Reviewer’s report

Title: Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines

Version: 2 Date: 12 May 2017

Reviewer: Allan Tucker

Reviewer's report:

This paper explores a number of existing techniques for dealing with imbalanced data prior to feature selection and classification using Support Vector Machines and Random Forests. These were applied on an immunoassay dataset in order to predict Hepatitis B and C. The paper is clearly written with a reasonable coverage of the background of imbalanced data analysis. What is more the empirical analysis is clear with the results explained in detail.

However, I have some concerns with the extreme nature of the imbalance. It is stated that only 2% of the cases are positive for HBV and 7% for HCV. This massively reduces the size of the dataset for the downsampling and risks some major overfitting issues due to a biased set of positive cases for SMOTE. The results show limited success for all methods and this is likely to be due to the limitations of the dataset.

In addition, I am sure why other approaches to dealing with small positive samples were not explored. For example, there are methods that can focus on learning the normal behaviour and explore deviation from normality. This is known as "one class classification" - these may have been better for such an extreme dataset. As a bare minimum, I would expect a discussion of these techniques.

Why was only one feature selection method explored. There is probably a lot interaction between the balancing method and feature selection and I would have liked to see some analysis of these interactions.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Unable to assess

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal