Applications of Machine Learning to Warfarin Dosing

Literature Review

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

ABSTRACT

Determining the correct Warfarin dosing regimen is a process which relies on a multitude of factors individual to each patient. Incorrect dosage can lead to life threatening adverse effects; and thus the development of a reliable prediction model could have profound impact. Here we analyse several prior attempts to build such a model using machine learning. The superiority of ensemble methods and importance of correct feature selection form the key observations made by the end of the survey.

KEYWORDS

Automated Warfarin dosing, machine learning, artificial neural network, support vector regression, genetic programming, ensemble model

1 INTRODUCTION

Warfarin is a widely used anticoagulant, primarily administered in order to thin blood when treating thrombosis and pulmonary embolisms.[3] While it remains one of the most popular and effective medications for this purpose, the use of Warfarin is accompanied by considerable risk. Too conservative a dose; and the desired thrombolytic effects may not take place. On the other hand, an overdose easily causes excessive bleeding and can even lead to more severe complications such as skin necrosis. [9]

The situation is further worsened by the fact that the therapeutic range (TR) which lies between over- and underdosing is comparatively small and can vary significantly based on both clinical and genetic factors. Clinicians thus have to make very careful decisions when prescribing an initial dose. Once this has been made, the patient's prothrombin time (a measure of blood's tendency to clot) is measured and a standardised metric - known as the International Normalised Ratio - can be calculated.[5] This can be used by the clinician to inform their next dose and, after repeating the process over several days, can determine a "maintenance" dosage which will keep the patient within the therapeutic range. This trial-and-error based strategy begs for a more systemic approach. While manually-followed algorithms exist, their convenient use relies on simplifying the interdependence between the many variables involved.

Machine Learning techniques, however, excel at tackling high-dimensional data and are thus the natural choice if we wish to confront the problem without sacrificing any its complexity.[3] In this review, we survey representative examples of machine learning applied to the problem of determining an appropriate Warfarin dose given a set of patient data. In the following section, give descriptions of the various machine learning techniques used and consider the advantages and disadvantages thereof.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 MACHINE LEARNING TECHNIQUES

2.1 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).

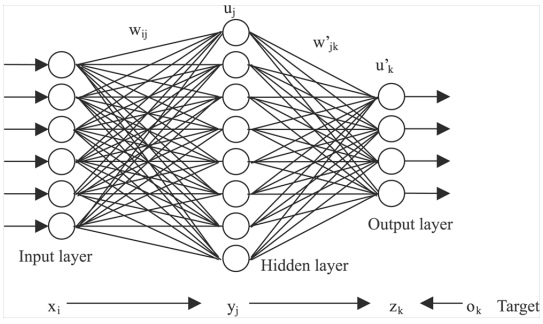

Figure 1: Example of An ANN Architecture

Hu et al. [3] experimented with a variety of ANN parameters as well as with which subsets of patients' dosages could be predicted best by this technique, however the transfer function used was not specified. The population was split according to whether Warfarin was begin taken concurrently with at least one of the forty medicines which are known to interfere severely with Warfarin's effects (Drug-to-drug interaction, or, DDI). In this case, the ANN performed best on patients without DDI. Furthermore, an architecture consisting of one five node hidden layer, a learning rate of 0.1 and an upper bound on training epochs of 600 proved optimal.

,

We note that this implementation may have been improved with a learning rate that decreases over the course of the training process.

Pavani et al. [6] also experimented with several architectures but also varied the transfer function. The result of the optimal architecture is not reported; but it is noted that the hyperbolic transfer function worked best. The network's training set was comprised of patients who were already in their TTR; and as a result it performed well at predicting maintenance doses for patients in the same category. Its performance suffered when predicting doses for patients outside of their TTR however; which points to the possibility of overfitting.

Solomon et al. []

2.2 Support Vector Regression

Support Vector Machines are a well established Machine Learning classification technique. The basic concept (in the linear case) is to determine a hyperplane which maximises its distance from the closest point in the feature space. Support Vector Regression aims to turn this model into a predictive one by learning a vector function

$$f(\mathbf{x_i}) = \mathbf{w} \cdot \mathbf{x_i} + b \tag{1}$$

which best estimates the training data, subject to the minimisation of $\mathbf{w} \cdot \mathbf{w}$.

Hu et al. [3] made use of a "ϵ-insensitive" loss function which penalises the system only if a data point occurs a distance greater than *epsilon* from the hyperplane. To make the problem soluble, slack variables $\xi_i, \xi_i^* \geq 0$ are incorporated which allow point i to deviate further than ϵ in each direction. In other words, the model is now subjected to minimising the sum of the squared norm of $\mathbf{w}$ and the weighted slack variables, as well as to the constraints

$$y_i - \mathbf{w} \cdot \mathbf{x_i} - b \leq \epsilon + \xi_i \tag{2}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b - y_i \leq \epsilon + \xi_i^* \tag{3}$$

where y_i is the target value for $\mathbf{x_i}$. Of the nine other methods they tried, this SVR implementation was the second most successful when applied to a population which contained patients both with and without DDI - only bested by the model which used an ensemble of SVRs.

Cosgun et al. [1] also used an ϵ-SVR except instead of the dot product $\mathbf{w} \cdot \mathbf{x_i}$, they used a Gaussian kernel

$$K(\mathbf{w}, \mathbf{x_i}) = e^{-\gamma ||\mathbf{w} - \mathbf{x_i}||^2} \tag{4}$$

The constant γ and the weight of the slack variables during minimisation where determined empirically with five-fold cross validation. This study took a combination of clinical and pharmacogenetic features in the form of single nucleotide polymorphisms (SNPs). The feature set was varied to use between 0 and 500 SNPs. SVR performed best with a feature set containing clinical factors as well as 200 SNPs. Performance was greatly sabotaged when SNPs were excluded from the feature set. Compared to the other methods used in the paper, SVR outperformed Boosted Tree Regression on average; but yielded slightly worse results than Random Forest Regression.

Liu et al. [4] compared the use of various ML techniques on patients of different ethnicities. SVR performed in the top three (of nine) techniques across all three of the different ethnicities considered. A notable observation was that it outperformed all the other methods when the patient required a lower than normal dose and was the third best at predicting a higher than normal dose. Its predictions of intermediate dosages were comparatively its weakest point.

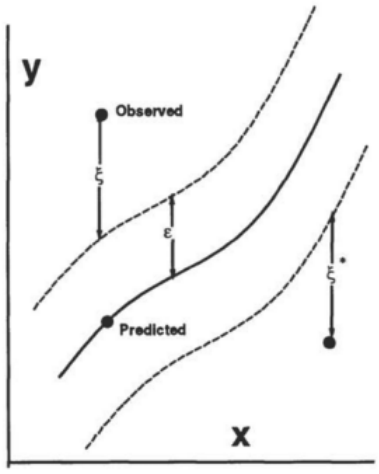

Figure 2: Low Dimensional Representation of SVR

2.3 Ensemble Methods

Ensemble methods involve the combination of results from several individual Machine Learning methods. Two common variations are given here.

2.3.1 Bagging. *Bagging* is a homogeneous ensemble method in that relies on a single base ML technique. The data set is bootstrapped (randomly sampled with replacement) repeatedly to generate several derivative data sets. The base method is then applied to learn a model for each of these data sets and, once trained, prediction is made by averaging their outputs.

Hu et al. [3] applied bagging to four different base algorithms: K-nearest neighbours (kNN), SVR, ANN, and a decision tree based method dubbed M5. In all instances the bagged variant outperformed its base algorithm. The algorithm which saw the greatest improvement with the use of bagging was ANN and the best method overall (when applied to the whole data set) was Bagged SVR. When the population was limited to patients without DDI, Bagged kNN proved to be the most effective method with single kNN coming second. This is interesting as kNN is the simplest of all methods described here.

Tao et al. [9] considered Bagging as applied to Genetic Programming (GP) and Genetic Algorithms (GA). For each chromosome c which encodes parameters such as the population size of the GP, the probability of crossover and the maximum tree depth; a multivariable nonlinear regression model is built using GP. This entailed initialising a set of sub trees whose nodes could either be (binary or unary) operations and features from the data set. A function is then built by composing the nodes of the tree and its fitness is determined by its prediction accuracy. Fitter trees have a higher chance of surviving to the next generation. The fittest trees will

then swap branches in the process of crossover and the process repeats until a baseline fitness is achieved across the population. The surviving subtrees then each form an element in a vector f(x). The regression model then takes the form:

$$F(\mathrm{x}) = \beta \cdot \mathrm{f}(\mathrm{x}) \tag{5}$$

where β is a vector containing the weight associated with each function.

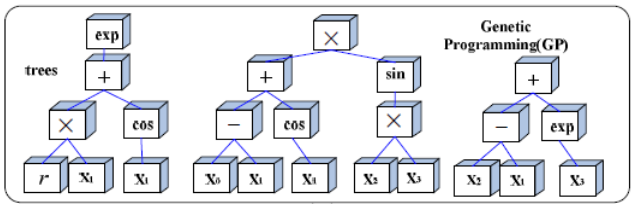

Figure 3: Example of Expression Trees Generated by GP

The success of $F(\mathrm{x})$ at predicting doses determines c's fitness and the entire evolutionary algorithm is iterated over at this level until a set of successful yet varied regression models are converged upon. These form the base models which the Bagging algorithm averages over to determine a final result.

The authors compared this technique to two different variants of SVR and ANN; and also across different feature sets. A comparison of the mean-square-error showed indicated that while the GP approach would underperform on the training set compared to the other methods, it performed best on the testing set in all cases but one (of nine).

2.3.2 Stacking. Stacking refers to using the output of one ML model as the input of another. This generalises to stacks of arbitrary size.

Sharabiani et al. [7] made use of a Relevance Vector Machine which classified patients into one of two groups and based on this classification would apply one of two linear regression models to determine the exact dose to be prescribed. This approach proved successful compared to non-stacked techniques used in previous studies.

Martin et al. [5] however, reported no benefit from using a stacked ensemble as opposed to a single base method. This may be as a result of their relatively small sample size (< 70).

3 DISCUSSION

Table 1 provides a representative summary of results. We note that the majority of these found an ensemble method to outperform the other techniques considered. The majority also made use of pharmacogenetic features. This comes with its own cost, however, as accurate genotyping remains a relatively time consuming and expensive process. Many of the techniques also used feature optimisation strategies in conjunction with the employed ML algorithms. For example, Sohrabi et al. [8] made use of Multi-Objective Particle Swarm Optimisation to choose the best possible feature set. Given the sheer number of factors which can affect Warfarin uptake, opting for the extra computation involved in choosing an optimal selection of features appears to be a worthwhile decision.

Unfortunately many of the studies used different metrics for evaluating their models and thus a direct comparison is not always possible. Many employed the Mean Absolute Error however further standardisation of evaluation metrics would be a beneficial convention.

Another observation is that the average error generally increased with the population size. This implies that the true complexity of the problem only becomes apparent once there is sufficient variability in the population.

4 CONCLUSIONS

We have see a wide variety of Machine Learning techniques applied to Warfarin dosage prediction. Across the variety of studies surveyed several factors stand out: ensemble methods tend to outperform their simpler counterparts, feature optimisation can play a significant role in a given model's performance, and that the specifics of the training population (such as DDI and ethnicity) favour particular models.

REFERENCES

[1] Erdal Cosgun, Nita A. Limdi, and Christine W. Duarte. 2011. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics* 27, 10 (mar 2011), 1384–1389. https://doi.org/10.1093/bioinformatics/btr159

[2] Enzo Grossi, Gian Marco Podda, Mariateresa Pugliano, Silvia Gabba, Annalisa Verri, Giovanni Carpani, Massimo Buscema, Giovanni Casazza, and Marco Cattaneo. 2014. Prediction of optimal warfarin maintenance dose using advanced artificial neural networks. *Pharmacogenomics* 15, 1 (jan 2014), 29–37. https://doi.org/10.2217/pgs.13.212

[3] 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

[4] 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 (aug 2015), e0135784. https://doi.org/10.1371/journal.pone.0135784

[5] Brent Martin, Marina Filipovic, Lara Rennie, and David Shaw. 2010. Using Machine Learning to Prescribe Warfarin. In *Artificial Intelligence: Methodology, Systems, and Applications*. Springer Berlin Heidelberg, 151–160. https://doi.org/10.1007/978-3-642-15431-7_16

[6] Addepalli Pavani, Shaik Mohammad Naushad, Rajasekar Manoj Kumar, Murali Srinath, Amaresh Rao Malempati, and Vijay Kumar Kutala. 2016. Artificial neural network-based pharmacogenomic algorithm for warfarin dose optimization. *Pharmacogenomics* 17, 2 (jan 2016), 121–131. https://doi.org/10.2217/pgs.15.161

[7] 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), 1–9. https://doi.org/10.1155/2015/560108

[8] Mohammad Karim Sohrabi and Alireza Tajik. 2017. Multi-objective feature selection for warfarin dose prediction. *Computational Biology and Chemistry* 69 (2017), 126 – 133. https://doi.org/10.1016/j.compbiolchem.2017.06.002

[9] Yanyun Tao, Yenming J. Chen, Xiangyu Fu, Bin Jiang, and Yuzhen Zhang. 2018. Evolutionary ensemble learning algorithm to modeling of warfarin dose prediction for Chinese. *IEEE Journal of Biomedical and Health Informatics* (2018), 1–1. https://doi.org/10.1109/jbhi.2018.2812165

,

Table 1: Comparison of Studies

Author(s)	Techniques Compared	Best Technique	Population Size	Metric	Value
Hu et al. [3]	10	Bagged SVR	587	MAE/dose (mg)	0.210
Pavani et al. [6]	1	ANN	207	MAE/week (mg)	1
Tao et al. [9]	4	Evolutionary Ensemble Model	289	MSE/dose($*10^{-2}$)	1.71
Cosgul et al. [1]	3	Random Forest Regression	290	R^2	66.4
Grossi et al. [2]	4	ANN	377	MAE/week (mg)	3.
Sharabiani et al. [7]	5	Stacked SVM + Regression	2274	RMS/week (mg)	8.4
Liu et al. [4]	9	Multivariate Adaptive Regression Splines	960	MAE/week (mg)	8.84