The Journal of Open Source Software

tcherry: Learning the structure of tcherry trees

Katrine Kirkeby[1], **Maria Knudsen**[1], **and Ninna Vihrs**[1]

1 Department of Mathematical Sciences, Aalborg University

DOI: 10.21105/joss.01480

Software

- Review
- Repository
- Archive

Submitted: 26 April 2019
Published: 18 July 2019

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

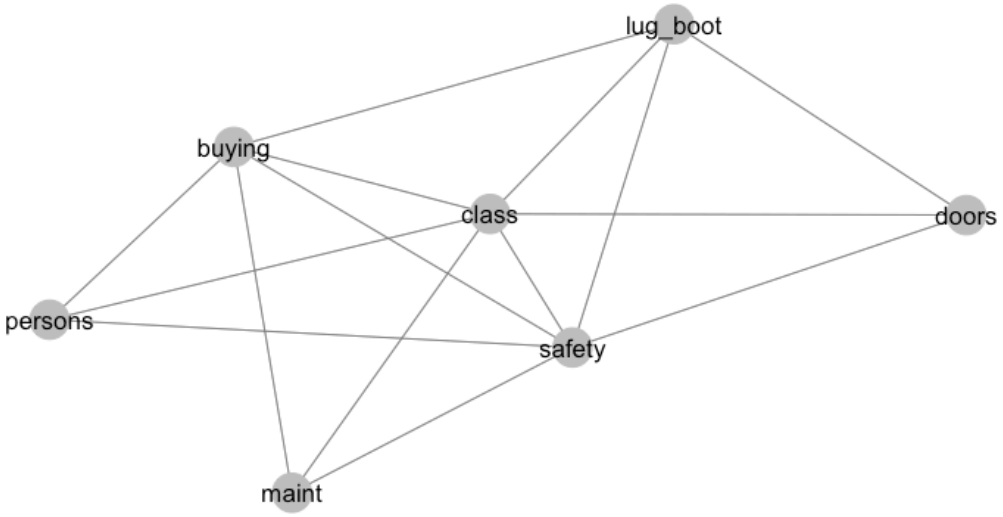
Summary

The R (R Core Team, 2018) package `tcherry` contains a variety of functions for learning the structure of a k'th order t-cherry tree from given categorical data, see for instance Kovács & Szántai (2012) for an explanation of this concept. This is a graphical model extending what is known as a Chow-Liu tree (Chow & Liu, 1968). Chow-Liu trees have for instance been used to estimate population frequencies of Y-STR haplotypes in Andersen, Curran, Zoete, Taylor, & Buckleton (2018) and t-cherry trees have been used to model relationships in social networks in Proulx (2015). The functions attempt to find a t-cherry tree structure of maximal likelihood. To do this exactly, it is necesarry to investigate all possible t-cherry tree structures of the given order. This is in most cases too time-consuming and therefore most of the functions use greedy search algorithms. Some implementations are inspired by algorithms in Kovács & Szántai (2010), Kovács & Szántai (2013) and Proulx (2015), but the package also contains some new algorithms and extensions. The package is only for structure learning and only categorical data is supported.

The package can be used as a tool to analyse problems exploring dependencies between any kind of categorical variables. The fitted t-cherry structure can be used to make statements about conditional dependencies and independencies. The structure can also be used for pattern recognition and independence statements can be used for variabel selection for a prediction problem (Szantai & Kovacs, 2010). If the structure is used in combination with packages such as `gRain` (Højsgaard, 2012), it may also be used to estimate probability distributions of the variables or for prediction. This makes it possible to use the structure as an expert system.

The t-cherry tree structure can be used in a variety of scientific fields such as biostatistics and artificial intelligence. The audience of the package is anyone who wants to model dependencies between categorical variables, approximate their probability distribution or solve classification problems with categorical variables.

The following figure shows an example of a fourth order t-cherry tree learned from the car evaluation data set from UCI Machine Learning Repository (Dua & Graff, 2017).



The R package `tcherry` is available on GitHub.

References

Andersen, M. M., Curran, J., Zoete, J. de, Taylor, D., & Buckleton, J. (2018). Modelling the dependence structure of y-str haplotypes using graphical models. *Forensic Science International: Genetics*, *37*, 29–36. doi:10.1016/j.fsigen.2018.07.014

Chow, C. K., & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions On Information Theory*, *IT-14 no. 3*, 462–467.

Dua, D., & Graff, C. (2017). UCI machine learning repository. University of California, Irvine, School of Information; Computer Sciences. Retrieved from http://archive.ics.uci.edu/ml

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1–26. doi:10.18637/jss.v046.i10

Kovács, E., & Szántai, T. (2010). On the approximation of a discrete multivariate probability distribution using the new concept of t-cherry junction tree. *Lecture Notes in Economics and Mathematical System*, *633*, 39–56. doi:10.1007/978-3-642-03735-1_3

Kovács, E., & Szántai, T. (2012). Hypergraphs as a mean of discovering the dependence structure of a discrete multivariate probability distribution. *Ann Oper Res*, *193*, 71–90. doi:10.1007/s10479-010-0814-y

Kovács, E., & Szántai, T. (2013). Discovering a junction tree behind a markov network by a greedy algorithm. *Optim Eng*, *14*, 503–518. doi:10.1007/s11081-013-9232-8

Proulx, B. (2015). *Impact of social structure on wireless networking: Modeling and utility*. Atizona State University.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Szantai, T., & Kovacs, E. (2010). Application of t-cherry junction trees in pattern recognition. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *1*(0), 40–45. Retrieved from https://www.edusoft.ro/brain/index.php/brain/article/view/103