A Randomness Based Analysis on the Data Size Needed for Removing Deceptive Rules

Kazuya Haraguchi, Mutsunori Yagiura, Endre Boros, and Toshihide Ibaraki

We consider a data set in which each example is an *n*-dimensional Boolean vector labeled as true or false. A pattern is a co-occurrence of a particular value combination of a given subset of the variables. If a pattern appears frequently in the true examples and infrequently in the false examples, we consider it as a good rule. In this paper, we discuss the problem of determining the data size needed for removing "deceptive" good rules; in a data set of a small size, many good rules may appear superficially, simply by chance, independently of the underlying structure. Our hypothesis is that, in order to remove such deceptive good rules, the data set should contain more number of examples than that at which a random data set contains few good rules. We justify this hypothesis by showing computational studies. We also derive an upper bound on the needed data size in view of our hypothesis.

Keywords: frequent/infrequent item sets, association rules, knowledge discovery, probabilistic analysis