Boosting stumps to determine the genes involved in cell proliferation due to ascorbic acid

Joaquim Pinto da Costa and Filipe Sousa

University of Porto, Portugal

Abstract

Boosting is an iterative algorithm that performs a linear combination of multiple predictions (function estimates) and can be applied both to regression and classification. Usually the individual predictors have mediocre performance in terms of misclassification error rates, like for instance stumps which are decision or regression trees with just one split. Nevertheless, the final linear combination found by boosting improves the performance of base classifiers, both theoretically and empirically, and produces a highly accurate classification rule. In our application we have a very small number of observations (9) and thousands of variables (genes) and so it will be very difficult to find a single accurate classifier. For that reason we decided to use stumps, as the data is not enough to build larger decision trees. However, by building many stumps, which will model different aspects of the data, and then combine them linearly with boosting, we hope to find a good classifier.

Our individual classifiers, stumps, are expected to present a large bias, due to their simplicity. Boosting is a bias reduction technique and typically improves the performance of a single (simple) tree model. The procedure starts by building in the first iteration a stump to predict if a certain example belongs to class ω_1 (in our application ascorbic acid) or not. Those observations which were not correctly classified by this stump will have a larger weight in the next iteration; the opposite happens with the ones correctly classified. In the end we have a set of M models (stumps in our case) and a weighted linear combination of them is found as the final model.

The ada package (Culp et al. 2006) implements the original AdaBoost algorithm (Freund and Schapire 1996, 1997) with other extensions. Some important features incorporated in this package include the use of both regression and classification trees for boosting and also various useful plots that aid in assessing variable importance and relationships between subsets of variables. This last feature is particularly interesting for us because, as we have thousands of variables (genes), we want not only an accurate final prediction model but mainly a way of choosing amongst those thousands of genes, the ones that matter.

The motivation for this work is to identify the genes involved in cell proliferation due to ascorbic acid (AA) and its stable form ascorbic acid 2-phosphate (AA2P). Our skin, besides providing cover for the underlying soft tissues, it also performs many additional functions, including protection against injury, bacterial invasion and desiccation. Obviously any lesion must be rapidly and efficiently repaired in order to keep homeostasis. Skin consists of two layers: an outer epidermis and a deeper connective tissue layer, the dermis. Fibroblasts, the most abundant cell type in the connective tissue, are responsible for the synthesis of almost the entire extracellular matrix and thus contribute to skin regeneration. There is evidence that fibroblasts found at sites of lesions proliferate more and more actively secrete extracellular matrix.

Ascorbic acid (AA), also known as vitamin C, is a sugar acid with antioxidant properties. ROS (reactive oxygen species that appear due to lesions and cytotoxic molecules) which contain unpaired electrons may interact with nucleic acids, proteins or lipids destroying them. Ascorbic acid can terminate these chained radical reactions by being a stable electron donor in interactions with free radicals.

Ascorbic acid 2-phosphate (AA2P) is a stable form of vitamin C and, as ascorbic acid, it is involved in enhanced cell proliferation and relative rate of extracellular matrix synthesis.

Our aim is to determine the genes involved in cell proliferation due to AA and AA2P treatments and we use the same dataset that was used in (Duarte & al. 2009). The dataset is a microarray matrix where the whole human genome is scanned to extract the expression profiles. Both treatments and controls (scorbutic cells) were analyzed in triplicate.

Important genes are those where expression profiles of AA and AA2P treated cells are significantly higher than scorbutic cells. This means a particular gene over expression being transcriptional products more abundant and available to cell proliferation.

Keywords

Boosting, Decision trees, Supervised classification, Microarrays.

References

Culp, M., Johnson, K., and Michailidis, G. (2006). ada: An R package for stochastic boosting. *J. Statist. Software 17*, 1–27.

Freund, Y., and Schapire, R. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 148–156.

Freund Y, and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci. 55(1)*, 119–139.

Duarte, T.L., Markus, S.C., and George, D.D. (2009). Gene expression profiling reveals new protective roles for vitamin C in human skin cells. *Free Radical Biology & Medicine 46*, 78–87.