Errors & Clarifications

Introduction

This document serves to correct typographical errors and clarify potentially misleading aspects of the text of the following PhD thesis.

"Computational approaches to predicting drug induced toxicity"

Richard Liam Marchese Robinson, King's College

University of Cambridge, 2013

http://www.dspace.cam.ac.uk/handle/1810/244242

Where appropriate, corrections are proposed to the existing text, as indicated in blue font.

The issues documented in this report are subdivided into the following sections (Table of Contents).

N.B. This document has not undergone any form of peer review.

Richard Marchese Robinson

Table of Contents

Typographical errors in the main text

1. On p.49, the definition of the covariance matrix is incomplete. The following sentence should be replaced.
    a. **Old sentence**: "The principal components (PCs) are the M eigenvectors of the covariance matrix ($X^T X$), computed from the $N \times M$ matrix ($\underline{X}$) with elements x_{im} denoting the value of the mth descriptor for the ith molecule."
    b. **New sentence**: The principal components (PCs) correspond to the M eigenvectors of the matrix $\underline{\underline{X^T X}}$, computed from the $N \times M$ matrix ($\underline{X}$) with elements x_{im} corresponding to the mean centred value of the mth descriptor for the ith molecule.
2. On p.49, "Shusko *et al.*" should read "Sushko *et al.*"
3. On p.68, "See Chapter 2, section 2.5" should say "See Chapter 2, section 2.3."
4. On p.90, the following sentence should have read as follows – reflecting the fact that the results obtained with some modelling approaches were summarised as the mean performance, in terms of the MCC, across multiple different models obtained using a given modelling approach with different random selections carried out during the model building phase. N.B. In some cases, for short hand, the text may discuss the performance of "models" in terms of the mean MCC, whereas this should – strictly speaking – refer to the performance of "modelling approaches".
    a. **Old sentence**: "The range of MCC values obtained on the 'external' validation sets, with either feature selection protocol, is comparable with those previously reported in the literature (Appendix B, Table B.1)."
    b. **New sentence**: "The range of (mean) MCC values obtained on the 'external' validation sets, with either feature selection protocol, is comparable with those previously reported in the literature (Appendix B, Table B.1)."
5. On p.93, the caption for Table 4.3 should have read as follows.
    a. **Old**: "MCC values obtained on 'external' test sets for all partitions of the Thai-313 and Dubus-203 datasets used to evaluate this author's modelling procedures. See Table 4.2 for presentational details."
    b. **New**: "(Mean) MCC values obtained on 'external' test sets for all partitions of the Thai-313 and Dubus-203 datasets used to evaluate this author's modelling procedures. See Table 4.2 for presentational details."
6. On p. 130, "corresponding to the geometric mean of their descriptor vectors" should state "corresponding to the arithmetic mean of their descriptor vectors".

Errors in citations

(1) The wrong reference is cited in the following footnote on p.42.
    a. **Old**: "Alternatively, this matrix being an example of a "contingency table",[94] the "contingency matrix".[230]"
    b. **New**: "Alternatively, this matrix being an example of a "contingency table",[233] the "contingency matrix".[230]"

(2) The wrong reference is cited for Toxtree on p.55.
    a. **Old**: "…freely available Toxtree (version 1.51)[159–161]"
    b. **New**: "…freely available Toxtree (version 1.51)[162,163]"
(3) The wrong reference is cited for Toxtree on p.183.
    a. **Old**: "In Chapter 3, a consensus model for mutagenicity/carcinogenicity was developed by combining the output generated by two commonly employed predictive toxicology programs: Toxtree[160] and Derek for Windows™.[271]"
    b. **New**: "In Chapter 3, a consensus model for mutagenicity/carcinogenicity was developed by combining the output generated by two commonly employed predictive toxicology programs: Toxtree[162] and Derek for Windows™.[271]"
(4) A reference was missing from the following sentence on p.152.
    a. **Old:** "A variety of Machine Learning methods, and feature selection strategies were considered in these studies - although, in most cases, the same descriptors were employed as per the original publications.[124–127,199]"
    a. **New:** "A variety of Machine Learning methods, and feature selection strategies were considered in these studies - although, in most cases, the same descriptors were employed as per the original publications.[123–127,199]"

Errors in references

1. Reference (405).
    a. **Old**: "Shanle, E. K.; Xu, W. *Chem. Res. Toxicol.* **2010**, *24*, 6–19."
    b. **New**: "Shanle, E. K.; Xu, W. *Chem. Res. Toxicol.* **2011**, *24*, 6–19."

Errors in the analysis

1. The following explanation of the meaning of the p-values presented in the corresponding text *in the footnote* starting on p.90 and ending on p.91 is incorrect. However, this does not affect the p-values presented or the description of how they were calculated. In part, this footnote contains a typographical error (blue text) and, in part, the reasoning employed here is problematic (red text).
    a. **Old footnote**: "The uncorrected p-values were calculated (see Chapter 2, section 2.6.4.1) from the MCC value (or mean value for the pseudo-stochastic methods – Winnow, RF, QuaSAR-Classify) obtained when training and testing the selected modelling approaches on a given train/test partition. They denote the conditional probability of obtaining a (mean) MCC value with at least the magnitude observed on a given test set, supposing a random predictor had been built on the corresponding training set. The Bonferroni correction provides an upper bound, equal to the value below which the corrected p-values are deemed statistically significant, to the conditional probability, supposing an unknown number of approaches performed like random predictors on average, of erroneously declaring that any model (or random selection from a set of corresponding models for the pseudo-stochastic algorithms) built, in this work, on one of the training sets would perform no differently, in terms of the mean MCC across various test sets, to a random predictor - based on the (mean) MCC value observed on the single corresponding test set."

b. **Typographical correction**: "The uncorrected p-values were calculated (see Chapter 2, section 2.6.4.1) from the MCC value (or mean value for the pseudo-stochastic methods – Winnow, RF, QuaSAR-Classify) obtained when training and testing the selected modelling approaches on a given train/test partition. They denote the conditional probability of obtaining a (mean) MCC value with at least the magnitude observed on a given test set, supposing a random predictor had been built on the corresponding training set. The Bonferroni correction provides an upper bound, equal to the value below which the corrected p-values are deemed statistically significant, to the conditional probability, supposing an unknown number of approaches performed like random predictors on average, of erroneously declaring that any model (or random selection from a set of corresponding models for the pseudo-stochastic algorithms) built, in this work, on one of the training sets would perform differently, in terms of the mean MCC across various test sets, to a random predictor - based on the (mean) MCC value observed on the single corresponding test set."

c. **Errors in this analysis:**

   i. For a single MCC value corresponding to a given train/test partition, obtained using a single model rather than the average across multiple runs of a pseudo-stochastic modelling approach, the corresponding uncorrected p-value denotes the conditional probability of obtaining an MCC value - on any test set - with at least the magnitude observed on the given test set, supposing a random predictor had been built on the corresponding training set. If the p-value is less than a pre-defined limit, the observed performance may be declared to be statistically significantly different to a random predictor. However, it is suggested in this thesis that the p-value calculated from the mean MCC value obtained for a single train/test partition for a pseudo-stochastic modelling approach, which generates different models for a given training set, would allow the average performance of that modelling approach, applied to that training set, to be declared statistically significantly different to the performance expected with an approach which only generated random predictors on that training set. This reasoning may not be mathematically valid.

   ii. In general, for a collection of models, the Bonferroni correction provides an upper bound, equal to the value below which the corrected p-values are deemed statistically significant, to the conditional probability, supposing an unknown number of models were random predictors, of erroneously declaring that any model built on one of the training sets would perform differently to a random predictor - based on the single MCC value observed on the corresponding test set. In the current context, this upper bound was also supposed to hold for the conditional probability of erroneously declaring that one of the pseudo-stochastic methods – when trained repeatedly on a given training set – would perform differently to a method which only generated random predictors on that training set. Again, this reasoning may not be mathematically valid.

2. The statistical analysis carried out in chapters 5 and 6 (see sections 5.3.5 and 6.4.3) was flawed and this *may* affect the remarks regarding statistical significance and the lack of clear differences in method performance commented upon in the abstract of this thesis. The most appropriate means for testing the two null-hypotheses presented in section 5.3.5 remains unclear, as was originally noted in section 5.3.5. However, in hindsight, the following additional comments may be made:
    a. The proposed *ad hoc* approaches for evaluating the statistical significance of (differences between) overall mean values of performance metrics, averaged across different possible train/test partitions and RNG seeds, are logically flawed e.g.
        i. The variability associated with different train/test partitions, obtained using a single RNG seed, should not have been evaluated as the statistical significance associated with the overall mean performance was of interest.
        ii. By generating multiple p-values, in a stepwise approach, when nominally evaluating the statistical significance associated with a single overall mean performance metric value– or a single pairwise difference between two overall mean performance metric values – this will have artificially inflated the number of p-values and hence skewed the corresponding p-value adjustments according to Benjamini and Yekutieli's method.
    b. Multiple hypothesis corrections for Chapter 5 should not have double counted pairwise comparisons between 2D descriptors on different hERG datasets that were only different due to different conformations.