Multi-Relational Data Mining of Time-Oriented Biomedical Databases

Rashmi Raj, Martin J. O'Connor, and Amar K. Das
Stanford Medical Informatics, Stanford University
{rashmisu, das}@stanford.edu

Background

Multi Relational Data Mining (MRDM) extends association rule mining to search for interesting patterns among data in multiple input tables (relations) rather than in one input table. Researchers have successfully applied MRDM in bioinformatics [1], but MRDM is limited in handling time-course and longitudinal data, which are commonly found in biomedical databases. MRDM cannot search for patterns involving the comparison or classification of temporal data, which are needed to study causal or dynamic phenomena.

Methods

To address these issues, we have developed a new MRDM method called ChronoMiner. The underlying algorithm uses as input a hierarchical view of relations, instead of the set view used in standard MRDM. In the hierarchical view, attributes of relations are related through "parent" and "child" relationship. ChronoMiner searches for interesting multi-relational patterns by partial or complete traversal of the virtual tree structure of the database relations. The hierarchical search facilitates the coupling and decoupling of new attributes for temporal pattern discovery. The mining of interesting patterns starts from the root and proceeds using top-down induction, allowing for comparison along the time dimension at every level of abstraction. We evaluated the algorithm by applying it to Stanford HIV Database (hivdb.stanford.edu) to mine associations between newly arising mutations in the HIV genome and past drug regimens containing protease inhibitors (PI).

Results

The database contained 4271 subjects who had a regimen containing protease inhibitors. In searching for new mutations that arose after the administration of a drug or drug category, ChronoMiner confirmed previously known associations. At the drug category level, for PI, it found 63P, 36I, 41K, 93L, 35D as the most frequent mutation occurrences. Traversing one level deeper, at each drug level, it could verify 41L, 67N, 70R, 210W, 215Y as the most frequent mutations for the drug AZT. We also found mutations, such as 122E for AZT, which our domain expert (Dr. Bob Shafer) viewed as novel and clinically meaningful.

Conclusions

Our research extends MRDM to include temporal comparisons and hierarchies in searching for patterns of interest in a biomedical database. The initial evaluation of the ChronoMiner algorithm provides promising results for the HIV drug resistance research domain. After further testing, we plan to extend this work to the discovery of novel time-oriented patterns in other biomedical genomics databases.

Reference

[1] Page D, Craven M. Biological applications of Multi-Relational Data Mining. *ACM SIGKDD Explorations Newsletter*, 5(1):69-79, 2003.