Chapter 6

Conclusion and Future Work

In this thesis, we had a closer look at the problem of classification in context of prototype based vector quantization and the respective model learning. The motivation of this work was to consider problems around evaluation of classifiers beyond the simple accuracy and their incorporation into the model description. Thereby, one of the key aspects was the integration of expert knowledge to obtain problem specific models. In this view, classification can be regarded as an *ill-posed* problem as pointed out in Section 3.4.

The aspects dealt with in this thesis, can be subdivided into the following main topics:

1. misclassification costs beyond accuracy
2. integration of structural information about data into model learning exemplified by processing functional data
3. semi-supervised learning and integration of uncertainty in labeling

Of course, these topics are neither independent nor complete. Usually they interact and are additionally influenced by further aspects which, however, may be unknown for the user. In this sense this thesis is a contribution for those classification tasks, where explicit knowledge about the data, classification costs or structural information is available.

One of the most powerful and also intuitive prototype based classifiers is the LVQ as introduced by Kohonen. We restricted us to the cost variant Generalized Learning Vector Quantization. In particular we focused on GLV0Q to consider other evaluation statistical classification measures beside accuracy and integrated knowledge about classification weights or misclassification costs. One essential outcome of the thesis is the interpretation of the GLVQ cost function as an approximation of counting misclassifications: the border sensitive GLVQ (see Sec. 5.1). This perception allows us to modify the GLVQ such that the direct optimization of statistical evaluation measures based on the confusion matrix by keeping the principle of GLVQ becomes possible. This idea was demonstrated with the F_β-GLVQ, where the cost function approximated the F-measure (see Sec. 5.2.2). Another option is the modification of the GLVQ cost function to regard asymmetric misclassification costs like they often occur in medical environments (see Sec. 5.2.1).

Another topic of this thesis was the integration of structural information about the data into the relevance learning scheme of GLVQ. We exemplified these ideas for functional data leading to the functional relevance learning (GFR/MLVQ, see Sec. 4.1.1) and the enhanced GR/MLVQ (eGR/MLVQ, see Sec. 4.1.2). Both algorithms integrate the lateral dependencies of the neighboring dimensions of functional vector data into the learning scheme. In the GFR/MLVQ instead of learning each dimension independently, the relevance profile is composed by a linear combination of non-linear and smooth basis functions and the number of parameters to adapt was decreased significantly. Yet, in the experiments the GFR/MLVQ shows difficulties learning these parameters. In contrast to GFR/MLVQ, in eGR/MLVQ the number of free parameters was indirectly reduced by adding a neighborhood function for relevance learning. The eGR/MLVQ leads to more stable results compared to the GR/MLVQ and experiments show a speed up in learning. Further, the eGR/MLVQ can be applied also to other kinds of data with structural information about the dependencies in the data.

The third topic was the integration of label information in unsupervised vector quantization principles to obtain semi-supervised methods. Therefore, the distance measures in Neural Gas and Self Organizing Maps are extended to incorporate label information in a multiplicative manner (FSSOM/FSNG, see Sec. 4.2.1 and 4.2.2). The resulting FSNG/FSSOM can handle labeled as well as unlabeled data where the label can be either crisp or fuzzy. The obtained models including of prototypes with fuzzy labels provide a broad range of applications. The FSNG/FSSOM model can be extened to handle also uncertainty in data labeling as explained in Section 5.3.

Summarizing all topics, as mentioned above, the considered model extensions and modifications of vector quantizers are neither comprehensive nor covering all aspects of classification learning. Although many real classification tasks and applications of vector quantization can be tackled in this manner, there are still opportunities for further improvements and perspectives. For example, we could think about other evaluation measures like the AUROC-measure as optimization objective for GLVQ. Another possibility would be to identify outliers or doubtful data samples in classification and to sensitize the GLVQ model for those data.

Further, so far only binary classification evaluation measures were covered in this thesis, when alternative misclassification costs were considered. There also exist generalizations of statistical measures to evaluate models for multi-class problems. An extension of the GLVQ for such problems might be feasible and desired.

Although this thesis concentrates on the standard GLVQ, a transfer to Relational or Median variants can be considered.