Reviewer's report

Title: Optimally splitting cases for training and testing high dimensional classifiers

Version: 1 Date: 14 September 2010

Reviewer: Shuangge Ma

Reviewer's report:

The topic investigated is of significant interest. In the literature, there are many different ways of splitting, but there is a lack of research on the splitting rule. I have the following concerns that I hope the authors can address.

1. Just by looking at the numerical results, I am not quite convinced about the conclusion of the 2/3 rule. The range of proportions in Table 1 is quite wide. Just by looking at Table 2, I would conclude the ½ (as opposed to 2/3) rule. More convincing numerical studies are needed.

2. In data analysis, the first three datasets are well known for being “easy to classify”. Experiments with data that are not as easy (with lower full dataset accuracy) are needed.

3. The simulation settings seem too simple. The real gene expression data has much more complicated distributions. Revision of the simulation settings (so that they better mimic real data) is needed.

4. How much does the result depend on the classification rule used? With the settings described in the article, there are a large number of classification rules. Is it reasonable to expect the obtained result to be applicable to a large number of classification rules? If yes (or no), why is that?

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests