Effect of data discretization on the classification accuracy in a high-dimensional framework

Annika Tillander and Tatjana Pavlenko

Stockholm University, Sweden

Abstract

Computionally efficient methods for classification are known to preform better when they are based on the discrete data, and this is especially important in a high dimensional setting, i.e. when the number of feature variables, p, is comparable to or greater than the number of observations, n. The goal of this on-going study is to empirically evaluate discretization of the continuous features and explore the effect of this procedure on the performance properties of a high dimensional classifier.

In order to represent a of variety discretization measures we focus on the hierarchical structure in Peng et al. (2009) and embed it into a high dimensional framework. We then explore six different discretization methods that can be seen in terms of *supervised* or *unsupervised*. Supervised discretization methods, also known as class-driven discretization technique, utilize the class membership, whereas unsupervised methods are known as thresholding as they mainly discretize the feature variables with regard to interval width or interval frequency.

As the goal of classification is to correctly predict a class membership of an observation we optimize the discretization procedure using the misclassification probability as a measure of the classification accuracy. To compare the classification performance between continuous feature and discretized ones we consider three supervised methods, k-nearest neighbor, Naive Bayes and a type of C4.5. To capture the effect of high dimensionality we investigate a variety of p for a fixed n, which makes it possible to evaluate the combined effect of the discretization and growing dimensionality.

Since the discretization is a data transformation procedure another aspect of this step is to investigate how the dependence structure is affected by the discretization. Accounting for such structures can improve accuracy and lead to models that are more interpretable (see Deng and Yuan 2009). For comparison of the dependence structure for discretized features with the original ones, two different performance measures were used; Frobenius norm and Kullback-Leibler loss. Various types of covariance were considered in the data generating process to explore the effect of discretization on the data dependence structure.

Current results reveal that a supervised entropy-based measure gives the over all lowest misclassification probabilities, whereas it alters the dependence structure mostly. The discretization method that best retain the dependence structure is an unsupervised binning measure.

Keywords

Supervised classification, Discretization, High dimensionality, Misclassification probability.

References

Peng, L., Qing, W., and Yuija, G. (2009). Study on comparison of discretization methods. *4th International Conference on Artificial Intelligence and Computional Intelligence*, 380–384.

Deng, X. and Yuan, M. (2009). Large Gaussian covariance matrix estimation with Markov structures. *J. Comput. Graph. Statist. 18(3)*, 640–657.