A Data-Mining Approach Applied to Life Sciences

Marcus-Kalish, M.[1], Bonne-Tamir, B.S.[2], Kalid, O.[3] and Freeman, A.[3]
[1] Interdisciplinary Center for Technology Analysis and Forecasting, Tel Aviv University,
[2] Medical School, Tel Aviv University,
[3] Department of Biotechnology, Tel Aviv University

A data-mining tool for analysing data and issuing predictions will be presented.

The motive behind WizWhy, the data-mining algorithm, developed with Abraham Meidan, was the need to reveal, especially in life sciences, the underlying rules behind specific phenomena.

Unlike other available tools for data mining or tools for prediction (such as neural networks, decision trees or genetic algorithms), the aim of this algorithm is to reveal ALL inter-variable relationships in order to construct a theorem and unravel the rules behind the inspected phenomena.

The association rules approach, used in this algorithm, is the only one which is committed to revealing all the if-then rules (that meet predefined thresholds), in regard to the rule's confidence level (i.e., probability) and support level (i.e., number of cases).

Further more, the if-then rules are used, in our case, for revealing a set of If-and-only-if rules, as well as rules having several conditions that are unexpected relative to simpler rules.

The presented data-mining algorithm is proven to reveal –

1. All the IF-THEN rules that meet user-predefined thresholds.

2. All the IF-THEN-NOT rules (i.e., if the condition holds the result does not hold)

3. A set of optimal IF-AND-ONLY-IF rules (i.e., necessary and sufficient conditions)

Accordingly, it calculates the confidence level (i.e. error probability or significant level) for each rule. The algorithm analyses the UNEXPECTED RULES, presenting interesting and rare cases, which might be very important in some of the Biology applications.

The algorithm has been applied successfully, to different areas in life sciences. We shall present two highly diverse applications:

- The analysis of anthropometrical and blood screening data for ethnic group classification (joint work with Bat-Sheva Bonne-Tamir)
- The analysis of small-molecule binding-profiles for applications in nanobiotechnology and biochemistry (joint work with Ori Kalid and Amihay Freeman). The data-mining algorithm was used to define the necessary and sufficient conditions for a cavity on the molecular surface to bind a small molecule. This application was used as part of a screening system that selects protein building blocks for nanostructures with pre-designed geometry.

e-mail: miriam@post.tau.ac.il