Mining Gene Expression Data using Rough Set Theory

S. Bulashevska, W. Dubitzky, R. Eils*
Division “Intelligent Bioinformatics Systems” (H0900), German Cancer Research Center, 69120 Heidelberg, Germany
*corresponding author: r.eils@dkfz-heidelberg.de

Background: Classification of human tumors into distinguishable entities is traditionally based on clinical, pathohistological, immunohistochemical and cytogenetic data. This classification technique provides classes containing tumors that show similarities but differ strongly in important aspects, e.g. clinical course, treatment response, or survival. New techniques like cDNA microarrays have opened the way to a more accurate stratification of patients with respect to treatment response or survival prognosis, however, reports of correlation between clinical parameters and patient specific gene expression patterns have been extremely rare. One of the reasons is that the adaptation of machine learning approaches to pattern classification, rule induction and detection of internal dependencies within large scale gene expression data is still a formidable challenge for the computer science community.

Method: Here we applied a technique based on rough set theory and Boolean reasoning [1,2] implemented into the Rosetta software tool. This technique has already been successfully used to extract descriptive and minimal ‘if-then’ rules for relating prognostic or diagnostic parameters with particular conditions. The basis of rough set theory is the indiscernibility relation describing the fact that some objects of the universe are not discerned in view of the information forming a concept. Rough set theory deals with the approximation of such sets of objects – the lower and upper approximations. The lower approximation consists of objects, which surely belong to the concept and the upper approximation contains objects which possibly belong to the concept. The difference between the upper and lower approximations - boundary region – consists of objects which cannot be properly classified by employing the available information.

In this study, objects are the AML/ALL cases [3]. The data was presented as a table with columns (attributes) presenting genes and expression data as attribute values. The goal was to discover the attributes – genes – which allow to discern between the two classes acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), while the cases within each class must not be discerned. The Boolean function supporting this discernibility notion was calculated. Thereafter, a set of decision rules was derived relating attribute value combinations with AML or ALL classes. The quality of the rules was defined by classification accuracy and completeness, and was estimated by an algorithm following the concept of Michalski [4] to compute a single value for rule quality.

Result and conlusions: With the above described rough set theory based approach we obtained 1140 rules which were filtered with respect to their quality. 33 rules describing ALL-cases and 19 rules for ALL remained after filtering [5]. In the final paper we will present the most informative rules. Furthermore, we will assess the quality of the rules with respect to their ability to predict therapy response for AML/ALL patients.

References:

1. Z.Pawlak, Rough Sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991

2. Ed. L.Polkowsky, Rough sets and current trends in computing, Proc. RSCTC '98, Warsaw, 1998

3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Science 286(5439): 531-537, 1999.

4. I.Bruha, Quality of Decision Rules: Definitions and Classifications, in Machine Learning and Statistics, ed. G.Nakhaeizadeh, C.C.Tailor, 1999

5. T.Agotnes, J.Komorowski, A.Ohrn, Finding high performance subsets of induced rule sets: Extended summary, in Proc. Seventh European Congress on Intelligent Techniques and Soft Computing (EUFIT'99), Aachen, ed. H.-J.Zimmermann, K.Lieven,99

We thank Prof. Jan Komorowski for his inspiration to use rough set theory.