Using Machine Learning to Prescribe Warfarin

Brent Martin, Marina Filipovic, Lara Rennie, and David Shaw

Computational Intelligence Research Lab
University of Canterbury, Christchurch New Zealand
brent.martin@canterbury.ac.nz,
marina@projectexperts.co.nz,
lara.rennie@gmail.com,
dpshaw@ihug.co.nz

Abstract. Predicting the effects of the blood thinner warfarin is very difficult because of its long half-life, interaction with drugs and food, and because every patient has a unique response to a given dose. Previous attempts to use machine learning have shown that no individual learner can accurately predict the drug's effect for all patients. In this paper we present our exploration of this problem using ensemble methods. The resulting system utilizes multiple ML algorithms and input parameters to make multiple predictions, which are then scrutinized by the doctor. Our results indicate that this approach may be a workable solution to the problem of automated warfarin prescription.

1 Introduction

Predicting the effect of the anticoagulant drug warfarin is a difficult and risky task. In this problem the history of dosages for cardiac and vascular patients is known, along with the corresponding International Normalised Ratio (INR), which measures the time the blood takes to clot and can be compared both between patients and across different countries. Noise is inherent in the data-set as it is impossible to control a patient's life-style, so confounding factors, including non-compliance (i.e. taking the wrong amount), are extremely likely. Other errors or missing data arise from the fact that the data have been obtained from hand-written doctor records of a patient. Errors when performing data input are also possible. It is therefore ideally suited as a data set to examine time-series data with machine learning. It is hoped some machine learning algorithm will be able to predict either the effect of the next dosage on the INR, or the optimal next dosage for a particular INR reading, with some degree of accuracy.

In this paper we report on studies into both of these aspects. Section 2 introduces the problem of warfarin prediction in more detail. In Section 3 we report on investigations into the ability of machine learning to predict the outcome of a warfarin dose. Based on these results we developed a prototype decision support system that presents predictions to a Doctor for a range of doses; this work is described in Section 4, which also reports a preliminary evaluation. We conclude and indicate future directions in Section 5.

D. Dicheva and D. Dochev (Eds.): AIMSA 2010, LNAI 6304, pp. 151–160, 2010.

2 Warfarin Prediction

Determining the appropriate dosage of warfarin is extremely difficult for many reasons. One of the most important issues is the way in which the drug effect is measured. Warfarin has a half-life of 36 hours [1], so it takes approximately 4-6 days to achieve new steady-state plasma concentrations after dose adjustments. As a result the maximum response to a dose is not visible for at least one or two days after ingestion, making it difficult to accurately adjust the last dosage given after a worrying INR reading. There is a large variance in individual responses to the drug due to its complex pharmacology [3], which is affected by many factors, including an individual's age, weight and gender, lifestyle habits such as alcohol and tobacco consumption, and even environmental factors such as diet. Patient compliance is also an issue [2]. The general health of the patient can also affect one's response to warfarin [1].

The simplest mechanism for prescribing warfarin is a “loading table" of rules specifying what dosage is needed following a given INR reading. This, however, does not acknowledge individual differences in warfarin response. Current advice is to adjust the dosage by 5-20% of the total weekly dose, depending on the current INR, the previous dose, and any reasons identified that might explain the undesirable current INR reading [4]. Graphical “nomograms" have also been developed to help with dose adjustment, such as that by Dalere, which is based on a model of warfarin activity [5]. In general success rates for physicians using nomograms, dosage adjustment tables or their own experience have not been particularly high. Schaufele, Marciello & Burke [6] demonstrated this by analysing 181 patients receiving warfarin treatment over a four-month period through a rehabilitation centre. Only 38% of all INR readings were found to be within the target range. Most physicians, however, achieve a 50-75% success rate for a particular patient [1]. This is still relatively low, especially when combined with the fact that between 1.1% and 2.7% of patients managed by anti-coagulant clinics suffered major bleeding [1].

There have been some attempts to utilise the machine learning capabilities of computers in warfarin treatment. The worth of machine learning as a potential approach has been demonstrated in the prescription of other drugs, as shown in a study by Floares et al [7], where neural networks were used successfully to compute an optimal dosage regimen for chemotherapy patients. These produced better results than the other conventional or artificial intelligence approaches reported in the study. Narayanan & Lucas [8] also attempted a machine learning solution to predicting INR levels after a given dosage, by using a genetic algorithm to select variables with which to train a neural network. However, no comparisons were offered to other solutions, examining only the benefits of the genetic algorithm in addition to the existing neural network. Neural networks have been also investigated by Byrne et al [9] and found to be twice as accurate as physicians at predicting the result of a given dosage. The benefits of extracting rules for warfarin dosage from ensemble learning have also been researched [10].

3 Applying ML to Warfarin

In our first study we explored the feasibility of using machine learning by trying to predict the effects of a given dose. Two options immediately presented themselves

when considering the form that the output from the machine learning algorithms should take. The first of these was a numerical value of the expected INR reading after a given warfarin dosage. However, this severely restricted the number of machine learning algorithms that can be used, because most produce a nominal classification as their output. Since the actual value of INR is less important than its relative position to the therapeutic index, it was decided that dosage result could be usefully classified as either "low", "in range" or "high". We similarly binned the input INR values and added this as an extra set of attributes.

Several approaches were proposed as possible solutions, within which different algorithms and attributes could be used. These approaches principally differ in the source of data used for training the system as well as whether some form of ensemble learning is to be used. The simplest solution predicts a patient's response to warfarin solely by examining their history of interaction with the drug. This has the advantage that any model of the patient built by the machine learning algorithm would be individualised as much as possible. However, this solution would obviously not be ideal if the patient did not have a long history on which the algorithm could be trained. Alternatively, data from all patients could be used to train an algorithm, from which predictions for an individual patient could be made. In such a system all data points are used by the algorithm, no matter to which patient they belong.

We also trialled a "two layer" ensemble approach. Ideas from two popular ensemble learning techniques, "bagging" [11] and "stacking" [12], were combined. Each patient's history, including that of the patient currently under study, was modeled separately using just their own data and selecting the ML algorithm that best predicted their INR. Each patient model was then given the same data point, and their predictions of the resultant INR level for this data point fed into the "second-layer" algorithm, which was trained to use these predictions as a means for predicting the final INR value for this patient. The second-layer algorithm was varied to try to maximise performance.

Many different combinations of attributes were trialled to represent the patient's history for a given data-point. When learning from only one patient's data, up to three previous dosages and INR values were provided as input. More data was provided when working with a training set based on multiple patients, with anywhere from one to twelve previous dosages used; it was hypothesised that increasing the number of attributes would help the algorithms to select only relevant data points from other patients' histories.

The patient data was provided by volunteers and collated by the fourth author (Mr David Shaw, a cardiothoracic surgeon). Although data for over 70 patients were initially provided, not all of these could be used. First, patients could be divided into two groups: initial-state patients (within two years of their heart valve operation) or steady-state patients. The problem of predicting their response to warfarin differs significantly between groups, and because of time constraints it was decided to restrict it to steady-state patients only. Some patients also had to be excluded because their history was of insufficient length. Furthermore, for systems based on data from more than one patient, all patients studied had to have the same therapeutic range. Ten patients were selected at random on which to perform experiments. Dosages recorded are those given over a week. The performance of the various systems was compared to base accuracies provided by Naïve Bayes and the graphical nomogram detailed in

Dalere's study, the latter approximating what a physician might have prescribed [5]. For each prediction all history up to (but excluding) the dose in question was used.

Results when learning from individual histories showed different algorithms and attributes suited different patient histories. Accuracies achieved varied from patient to patient, and a significant improvement could be noted over time for patients with long histories. For example, the accuracy for patient 31 reached a maximum of 53% over the whole data-set, but when only the accuracy over the last 12 data-points was considered, 67% was observed. The best performing algorithm and attributes was not necessarily the same for these two situations, however. When learning from multiple patient histories the most predictive attributes and most accurate algorithm again varied between patients. However, there was less variance here than for individual histories. The success rate of the best solution varied between 61% and 90%. Most patients, however, needed a larger history than when trained on only their own data. There was also more consensus on the most successful algorithms, with Ridor [13], NNge [14] and a Bayesian Network (created by genetic search) the most successful algorithms on more than one occasion each. This is potentially a more useful solution, in that theoretically a patient does not need to have an extensive history to make predictions. Finally, we evaluate the 'two-layer' approach. Results were not as promising as initially hoped. Accuracies of the best second-layer algorithm ranged from 44% to 85% depending on the patient. Many machine learning algorithms for the second layer machine were trialled, yet the best result that could be achieved for a particular patient was usually achieved by many different algorithms. This suggests that the choice of algorithm for this second machine may not be important.

Comparisons were made between different ways of predicting the effect of a given warfarin dosage by examining the percentage of correct predictions made over the datapoints. However, predictions are only made from at least the 7th data-point, and on some occasions slightly later than this, if the attributes required dictate this. This allows sufficient history to be used to make predictions. Table 1 shows a comparison of the best machine learning solutions with graphical nomograms, the accuracy of the physician and a Naïve Bayesian prediction. The machine learning solutions compared are the best algorithm combination when learning on the patient's own history, the best when learning on multiple histories, and the best two-layer result obtained. The "base accuracy" to which each solution is compared is the percentage of the most popular class over the patient's history, or the accuracy that would be achieved if the most common result was predicted in every case.

One-way ANOVA was performed on the raw percentage success for each patient of the five different prediction methods. A significant difference was found ($F=7.198$; $p<0.001$). Post-hoc analysis using paired t-tests with a Bonferroni correction was hence applied. This showed a weakly significant difference between Naïve Bayes and the machine learning solutions, apart from the case of the two-layer approach from which Naïve Bayes was not significantly different. There was a weakly significant difference in accuracy between the two-layer approach and the nomogram (66.5% versus 52.5% respectively), $t=4.5$; $p<005$. More crucially, strongly significant differences were observed between the machine learning solution using only the individual histories and the nomogram (66.5% versus 52.5%), $t=8.2$, $p<0.001$. There was also a significant difference between the machine learning solutions using the patient's own

Table 1. Comparison of the success in predicting the effect of warfarin dosages of the best solution for each approach (percentage improvement over the base accuracy)

Patient ID	Individual patient history	Multiple histories	Two-layer	Nomogram	Doctor	Naïve Bayes
2	10	**17**	0	-18	0	4
5	0	8	0	-19	-64	**9**
31	5	**23**	-4	-8	0	-4
39	23	**32**	31	7	0	15
40	0	**3**	0	-20	-60	-13
44	13	**19**	5	3	-9	12
48	22	**36**	17	4	0	17

history versus learning on multiple patients, which achieved a mean accuracy of 79%, $p<0.003$. When learning on the history of multiple patients, the best machine learning solution was significantly superior to both the nomogram ($t=11.2$, $p<0.0001$) and the doctors' accuracy (41%), $t=4.9$; $p<0.003$.

4 A Decision Support Approach

The first study suggested that ML can potentially predict warfarin's effect. However, this is a more tightly constrained problem than the one we are trying to solve; in general we want to produce a model that indicates the *best* dose. Machine learning is not sufficiently accurate to do this directly owing to the lack of information about other factors (such as lifestyle) and the strong interaction between attributes, but it might prove useful in giving a doctor some guidance on the likely effect of a range of doses, such that the doctor could make a more informed decision. The main goal of this second study was to build on what was learned in the first study and develop a prototype for a web-based predictor that provides precisely this information. This prototype allows the doctor to have enough flexibility to create and combine different attributes in order to create a model that will predict a dose's effect on the patient with the greatest accuracy, by allowing them to select/create the input attributes and combine them in order to make the best predictive model on a per-patient basis.

When creating an attribute there are a number of different functions that can be applied to any numeric data field, such as sum, average and delta. Each function can be applied to different length time periods. Calculated attributes can be repeated up to four times per example (i.e. for the current and previous three dosage intervals). General attributes (e.g. gender) can also be added or removed from the final group of attributes. Further, each patient's model can be created using their history only or by utilizing that of all patients in the database. Finally, there are potentially many different ML algorithms that can be used to create a model; the prototype uses the WEKA code base [13] to provide ML algorithms, and the doctor can select which ones they wish to use for a given patient.

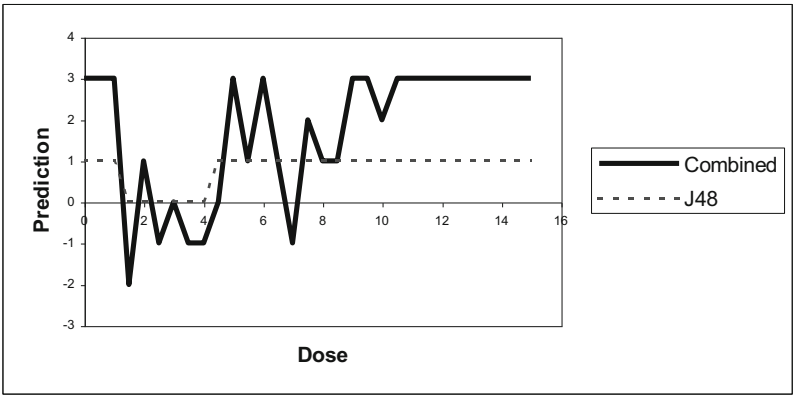
The output of the system is a graph that plots the predicted effect of all doses from 0 to 15mg in 0.5mg increments. Two kinds of output are possible: nominal or numeric (as previously described). For numeric outputs the graph for each model is presented, while for nominal attributes we also add a simple ensemble learner: each model's output is translated into a number (low=-1; in-range=0; high=1) and we sum these values to get the combined prediction; our expectation was that this combined output would prove superior to any individual model and overcome the problem previously identified whereby every patient requires an individual model. For the purpose of this evaluation we used J48 (a variant of C4.5 [15]), NNGE [14] and IBk [16] for nominal predictions and IBk and SMO regression for numeric prediction, for all patients.

The patient data used in this research came from the same source as the first study. Among initial state patients there are some that show a steady reaction to warfarin from the early days (months); these patients should be relatively easy to predict. Others have an inconsistent history even after a year of taking the drug, possibly because they have made frequent lifestyle changes. Patients that have only recently started taking warfarin present the biggest challenge. To explore the feasibility of creating predictive models for these three groups of patients (initial, steady-state, varying), three patients were examined: one with a long and steady history (over 2 years), one with a moderate-length, inconsistent history and one with almost no history at all. The latter was in fact a copy of a patient that we do have history for, with most of the historic data excluded during training.

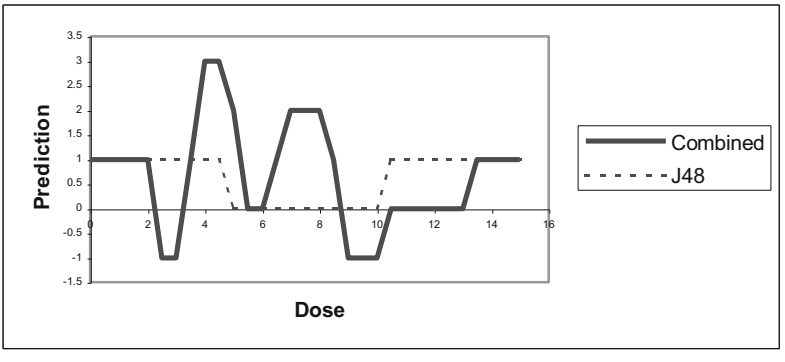
After some initial regression testing we concentrated on the sum, average and delta functions when calculating attribute values since they appeared to have some influence over INR readings. It has been recognized that a patient's lifestyle data would be of great value in predicting the right dose of warfarin, but the only data we have available at the moment is the gender (which is known for only some of the patients). This is specified as being female, male or missing, and was used as an attribute when learning on multiple patient histories to see if this improved accuracy. Regarding the interval over which to sum/average doses, it should be noted that the points in a patient's data set do not have the same temporal gap. However, almost all readings are 1-5 weeks apart, and the warfarin from a specific dose is processed within a week. We therefore used one week as the period for our testing purposes.

To evaluate the flexibility of the approach we developed five test conditions covering a range of attribute options. Test 1 used a single attribute only, consisting of the warfarin dose summed over the last seven days. Test 2 added the average and delta doses, and test 3 added gender. Test 4 investigated the efficacy of adding previous historic data by applying the same attributes as test 2 (sum, average and delta doses) for each of the *last three* dosage periods. Finally, test 5 used all of the attributes of test 3 plus the average INR and dose over the entire history period. Note that these tests do not necessarily capture the best way to predict for each of these patients. Rather, they illustrate how much variation can be quickly achieved using the system and show how it facilitates building useful models.

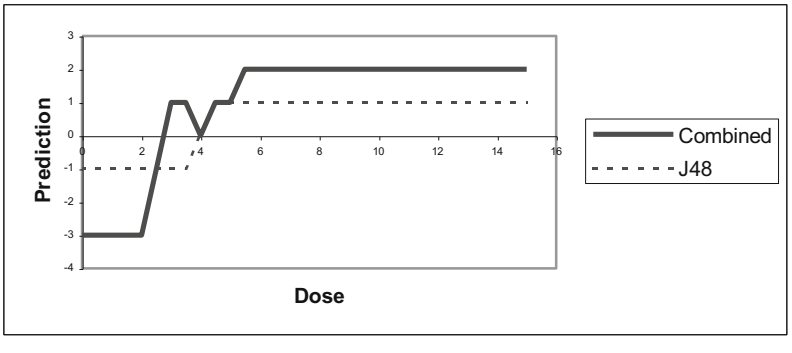
Fig. 1. shows some of the output graphs produced, representing tests 1, 4 and 5 for patient 2 (long history, varying response) using multiple patient models, for the output from the combined models and J48. We chose J48 for comparison because the results appeared reasonably useful (this was not true for the other algorithms). Test 1 is a



Test 1



Test 4



Test 5

Fig. 1. Sample outputs for patient 2 (long history, varying response)

baseline, which we would not necessarily expect to perform very well, and this is clearly the case; both the combined result and that of J48 are of little help for this patient. Adding further history (test 4) is of itself not very helpful because, as described previously, the other patients' response to warfarin may not match that of this patient, so the additional data is still of limited use. In contrast, adding the average

INR and dose for each patient improves the models considerably; for test 5 J48 produces a very clean model that predicts the correct dose to be 4.0, which agrees with this patient's history. The combined model, whilst messier, predicts the same dose if the mid-point between the –3 and +2 plateaus is used. In contrast there is no significant effect from adding this information for patient 1 (none of the models were very useful), whilst patient 3 is easily predicted from single patient data owing to the steady nature of their response.

Overall the models in test 1 were of limited use. Patient 1 (short history) has only a couple of days worth of history and therefore the only prediction possible would be by using other patient's history. The results obtained for each model showed that INRs are spread over the range of 2-8.5mg, which is of no use. For patient 2 (long history, varying response to warfarin) the single patient model learned by J48 gave a clean prediction of 4.5-5, while the combined model was noisier but nonetheless predicted the same value; this matches the patient's data. The multiple-patients model however was very noisy and did not produce any useful result. Patient3 (steady history) also produced good models for single patient data, with J48 predicting a dose of 8.5-10, and the combined model suggesting a dose of 8.5-9, both of which appear to be a good match to the data. Again the multiple patient model was too noisy to be of use. Test 2 (sum, average and delta over the last 7 days) showed similar results to the test 1 but with little bit more consistency. In test 3 the gender was added, but this didn't result in significant improvements to any of the models.

Test 4 also used the sum, average and delta values for the last 7 days, however it then repeated this calculation for the *previous two doses.* The purpose of this test was to illustrate how including additional history (as used in the first study) affects the predictions produced. Patient 1 (short history) was not included in this test since the history was not long enough and therefore the calculations on the previous intervals could not be performed. For patient 2 (varying response) this test produced more consistent results between the algorithms used. However, the suggested dose by all 3 nominal algorithms increased from 4-5 to to 5.5mg, which is not supported by the data. This is probably the result of using the delta information; all of our tests suggested this attribute is not useful. The accuracy of the model decreased for patient 3 (steady state) for both single and multiple patient models, even though this patient was the easiest to predict in previous 3 tests. This suggests additional dose history by itself is not useful for such patients.

Finally, test 5 added average totals of INR and dose. These attributes might be expected to have two potential benefits. First, for single patient models they would smooth out temporary fluctuations in response. Second, for multiple patient models they might eliminate patients that do not share a similar warfarin response. For all three patients the multiple-patient models improved. In the case of patient 1, this was the only set of parameters that produced a usable model; the predicted in-range dose was 3-4 OR 6-10. The latter is spurious and suggests the model is still not sufficiently selective. However, if we assume the doctor has a rough idea of the ideal dose then they could usefully use the 3-4 value as a guide. For the other two patients the multiple-patient model improves dramatically with both J48 and the combined models giving a clean response at around the dose expected (4.0 and 8.5-10 for patients 2 and 3 respectively). This suggests that with the right selection of attributes a multiple-patient model can be learned that produces useful results, which bodes well for dealing with patients whose output response is highly variable or whose history is fairly short.

5 Conclusion

The goal of this research is to develop an online warfarin prediction application that will initially be used by doctors, but which may in time become suitable for patient self-prescription. In this study we first investigated how well various machine learning approaches compare to more traditional ways of predicting the effect of a dose of warfarin in heart valve patients. A new "two-layer" approach was tried to test the hypothesis that the warfarin problem consists of multiple, potentially related data sets. Many different attribute combinations were attempted to provide the best representation of the data and any temporal patterns observed that could help with prediction. When tested on the data of heart valve patients it was found that the effect of a warfarin dosage could be predicted with the most accuracy by machine learning algorithms learning on the history of multiple patients. However, the best performing algorithm and attributes differed from patient to patient, making a one-fits-all solution unlikely.

Despite the shortcomings identified, in the second study we investigated whether machine learning is sufficiently accurate to be useful in guiding a doctor to make an informed prescription. We developed a prototype that outputs a prediction for a range of doses, from which the correct dose might be inferred. Overall we were able to produce plausible models for three test patients: one with very little history, one with a long history of varying warfarin response, and one with a steady response. For patients with little or varying history a multiple-patient model that included their average INR reading and dose performed best, while for the steady-state patient an excellent model was created based only on their most recent dose. We also investigated using a model that reported the combined results of several algorithms. In general this model outperformed any single ML algorithm for all three patients, suggesting an ensemble approach may overcome some of the problems of the highly individual responses to warfarin observed. The biggest challenge is predicting warfarin response accurately for new patients: such patients can only be predicted using the multiple-patient model, and its accuracy is directly proportional to the number of patients for which we have history. For the purpose of this research we had a small number of histories available (19) but still we were able to predict the right dose for our test Patient 1 to a reasonable degree of accuracy, which is a promising achievement.

Overall the results are sufficiently positive to encourage us to continue to explore this approach. We will further investigate combinations of attributes and parameters to determine the extent to which we can narrow the search. In particular, this study made very little use of previous INR results, which are likely to be useful. We will also investigate more sophisticated ensemble methods to try to increase the reliability of the combined model. Finally, we will add more data and trial the system with doctors who prescribe warfarin to evaluate its utility.

The prescription of warfarin remains a difficult problem with life-and-death implications. This research is a promising step towards better administration of this important therapy.

References

1. Gallus, A., Baker, R., Chong, B., Ockelford, P., Street, A.: Consensus guidelines for warfarin therapy. Medical Journal of Australia 172, 600–605 (2000)
2. Roland, M., Tozer, T.: Clinical Pharmacokinetics: Concepts and Applications. Williams and Wilkins, Philadelphia (1995)
3. Gage, B.F., Fihn, S.D., White, R.H.: Management and dosing of warfarin therapy. The American Journal of Medicine 5, 211–228 (2000)
4. Jaffer, A., Bragg, L.: Practical tips for warfarin dosing and monitoring. Cleveland Clinic Journal of Medicine 70, 361–371 (2003)
5. Dalere, G.: A graphical nomogram for warfarin dosage adjustment. Pharmacology 19, 461–467 (1999)
6. Schaufele, M.K., Marciello, M.A., Burke, D.T.: Dosing practices of physicians for anticoagulation with warfarin during inpatient rehabilitation. American Journal of Physical Medication and Rehabilitation 79, 69–74 (2000)
7. Floares, A.G., Floares, C., Cucu, M., Marian, M., Lazar, L.: Optimal drug dosage regimens in cancer chemotherapy with neural networks. Journal of Clinical Oncology 22, 2134 (2004)
8. Narayanan, M., Lucas, S.: A genetic algorithm to improve a neural network to predict a patient's response to warfarin. Methods of Information in Medicine 32, 55–58 (1993)
9. Byrne, S., Cunningham, P., Barry, A., Graham, I., Delaney, T., Corrigan, O.I.: Using neural nets for decision support in prescription and outcome prediction in anticoagulation drug therapy. In: Lavrac, N., Miksch, S. (eds.) Fifth Workshop on Intelligent Data Analysis in Medicine and Pharmacology, Berlin (2000)
10. Wall, R., Cunningham, P., Walsh, P., Byrne, S.: Explaining the output of ensembles in mdeical decision support on a case by case basis. Artificial Intelligence in Medicine 28, 191–206 (2003)
11. Breiman, L.: Bagging predictors. Machine Learning 26, 123–140 (1996)
12. Wolpert, D.H.: Stacked generalisation. Neural Networks 5, 241–259 (1992)
13. Witten, I.H., Frank, E.: Data mining. Morgan Kaufman, San Francisco (2000)
14. Martin, B.: Instance-based learning: nearest neighbour with generalisation. In: Computer Science. University of Waikato, Hamilton (1995)
15. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
16. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6, 37–66 (1991)