Author’s response to reviews

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

Authors:

Alice Richardson (alice.richardson@anu.edu.au)

Brett Lidbury (brett.lidbury@anu.edu.au)

Version: 1 Date: 23 Feb 2017

Author’s response to reviews:

Thankyou for the opportunity to further revise this paper in response to the referees' comments.

Associate Editor's comments:

1. The data description section needs improvement, as for example it discusses the statistics about positive and negative cases before defining what is positive and what is negative.

The relevant paragraphs in the Data section have been moved to the Results section (see Referee 2 comment 15).

2. I think that adding a Table with all the numbers and statistics would help to clarify this section.

Table 1 has been moved into the Data section: it contains the descriptions of the assay variables. Referee 2 (comment 15) asks for the first paragraph of page 6 to move to the Results section, and it has been converted into Table 2.
3. Another point that still needs clarification is the one related to the analysis of variance. Besides the comments specified by Reviewer 1, it is important to better explain how this analysis was carried out in terms of data and statistics that were used.

Paragraphs on the analysis of variance in the Results section have been pulled back into the Methods section just under Renumbered Table 3.

Reviewer 1

Abstract:

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

The Background of the Abstract has been altered, in line with comments 6 and 7, to reflect the tighter aim of the paper.

Background:

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

Table 1 has been moved to the Methods section.

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".

The term “resulting machines” has been removed. The language around SVMs has also been tightened up throughout the paper, mostly by simply referring to the SVM without further qualification.

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.

All references have been checked and where grammatically sensible have been moved to the end of the sentence.

5. "Recent developments" should be removed - active learning, for example, has been around for about ten years I think?

This phrase has been removed.

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.
This revision of the paper has taken the suggestion to tighten the focus on comparing three sampling or balancing methods (simple downsizing, multiple downsizing, and SMOTE) and one feature selection method (random forest). The literature review now addresses past work on these topics.

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.

As the referee notes, this comment is tied to the previous one, where we note that our tightened aim is to compare three sampling methods (simple downsizing, multiple downsizing, and SMOTE) and one feature selection method (random forest). The literature review now identifies key work in each of those areas, along with pathology data analysis in general. Some new references have been identified, some previous ones retained, and some removed. Some of the literature referred to in the Methods section has also been consolidated into this section in the Introduction.

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.
We have replaced the phrase “pre-processing” with “balancing” as that is the specific aspect of pre-processing investigated in this paper.

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

“Very” has been removed from the paper.

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.

Reference [20] appears to be one of the first reports of multiple downsizing, and we acknowledge that it concentrates on microarray data (very high-dimensional). Our data has around 30 dimensions, and we have been unable to find reference to multiple downsizing being used on data of comparable dimension. The variable selection that brings the 30 variables down to 5 happens in the second step of the analysis. We used multiple downsizing in our paper to tackle the imbalance in a data set rather than the high dimensionality.

A comment regarding the focus of reference [20] has been included in the Methods section.

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.

The structure of the Results section was previously revised to focus on the performance measures, but now has been revised to focus on the aims (the balancing method and the feature selection). The interpretation of interactions has been teased out separately for each virus and for each performance measure, to make it clear that the interactions are predictable.

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.

The Comparison section now includes the example which was in the Limitations section.

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).

The original Limitations section has been moved to the Comparison to current literature section. A new limitations sections has been written to reflect the referee’s suggestions.
Conclusion:

14. Try a conclusion that just summarises the contributions of the paper and very briefly (one sentence) explains the implications for pathology practice.

The Conclusion has been changed to better reflect the contribution of the paper, and provide a short implication for pathology practice.

Reviewer 2:

15. The paragraph starting at the top of page six belongs in the results section.

The table has been moved to the beginning of the Results section and a table provided (refer to Editor’s comment 2).

16. Figure 2 requires a better explainer to the lay reader.

Several sentences have been added or altered beneath Figure 2 to clarify its interpretation.