P09

AucPR: An AUC-based approach using penalized regression for disease prediction with high-dimensional omics data

Wenbao Yu 1 and Taesung Park1,2

1 Department of statistic, Seoul National University, Korea

2 Interdisciplinary Program in Bioinformatics, Seoul National University, Korea

Abstract

Motivation

It is common to get an optimal combination of markers for disease classification and prediction when multiple markers are available. Many approaches based on the area under the receiver operating characteristic curve (AUC) have been proposed. Existing works based on AUC in a high-dimensional context depend mainly on a non-parametric, smooth approximation of AUC, with no work using a parametric AUC-based approach, for high-dimensional data.

Results

We propose an AUC-based approach using penalized regression (AucPR), which is a parametric method used for obtaining a linear combination for maximizing the AUC. To obtain the AUC maximizer in a high-dimensional context, we transform a classical parametric AUC maximizer, which is used in a low-dimensional context, into a regression framework and thus, apply the penalization regression approach directly. Two kinds of penalization, lasso and elastic net, are considered. The parametric approach can avoid some of the difficulties of a conventional non-parametric AUC-based approach, such as the lack of an appropriate concave objective function and a prudent choice of the smoothing parameter. We apply the proposed AucPR for gene selection and classification using four real microarray and synthetic data. Through numerical studies, AucPR is shown to perform better than the penalized logistic regression and the nonparametric AUC-based method, in the sense of AUC and sensitivity for a given specificity, particularly when there are many correlated genes.

Conclusion

We propose a powerful parametric and easily-implementable linear classifier AucPR, for gene selection and disease prediction for high-dimensional data. AucPR is recommended for its good prediction performance. Beside gene expression microarray data, AucPR can be applied to other types of high-dimensional omics data, such as miRNA and protein data.