Automated Warfarin Dosing for South Africans with Evolutionary Ensemble Learning

Neville Varney-Horwitz
VRNNEV001
University of Cape Town
vrnnev001@myuct.ac.za

ABSTRACT

In this paper, we investigate the use of ensemble machine learning techniques in tandem with evolutionary strategies for hyperparameter tuning for the application of predicting Warfarin doses based on patient data.

CCS CONCEPTS

• **Applied computing → Consumer health**; *Health care information systems*; Genetics; • **Computing methodologies → *Supervised learning***; *Machine learning approaches*;

1 INTRODUCTION

Modern healthcare practices often rely on the use of substances that, in inappropriate amounts, can cause serious damage to the human body. Warfarin, a drug prescribed to patients with illnesses related to excessive blood coagulation, is no exception. In fact, Warfarin is considered to have an abnormally narrow therapeutic range: healthcare professionals are required to determine the accurate dose for a given patient with very little room for error. This problem is compounded by the fact that a given patient's ideal Warfarin dose is dependent on a multitude of factors; as well as highly sensitive to changes in diet and other medication[7, 14].

In the search for a reliable means of determining the correct dose given available medical data, a range of approaches have been tried over the decades. With the rising popularity of machine learning, several such techniques have been employed with promising results. Pathcare, a laboratory testing services pathology practice, initiated this project as a proof of concept for further research in this direction. Given the individual success of Multi-Layer Perceptron models, Bagging ensembles and Evolutionary Algorithms in the past[6, 9, 12, 13, 15], we aim here to apply these three methods in tandem.

The two broad aims of this paper are as follows:

- Determine whether ensemble models perform better on local versus international data.
- Determine to what extent additional features in the data affect the performance of ensemble models.

2 BACKGROUND

2.1 Warfarin

Warfarin serves as a highly effective anticoagulant; primarily used to treat thrombosis. Such is its effectiveness and widespread use, the World Health Organisation included it in its List of of Essential Medicines. Warfarin decreases the availability of vitamin K_1 which directly reduces the efficacy of the major clotting factors in the body. An effective dose of Warfarin is thus dependent on all variables that affect the presence of vitamin K_1, including diet, enzyme-inducing comedications, as well as relevant genotypes. The genotypes for CYP2C9 and VKORC1 (along with their polymorphisms) encode for enzymes that are responsible for Vitamin K availability and are thus of particular interest.

Currently, standard practice involves administering an initial dose and changing it over the course of several frequent checkups until the patient's condition is stable. The main indicator used by clinicians to determine blood clotting tendency is the internationalised normalised ratio (INR) value[8]. In general terms, this measurement involves timing how long a sample of blood takes to clot. In most cases, the target INR range is between 2 and 3[2], with less indicating a higher dose being administered and vice versa.

2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) seek to emulate the biological process of learning. A set of input nodes (or "neurons") feeds into one or more layers of internal, "hidden", nodes, each with an associated weight. At each node, the input signal combined with the weight and, depending on the implementation, may be cut off if it does not reach a chosen threshold. The output of the network is given by the value(s) of the node(s) in the final layer. In the training phase, the output will be compared with the expected output and the magnitude of the difference will then be used to alter the weights of the hidden nodes through a process known as *back-propagation*. This is repeated until the network meets the desired specifications. Numerous parameters affect the performance of an ANN: in particular its architecture (number and size of the hidden layers) as well as its learning rate (how aggressively the weights are altered at the end of each training iteration)[5].

2.3 Ensemble Methods

Ensemble methods in the context of machine learning involve the aggregation of multiple "base" machine learning models to achieve better results than their individual use. This definition is purposefully broad in that this can occur in a multitude of configurations. The most notable possibilities for variation are as follows[1]:

- The type of base model can either remain uniform or heterogeneous.
- Each of the base models can be trained with the entirety of the dataset or a sample thereof.
- Each of the base models can be trained using all or some of the features included in the dataset.

UCT, CSC4000W, May 2018
2018.

- The sampling of data points or features can be done with or without replacement.
- The final output of the ensemble can either entail a simple averaging of the base models' predictions, a weighted average or the result of passing the input through a sequence of estimators.

Here we will focus on homogeneous ensembles that make use of simple averaging. In the case that this is done alongside sampling the dataset with replacement, this is known as bootstrap aggregating (Bagging). When features are sampled with replacement, this is referred to as the Random Subspaces method. Finally, a combination of these two is known as the Random Patches method.

2.4 Evolutionary Algorithms

Evolutionary Algorithms are a family of optimisation methods which mimic the process of natural selection. Individuals take the form of potential solutions to a given problem and a population consists of a collection of these. The population undergoes successive generations via mating, mutation and selection[10].

- Mating/Crossover: this operator involves blending the characteristics of two or more individuals to form a new one.
- Mutation: a unary operator that makes a minor change to the individual's position in the solution space.
- Selection: The driving force of evolution, this algorithm selects candidates for the next generation based on their fitness. A common selection strategy is that of the k-tournament: k individuals are randomly selected from the population and the fittest between them is selected.

The composition of successive generations can also be chosen. The (μ, λ) method entails generating λ children from the parents and then selecting μ of them to form the next generation; whereas the $(\mu + \lambda)$ method also generates λ children but the selection process involves the parents as well.

A subset of the EA family is known as Evolutionary Strategies. ES are primarily used for numerical optimisation and differ from typical Evolutionary Algorithms in that each individual is made up not only of a candidate solution vector; but also a "strategy" vector which affects the magnitude by which the solution vector mutates[3].

3 DATA PREPARATION

The first dataset, supplied to us by PathCare, consisted of approximately 8900 records of South African patient visits. Each record consisted of the patient's sex, date of birth, previous dose, current INR, target INR, prescribed INR as well as their particular conditions and other medications.

The second dataset available to us was that of the International Warfarin Pharmacogenetics Consortium (IWPC) which contained the combined collected data of several international pharmacogenomics research centers and made freely available online. This dataset consisted of 4798 records containing a patient's sex, race, age (given as a 10 year range), height, weight, comorbidities, medications, smoking habits, stable INR, CYP2C9 and VKORC1 genotypes, and finally their stable dosage.

Considering that the performance of statistical models on each dataset was going to be compared, several adjustments needed to be made for a fair experiment. Firstly, the IWPC dataset only contained records for patients who had already reached their therapeutic INR range. Thus the PathCare dataset was filtered, leaving approx. 4600 records wherein the INR was therapeutic. This result was serendipitous as now both datasets were of roughly equal size, removing another variable.

Secondly, the IWPC dataset was altered such that the age reflected the middle value of the decade range originally given and the PathCare dataset altered to transform date of birth to current (2018) age in years. The latter change is indeed flawed as the age given should be the age at the time of dosage, however this was chosen option considering that no data was available for when the check-up was made.

Thirdly, since the PathCare columns relating to comedicaions and comorbities were highly irregular (consisting of multiple values in each entry, spelling mistakes and high variability), it was decided that for this purpose the only usable input features were age, sex and INR value. We will dub this feature set 1.

The experimental design was thus amended such that a variety of models be tested on both datasets, each limited to these features. The models would then be applied to the IWPC dataset with the feature set expanded to include race, height, weight, and whether or not the patient was a current smoker (feature set 2). The significance of this dataset is that it consists of measurements that can easily be made in a clinical context. The models would also be evaluated on the IPWC dataset in almost its entirety (feature set 3). This would include feature set 2 as well as: use of Amiodarone (an enzyme inducer), CYP2C9 genotype, VKORC1 (along with its polymorphisms) genotypes where fewer than 20 percent of the entries were empty. The other features were dismissed for containing a majority of empty or non-applicable values.

The primary base model to be used was the Artificial Neural Network. In the case of ANNs, input values which vary greatly in range can lead to undesirable behaviour - such as particular features with large values desensitising neurons to features with typically smaller values. To prevent this, all columns in each dataset were regularised by subtracting each value from the column's mean and dividing the result by the standard deviation. All data manipulation was implemented using the Python Pandas and Numpy libraries.

The IWPC and PathCare datasets were then each split into training and testing sets with an 80:20 ratio. The models would be experimented with and optimised using the training set and then a final evaluation made by comparing predictions with the test set. The two evaluation metrics used were Mean Absolute Error (MAE) and PW20: the percentage of predictions made within 20% of the actual values. The latter is of particular significance considering that a patient's INR value only need reach a given range and therefore a range of dosages would suffice for this purpose.

4 EXPERIMENT IMPLEMENTATION

Due to its multiplicity of relevant packages and use in the literature, Python 3.5 was chosen as the language of implementation. The BaggingRegressor, RandomForestRegressor, GradientBoostingRegressor, MLPRegressor and LinearRegression classes were used from the sklearn[11] package. Evolutionary Algorithm related code was written with use of the DEAP[4] package.

The three ensembles to be optimised using EAs were Bagging, Random Subspaces and Random Patches, all of which using a Multilayer Perceptron as the base model. In each case, individuals consisted of vectors containing the number of neurons in the MLP's hidden layer, the number of estimators in the ensemble, the percentage of features to use when training each base model (where applicable) and the percentage of data points to use when training each base model (where applicable). Three EA techniques were attempted: a traditional (μ, λ) Evolutionary Strategy, Covariance Matrix Adaptation Evolution Strategy and a $(\mu + \lambda)$ Evolutionary Algorithm.

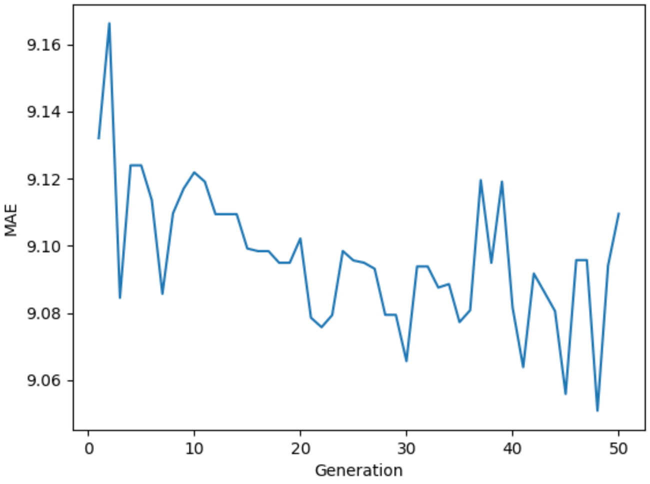

Figure 1: Minimum MAE for each generation using (μ, λ) ES

In each case individuals were evaluated by training an ensemble with the corresponding parameters and running a 10-fold cross-validation on the training data. The (μ, λ) Evolutionary Strategy proved most successful as it most easily kept parameters within a natural range. For this implementation, the values of μ and λ were 5 and 20 respectively; and the number of generations, 50. These numbers had to be kept relatively low due to extended evaluation times for individuals. Crossover occurred with a probability of 0.6 and and the cxESBlend routine from the DEAP library was used. Mutation occurred with a probability and DEAP's mutESLogNormal was used. Other variations which did not yield better results include framing the problem as a multiple-objective optimisation one wherein PW20 was maximised while MAE simultaneously minimised as well as allowing ensembles to be made up of MLPs with differing numbers of neurons.

Model	PathCare F1		IWPC F1		IWPC F2		IWPC F3	
	MAE	PW20	MAE	PW20	MAE	PW20	MAE	PW20
Linear Regression	10.83	0.36	12.34	0.32	11.37	0.36	11.12	0.39
Multilayer Perceptron	11.00	0.37	12.23	0.33	12.37	0.35	14.05	0.26
MLP Bagging	10.97	0.36	12.62	0.33	11.85	0.35	14.44	0.24
MLP Random Subspaces	10.98	0.35	12.48	0.33	12.23	0.35	13.23	0.27
MLP Random Patches	11.10	0.34	12.53	0.33	12.73	0.33	15.37	0.22
Random Forest	12.73	0.29	13.63	0.31	12.70	0.32	14.67	0.24
Gradient Boosting	11.02	0.35	12.33	0.33	11.74	0.36	13.45	0.27

5 RESULTS AND CONCLUSION

Unexpectedly, the MLP ensembles' performance on the testing set was inversely proportional to their validation scores on the training set (their validation scores had increased with the number of features available). We see here that many of the ensembles performed successively worse on feature sets 1, 2 and 3. An explanation for this may be that the models were overfitted, or that the data between the training and testing splits was insufficiently balanced.

The fact that Linear Regression outperformed various complex Machine Learning techniques in many cases indicates that the problem may not require as sophisticated methods to yield satisfactory results.

We also see that despite considerable efforts being made to optimise the MLP ensembles, they do not significantly outperform other (unoptimised) techniques such as Gradient Boosting. It may be the case that Evolutionary Algorithms are not best suited to this problem domain.

Between the PathCare and IWPC datasets using feature set 1, however, we see a definite increase in performance on the PathCare dataset across the board. This may be the case due to the fact that the population from which this dataset was collected is more genetically correlated; whereas the IWPC dataset contained data from individuals of different nationalities.

REFERENCES

[1] [n. d.]. 1.11. Ensemble methodsÂű. http://scikit-learn.org/stable/modules/ensemble.html

[2] G. W. Albers, J. E. Dalen, A. Laupacis, W. J. Manning, P. Petersen, and D. E. Singer. 2001. Antithrombotic therapy in atrial fibrillation. *Chest* 119, 1 SUPPL. (2001). https://doi.org/10.1378/chest.119.1

[3] Hans-Georg Beyer and Hans-Paul Schwefel. 2002. Evolution strategies – A comprehensive introduction. *Natural Computing* 1, 1 (01 Mar 2002), 3–52. https://doi.org/10.1023/A:1015059928466

[4] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research* 13 (jul 2012), 2171–2175.

[5] Simon Haykin. 1998. *Neural Networks: A Comprehensive Foundation* (2nd ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.

[6] Ya Han Hu, Fan Wu, Chia Lun Lo, and Chun Tien Tai. 2012. Predicting warfarin dosage from clinical data: A supervised learning approach. *Artificial Intelligence in Medicine* 56, 1 (2012), 27–34. https://doi.org/10.1016/j.artmed.2012.04.001

[7] Daniel E. Jonas and Howard L. McLeod. 2009. Genetic and clinical factors relating to warfarin dosing. *Trends in Pharmacological Sciences* 30, 7 (2009), 375–386. https://doi.org/10.1016/j.tips.2009.05.001

[8] T B L Kirkwood and S M Lewis. 1983. Requirements for thromboplastins and plasma used to control oral anticoagulant therapy. *WHO Tech Rep Ser* 687 (1983), 81–99.

[9] Rong Liu, Xi Li, Wei Zhang, and Hong Hao Zhou. 2015. Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database. *PLoS ONE* 10, 8 (2015). https://doi.org/10.1371/journal.pone.0135784

[10] Melanie Mitchell. 1998. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, USA.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[12] Ashkan Sharabiani, Adam Bress, Elnaz Douzali, and Houshang Darabi. 2015. Revisiting warfarin dosing using machine learning techniques. *Computational and Mathematical Methods in Medicine* 2015 (2015). https://doi.org/10.1155/2015/560108

[13] Ashkan Sharabiani, Houshang Darabi, Adam Bress, Larisa Cavallari, Edith Nutescu, and Katarzyna Drozda. 2013. Machine learning based prediction of warfarin optimal dosing for African American patients. In *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*. 623–628. https://doi.org/10.1109/CoASE.2013.6653999

[14] P S Wells, A M Holbrook, N R Crowther, and J Hirsh. 1994. Interactions of warfarin with drugs and food. , 676–683 pages. https://doi.org/10.7326/0003-4819-121-9-199411010-00009

[15] Qin Zhou, Joey Kwong, Jie Chen, Wenzhe Qin, Jin Chen, and Li Dong. 2014. Use of artificial neural network to predict warfarin individualized dosage regime in Chinese patients receiving low-intensity anticoagulation after heart valve replacement. *International Journal of Cardiology* 176, 3 (oct 2014), 1462–1464. https://doi.org/10.1016/j.ijcard.2014.08.062