

Analyzing multi-dimensional fuzzy scientific data through flexible discretization, association, and density-based clustering

Presented by Hongyuan Li

As high throughput technologies continue to produce large amounts of scientific data, the ability to extract useful, practically applicable information from these data becomes an increasingly important facet of scientific research. However, scientific data are usually multi-dimensional, fuzzy, and irregular in structure - making data mining very challenging. Herein, a level-wise association and density-based pattern discovery framework was designed. The structure of the framework is simple: besides a pattern frequency parameter, minimal support, only one user defined interface, IAdjacency, needs to be implemented to incorporate domain-specific knowledge. The effectiveness of this framework was demonstrated through the example of protein structural motif mining. MotifMiner, the protein structure data mining tool set based on this framework, successfully derived thousands of protein structural motifs organized by number of atoms, which includes known structures such as alpha-helices, beta-sheets as well as uncharacterized substructures. The structure motifs derived were used to fingerprint protein molecules and to establish a fast 3D-search. Results showed that a thorough search of 6513 non-redundant protein structures by protein molecule fingerprints can be accomplished within 1 second. Recursive application of this framework on the protein motif fingerprints resulted in clustering of thousand of protein molecules into structurally and functionally related groups in minutes. This framework has built-in fuzzy data handling features and allows flexible search. Results showed that MotifMiner can easily detected twisted alpha-helical and beta-sheeted substructures as well as molecule sidechain interaction patterns and active sites. Flexible search enables thorough search of motifs (partial molecule structure such as active sites) or molecule of interested (particularly small molecules such as potential drugs) with user specified similarity, which could be invaluable to pharmaceutical discovery. In addition to algorithm design, pattern search, pattern interestingness and pattern characteristics were also tested using protein structural motif mining as example. Rather than trying to define universal parameters for simplified data analysis, this framework allows easy incorporation of domain-specific knowledge to handle multi-dimensional fuzzy scientific data and customizable filtering and abstraction to extract different aspects/levels of information. Ultimately, this framework may be applied to any multi-dimensional scientific data analysis such as microarray data analysis to establish knowledge bases for biological scientists searching for specialized information.