Randomization Methods for Assessing the Significance of Data Mining Results

Heikki Mannila[1,2]

[1] Helsinki Institute for Information Technology HIIT
[2] University of Helsinki and Helsinki University of Technology
Heikki.Mannila@tkk.fi

Abstract. Data mining research has developed many algorithms for various analysis tasks on large and complex datasets. However, assessing the significance of data mining results has received less attention. Analytical methods are rarely available, and hence one has to use computationally intensive methods. Randomization approaches based on null models provide, at least in principle, a general approach that can be used to obtain empirical p-values for various types of data mining approaches. I review some of the recent work in this area, outlining some of the open questions and problems.

J. Rauch et al. (Eds.): ISMIS 2009, LNAI 5722, p. 1, 2009.