CHAPTER 8

CONCLUSIONS AND OUTLOOK

8.1. Conclusions

This thesis has presented several aspects of data mining research using P-trees. Chapter 2 motivated the use of bit-column-based data structures, in general, and discussed some of the theoretical underpinnings. Chapter 3 more specifically described P-trees, which use hierarchical compression on bit-columns. A new, generalized ordering scheme was introduced that improves compression for data that do not have the continuity properties of spatial data for which P-trees were originally developed. Implementation choices were discussed with the goal of motivating the representation that was developed for this thesis, which allows particularly fast ANDing. An application programming interface that was developed to allow reuse of P-tree-related software components written for this thesis and other research work was introduced. All programs described in this thesis used the P-tree implementation and a previous version of the P-tree API.

Kernel-density estimates are the unifying concept for the data-mining techniques in this thesis. Chapter 4 introduced kernel functions and put different algorithms into context. Chapters 5 through 7 represented papers that either have been published or are intended for publication. Chapters 5 and 6 introduced two fundamentally different classification algorithms both of which are new, not only in their use of P-trees, but also in their kernel-based treatment of continuous attributes. In the case of the Semi-naive Bayesian algorithm of Chapter 6, the kernel-based representation made an algorithm successful that had previously been discarded as unsuccessful by researchers who treated continuous attributes

through intervalization. The kernel-based viewpoint leads to an elegant and general mathematical description. Whereas previous researchers had generalized the Naive Bayes classifier by joining up to two categorical attributes, our formalism and implementation allow joining arbitrarily many attributes that can be continuous or categorical.

In the rule-based algorithm of Chapter 5, the kernel viewpoint led to a neighborhood-based treatment of continuous attributes that differs significantly from the standard intervalization in decision tree induction. It resulted in an algorithm that is competitive with existing decision tree induction programs despite its relative simplicity compared with standard decision tree induction algorithms that have been developed for a long time. Two further ideas were pursued in this paper, one being that "information gain," as it is commonly used in decision tree induction, should really be seen as an approximation to an exact expression. This exact expression follows both from the analogy with physical entropy that Shannon [1] noted as well as from a probability argument that is given in Appendix B. The practical success could, unfortunately, not be demonstrated which may either be due to the requirements of our specific algorithm or may indicate a problem with the concept of using entropy as a criterion. A second idea within the rule-based approach was to generate multiple rules by starting the rule-generation process multiple times with a different starting attribute each time and combining results. This strategy led to a further increase in prediction accuracy.

Chapter 7, finally, motivated a kernel-density-based clustering approach from its relationship to a well-known partitioning algorithm. Combination with hierarchical clustering ideas and a P-tree implementation led to a versatile and competitive clustering technique.

8.2. Outlook

The work documented in this thesis has drawn on a wide range of inter-related classification and clustering techniques to produce new algorithms and efficient P-tree-based implementations. Data sets from traditional machine learning domains were used as well as data mining relevant data sets, in particular from genomics and remotely sense imaging domains. Problems that are typical in data mining tasks, such as dealing with a large number of attributes, were addressed. Biological data with many unknown values and multi-valued attributes were also used ("gene-function" data set).

Independent work [2] highlighted the need to deal with several additional properties of data that are common in data-mining problems. Examples are a hierarchical structure of categorical attributes, a very small probability of the minority class label, and data with a graph structure. A graph structure arises when relationships exist between individual data points, such as protein-protein interactions for genomic data or links between web pages for web data. Whereas graph structure itself is the subject of much research, little is known about modifications to classification, clustering, and association rule mining that have to be made to account for a graph.

Work on this thesis and [1] have shown a need for visualization. Users of classification programs commonly request that predictions be interpretable. The classification algorithms described in this thesis have the potential of providing such information in the form of a visualization of those data points that contribute to a decision. The need for visualization becomes even clearer when unconventional data are concerned. Graph visualization formed an essential part of the successful completion of [2].

Visualization of the node information within the graph structure of protein-protein interactions strongly directed our investigation.

In summary, it can be said that this thesis forms a solid foundation of P-tree-based classification and clustering techniques to move on to unconventional and exciting problems.

8.3. References

[1] C. Shannon, "A Mathematical Theory of Communication," Bell Systems Technical Journal, Vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[2] A. Perera, A. Denton, P. Kotala, W. Jockheck, W. Valdivia Granda, and W. Perrizo, "P-tree Classification of Yeast Gene Deletion Data," SIGKDD Explorations, Vol. 4, No 2, pp. 108-109, Dec. 2002.