Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220496

Optimization of Performance by Combining Most Sensitive and Specific Models in Data Science Results in Majority Voting Ensemble

Katoo M. MUYLLE[a], Pieter CORNU[a], Wilfried COOLS[b], Kurt BARBÉ[b], Ronald BUYL[b] and Sven VAN LAERE[b,1]
[a] *Centre for Pharmaceutical Research (CePhar), Vrije Universiteit Brussel, Belgium*
[b] *Department of Public Health, Vrije Universiteit Brussel, Belgium*

Abstract. Ensemble modeling is an increasingly popular data science technique that combines the knowledge of multiple base learners to enhance predictive performance. In this paper, the idea was to increase predictive performance by holding out three algorithms when testing multiple classifiers: (a) the best overall performing algorithm (based on the harmonic mean of sensitivity and specificity (HMSS) of that algorithm); (b) the most sensitive model; and (c) the most specific model. This approach boils down to majority voting between the predictions of these three base learners. In this exemplary study, a case of identifying a prolonged QT interval after administering a drug-drug interaction with increased risk of QT prolongation (QT-DDI) is presented. Performance measures included accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Overall performance was measured by calculating the HMSS. Results show an increase in all performance measure characteristics compared to the original best performing algorithm, except for specificity where performance remained stable. The presented approach is fairly simple and shows potential to increase predictive performance, even without adjusting the default cut-offs to differentiate between high and low risk cases. Future research should look at a way of combining all tested algorithms, instead of using only three. Similarly, this approach should be tested on a multiclass prediction problem.

Keywords. Artificial Intelligence, Performance Measures, Majority Voting, Ensemble Learning, Drug Interactions

1. Introduction

In the field of data science, it is common practice to use the ensemble modeling technique. In that technique, the word *ensemble* literally means together, where the data scientist uses different base models that work together to get a final outcome.[1] In ensemble modeling, literature often refers to two basic techniques, i.e. bagging and boosting.[2-4]

In bagging, the researcher first selects a random sample that serves as input to a certain model to train that model, and repeats this process multiple times. In the end, all different trained models are aggregated to reduce the variance, for example by using the majority voting technique.[4, 5] In boosting, different models are build sequentially,

[1] Sven Van Laere (Corresponding author), Department of Public Health, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Jette, BELGIUM; E-mail: Sven.Van.Laere@vub.be.

where the output of the previous model is used as input for the next model in order to reduce bias in the outcome.[6] Eventually, all models are then used together to obtain an overall prediction.

In this manuscript, the authors did not prespecify what is meant by 'models' in the explanation of bagging and boosting written above. The simplest forms of these models are a decision tree (in case of a categorical outcome) or a decision tree regressor (in case of a continuous outcome), but other modelling techniques (e.g. linear or logistic regression models) can also be used in these ensemble modelling techniques.

The performance of a final model is often expressed in terms of accuracy, precision, recall, or another performance characteristic.[7] In general, it is difficult to decide which model is the best model on overall, since different performance characteristics tend to focus on different desirable characteristics (e.g. having few false negatives). Therefore, researchers invented performance measures like the F1-score that combines both the performance measures precision (positive predictive value, PPV) and recall (sensitivity) by taking its harmonic mean to be able to compare models.

In a clinical setting, such models can be incorporated into clinical decision support systems (CDSS) to determine when an alert to a healthcare professional should be triggered to support decision-making. The developer of such a CDSS is not solely interested in identifying positive cases (i.e., patients with a certain disease or symptom), but should also focus on negative cases (i.e., patients that do not have a certain disease or symptom).[8] This can be explained by the fact that a major drawback of improving the sensitivity of a CDSS is that it becomes less specific, potentially leading to alert fatigue of that healthcare professional.[9]

In this work, the authors are interested in increasing overall performance by complementing the best performing model (selected based on its harmonic mean of sensitivity and specificity (HMSS)) with models having the highest sensitivity and the highest specificity.

2. Methods

In an effort to increase model performance on a dataset for predicting QT prolongation after exposure to a drug-drug interaction with increased risk of QT prolongation (QT-DDI), the authors used several algorithms (logistic regression, Gaussian Naive Bayes classifier, support vector machine, random forest, gradient boosting, etc.) to predict high and low risk patients. Prevention of drug-induced QT prolongation can in turn prevent a potentially lethal type of ventricular tachycardia called Torsades de Pointes (TdP)). In a systematic review, Arunachalam et al.[10] reported an incidence of drug-induced QT prolongation of 6.3%, while the incidence of TdP was 0.33% in patients exposed to drugs that prolong the QT interval.

In an internal validation phase, 350 patients were modelled using 10-fold stratified cross-validation. Subsequently, an external validation was performed on 110 new patient cases to find a best overall model. Performance measures were reported in terms of accuracy, sensitivity, specificity, PPV, and negative predictive value (NPV). Overall performance was measured in terms of HMSS. The authors combined the best overall performing algorithm, with the most sensitive and the most specific model as demonstrated in Table 1.

This new approach of combining prediction results can be viewed as using one default prediction outcome (algorithm 1), with two complementary prediction outcomes

of which one focusses on predicting positive cases (i.e., high risk cases – algorithm 2) and one focusses on predicting negative cases (i.e., low risk cases – algorithm 3). When algorithm 2 predicts a high risk case, the combined prediction will predict a high risk case as outcome. Only when algorithm 3 predicts a low risk case at the same time, the combined outcome falls back to the default algorithm 1. Similarly, when algorithm 3 predicts a low risk case, the combined algorithm will return this prediction. Also here, the combined algorithm falls back to the default algorithm 1 when algorithm 2 predicts to be a high risk case and algorithm 3 predicts to be a low risk case. In all other cases, the default prediction outcome of the best overall algorithm 1 was used.

Table 1. Prediction strategy combining the best overall, sensitive (precision) and specific (recall) algorithm.

	Prediction algorithm 1 *(best overall)*	*Prediction algorithm 2* *(best sensitivity)*	*Prediction algorithm 3* *(best specificity)*	*Combined prediction*	*Algorithm to follow*
Case 1	Low risk	Low risk	Low risk	Low risk	3
Case 2	Low risk	Low risk	High risk	Low risk	1
Case 3	Low risk	High risk	Low risk	Low risk	1
Case 4	Low risk	High risk	High risk	High risk	2
Case 5	High risk	Low risk	Low risk	Low risk	3
Case 6	High risk	Low risk	High risk	High risk	1
Case 7	High risk	High risk	Low risk	High risk	1
Case 8	High risk	High risk	High risk	High risk	2

3. Results

In Table 2, the results of the external validation were reported. The performance increased by combining the best overall performing algorithm (based on HMSS) with the best sensitive and the best specific algorithm.

The overall performance, measured by the HMSS, of that combined algorithm increased with about 3% (i.e., from 63.49% in the best overall performing model (algorithm 1) to 66.57% in the new combined form). The accuracy increased with about 5.5% (from 63.64% to 69.09%), the sensitivity increased with about 6% (from 62.50% to 68.75%), the specificity remained the same (64.52%), the PPV increased with about 2.5% (from 57.69% to 60.00%) and the NPV increased with about 4% (from 68.97% to 72.73%). (see Table 2)

This combined form of prediction can – in a way – be interpreted as a special form of majority voting ensemble. The *ensemble* part in that wording refers to the three algorithms that are used to construct a new combined prediction and the *majority voting* part in that wording refers to the combined prediction that is in fact based on the majority of votes over the three algorithms (see Table 1). The majority vote resulted in a positive or negative case (i.e. respectively high or low risk for QT prolongation after administering a QT-DDI) when respectively two or three of these selected models predicted a positive or negative case.

Table 2. Summary of performance characteristics obtained after external validation. Performance characteristics increased by applying ensemble modeling of the best performing algorithms.

Algorithm number	1	2	3	4	…	Combined
Accuracy	0.6364	0.4818	0.6909	0.6182		0.6909
Sensitivity	0.6250	**0.9375**	0.4167	0.5833		0.6875

Specificity	0.6452	0.1290	**0.9032**	0.6452		0.6452
PPV	0.5769	0.4545	0.7692	0.5600		0.6000
NPV	0.6897	0.7273	0.6667	0.6667		0.7273
Harmonic mean	***0.6349***	*0.2268*	*0.5703*	*0.6127*		*0.6657*

Algorithm 1: Logistic regression with forward model building (best overall); Algorithm2: Gaussian Naive Bayes classifier (best sensitivity); Algorithm 3: Random forest with feature selection (best specificity); Algorithm 4: Random forest with feature selection; PPV: positive predictive value; NPV: negative predictive value; Harmonic mean: harmonic mean of sensitivity and specificity (HMSS). Only algorithm 1, 2 and 3 were used to construct the new combined prediction outcome.

4. Discussion

This study presents a way of increasing predictive performance by applying majority voting ensemble by combining the three most performant algorithms in terms of overall performance, sensitivity and specificity. The reason for using sensitivity and specificity stems from the fact that CDSS should both focus on identifying the positive and negative cases in order to keep alert fatigue as low as possible. This paper presents a case of finding an optimized prediction of a high risk for QT prolongation in patients exposed to QT-DDIs. It was shown that, by combining these three models or algorithms, all performance characteristics increased or remained equal to that of the original best performing prediction model.

The strength of this approach is that it is fairly simple to obtain this increased performance: (a) consider a number of prediction algorithms, (b) conduct the training and evaluation of these models based on the internal validation dataset, (c) test all algorithms with an external validation dataset, and then (d) constitute a new combined model based on the performance characteristics of all models. A limitation of this approach is that it requires many different models before being able to select a model with a high enough sensitivity and a high enough specificity in order to find a possibility to increase overall performance. Another limitation to this approach is that it depends on only two measures (i.e., sensitivity and specificity). Moreover, these two performance measures can be increased or decreased according to the chosen cut-off that the researcher applies. For example, in logistic regression, a traditional cut-off is 0.50 for the predicted probability of a certain event (e.g., QT prolongation) in order to predict whether an event will occur or not.[11] Sensitivity analysis might find a more optimal cut-off in order to optimize the sensitivity-specificity trade-off. Another limitation is that algorithms that have 100% sensitivity or 100% specificity, e.g. the case when the algorithm either classifies each data instance as positive or negative, will not increase the performance.

Therefore, future work will additionally implement a sensitivity analysis on all algorithms when performing the external validation in order to find an appropriate cut-off before combining classifiers. This will prevent algorithms from having an utterly high/low sensitivity or an utterly high/low specificity. In this paper, we showed an example of a binary outcome (high or low risk for QT prolongation) by testing the performance of different classifiers. Future work should also assess the potential of using this technique in a multiclass setting. Moreover, to prevent this new approach from only taking the three best performing algorithms (respectively in terms of overall performance, sensitivity and specificity), future work might exist in finding a way to combine the information of all different models to boost the overall performance in an optimal way.

5. Conclusion

In this study, the authors showed that combining the best performing overall classifier with the classifiers having the best sensitivity and best specificity, increased predictive performance. The approach is fairly simple once having multiple trained classifiers ready to be tested on an external dataset. The presented method boils down to a type of majority voting ensemble where three algorithms are combined using the most present predicted outcome as final outcome.

References

[1] Zhou ZH, "Ensemble Learning," in Machine Learning, Zhou ZH, Ed. Singapore: Springer Singapore, 2021, pp. 181-210. doi: https://doi.org/10.1007/978-981-15-1967-3_8.
[2] Breiman L, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996/08/01 1996, doi: https://doi.org/10.1007/BF00058655.
[3] Freund Y, Schapire RE. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997/08/01/ 1997, doi: https://doi.org/10.1006/jcss.1997.1504.
[4] Kuncheva LI, Skurichina M, Duin RPW. "An experimental study on diversity for bagging and boosting with linear classifiers," Information Fusion, vol. 3, no. 4, pp. 245-258, 2002/12/01/ 2002, doi: https://doi.org/10.1016/S1566-2535(02)00093-3.
[5] Altman N, Krzywinski M. "Ensemble methods: bagging and random forests," Nature Methods, vol. 14, no. 10, pp. 933-934, 2017/10/01 2017, doi: https://doi.org/10.1038/nmeth.4438.
[6] Webb GI, Zheng Z. "Multistrategy ensemble learning: reducing error by combining ensemble learning techniques," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 8, pp. 980-991, 2004, doi: https://doi.org/10.1109/TKDE.2004.29.
[7] Sokolova M, Japkowicz N, Szpakowicz S. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in AI 2006: Advances in Artificial Intelligence, Berlin, Heidelberg, A. Sattar and B.-h. Kang, Eds., 2006// 2006: Springer Berlin Heidelberg, pp. 1015-1021. doi: https://doi.org/10.1007/11941439_114.
[8] Payne TH, et al. "Recommendations to improve the usability of drug-drug interaction clinical decision support alerts," (in eng), J Am Med Inform Assoc, vol. 22, no. 6, pp. 1243-50, Nov 2015, doi: https://doi.org/10.1093/jamia/ocv011.
[9] Légat L, Van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. "Clinical Decision Support Systems for Drug Allergy Checking: Systematic Review," (in eng), Journal of medical Internet research, vol. 20, no. 9, pp. e258-e258, 2018, doi: https://doi.org/10.2196/jmir.8206.
[10] Arunachalam K, Lakshmanan S, Maan A, Kumar N, Dominic P. "Impact of Drug Induced Long QT Syndrome: A Systematic Review," (in eng), J Clin Med Res, vol. 10, no. 5, pp. 384-390, May 2018, doi: https://doi.org/10.14740/jocmr3338w.
[11] Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression, 3rd Edition ed. Wiley, 2013.