

Empirical Bayes Methods for Massive Transaction Data Sets

William DuMouchel (AT & T Research)
dumouchel@research.att.com
Daryl Pregibon

Abstract

This paper considers the framework of the so-called “market basket problem”, in which a database of transactions is mined for the occurrence of unusually frequent item sets. In our case, “unusually frequent” involves estimates of the frequency of each item set divided by a baseline frequency computed as if items occurred independently. The focus is on obtaining reliable estimates of this measure of interestingness for all item sets, even item sets with relatively low frequencies. For example, in a medical database of patient histories, unusual item sets including the item “patient death” (or other serious adverse event) might hopefully be flagged with as few as 5 or 10 occurrences of the item set, it being unacceptable to require that item sets occur in as many as 0.1 percent of millions of patient reports before the data mining algorithm detects a signal. Similar considerations apply in fraud detection applications. Thus we abandon the requirement that interesting item sets must contain a relatively large fixed minimal support, and adopt a criterion based on the results of fitting an empirical Bayes model to the item set counts. The model allows us to define a 95 percent Bayesian lower confidence limit for the “interestingness” measure of every item set, whereupon the item sets can be ranked according to their empirical Bayes confidence limits. For item sets of size $J > 2$, we also distinguish between multi-item associations that can be explained by the observed $J(J - 1)/2$ pairwise associations, and item sets that are significantly more frequent than their pairwise associations would suggest. Such item sets can uncover complex or synergistic mechanisms generating multi-item associations. This methodology has been applied within the U.S. Food and Drug Administration (FDA) to databases of adverse drug reaction reports and within AT and T to customer international calling histories. We also present graphical techniques for exploring and understanding the modeling results.