On the Discovery of Patterns in Medical Data

Jorge C.G. Ramirez[1,2], Lynn L. Peterson[1], Dolores M. Peterson[2], and Gretchen K. Cormier[3]

[1] Dept of Computer Science & Engineering, Univ of Texas at Arlington, PO Box 19015, Arlington, TX 76019-0015
{ramirez, peterson)@cse.uta.edu
[2] Dept of Internal Medicine, Univ of Texas Southwestern Medical Ctr, 5323 Harry Hines Blvd, Dallas, TX 75235-9103
{jramir, dpeter}@mednet.swmed.edu
[3] Computer Consulting Specialists, PO Box 117, Dripping Springs, TX 78620-0117
gkateri@aol.com

There has been a recent proliferation in the literature on knowledge discovery in databases (KDD) and data mining. Most data mining techniques require that the data be in some standard form. However, many databases, especially medical databases, have features that make them different from the data collections used with most data mining methods.

Specifically, medical data can be any combination of binary, numeric, symbolic, text and/or image data. In addition, the data is temporal, with different significance to the temporal aspect, depending on the specific data (e.g., a specific blood test result vs. a diagnosis event). Furthermore, the data fields themselves typically are not the same at each collection point (i.e., different doctors may have different sets of tests run on different patients, despite the current set of diagnoses being the same). The purpose of this in progress research is to develop a methodology for the discovery of patterns in medical data that span the course of disease.

We have first looked at the KDD field for techniques that might be useful in our domain. As a result of our investigation to date, we have concluded first that KDD is pervaded by human intervention and verification of hypotheses, as opposed to discovery. Second, data mining techniques require that data be in standard form, whether it be in the form of a training set or preprocessed to meet the techniques' input requirements. Finally, our perspective of spanning the course of disease does not well fit with any of the "traditional" data mining techniques. In view of these conclusions, it seems that the temporal issue is the dominant factor in our problem domain and therefore, we must attempt to adapt or create new techniques to accomplish our goal.

(Srikant & Agrawal 1996) outline a methodology for discovery of generalized sequential patterns in a database of sequences. While strong for our domain in some respects, the sequences involved are transactions that are made up of item sets. The presence of an item in the set or lack thereof is how matches are made between sequences of transactions. (Mannila & Toivonen 1996) outline a methodology for discovery of frequent generalized episodes from a sequence of events, where each event has attributes associated with it. It even goes as far as considering events in parallel. However, the appropriateness for our domain is unclear, since both approaches require the presence of items or events for sequence matches. The fact that some events (e.g. test results) are present or absent at a given point in time is not necessarily enough to eliminate it from consideration as a match.

Although Srikant & Agrawal do address the issues of time constraints on adjacent elements in the sequence, it is a uniform constraint across the sequence. Further, it does not address the issues associated with the specific temporal significance of any given event.

Based on this foundation, a good description of our domain is sequences of event sets. This would incorporate the use of the Srikant & Agrawal approach on the sequence level and the use of the Mannila & Toivonen approach on the event set level. However, in order to match patterns in the disease data, development will need to focus on how to find matches between event sets, such that exact attribute matches are not required, the presence or absence of an attribute does not prevent a match, and the temporal significance of an event is not overlooked.

Acknowledgments

Jorge Ramirez is supported in part by NSF grant GER-9355110.

References

Mannila, H. and Toivonen, H. 1996. Discovering generalized episodes using minimal occurrences. Proc Second Intnl Conf on Knowledge Discovery in Databases, 146-150.

Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. Proc Fifth Intnl Conf on Extending Database Technology.