Search for top-k consensus objects in multiple ranked lists: TopKInference versus other recent procedures

Vendula Švendová[1], Michael G. Schimek[1,2]
[1]Medical University of Graz, Graz, Austria
[2]Corresponding author's email: michael.schimek@medunigraz.at

Abstract

In recent years an increasing interest in the statistics of ranked lists can be seen, primarily stimulated by new biotechnologies as well as novel Web tools. Typically, such lists comprise tens of thousands of objects (e.g. genes in highthroughput analysis or URLs in Web search engines). However, only a comparably small subset of k top-ranked objects is informative. These objects are characterized by a strong overlap of their rank positions when they are assessed several times. The statistical task is to identify these top-k elements. Until recently there has not been a formal approach to this challenging task. Because of the high dimensionality of the data sets of interest and the associated severe multiplicity problem when classical inference procedures are adopted, it is hard to derive a methodologically convincing solution. Hall and Schimek (2012, JASA, 107, p. 661-672) have suggested a moderate deviation-based inference procedure implemented in the `R` procedure `TopKInference`. Because of the high practical relevance of the top-k ranked lists problem, recently also other approaches have been suggested, usually in the domain of genomics. We compare Hall and Schimek (2012) in detail with the most comprehensive method of Plaisier et al. (2010, Nucl. Acids Res., 38, 17, e169).

Keywords: Heatmap, hypergeometric distribution, inference, moderate deviation, rank order, top-k ranked list.