a trainable combiner which is inspired by the concept of generalized stacking. Consequently, the new approach is referred to as *stacking for Ensembles of local experts*. The basic idea is to use the output of local classifiers that represent statistical models for designated parts of NMR spectra as input data for a second classifier. The parametrization of the level-1 generalizer is adjusted by cross-validation. Thereby, rules for classification according to the local experts' predictions are

automatically derived by the stacked classification algorithm. Figure 3 gives an overview of the system.

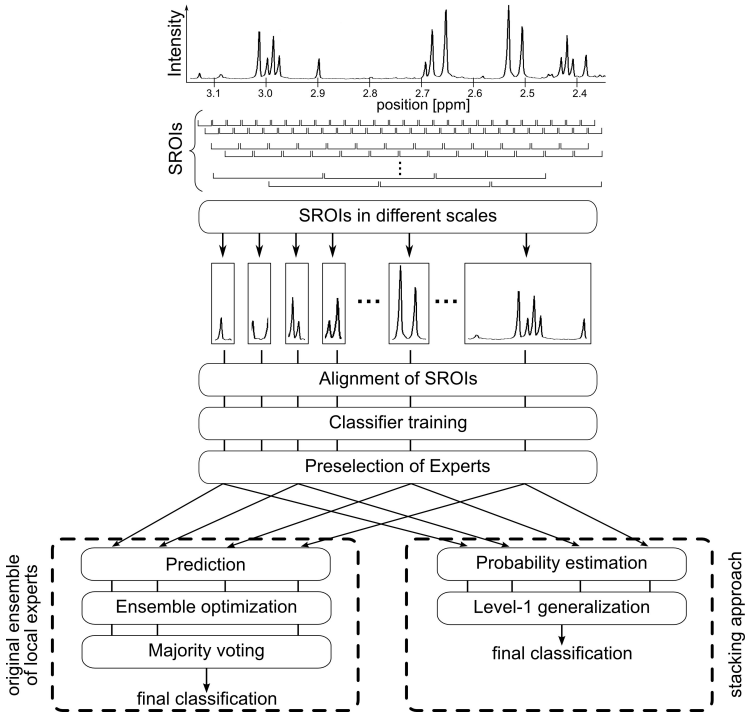

Fig. 3. Classification system for automatic analysis of NMR data based on an Ensemble of local experts and stacking for improved generalizability (see text for description).

The new approach of stacking for local experts in Metabonomic applications extends our previous work on Ensembles of local experts for the automatic analysis of NMR data [11]. Local information of spectroscopic data is treated by classifiers that have restricted views on the data thereby covering very few peaks only. Spectral regions of interest (SROIs) are determined on NMR data by applying a sliding window approach and an alignment procedure, respectively, in order to compensate peak shifts induced by changes of physiochemical factors like pH or ion concentration. SVMs using a radial basis function kernel are trained for each SROI of spectral intensities and serve as local experts for specific regions. Initially experts exhibiting insufficient classification performance on a validation set are excluded from the Ensemble and selection of experts for Ensemble aggregation by majority voting is optimized in a final step in order to achieve a suitable classification accuracy.

Only very few substances (or combinations of substances) that indicate induced organ toxicities after drug application (by changes in their concentration)

are known in safety pharmacology. This fact supports the assumption that even though concentration information of numerous molecules is present in an NMR spectrum only a very small fraction is useful for the detection of drug-induced organ toxicities. Thus, the identification of a suitable combination of local experts for the final classification is the most crucial step in the Ensemble of local experts. Unfortunately, this process is prone to overfitting.

Dos Santos et al. proposed to use a genetic algorithm approach for (general) Ensemble optimization which reduces the tendency to overfitting [14]. They found that a cross-validation procedure has to be used in order to evaluate the classification performance on an independent validation set. However, analyzing our task it becomes clear that cross-validation is not generally applicable to majority voting. Thus, we focus on an alternative aggregation approach aiming at improved generalizability while retaining high classification performance.

According to the terminology used in the stacked generalization literature in our approach SVMs, serving as local experts on NMR data, are used as level-0 models. According to [13] for best generalization the level-1 generalizer should have probabilistic input, i.e. level-0 classification have to provide probabilities rather than binary decisions. The latter is, however, the case for standard Support Vector Machines as they generate -1 or 1 decisions, respectively, depending on the position of the test sample w.r.t. the hyperplane that separates the particular classes. In order to achieve probabilistic SVM decisions we integrate an additional post-processing step into SVM classification. As developed by Platt [15], and further refined by Lin and coworkers [16] probabilities can be generated from SVM decisions by fitting a parametrized sigmoidal function to the distances of the samples from the labeled training set to the separating hyperplane. The rationale is based on the assumption that greater distances indicate higher confidences for the classification. Analogously smaller distances correspond to smaller confidences. In our approach SVM probabilities derived in this way represent the output of the local experts which is fed into the level-1 classifier.

To sum up, the original Ensemble of local experts approach is extended using a stacked classifier on the outputs of the local experts for final classification. Furthermore, probability estimates of SVMs are integrated in the Ensemble system and serve as input for the level-1 generalizer. These enhancements are supposed to increase the generalization of the Ensemble of local experts, which will be shown by an experimental evaluation and comparison to previous results.

4 Experimental Evaluation

In order to evaluate the effectiveness of the new approach with respect to the addressed improved generalizability we performed various practical experiments. As in our previous work they are related to the detection of drug-induced organ toxicities based on a challenging real-world data set from pharmaceutical

industry[1]. This set contains 896 ^{1}H NMR spectra of urine samples from rats treated with one of 53 pharmaceuticals. According to literature investigations and histological judgments induced organ toxicity regarding proximal tubulus (kidney) is present in 259 samples (= 18 pharmaceuticals). Details on spectra measurements, data treatment and histological judgment are given in [11].

The presentation of the results is structured as follows. First we determine the optimal configuration of the proposed approach by evaluating the suitability of various level-1 generalizers, measuring the impact of probabilistic level-0 model outputs for our task. Subsequently the classification capabilities are directly compared with those that have been achieved using the original Ensemble of local experts thereby clearly indicating the improved generalizability of the proposed method for the particular test sets. For all experiments the selection of SROIs, training of local experts and their preselection for Ensemble generation, resulting in 147 local experts for the final Ensemble, were performed as previously [11].

We performed a five-fold cross-validation and test procedure for training, parameter optimization and final test. Samples were grouped according to target and indication of their corresponding pharmaceutical. These groups of samples are sub-divided into five sets while trying to keep ratios of non-toxic and toxic samples approximately equal. Training was pursued using three fifths of the sets, parameter optimization by one fifth and final testing on the remaining set in every possible configuration. Final classification rates are given as averages over the results on the particular sets. In addition to focusing on specific samples a further goal is the classification of *pharmaceuticals* as being toxic or non-toxic. Thus, results for analyzed samples that have been collected at different time-points are aggregated to final classifications of the corresponding pharmaceutical (maximum mean value) – referred to as group-classification. Due to its robustness to imbalanced data-sets the *Matthews Correlation Coefficient* [17] – MC – (normalized to $[-1 \dots 1]$) has been used as primary evaluation criterion for all training and optimization procedures. For completeness also classification accuracy (acc), specificity (spec) and sensitivity (sens) values are shown.

Table 1 summarizes the classification results achieved for different level-1 generalizers based on probabilistic level-0 model outputs. Results for cross-validation and test using either a k nearest neighbor classifier (kNN – with $k = 7$ optimized according to a fixed grid from one to 31), grid-search optimized SVMs with linear (LSVMs) or radial basis kernel functions (RSVMs), respectively, and random forests (RFs) [18] are shown. RFs are parametrized depending 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. Analyzing the results (level of significance for all sample-based experiments: $\approx \pm 2.5\%$) it can be seen that RSVMs outperform all other techniques. Furthermore, it becomes clear that improved stability of classification results when turning from cross-validation towards test is gained independently of the particular choice of level-1 generalizer.

[1] The presented evaluation is restricted to this data set due to the lack of publicly available data sets of NMR spectra. The presented approach is, however, not dependent on the type of data used and generally applicable.

Table 1. Classification performance of different level-1 generalizer algorithms based on probabilistic outputs of RSVMs as level-0 models in the Ensemble of local experts.

	cross-validation				test			
Measure	***k*NN**	**RF**	**LSVM**	**RSVM**	***k*NN**	**RF**	**LSVM**	**RSVM**
acc [%]	82.7	77.0	82.4	**83.5**	80.7	76.7	80.7	**81.0**
spec [%]	**95.4**	92.5	94.5	94.4	**93.9**	92.8	93.3	92.0
sens [%]	51.4	39.0	52.5	**56.8**	48.3	36.7	49.8	**54.1**
MC	0.551	0.383	0.542	**0.575**	0.494	0.366	0.496	**0.510**

Table 2. Classification performance using probabilistic or prediction outputs of RSVMs as local experts and RSVM for stacked classification.

	cross-validation		test	
Measure	**probabilistic**	**prediction**	**probabilistic**	**prediction**
acc [%]	**83.5**	76.9	**81.0**	61.7
spec [%]	**94.4**	91.1	**92.0**	73.8
sens [%]	**56.8**	42.1	**54.1**	32.1
MC	**0.575**	0.387	**0.510**	0.058

In order to validate the necessity of probabilistic outputs of level-0 models for a generalizing stacking system (as claimed by Ting et al. [13] – cf. section 3) the new approach was also evaluated using binary predictions of local experts. By means of the results presented in table 2 the assumption of Ting et al. can clearly be confirmed also for the Metabonomic application case. Using Ensembles with probabilistic level-0 outputs better classification accuracies together with improved generalizability can be achieved.

Although the preselection of local experts already reduces the set of experts used for stacked classification to those with a reasonable classification accuracy, a further selection of experts is beneficial as shown in the original Ensemble of local experts approach. A well-known method for variable weighting of a labeled multidimensional data set according to the relevance of the variables for class separation is the projection to latent structures (PLS, also referred to as partial least squares) [19]. PLS transformation is comparable to principal component analysis (PCA) differing, however, in the optimization criterion. For PLS it is not the explained variance of the new coordinate system that is optimized but the covariance between the data variables and class labels. Thus, PLS focuses on variables (in this case local experts) relevant for class discrimination, thereby achieving an implicit weighting of experts. The application of the PLS transformation on the probabilistic outputs prior to classification by a RSVM leads to an improved classification accuracy on the cross-validation set and only a slight decrease on the test set can be observed (cf. table 3).

The final part of the evaluation addressed a direct comparison of the results achieved using either the original Ensemble of local experts approach or the enhanced version integrating the proposed stacking technique. The results presented in table 4 clearly indicate the improved generalization of the latter. The

Table 3. Changes in classification performance using a PLS transformation prior to classification by RSVM.

	cross-validation		test	
Measure	**RSVM**	**PLS + RSVM**	**RSVM**	**PLS + RSVM**
acc [%]	**83.5**	**83.5**	81.0	**82.6**
spec [%]	**94.4**	90.0	**92.0**	88.2
sens [%]	56.8	**67.6**	54.1	**68.7**
MC	0.575	**0.590**	0.510	**0.574**

Table 4. Evaluation of the original Ensemble of local experts approach and the proposed stacking modification on cross-validation (xval) and test. Bold numbers indicate the best performance of the two methods on the respective set, while italic numbers highlight the best relative change (Δ) when comparing cross-validation and test.

	Local Experts			**Local Experts + Stacking**		
Measure	xval	test	Δ	xval	test	Δ
sample classification						
acc [%]	**86.3**	77.8	-9.9 %	83.5	**82.6**	*-1.1 %*
spec [%]	**99.2**	**94.2**	-5.0 %	90.0	88.2	*-2.0 %*
sens [%]	54.4	37.5	-31.1 %	**67.6**	**68.7**	*+1.6 %*
MC	**0.659**	0.402	-39.0 %	0.590	**0.574**	*-2.7 %*
group classification						
acc [%]	**98.1**	88.5	-9.8 %	92.3	**90.4**	*-2.1 %*
spec [%]	**100**	94.1	-5.9 %	**100**	**97.1**	*-2.9 %*
sens [%]	**94.4**	**83.3**	-11.8 %	77.8	77.8	*±0 %*
MC	**0.958**	0.785	-18.1 %	0.834	**0.786**	*-6.1 %*

classification accuracies on cross-validation slightly decrease or remain almost the same (level of significance for group classification: $\approx \pm 8.2\%$) but the effect of overfitting as it was observed for the original Ensemble of local experts can almost be eliminated when using the new approach. This improved generalizability is of major importance for safety pharmacology where unknown samples need to be classified reliably.

5 Conclusion

Generalizability of an automatic classification system is a major prerequisite for its application to real-word problems. In this paper we presented a new model combination approach that does not suffer from the problem of overfitting during optimization as observed for our previously developed Ensemble of local experts approach for Metabonomic applications. Motivated by the concept of generalized stacking outputs of local NMR experts are aggregated for final Ensemble estimation. Improved generalizability is achieved by parameter optimization using a cross-validation approach in a hierarchical classification framework. The decisions of local classifiers – level-0 models – serve as input for an additionally

subsequent level-1 generalizer. Generally, stacked generalization works best when integrating level-0 models that generate probabilistic outputs. Consequently, in our approach decisions of Support Vector Machines are transformed from binary towards pseudo-probabilistic by means of a sigmoidal fitting function.

The effectiveness of stacking for Ensembles of local experts for Metabonomic applications was demonstrated in an experimental evaluation. Analyzing a challenging real-world set of NMR spectra from industrial drug design shows improved generalizability of the classification performance when turning from cross-validation towards test. In the typical use-case of an automatic classification system in safety pharmacology toxicities need to be predicted reliably for unknown samples. Thus, reducing the effect of overfitting is of major importance which clearly emphasizes the practical relevance of the proposed approach.

Acknowledgments Parts of this work have been funded by a grant from Boehringer Ingelheim Pharma GmbH & Co. KG. (BI), Genomics group. We would like to thank the General Pharmacology Group of BI for providing the data.

References

1. Nicholson, J.K., et al.: Metabonomics: a platform for studying drug toxicity and gene function. Nature Reviews Drug Discovery **1** (2002) 153–161
2. 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** (2008) 41–49
3. Wolpert, D.H.: Stacked generalization. Neural Networks **5** (1992) 241–259
4. 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** (1998) 235–244
5. Fieno, T., Viswanathan, V., Tsoukalas, L.: Neural network methodology for ^{1}H NMR spectroscopy classification. Proc. Int. Conf. on Information Intelligence and Systems (1999) 80–85
6. Beckonert, O., et al.: NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. Analytica Chimica Acta **490** (2003) 3–15
7. Ebbels, T., et al.: Toxicity classification from metabonomic data using a density superposition approach: CLOUDS. Analytica Chimica Acta **490** (2003) 109–122
8. 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
9. Specht, D.F.: Probabilistic neural networks. Neural Networks **3** (1990) 109–118
10. 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: Multiple Classifier Systems. Number 4472 in LNCS (2007) 42–51
11. Lienemann, K., Plötz, T., Fink, G.A.: Automatic classification of NMR spectra by ensembles of local experts. In: Structural, Syntactic, and Statistical Pattern Recognition. Number 5342 in LNCS (2008) 790–800
12. Breiman, L.: Stacked regressions. Machine Learning **24** (1996) 49–64

13. Ting, K.M., Witten, I.H.: Issues in stacked generalization. Journal of Artificial Intelligence Research **10** (1999) 271–289
14. dos Santos, E.M., et al.: Overfitting in the selection of classifier ensembles: a comparative study between PSO and GA. In: Proceedings of the 10th annual conference on Genetic and evolutionary computation, ACM (2008) 1423–1424
15. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, MIT Press (1999) 61–74
16. Lin, H.T., Lin, C.J., Weng, R.: A note on platt's probabilistic outputs for support vector machines. Machine Learning **68** (2007) 267–276
17. Matthews, B.W.: Comparison of the predicted and observed secondary structure of the T4 phage lysozyme. Biochimica et Biophysica Acta **405** (1975) 442–451
18. Breiman, L.: Random forests. Machine Learning **45** (2001) 5–32
19. Wold, H.: Estimation and Prediciton. In: Estimation of Principal Components and Related Models by Iterative Least Squares. New York: Academic Press (1966) 391–420