Preface

This dissertation develops, analyzes, and evaluates focusing solutions for data mining. Data mining is a particular phase in knowledge discovery that applies learning techniques to identify hidden information from data, whereas knowledge discovery is a complex, iterative, and interactive process which covers all activities before and after data mining. Focusing is a specific task in the data preparation phase of knowledge discovery. The motivation of focusing is the existence of huge databases and the limitation of data mining algorithms to smaller data sets. The purpose of focusing is data reduction before data mining, either in the number of tuples, the number of attributes, or the number of values. Then, data mining applies techniques to the reduced data and is still able to achieve appropriate results.

In this dissertation, we first analyze the knowledge discovery process in order to understand relations between knowledge discovery tasks and focusing. We characterize the focusing context which consists of a data mining goal, data characteristics, and a data mining algorithm. We emphasize classification goals, top down induction of decision trees, and nearest neighbor classifiers. Thereafter, we define focusing tasks which include evaluation criteria for focusing success. At the end of the first block, we restrict our attention to focusing tasks for the reduction of the number of tuples.

We start the development of focusing solutions with an analysis of state-of-the-art approaches. We define a unifying framework that builds on three basic techniques: Sampling, clustering, and prototyping. We describe instantiations of this framework and examine their advantages and disadvantages. We follow up the unifying framework and establish an enhanced unified approach to focusing solutions which covers two preparation steps, sorting and stratification, and the application of sampling techniques. We reuse random sampling and systematic sampling from statistics and propose two more intelligent sampling techniques, leader sampling and similarity-driven sampling. We implement the unified approach as a generic sampling algorithm and integrate this algorithm into a commercial data mining system.

Thereafter, we analyze and evaluate specific focusing solutions in different domains. We exemplify an average case analysis to estimate expected average

classification accuracies of nearest neighbor classifiers in combination with simple random sampling. We further conduct an experimental study and consolidate its results as focusing advice which provides heuristics for appropriate selections of best suited focusing solutions. At the end, we summarize the main contributions of this dissertation, describe more related work, raise issues for future work, and state some final remarks.