Title: A study of applying copulas in data mining
Author: Martin Ščavnický
Department: Department of Theoretical Computer Science and Mathematical Logic
Supervisor: RNDr. Ing. Martin Holeňa CSc., Department of Theoretical Computer Science and Mathematical Logic

Abstract: Copulas are functions that describe the relationship between a multivariate distribution function and its marginals. They provide a way to model multivariate distribution functions, and are extensively used in finance and studied in data mining. In practice, there are many different copula families and no standard way for choosing the right one. In our work, we compare suitability of different copula families in data mining. We fit classification data using 8 copula families and compare them using 3 measures of fit. We also use a classification algorithm based on copulas and compare its accuracy for different copula families. The results indicate that elliptical copulas fit our data better, but hierarchical Archimedean copulas give comparable accuracy in the classification. We also propose and test a modified method for modelling data using hierarchical Archimedean copulas, which fits some datasets with negative dependence between attributes better. Based on this modified method, we propose a visualization of dependence in data and observe wheter it captures differences between classes.

Keywords: data mining, relationships between attributes, probabilistic relationships, copulas, kinds of copulas