CHAPTER 1. INTRODUCTION

Modern storage and processor technology makes the accumulation of data increasingly easy. Prediction and data analysis tools must be developed and adapted rapidly to keep up with the dramatic growth in data volume. Data mining pursues the goal of extracting information from large databases with consideration for the storage structure. Some data mining techniques have been developed recently with the exploration of large databases in mind, such as Association Rule Mining, while others are adaptations of traditional methods that have long been used in statistics and machine learning, such as classification and clustering. Classification is the process of predicting membership of a data point in one of a finite number of classes. The attribute that indicates the class membership is called the class label attribute. Clustering, in contrast, has the goal of identifying classes in data without a predefined class label.

This thesis focuses on classification and clustering techniques that can be formulated in terms of kernel-density estimates, i.e., algorithms in which the density of all or a subset of data points is calculated from the data set through convolution with a kernel function. Many statistics and machine learning techniques can be viewed in this fashion. The algorithms in this thesis were developed to suit the concept of P-trees [1], which can be seen as representing the limiting case of a database model called vertical partitioning. Most database systems are organized in record-oriented data structures, referred to as horizontal partitioning. Vertical partitioning, or the organization of tables by column, however, is recognized by some researchers as a solution to serious problems such as input/output [2]. P-trees take the vertical partitioning model to the limit by not only breaking up relations into their attributes, but also further decomposing attributes into

individual bits. The resulting bit-vectors are compressed in a hierarchical fashion that allows fast Boolean operations. The optimization of Boolean operations on P-trees is essential for the purpose of most data mining tasks that require knowledge of the number of data points with attributes of a particular value or within a particular range.

1.1. Organization of the Thesis

This thesis is organized according to the format that is recommended when scholarly journal manuscripts are included. Several current problems in data mining are addressed. All algorithms show a scaling with respect to the number of data points that is better than linear due to the use of the P-tree data structure. The algorithms are designed to use this data structure efficiently. In contrast to many traditional algorithms, the techniques in this thesis can, furthermore, successfully be used for problems with a large number of attributes that are common in many current data mining settings.

Chapter 2 motivates the bit-column-based data organization that is the basis for the P-tree data structure and analyses the theoretical implication for the representation of attributes of various domains. Sections 2.3-2.5 draw on a multi-authored paper that is not yet completed, but available in draft form [3].

Chapter 3 describes the P-tree data structure in detail and introduces a new sorting scheme, generalized Peano-order sorting, that significantly improves storage efficiency and execution speed on data that show no natural continuity. The logical structure of P-trees is described, and representation choices are highlighted with focus on an efficient array-converted tree-based representation that was developed specifically for this thesis. Appendix A formalizes the development of this representation. A full implementation of

P-tree construction and ANDing in Java was completed as part of the thesis and was used for the data mining applications of the remaining chapters. Chapter 3 describes an application programming interface (API) that was developed in conjunction with other group members to allow reuse of components developed for this thesis and other research work. All new P-tree code development is being based on a revised C++ version of the API.

Chapter 4 introduces kernel functions and develops the mathematical framework for the following two chapters. Generalizing previous work on kernel-based classifiers using P-trees [4], this chapter presents an implementation of the kernel concept that can be seen as the basis for modified approaches in the following chapters. Chapter 4 also includes some remarks to assist in the transition between later chapters.

Chapter 5 represents a paper that can be seen within the kernel framework, but that also has roots in decision tree-induction and rule-based techniques. It makes use of information gain as well as statistical significance to determine relevant attributes sequentially using a step-shaped kernel function. The fundamental relationship between information gain and significance is demonstrated, where the standard formulation of information gain [5] emerges as an approximation of an exact quantity that is derived in Appendix B.

Chapter 6 represents a paper that introduces a Semi-Naive Bayes classifier using a kernel-based representation. Naive Bayes classifiers in which single-attribute probabilities are evaluated based on one-dimensional kernel estimates are well known in statistics [6]. This technique has not, however, previously been recognized by researchers who generalize the Naive Bayesian approach to a Semi-Naive Bayes classifier by combining

attributes [7,8]. We demonstrate that the resulting lazy algorithm shows significant improvements in accuracy compared with the Naive Bayes classifier and can be efficiently implemented using P-trees and the HOBbit distance function.

Chapter 7 represents a multi-author paper that has been published [9]. This paper demonstrates that kernel-density-based clustering [10] can be motivated as an approximation to two of the best-known clustering algorithms, k-means and k-medoids [11]. The paper integrates hierarchical clustering ideas. Dr. William Perrizo guided the work; the performance was compared with results that Qiang Ding achieved for his implementation of a k-means algorithm, and Qin Ding helped with writing and formatting. Chapter 8 concludes the thesis.

1.2. References

[1] Q. Ding, M. Khan, A. Roy, and W. Perrizo, "P-tree Algebra," ACM Symposium on Applied Computing (SAC'02), Madrid, Spain, 2002.

[2] M. Winslett, "David DeWitt Speaks Out," SIGMOD Record, Vol. 31, No. 2, pp. 50-62, 2002.

[3] A. Denton, Q. Ding, W. Jockheck, Q. Ding, and W. Perrizo, "Partitioning - A uniform model for data mining," draft available at

http://www.cs.ndsu.nodak.edu/~datasurg/papers/foundation_correct_version.pdf

created and accessed Nov. 18, 2002.

[4] W. Perrizo, Q. Ding, A. Denton, K. Scott, Q. Ding, and M. Khan, "PINE - Podium Incremental Neighbor Evaluator for Spatial Data using P-trees," Symposium on Applied Computing (SAC'03), Melbourne, Florida, 2003.

[5] C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, Vol. 27, pp. 379-423 and 623-656, 1948.

[6] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, New York, 2001.

[7] I. Kononenko, "Semi-Naive Bayesian Classifier," in Proceedings of the Sixth European Working Session on Learning, Berlin, 206-219, 1991.

[8] M. Pazzani, "Constructive Induction of Cartesian Product Attributes," Information, Statistics, and Induction in Science, Melbourne, Australia, 1996.

[9] A. Denton, Q. Ding, W. Perrizo, and Q. Ding, "Efficient Hierarchical Clustering of Large Data Sets Using P-trees," 15th International Conference on Computer Applications in Industry and Engineering (CAINE'02), San Diego, California, 2002.

[10] A. Hinneburg and D. A. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," in Proceedings 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), New York, 58-65, Aug. 1998.

[11] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," John Wiley & Sons, New York, 1990.