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: 0 Date: 31 Dec 2016

Reviewer: Adam Dunn

Reviewer's report:

Thank you for the opportunity to provide a further review for this manuscript. I have read the updated (highlighted) manuscript and the response to the reviewers. The authors have clarified the aims of the manuscript, which is to examine how different methods for sampling data affect the performance of classifiers constructed using support vector machine (SVM) methods and by selecting five features. The manuscript still has some major issues that could be resolved with a bit more effort spent on literature review, writing, and argument structure.

Abstract:

1. Consider including the aim of the manuscript as the final sentence in the background so the reader knows what is being done.

Background:

2. Table 1 probably doesn't belong in the same spot any more.

3. I still wouldn't call SVM a "pattern recognition" technique. I understand that the jargon people use for machine learning techniques in different fields can be different but stick to calling the SVM method an algorithm for producing a binary classifier and the SVMs themselves as classifiers (or models, if you prefer). Otherwise you risk alienating many of your readers. Also, avoid using the term "resulting machines".

4. The references are now not used as nouns in the sentence (good) but still sit in the middle of sentences - moving them to the end of the sentence or at least next to punctuation wherever possible will enhance readability.
5. "Recent developments" should be removed - active learning, for example, has been around for about ten years I think?

6. The literature review in the background has either one of the two following problems: (a) it seems like a random selection of partially relevant articles have been found and added in retrospectively, which is a common problem for reviews that try to cover a very broad range of topics, or (b) it is too general, and should focus on the specific articles that address the aims of the paper. In this case, I think the aims of the paper could be made tighter (see below) to avoid the need to review literally hundreds of articles that deal with sampling, boosting, feature selection, etc. in imbalanced sets, or the authors should spend the time searching for state-of-the-art methods in sampling and implement them. I think perhaps the former is less time-consuming.

7. The aims of the research, as they stand, are not addressed in the manuscript and would provide more than enough reason to dismiss the work entirely. There is no description of how the classifiers were used to enhance laboratory diagnosis in practice (and no measures of how they might have improved performance in the lab compared to what already exists), and what is being described as pre-processing (see below) actually covers a much broader set of methods (e.g. selecting five features consistently means that feature selection has not been explored). Can I suggest an aim that precisely covers the contributions of the research, which is to examine how two sampling methods affect the performance of SVM classifiers in imbalanced pathology data. I would also suggest putting that aim into the abstract directly. There is absolutely nothing wrong with aiming to test two sampling methods and a feature selection method in unbalanced pathology data - and there is no need to try to sell the research as much more than it actually is. Precise descriptions are much better than bad generalisations.

8. I wouldn't consider feature selection and sampling as part of pre-processing personally. This may be my own misunderstanding, but I think of pre-processing as handling the data and dealing with data validation and missing values, and things like feature selection and sampling as part of the process for producing the most robust and accurate classifiers possible given the data.

9. Avoid using words like "very" and statements that cannot be quantified.

Methods:

10. The reference used for multiple downsizing is one that suggests that this is useful for high dimensional datasets (tens of thousands) while the actual problem considers only 5, which is not high-dimensional.
Results:

11. The structure of the results could be improved. Structure the results to address the aims, not the outcome measures. I suggest one paragraph that *reports* how the performance (precision, recall, F1-score) were affected by the choice of sampling method in both HBV and HCV, and then one paragraph that *reports* how the performance (precision, recall, F1-score) were affected by using feature selection (or not). Also be very careful with how you report their interaction unless you have a strong explanation for why the interaction might have influenced the results. Otherwise it might appear to the reader as if the methods had an entirely unpredictable effect on the performance of the classifiers that related more to the specific nature of the data than the choice of methods.

Discussion:

12. The comparison to existing literature is only to one paper (the authors' own paper), and does not include another example which is in the limitations section, and this is entirely inadequate. If the aim of the manuscript is to compare two sampling methods and a feature selection method in SVM classifiers for unbalanced pathology data, then it should compare the results (i.e. which method worked better and why?) with other papers that have done the same. If there really are no papers in pathology, then the comparison should be to other application domains in which sampling methods have been used in concert with SVMs when applied to low-dimensional unbalanced data. There should be no case where *only* the authors' previous paper is the only relevant literature.

13. The limitations describe a comparison to another paper and the reason the results were different - that should be in the comparison not in the limitations. The limitations (in this case) should be about why the results may not generalise (only two datasets examined; maybe refer to the synthetic data examples), the fact that only a limited set of feature selection and sampling methods were tested (and that there are many others).

Conclusion:

14. Try a conclusion that just summarises the contributions of the paper and very briefly (one sentence) explains the implications for pathology practice.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

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

Yes

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

No

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:

Needs some language corrections before being published

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