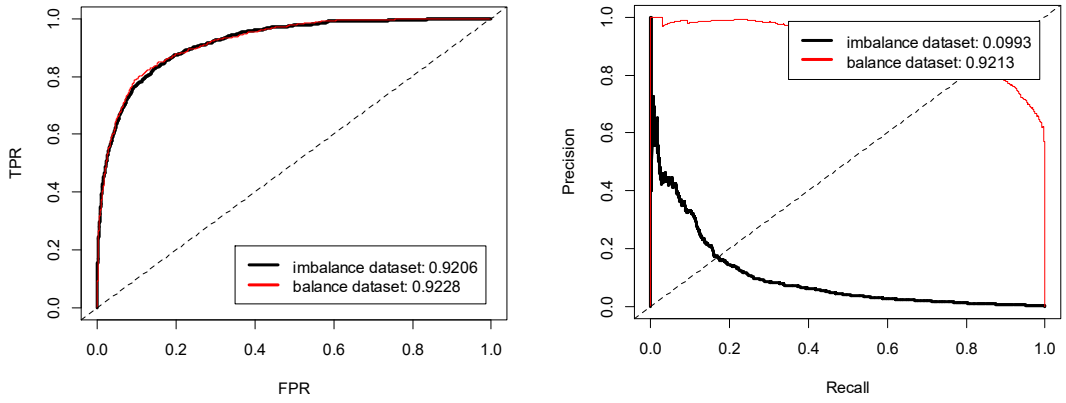
Influence of imbalance

Let the predictive scores of positive samples follow the normal distribution that the mean is 1 and the standard deviation is 1, and the predictive scores of negative samples follow the normal distribution that the mean is -1 and the standard deviation is 1. In that way, there are 1000 positive samples and 1000 negative samples, as well as 1000 positive samples and 1000*400 negative samples in the imbalance dataset. Then, we plot ROC curves and PR curves in balance and imbalance dataset as follows:



(A) (B)

Figure 1. performance in terms of different metrics. (A) ROC curves. (B) PR curves.

As shown in Figure 1, we can see that ROC curves are similar in balance and imbalance dataset, while PR curves of balance dataset and imbalance dataset are quite different. In other words, comparing PR curves and ROC curves, we could observe that AUPR is very low at high AUC for imbalance dataset. Meanwhile, through observation of PR curve in imbalance dataset, the precision is very low at reasonably high recall. In general, extreme imbalance is a main reason of phenomenon that the precision is very low at reasonably high recall and AUPR is very low at high AUC.