CHAPTER EIGHT

CONCLUSION

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1 Summary</td>
<td>130</td>
</tr>
<tr>
<td>8.2 Contribution</td>
<td>131</td>
</tr>
<tr>
<td>8.3 Unresolved issues</td>
<td>132</td>
</tr>
</tbody>
</table>

We conclude by giving a brief summary of the work presented in this thesis in Section 8.1. We highlight our specific contribution to the field in Section 8.2 and conclude by mentioning currently unresolved issues in Section 8.3.

8.1 SUMMARY

In Chapter 4, we presented theoretical and empirical arguments for choosing SVM hyperparameter values efficiently on large datasets. In particular, we showed that the appropriate value of $C$ is expected to scale inversely with $N$ and that suitable values for the kernel width $\gamma$ only depend weakly on $N$. Choosing $C$ on small datasets was also shown to be sensitive to underfitting in cases where subsets are selected from problems that have little overlap between classes.

An interesting relationship between good values of $\gamma$ and the dimensionality of the data $d$ was also identified, in that the optimal $\gamma$ correlates well with $\frac{1}{d}$ across several problems (see Table 4.7). This choice of $\gamma$ followed by a line search over $C$ was empirically shown to be stable, hence we concluded that an efficient algorithm for choosing the SVM hyperparameter
values is to set $\gamma = \frac{1}{d}$ and to do a corresponding line search for $C$. In Chapter 7, this algorithm was indeed shown to be much faster than a normal grid search, and in most cases competitive in terms of accuracy (this can be seen in Figs. 7.1 – 7.14 and Tables 7.4 – 7.16, see SVM-100%-mono).

Our analysis of the hyperparameters also focused our attention on the fact that large $C$ is in general good from an accuracy point of view; this is certainly true when using a linear kernel, and for an RBF kernel, there are indications that this is also the case, although the evidence is not as overwhelming. The theoretical implications were shown to be significant, as very large values of $C$ cause the misclassification term to dominate the margin term in the SVM error function; this in turn casts doubt on the LMC tag associated with SVMs and the underlying SRM motivation.

The relationship between large $C$ and high SVM accuracy on non-separable datasets was used to argue that error functions employing only a misclassification term should be competitive in accuracy, and conceivably more efficient to optimize. Competitive accuracy was indeed achieved when using the BFGS algorithm to optimize a PK function, and for very large datasets, employing Rprop also provided promising results, both in terms of accuracy and time. However, we have not succeeded in developing an algorithm that competes with SVMs across all the problems we have considered – hence, SVM optimization along with efficient parameter selection is currently our preferred approach to large classification problems.

### 8.2 CONTRIBUTION

Using insights obtained by investigating efficient heuristics for choosing SVM hyperparameters, an algorithm was proposed which allows one to train SVMs in a fraction of the time required to train an SVM using the entire dataset and with conventional grid searches. The algorithm proposes choosing $\gamma = \frac{1}{d}$ and searching for $C$ on a small subset (10%) of the data. This algorithm is shown to be competitive on almost all datasets we have considered, and reduces the computer time required to select appropriate SVM hyperparameter values by several orders of magnitude compared to a regular grid search.

An important theoretical insight was achieved when we understood that the often superior performance of SVMs may not be related to the underlying SRM theory. This stems from the fact that the misclassification term appears to be more important than the margin term in the error function. Indeed, empirical observations have shown that, especially when using linear kernels, the margin is often negligible when compared to the misclassification term.

Algorithms that exploit the insight that the margin term is redundant on non-separable
datasets were also proposed (PK), and promising results were obtained when using both Rprop and the three-stream algorithm to optimize the PK.

On the application side, some of the tools used in this thesis were applied to address real world problems successfully, such as the DFKI age classification problem [10] and tree species classification using hyper-spectral images[81].

8.3 UNRESOLVED ISSUES

Although satisfying progress has been made with the main aim of this research, namely to develop an efficient approach to the classification of large data sets, our work has raised a number of issues that we have not been able to resolve yet. The most interesting of these issues are:

• We have presented several types of evidence to suggest that the margin term is superfluous with regard to classification accuracy in SVMs; however, that term is structurally crucial for the development of the SMO algorithm, which solves the KKT equations with great efficiency. The question thus remains whether it is possible to define an approach that does not employ a margin term, but is competitive with SVMs in all cases. We believe that options for finding such an algorithm include, among others, refinement of SGD algorithms, or an SMO equivalent to optimize error functions without a margin term.

• A better theoretical understanding of the region within which a good value of $C$ can be found is also still an unsolved problem. We have shown that a large $C$ is in general good from an accuracy point of view, but choosing an excessively large value makes the optimization process very expensive computationally. It is hoped that an appropriate theoretical understanding would enable us to select the mid-ground where good accuracy is achieved with a limited computational budget.

• Our SGD method is currently closely attuned to SVM optimization – choices such as the scaling of the initial learning rate and stopping criteria have only been considered within this framework. Given the promising performance achieved with this algorithm, it would be worthwhile to investigate whether it can be adapted to do general stochastic optimization efficiently.

• While we have shown that it is beneficial to search for the optimal hyperparameter values on less data, understanding the amount of data required for choosing reasonable hyperparameter values is unresolved. This is an important unanswered question,
given the substantial improvements in speed we have achieved with somewhat arbitrary choices of the subset used for hyperparameter value selection. Factors such as the apparent sensitivity of $C$ to underfitting on subsets of the data where there is little overlap will have to factor into an answer to this question.

We believe that the approaches described in the current thesis will suffice for training sets of up to tens of thousands of samples; for even larger training sets, resolution of some of the above-mentioned issues is crucial.