Building models of bacterial mutagenicity from biased training data

Chris G. Barber[1], Thierry Hanser[1], Naomi L. Kruhlak[2], Lidiya Stavitskaya[2], Jonathan D. Vessey[1], Stéphane Werner[1].

1 Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Holbeck, Leeds LS11 5PS UK

2. FDA Center for Drug Evaluation and Research (CDER), Silver Spring, MD, 20993, USA

Abstract 541
Poster Board 259

Introduction

Models for predicting bacterial mutagenicity are now widely used by pharmaceutical sponsors to assess the genotoxic potential of impurities in pharmaceutical products. Models built using machine learning (ML) techniques are commonly trained using balanced datasets where, in this case, equal numbers of compounds are positive and negative for mutagenicity. Building accurate models using ML from biased training data – unequal numbers of positive and negative compounds – can be a challenge. Sarah Nexus is a program for predicting bacterial mutagenicity that uses a self-organising hierarchical network (SOHN). Hitherto SOHN models have been built using data that have little bias; however, if models are built using biased training data, then there is a need to ensure that the model learns sufficiently well about the minor class. If the dataset is biased towards negative compounds, this would result in a model for mutagenicity with depressed sensitivity.

Approaches to dealing with training set data bias

In the figures below, green points represent negative data and red points represent positive data: the negatives outnumber the positives 2:1 (Figure 1). Two possible approaches to dealing with data set bias are: oversample the minor class (Figure 2) or under-sample the major class (Figure 3).

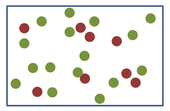

Figure 1: Dataset biased 2:1 negatives

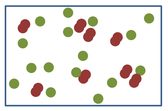

Figure 2: Oversample the minor class

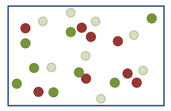

Figure 3: Undersample the major class

Three possible approaches to under-sampling the major class are: take the data closest to points in the minor class (Figure 4); take the most diverse set of data from the major class (Figure 5) or repeatedly sample at random from the major class (Figure 6).

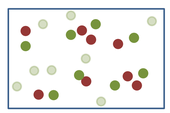

Figure 4: Major class data points closest to minor class data points

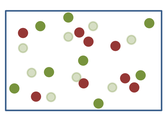

Figure 5: The most diverse set of data from the major class

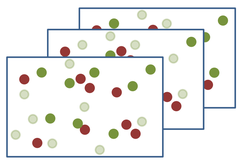

Figure 6: Repeatedly sample at random

Where multiple models have been built, their predictions must be resolved. Two ways of doing this are using a simple majority vote (Figure 7) or taking the single most confident prediction (Figure 8)

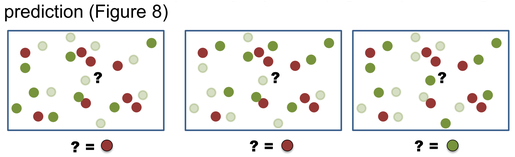

Figure 7: Using a simple majority vote between these three models, an unknown compound at ? would be predicted as positive by 2 votes to 1 (assuming that, in this illustration, the prediction is made from the single nearest neighbour).

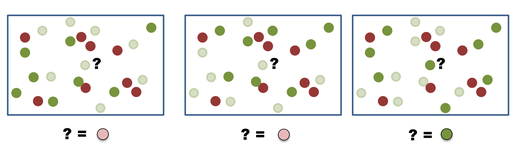

Figure 8: Using a single most confident prediction to decide between the three models in Figure 7 would give an overall prediction for an unknown compound at ? as negative (assuming that confidence in prediction is a function of distance in chemical space).

Oversampling *vs.* Undersampling *vs.* No correction

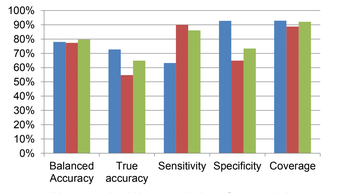

Figure 9: Comparison of performance metrics for SOHN models for Ames activity in S. typhimurium strain TA100 built from uncorrected, undersampled and oversampled training sets. The metrics are defined using the number of true or false positives (TP, FP), true or false negatives (TN, FN) and number of compounds in the test set (**N**). Sensitivity = TP/(TP + FN). Specificity = TN/(TN + FP) True accuracy = (TP + TN) /**N**. Balanced accuracy = (Sensitivity + Specificity)/2. Coverage = (TP + FP + TN + FN)/**N**

Figure 9 shows the prediction performance of Sarah SOHN models of Ames activity in Salmonella typhimurium strain TA100 for a set of 7007 compounds. The training set is biased 3.6:1 in favour of Ames negative compounds. As a result the models learn less about the Ames positive compounds unless the model training set is corrected. Either oversampling the Ames positives or undersampling the Ames negatives (using a single pass of selecting at random from the Ames negatives) leads to an increase in sensitivity at the expense of specificity, with little change in balanced accuracy or coverage.

The challenge then is to retain the improvements in sensitivity resulting from balancing the training set, while minimising the loss in specificity.

Different undersampling techniques

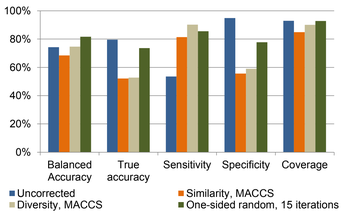

Figure 10: Comparison of performance metrics for SOHN models for Ames activity in S. typhimurium strain TA100 built from an uncorrected training set and training sets undersampled in the major class (Ames negatives) using two different techniques

When undersampling the major class to generate a balanced training set, the methods in Figures 4 – 6 gave different performance metrics as shown in Figure 10.

Chemical diversity can be measured using the presence or absence of chemical substructures or 'keys'; one common set of keys are MACCS[1] keys. Selecting data in the major class (Ames negatives) based on their chemical diversity as judged by MACCS keys produced the best sensitivity. However, good sensitivity balanced with only a modest reduction in specificity compared to the uncorrected training set was obtained by repeated random sampling of the major class.

Different methods of combining predictions

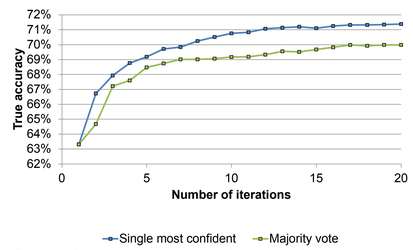

Figure 11: Variation in true accuracy of predictions for an all strains model of Ames mutagenicity with number of iterations of random sampling from the major class (negatives).

For ICH M7 it is important that a system produces a single, repeatable prediction rather than presenting the user with many different predictions. There is, therefore, an issue when multiple models are used to make a prediction – as is the case where a balanced training set is repeatedly generated by random undersampling of the major class: how to make a single prediction. In Figures 7 and 8 two possibilities are described: taking the single most confident prediction or taking a majority vote.

As more iterations of the model building from generated training sets takes place two things happen: firstly the coverage of the set of models increases – because there is an increase in the chance that a prediction for a given compound will be made by at least one model – and the confidence of the most confident prediction increases. Both of these changes reach an asymptote as the number of iterations increases.

Figure 11 shows how the true accuracy – which takes into account both accuracy and coverage – changes with increasing numbers of iterations. Note that the single most confident predictions tend to outperform overall predictions taken by a majority vote.

Different degrees of bias

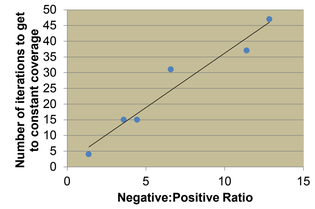

Figure 12: Scatterplot with correlation of the number of iterations of random sampling from the major class needed to produce a set of models whose coverage showed no further increase with additional iterations against the data set bias.

As shown in Figure 11, generating multiple models whose training set is balanced by random selection from the major class shows improvement in performance as more iterations of model building are undertaken.

One method of measuring this improvement in performance is coverage by the set of models – that is the number of compounds where a prediction is made by at least one of the models in the set.

Again as referred to above, the coverage tends to reach a constant value after a number of iterations. In Figure 12 the number of iterations needed to reach a constant level of coverage is plotted against the data bias for sets of models built for Ames mutagenicity against five different strains of Salmonella typhimurium and an all strains model.

Figure 12 shows that the relationship between the data bias and the number of iterations needed to get to a constant performance is approximately 3 - 5 times the bias, e.g. a bias of 11.4 :1 needs 37 iterations to get to constant coverage.

'Best of breed' changes in performance against uncorrected data

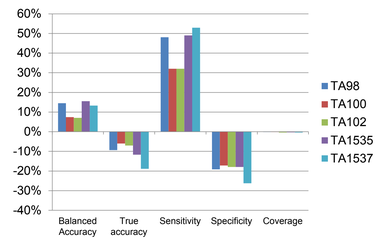

Figure 13: Variation in performance criteria for 'best of breed' models built by correcting data bias compared to models built without correcting for bias for prediction of Ames mutagenicity five different strains of Salmonella typhimurium. Changes are not relative measures, e.g. for TA98 the sensitivity of the model built without correcting for training set bias is 39%, whereas for the models built by balancing the training set the sensitivity is 87%, i.e. a difference of 48%. The bias towards negatives for each strain is as follows: TA98 6.6:1. TA100 3.6:1. TA102 4.4:1. TA1535 11.4:1. TA1537 12.8:1

Figure 13 shows the difference in performance measures that result in correcting for a biased training set when building SOHN models for Ames mutagenicity for five different strains of S. typhimurium.

When correcting for training set bias, the 'best of breed' methodology is to build an ensemble of models from training sets where the data from the major class is selected at random. The ensemble reports a single prediction, with a single explanation, which is the most confident prediction of those it has produced.

Figure 13 shows that large gains in sensitivity (prediction of the minor class) can be made although they incur a cost in specificity. Overall there are improvements in balanced accuracy, though the true accuracy (which is biased towards prediction of the major class) is reduced. Sufficient numbers of iterations result in minimal loss of coverage by correcting for training set bias.

Conclusions

Models with the best sensitivity are built from training sets balanced using a diversity approach. The cost of this sensitivity is very great however and this is reflected particularly in specificity and true accuracy. The more biased that dataset, the greater the cost in terms of coverage of building a single model with the major class selected by its diversity.

Balanced accuracy is relatively insensitive to dataset bias, but the effects of different approaches to selecting a balanced training set are increasingly apparent as the dataset from which the training set is drawn becomes more unbalanced.

True accuracy – which takes into account the prediction coverage – is maximised using an approach of repeatedly selecting from the major class and aggregating models by taking the single most confident prediction for each compound. In general, the number of iterations of the approach needed to reach the best performance is 3 – 5 times the ratio of the major class to the minor one.

References and Notes

1. Durant, J.L *et al*. J. Chem. Inf. Comput. Sci. 2002, **42**, 1273-1280.

This work was performed under a Research Collaboration Agreement (RCA) between Lhasa Limited and FDA/CDER. The findings and conclusions in this presentation have not been formally disseminated by the FDA and should not be construed to represent any agency determination or policy. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.