2017-06-30

Discussion on "Random-projection ensemble classification" by T.I. Cannings and R.J. Samworth

Stander, Julian

Journal of the Royal Statistical Society. Series B: Methodological
Royal Statistical Society

Discussion of paper: Random-projection ensemble classification by Timothy I. Cannings and Richard J. Samworth

Dr Julian Stander J.Stander@plymouth.ac.uk
and
Dr Luciana Dalla Valle luciana.dallavalle@plymouth.ac.uk
University of Plymouth

We congratulate the authors on their paper and R package that makes the RPEnsemble methodology readily applicable. Here we outline an experiment and ask two questions.

We worked with data discussed by Baldino (2016) comprising 120 Trip Advisor reviews and each reviewer's star classification. We combined 1, 2 and 3 stars into Class 1, and 4 and 5 stars into Class 2. Using R's `tm` package (Feinerer and Hornik, 2015), we computed the transpose of the Term-Document Matrix. This word count matrix had $n = 120$ rows corresponding to reviews and $p = 2644$ columns corresponding to words, with ninety-seven percent zeros. We normalized the rows by dividing by review lengths. We randomly selected 60 reviews as training data, with the remaining 60 being test data. We applied RPEnsemble to the normalized word count feature matrix. For comparison, using dictionaries of 2006 positive and 4783 negative words (Liu, Hu and Cheng, 2005), we calculated a sentiment score for each review as the difference between the number of positive and negative word matches. We normalized these scores by dividing by review length to obtain sentiment intensities. We then applied binary logistic regression with sentiment intensity as explanatory variable. Over 50 repetitions, with $d = 2$ we obtained a quite low average misclassification rate of 25.5% (standard deviation 1.2%) for the normalized word count RPEnsemble methodology using $B_1 = 500$, $B_2 = 50$, LDA, Gaussian projections and the leave-one-out test error. The average $\hat{\alpha}$ was 1.69 (0.0127). When QDA or the axis projection method was used, the RPEnsemble average misclassification rate was often considerably worse (non-overlapping confidence intervals), although RPEnsemble seemed quite robust to other choices including the value of $d = 3, \ldots, 9$. For our sentiment intensity logistic regression the average misclassification rate was lower at 12.2% (0.76%). We therefore conclude that RPEnsemble can be successfully used to classify hotels using only review word counts. Naturally, better classification results can be obtained by performing a sentiment analysis which makes use of information from positive and negative word dictionaries.

Can the proportion of the classifications $C^1(x), \ldots, C^{B_1}(x)$ in each category be used to quantify classification uncertainty?

The copula construction (Sklar, 1957) provides flexible multivariate models, with vine copulas (Aas *et al.*, 2009) being used when p is large. Could the use of a classifier defined using copula densities estimated on low dimensional projections of the original data, perhaps with robust marginal modelling, be a way of exploiting copula flexibility, while avoiding high dimension estimation problems?

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009) Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, **44**, 182–198.

Baldino, A. (2016) *Information Mining from Social Media*. Masters Thesis, University of Rome La Sapienza.

Feinerer, I. and Hornik, K. (2015) `tm`: Text Mining Package. R package version 0.6-2.
http://CRAN.R-project.org/package=tm.

Liu, B., Hu, M. and Cheng, J. (2005) Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th International World Wide Web conference* (WWW-2005), May 10–14, Chiba, Japan.

Sklar, A. (1957) Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.