Abstract

Data Mining can be seen as an extension to statistics. It comprises the preparation of data and the process of gathering new knowledge from it. The extraction of new knowledge is supported by various machine learning methods. Many of the algorithms are based on probabilistic principles or use density estimations for their computations. Density estimation has been practised in the field of statistics for several centuries. In the simplest case, a histogram estimator, like the simple equal-width histogram, can be used for this task and has been shown to be a practical tool to represent the distribution of data visually and for computation. Like other nonparametric approaches, it can provide a flexible solution. However, flexibility in existing approaches is generally restricted because the size of the bins is fixed—either the width of the bins or the number of values in them. Attempts have been made to generate histograms with a variable bin width and a variable number of values per interval, but the computational approaches in these methods have proven too difficult and too slow even with modern computer technology.

In this thesis new flexible histogram estimation methods are developed and tested as part of various machine learning tasks, namely discretization, naive Bayes classification, clustering and multiple-instance learning. Not only are the new density estimation methods applied to machine learning tasks, they also borrow design principles from algorithms that are ubiquitous in artificial intelligence: divide-and-conquer methods are a well known way to tackle large problems by dividing them into small subproblems. Decision trees, used for machine learning classification, successfully apply this approach. This thesis presents algorithms that build density estimators using a binary split tree to cut a range of values into subranges of varying length. No class values are required for this splitting process, making it an unsupervised method. The result is a histogram estimator that adapts well even to complex density functions—a novel density estimation method with flexible density estimation ability and good computational behaviour.

Algorithms are presented for both univariate and multivariate data. The univariate histogram estimator is applied to discretization for density estimation and also used as density estimator inside a naive Bayes classifier. The multivariate histogram, used as the basis for a clustering method, is applied to improve the runtime behaviour of a well-known algorithm for multiple-instance classification. Performance in these applications is evaluated by comparing the new approaches with existing methods.