Association rules and compositional data analysis: implications to big data

R. S. Kenett[1,2,3], J.A. Martín-Fernández[4], and M. Vives-Mestres[4]
1KPA Group, Raanana, Israel; 2University of Torino, Torino, Italy; 3Hebrew University, Jerusalem, Israel
4Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

Abstract

Many modern organizations generate a large amount of transaction data, on a daily basis. Transactions typically include semantic descriptors that require specialised methods for analysis. Association rule (AR) mining is a powerful semantic data analytic technique used for extracting information from transaction databases. AR was originally developed for basket analysis where the combination of items in a shopping basket are evaluated to determine prevalence. To generate an AR, the collection of more frequent itemsets—a set of two of more items—must be detected. Then, as a second step, all possible ARs are generated form each itemset. The ARs are then ranked using measures of association labelled, in this context, "measures of interestingness". The R package "arules" provides more than a dozen such measures including the relative linkage disequilibrium (RLD) which normalises classical Euclidean distances of the itemset from a surface of independence. In this work we study AR and RLD from compositional data (CoDa) perspective. It is well known that CoDa methodology provides nice properties such as subcompostional coherence and scalability. In this work we explore their implications to AR mining in big data analysis. The aim is to analyse if these properties ensure that the AR characteristic is not scale dependent and that if we consider a subset of the original items, we still keep similar behaviour. The talk will focus on such aspects, including the dynamic visualization of CoDa-AR measures on a simplex representation of the itemsets and its multidimensional extension.

References

Kenett, R.S. and Salini, S. (2008). Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks. *Trans. on Machine Learning and Data Mining*, Vol.1, No 2, pp. 83-96.

Kenett, R.S. and Salini, S. (2010). Measures of association applied to operational risks in *Operational Risk Management: a practical approach to intelligent data analysis* (Kenett R.S. and Raanan, Y, editors), John Wiley and Sons.

Martín-Fernández, J.A., Vives-Mestres, M. and Kenett, R.S. (2016). Understanding association rules from a compositional data approach. SIS 2016, 48th Meeting of the Italian Statistical Society, Università di Salerno, Italy, June 8-10th.

Hahsler, M., Buchta, C., Gruen, B. and Hornik, K. (2008). *arules R Package*, Version 0.6-6, *Mining Association Rules and Frequent Itemsets*, https://cran.r-project.org/web/packages/arules/index.html