Decision Trees for Probability Estimation: An Empirical Study

Han Liang, Harry Zhang, Yuhong Yan

Accurate probability estimation generated by learning models is desirable in some practical applications, such as medical diagnosis. In this paper, we empirically study traditional decision-tree learning models and their variants in terms of probability estimation, measured by Conditional Log Likelihood (CLL). Furthermore, we also compare decision tree learning with other kinds of representative learning: na¨ıve Bayes, Na¨ıve Bayes Tree, Bayesian Network, K-Nearest Neighbors and Support Vector Machine with respect to probability estimation. From our experiments, we have several interesting observations. First, among various decision-tree learning models, C4.4 is the best in yielding precise probability estimation measured by CLL, although its performance is not good in terms of other evaluation criteria, such as accuracy and ranking. We provide an explanation for this and reveal the nature of CLL. Second, compared with other popular models, C4.4 achieves the best CLL. Finally, CLL does not dominate another well established relevant measurement AUC (the Area Under the Curve of Receiver Operating Characteristics), which suggests that different decision-tree learning models should be used for different objectives. Our experiments are conducted on the basis of 36 UCI sample sets that cover a wide range of domains and data characteristics. We run all the models within a machine learning platform - Weka.