Chapter 8

Conclusions

I have provided a novel unified view on the seemingly diverse field of all-in-one multi-class SVMs. Albeit all popular all-in-one approaches reduce to the standard SVM for binary classification problems, they differ along three dimensions when applied to more than two classes. These are the presence or absence of a bias term in the classification functions, the use of a relative or absolute margin concept, and the way of combining margin violations in their loss functions.

The unified scheme pointed at a canonical combination of these features that had not been investigated. The missing machine, which can be viewed as marrying the approaches by Crammer & Singer (CS, [42]) and Lee, Lin, & Wahba (LLW, [112]), has been derived and evaluated. The new SVM named DGI SVM considers the maximum over the margin violations per variable in its loss function, and an absolute margin concept as proposed by LLW.

A fast training algorithm for WW, LLW and DGI SVMs is presented. By dropping the bias term—as done in the CS approach—the equality constraints in the dual problems for all machines have vanished. This makes decomposition methods easily applicable. A second order working set selection algorithm using working sets of size two for these problems has been proposed. Instead of choosing the smallest, irreducible working set size, to use a working set size of two whenever possible is proposed. This allows for a still tractable analytic solution of the sub-problem and as shown empirically this corresponds to a significantly better trade-off between iteration complexity (as, e.g., determined by the working set selection heuristic and the gradient update) and progress. That is, sequential two-dimensional optimization (S2DO) should be favored over the strict SMO heuristic. This is also supported by the findings in [149] for binary SVMs. The S2DO heuristic is not restricted to the SVMs considered in this study, but can be applied to machines involving quadratic programs without equality constraints in general.

The developed solver is applied to all multi-class SVM machines, and this made the empirical comparison fair enough to draw conclusions about the required training times. Another novel contribution of this thesis regarding SVM solvers is that for all-in-one multi-class machines a new caching technique, which needs only to
store a $O(d^2)$ matrix and a $O(\ell^2)$ matrix instead of $O(s^2)$ matrix where $s$ is $d \times \ell$, has been developed. This caching technique made possible to use WW and LLW in all data sets. According to my knowledge, S2DO is the only existing solver for LLW that is using decomposition algorithms. As a result, LLW method can now be used for much larger data sets.\footnote{The original solver proposed by the Lee et al. \cite{Lee2001} is based on interior point method and have complexity of $O(s^3)$ and a memory requirement of $O(s^2)$.}

An extensive empirical study has been accomplished. The new solver allows to apply better model selection procedures for all multi-class machines. This is, because of two reasons, a particularly important contribution of this study. First, until now the WW and LLW methods were often either ignored or not carefully analysed in empirical studies due to the lack of efficient solvers \cite{Li2001, Lee2001}. The second reason is that researchers did not make suitable model selection because of the computational requirement of these methods \cite{Lee2001, Zhang2003}. One-vs-all (OVA) and CS are considered the best machine for multi-class problems because of these reasons and also due to their high training speeds. However, the empirical analysis presented in this thesis showed that this common belief is at least not completely true. Empirical analysis revealed two important facts. First, LLW is better than all other methods in the sense of classification accuracy and the second best method is WW (see Section 7.2). The second insight is that WW is not slower than the CS method (see Section 7.2). Further, if one focuses only all-in-one class machines, the superior results supplied by LLW and WW implies that sum-loss machines are in general better than max-loss machines (i.e., CS and DGI).

The results of six multi-class methods on Bioinformatics data sets implied that the model selection very important when the data set at hand is small (i.e., we face a small sample problem, see Section 7.4.1 and Section 7.4.3). Finally, the results of the traffic sign recognition problem implied that neither using the good features nor using the good classifiers gives the best result. In order to solve real world problems we need to take into account both issues (see Section 7.3).

The extensive experimental comparison showed that the WW approach generated hypotheses with higher classification accuracy compared to the CS machine. Both approaches outperformed the one-versus-all method in this respect. Using S2DO, the original WW multi-class SVM now becomes at least as fast as the CS method trained with tailored, state-of-the-art second order working set selection. This indicates that the faster training times observed for the CS SVM compared to the WW formulation were not achieved by reducing the number of slack variables, but rather by dropping the bias term from the hypotheses (this is in accordance with the findings in \cite{van2005}, where training times increased drastically when adding bias parameters to the CS machine). The better generalization results are in accordance with newly derived risk bounds. These follow from a union bound on results for binary machines and are lower for the WW SVM compared to the CS machine. Given the empirical and theoretical results, there is no reason any more for a priori preferring the CS SVM to the original (WW) method. We hope that the results of...
this thesis makes the WW method more popular among practitioners, because it offers improved accuracy without additional costs in training time compared to CS.

From a theoretical point of view, the decisive property of the LLW multi-class SVM is the classification calibration of its loss function [154]. The efficient solver proposed in this thesis makes LLW training practical and thereby allowed for the first extensive empirical comparison of LLW with alternative multi-class SVMs. The LLW method is the only classification calibrated machine in this comparison [154] and showed the best generalization performance. This improved accuracy required considerably more training time. However, if training time does not matter, the LLW machine is the multi-class SVM of choice. This experimental result corroborates the theoretical advantages of the LLW machine.

In this study, I considered batch learning of multi-class SVMs. For binary classification, it has been shown that improved second-order working set selection derived for batch learning is even more advantageous when applied to on-line learning in LASVM [79]. Therefore, I am confident that the results in this study also carry over to the popular LaRank online multi-class SVM [19].