BayesClass: An R package for learning Bayesian network classifiers

Bojan Mihaljevic[1,⋆], **Pedro Larrañaga**[1], **Concha Bielza**[1]

1. Computational Intelligence Group, Departamento de Inteligencia Artificial Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo sn, 28660, Boadilla del Monte, Madrid

⋆Contact author: b.mihaljevic@alumnos.upm.es

Keywords: Bayesian network classifiers, supervised classification, machine learning

BayesClass implements ten algorithms for learning Bayesian network classifiers from discrete data. This includes score+search algorithms and those that maintain the structure of a naive Bayes but extend it with additional parameters. Many of the algorithms perform implicit feature selection. Most of the structure searching methods are based on the forward search heuristic while score functions include classifier accuracy, likelihood, and a function based on the significance of dependency between two sets of variables (see Blanco et al. (2005)). Implemented algorithms include: *tree augmented naive Bayes* (TAN) (Friedman et al. (1997)), *forward sequential selection* (Langley and Sage (1994)), *forward sequential selection and joining* (FSSJ) (Pazzani (1996)), and adaptations of those methods from (Blanco et al. (2005)); the *adjusted probability naive Bayesian classification* (Webb and Pazzani (1998)), the *attribute weighted naive Bayes* (Hall (2007)), and others. We also propose an adaptation of the TAN algorithm to operate on structures learned by FSSJ.

The assessment and use of induced classifiers are straightforward. An interface to the **caret** package allows for estimation of predictive performance by resampling. Several discretization procedures are implemented. Discretization learned from training data can be applied to test data by mapping real-valued unseen data to the intervals learned, so that the effect of discretization on classifier performance can be assessed. All algorithms can handle incomplete training data, approximating the sufficient statistics of a probability function corresponding to some node X by ignoring cases with missing values for either X or any of its parents. An interface to the **gRain** package provides inference and its integration with the **Rgraphviz** package, thus enabling graph plotting. We expect to publish **BayesClass** on CRAN during april of 2013.

References

Blanco, R., I. Inza, M. Merino, J. Quiroga, and P. Larrañaga (2005). Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics 38*(5), 376–388.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning 29*, 131–163.

Hall, M. (2007). A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems 20*(2), 120–126.

Langley, P. and S. Sage (1994). Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence (UAI-1994)*, pp. 399–406. Morgan Kaufmann.

Pazzani, M. (1996). Constructive induction of cartesian product attributes. In *Proceedings of the Information, Statistics and Induction in Science Conference (ISIS-1996)*, pp. 66–77.

Webb, G. I. and M. J. Pazzani (1998). Adjusted probability naïve Bayesian induction. In *Proceedings of the 11th Australian Joint Conference on Artificial Intelligence (AI-1998). Lecture Notes in Computer Science*, Volume 1502. Springer.