Zusammenfassung

The problem of feature selection is an important issue in Machine Learning and of high practical relevance, e.g. in bioinformatics. Similar to a human being a machine can improve its generalization capability by filtering out the most relevant features (attributes) of the data to learn. There are different possibilities to define what "relevant" actually means. But the common goal is to find the combination of features which induces the lowest estimated generalization error. This is a difficult combinatorial problem. Therefore Genetic Algorithms have been proposed before (e.g. Ferri, Kadirkamanathan and Kittler (1993), Brill, Brown and Martin (1992)) to solve this problem. A combination of features is evaluated by estimating the generalization performance of the learning machine with regard to the selected feature subset. Usually this is done by means of cross-validation on the training data. Especially for Support Vector Machines, however, there exist theoretical bounds on the generalization error. This allows the design of Genetic Algorithms for feature selection using these bounds. The advantages are on one hand a reduction of overfitting, and on the other hand lower computational costs. This is confirmed by experiments on toy data and on DNA micro array data. Having a similar performance in comparison to the well known RFE algorithm, which is especially designed for SVMs as well, one can in fact save time by using Genetic Algorithms in combination with bounds on the generalization error, if the number of features to select is not known beforehand.