

Preface

Graphs are mathematical constructs for representing objects or systems which contain structural (*i.e.* relationship) information. Graphs have been used in many domains, from software engineering to artificial intelligence. However, the use of graphs in machine learning has been limited, compared to the more prevalent vector model which does not capture structural information. This can be attributed to the fact that until recently we have not had suitable graph-theoretic tools for dealing with graphs, and that comparing graphs for the purpose of evaluating structural matching is often of high time complexity. For these reasons, most machine learning approaches that deal with graph-based data introduce either restrictions on the type or related attributes of the graphs used, or they require a totally new mathematical foundation for handling graphs (such as probability theory). An unfortunate drawback of these types of methods is that existing machine learning algorithms either cannot be applied or require extensive modifications.

In this book, we present methods for utilizing graphs with well-known machine learning algorithms, such as k -means clustering and k -nearest neighbors classification, with virtually no significant modifications; the extensions for allowing these algorithms to use graphs is direct and intuitive. Further, we show how we can achieve polynomial time complexity with the restriction that the graphs contain unique node labels. This is the only restriction we place on the graphs, and we can even discard this limitation if we use a sub-optimal approximation approach for determining graph distance or if the graphs are relatively small.

To demonstrate the effectiveness of these approaches for performing clustering and classification with graphs, we apply them to the domain of web content mining. We show how web document content can be mod-

eled with different graph representations, and how these graphs contain additional information not found in traditional vector representations of document content. We perform experiments comparing the clustering and classification performance of the graph-based approach to that of the vector approach. Other interesting results show how we can visualize the graph space through multidimensional scaling, and how we can create a multiple classifier system that contains both statistical (vector) and structural (graph) classifiers. We also present the details of a web search clustering system that we developed and show how the system was easily modified to use graph-based representations instead of vector representations; the graph representations have a noticeable benefit in that they allow the terms in the cluster labels to have the correct ordering, and also provide differentiation between isolated terms and multi-term phrases.

Acknowledgements

This work was supported in part by the National Institute for Systems Test and Productivity at the University of South Florida under the U.S. Space and Naval Warfare Systems Command Contract No. N00039-02-C-3244.

*A. Schenker
H. Bunke
M. Last
A. Kandel*