Natalia Becker
Dr. sc. hum.

Classification and Feature Selection Using Penalised Support Vector Machines: Development of a Flexible Classifier with Applications to High-Dimensional Breast Cancer Data

Promotionsfach: Molekulare Genetik (DKFZ)
Doktorvater: Prof. Dr. Peter Lichter

Classification and variable selection plays an important role in knowledge discovery in high-dimensional data. Although Support Vector Machine (SVM) algorithms are among the most powerful classification and prediction methods with a wide range of scientific applications, SVM does not include automatic feature selection and therefore a number of feature selection procedures have been developed to extend SVM. Regularisation approaches extend SVM to a feature selection method in a flexible way using penalty functions like L1, smoothly clipped absolute deviation (SCAD) and Elastic Net.

In this thesis, a novel feature selection method for SVM classification using a combination of two penalties, SCAD and L2 was proposed. The new method, called Elastic SCAD, should overcome the limitations of each penalty alone. The commonly used penalty functions were investigated in parallel with the new method using simulated and publicly available data.

Since SVM models are extremely sensitive to the choice of tuning parameters, an interval search algorithm is implemented, which in comparison to a fixed grid search finds rapidly and more precisely optimal tuning parameters. All penalised SVM methods as well as algorithms for the search of optimal tuning parameters were implemented in the freely-available R package ’penalizedSVM’.

A sensitivity analysis was performed on simulated data to investigate whether results depend on the specific simulation design. Our simulation study revealed that ’elastic’ approaches,
the Elastic Net SVM and the Elastic SCAD SVM, showed a similar behaviour independently of the simulation design. These classifiers provided small prediction errors, provided sparse predictors and discovered relevant features more frequently than the irrelevant ones. Moreover, the Elastic SCAD SVM outperformed the Elastic Net SVM by providing more sparse classifiers. Both ’elastic’ methods were able to consider correlation structures in the input data (grouping effect).

The identification of tuning parameters is commonly done based on cross-validation. In this thesis, the effect of using generalised cross-validation (GACV) instead of k-fold cross-validation was investigated. The GACV method is shown to have larger misclassification errors and provides inflexible sparse solutions constantly for all levels of model complexity. In contrast, the five-fold cross-validation classifiers responded

sensitively to the number of relevant features in the data and selected features in a more flexible way. Based on the comparison of simulation scenarios, which use five-fold cross-validation and generalised cross-validation, the five-fold cross-validation is recommended.

Moreover, the influence of search algorithms for tuning parameters on the resulting classifiers was investigated. Classifiers, which were built using the fixed grid algorithm, were compared with classifiers resulting from using the interval search. The interval search proved to be efficient: it found optimal tuning parameters quickly, comparably accurate and was also able to work successfully on large intervals without degenerating into arbitrariness or getting slower.

For simulated sparse data, the feature selection methods selected less features with increasing sample size. The classifiers recovered the true positives more accurately, while the levels of false positives decreased. From the simulation study can be concluded that for sufficiently large sample sizes, feature selection methods with combined penalties were more robust in changing the model complexity than using single penalties alone. As expected, SCAD SVM and L1 SVM showed very good performance in terms of prediction accuracy for sparse models, but did not show improvements for less sparse models. Combined penalty functions Elastic Net and Elastic SCAD performed well for both sparse and less sparse models. Moreover, Elastic SCAD provided sparser classifiers, in terms of number of features selected, than Elastic Net SVM.

Finally, when applied the existing and the novel penalised SVM methods to two publicly available breast cancer data sets, Elastic SCAD classifiers had the maximal Youden index in both applications in comparison to other investigated SVM methods with and without feature selection. For both real data sets and in almost all scenarios of the simulation study the four feature selection classifiers outperformed ordinary Support Vector Classification with respect to predictive accuracy.

Overall one can conclude, that the Elastic SCAD SVM performed very flexible in sparse and non-sparse situations providing reasonable sensitivity and specificity. Whereas the standard MammaPrint(R) signature showed neither a small test misclassification rate nor a reliable sensitivity for the selected validation set of the NKI data set. Therefore, an improvement of the MammaPrint(R) classifier is needed.

In conclusion, the proposed Elastic SCAD SVM algorithm provides the advantages of the SCAD penalty and at the same time avoids sparsity limitations for non-sparse data. The results from the simulation study and real data applications render Elastic SCAD SVM with automatic feature selection a promising classification method for high-dimensional applications. Furthermore, the application of k-fold cross-validation and interval search in tuning SVM parameters are highly recommended.