ASSOCIATION BUNDLE IDENTIFICATION FOR CATEGORICAL DATA *

Wenxue Huang[1], Milorad Krneta[1], Limin Lin[1,2], Jianhong Wu[2]

Abstract. We introduce the notion of association bundle identification over 2D categorical data set where each column represents a variable and each row an observation. Distinct from the association rule mining, the introduced association bundle has elements that have *pairwise* instead of *simultaneous* significant co-occurrence, and are not in "antecedent" and "consequent" query format. Comparing with categorical data clustering, the homogeneity of elements in an association bundle expresses the statistical linkage between *variable values* instead of the similarity between *observations*, and may span only subset of variables. We map association bundles into a set of *complete subgraphs* in order to identify these bundles by our proposed *complete subgraph detection* algorithm which is different from the traditional graph-partitioning algorithms. We show that the algorithm is efficient with a polynomial time cost to the number of variable values when the number of overlap layers among association bundles is relatively small.

Keywords: Association bundles, categorical data mining, complete subgraphs

*This work was partially supported by the Collaborative Research Development Program of Natural Sciences and Engineering Research Council of Canada and by Canada Research Chairs Program.

1. Generation5 Mathematics Technology Inc., 600-515 Consumers Road, Toronto, ON. M2J 4Z2 Canada.

2. Department of Mathematics & Statistics, York University, Toronto, ON. M3J 1P3 Canada.