CLASSIFICATION WITH UNSTRUCTURED PREDICTORS AND AN APPLICATION TO SENTIMENT ANALYSIS

Abstract:

Unstructured data refers to information that lacks certain structures and cannot be organized in a predefined fashion. Unstructured data often involves words, texts, graphs, objects or multimedia types of files that are difficult to process and analyze with traditional computational tools and statistical methods. This work explores ordinal classification for unstructured predictors with ordered class categories, where imprecise information concerning strengths of association between predictors is available for predicting class labels. However, imprecise information here is expressed in terms of a directed graph, with each node representing a predictor and a directed edge containing pairwise strengths of association between two nodes. One of the targeted applications for unstructured data arises from sentiment analysis, which identifies and extracts the relevant content or opinion of a document concerning a specific event of interest.

We integrate the imprecise predictor relations into linear relational constraints over classification function coefficients, where large margin ordinal classifiers are introduced, subject to many quadratically linear constraints. The proposed classifiers are then applied in sentiment analysis using binary word predictors. Computationally, we implement ordinal support vector machines and $\psi$-learning through a scalable quadratic programming package based on sparse word representations. Theoretically, we show that utilizing relationships among unstructured predictors improves prediction accuracy of classification significantly. We illustrate an application for sentiment analysis using consumer text reviews and movie review data. Supplementary materials for this article are available online. This is joint work with Junhui Wang, Xiaotong Shen and Yiwen Sun.