Abstract

The ability to learn from experience is a fascinating and distinguishing feature of intelligent life. In the course of evolution, increasingly more sophisticated strategies of adapting behavior to particular surrounding conditions have emerged. Inspired by the efficiency of learning in nature, supervised machine learning considers a formal model of learning specific input-output relationships and adopts the paradigm of *learning from examples* to induce functions which predict target objects associated with input patterns. In conventional machine learning, the process of learning is characterized by a unidirectional information flow between the two idealized protagonists, i.e., information is submitted only from the teacher to the learner. In contrast to this, natural learning processes typically are based on complex interactions between the learner and the teacher. Bidirectional communication - from learners to the teaching entity and vice versa - is a fundamental component of learning. As motivated by nature, active machine learning substitutes the conventional passive model by an extended model which incorporates a restricted form of interactive learning in order to expedite the artificial learning process. We study the concept of active learning, which facilitates the learning process by reducing the amount of representative examples required, in the prospering field of kernel machines for various categories of learning problems.

A comprehensive corpus of flexible techniques for constructing prediction systems with excellent generalization capabilities has evolved as the constructive outcome from the inspiring synergy of statistical learning theory, optimization theory and more applied areas of research in machine learning in recent years. Among the variety of methods which have been developed, support vector machines are the most popular. With increasing computational power being available today, optimized training algorithms are able to cope with large-scale learning problems. However, advances in computational speed and more efficient training algorithms do not solve the inherent problem which consists in the fact that conventional supervised machine learning relies on a set of input patterns which have to be assigned to the corresponding target objects. In many areas of application, the task of assigning target objects cannot be accomplished in an automatic manner, but depends on time-consuming and expensive resources, such as complex experiments or human decisions. Hence, the underlying assumption that a set of *labeled* examples is submitted to the learning algorithm disregards the labeling effort that is necessary in many cases.

The superordinate concept of active learning refers to a collection of approaches which aim at reducing the labeling effort in supervised machine learning. We consider the pool-based active learning model, where the essential idea is to select promising *unlabeled* examples from a given finite set in a sequential process in the sense that

the corresponding target objects contribute to a more accurate prediction function. In contrast to conventional supervised learning, pool-based active learning considers an extended learning model in which the learning algorithm is granted access to a set of initially unlabeled examples and provided with the ability to determine the order of assigning target objects with the objective of attaining a high level of accuracy without requesting the complete set of corresponding target objects.

The present thesis pursues the general objectives of improving and generalizing existing approaches in pool-based active learning with kernel machines to a broader field of learning problems and of providing a thorough analysis of underlying theoretical aspects. More precisely, our main contributions can be summarized as follows:

- *Improvement of Efficiency:* In the context of binary classification learning, we propose a novel strategy for the selection of batches of multiple examples. As the fundamental component of our approach, we incorporate a measure of diversity to provide a more efficient strategy in terms of the number of labeled examples necessary to attain a particular level of classification accuracy. Moreover, we suggest modified multiclass selection strategies for different binary decomposition methods which yield substantial improvements over previous research.
- *Generalization:* Label ranking forms a category of preference learning problems which has not been investigated in active learning research previously, despite the fact that the labeling effort is an even more essential matter of relevance here than in classification learning. Employing the constraint classification and the pairwise ranking techniques, we propose generalizations of pool-based active learning and demonstrate a substantial improvement of the learning progress.
- *Theoretical Foundations:* We investigate a linear learning setting to analyze theoretical drawbacks of volume-based selection strategies and show that the minimization of the volume of the version space can be viewed as a *necessary* precondition for the minimization of the proposed improved selection criterion. Moreover, we prove a novel convergence theorem for the volume-based SIMPLE selection strategy in the case of the maximum radius ball approximation.

To this end, we present a general view of the concept of supervised machine learning and formalize the considered learning model in order to establish a common basis in terms of notation. We focus on the hypothesis class of linear classifiers which can be extended by the elegant and flexible concept of kernels. Among the variety of kernel classifiers, support vector machines and Bayes point machines are well-studied examples of linear learning algorithms which exhibit an appealing geometric interpretation in the so-called version space model. Moreover, linear classifiers serve as fundamental components in solving problems of more complex categories in supervised learning. In order to improve numerical stability and accuracy, we propose modifications to a kernel billiard algorithm for approximating the Bayes point.

We embark on a detailed presentation of the class of active learning selection strategies which aim at reducing the volume of the version space by means of approximating the center of mass and discuss the underlying theoretical line of reasoning. Moreover, we analyze a linear learning setting where volume-based strategies exhibit some potential shortcomings and propose an improved binary selection strategy to address

this problem. In order to reduce the computational complexity of the envisaged selection strategy, we present a sophisticated subsampling technique which preselects a subset of unlabeled examples according to a less demanding volume-based strategy as candidates for our novel selection strategy. From a theoretical point of view, the minimization of the volume of the version space is a *necessary* precondition for the minimization of the improved selection criterion. As we anticipated on account of our theoretical analysis, experimental results demonstrate that the two-layered subsampling approach substantially decreases the computational complexity without a concomitant loss of classification accuracy. Apart from practical issues, the considered setting sheds light on potential shortcomings of volume-based strategies from a theoretical perspective.

The basic course of action in pool-based active learning consists in the sequential process of selecting *individual* examples from a set of unlabeled examples and requesting the corresponding target objects. However, both from a computational point of view and, more significantly, with regard to common characteristics of learning problems in practice, generalizing this scheme such that *sets* of unlabeled examples are selected and submitted to the labeling component yields beneficial effects. We propose a generalized selection strategy which incorporates a measure of diversity in the active selection of batches of multiple examples in binary classification learning. Furthermore, we present experimental results indicating that this novel approach provides a more efficient method in terms of the number of labeled examples necessary to attain a particular level of classification accuracy than a commonly employed extension of a volume-based selection strategy.

While most research on active learning in the field of kernel machines has focused on binary problems, less attention has been paid to the problem of learning classifiers in the case of multiple classes. We consider three common decomposition methods for expressing multiclass problems in terms of sets of binary classification problems and propose novel active learning selection strategies in order to reduce the labeling effort. A variety of experiments conducted on real-world datasets demonstrates the merits of our approach in comparison to previous research.

The effort necessary to construct sets of labeled examples in a supervised learning scenario is often disregarded, though in many applications, it is a time-consuming and expensive procedure. While the process of labeling constitutes a major issue in classification learning, it becomes an even more substantial matter of relevance in label ranking learning, which considers the more complex target domain of total orders over a fixed set of alternatives. We introduce a novel generalization of pool-based active learning to reduce the labeling effort based on both the pairwise ranking and the constraint classification techniques for representing label ranking functions.

The final part of this thesis is devoted to a more thorough theoretical analysis of volume-based selection strategies in binary classification learning. We derive a convergence theorem for a common volume-based selection strategy in the case of the maximum radius ball approximation of the center of mass and present a survey of related results.