A comparison of data mining approaches in the categorization of oral anticoagulation patients

Francesco Archetti[1], Ilaria Giordani[1], Enza Messina[1], Giulia Ogliari[2], Daniela Mari[2]
[1]*Disco - University of Milano-Bicocca*
{archetti,giordani,messina@disco.unimib.it}
[2]*IRCCS Istituto Auxologico Italiano and University of Milano*
{daniela.mari, g.ogliari@unimi.it}

Abstract

Oral anticoagulation therapy, largely performed by warfarin-based drugs, is commonly used for patients with a high risk of blood clotting which can lead to stroke or thrombosis. The state of the patient, with respect to anticoagulation, is captured by the index INR, which is to be kept within a therapeutic range. The patients' response is marked by high inter-individual and inter-temporal variability, which can lead to serious adverse events. Polymorphisms of two genes CYP2C9 and VKORC1, considered markers of lower dosage requirements, still account for a relatively minor part of this variability. In this work, authors show that classification methods can identify groups of patients homogeneous with respect to the dynamics of INR. In particular, authors use classification methods in order to characterize patients according to their warfarin metabolism and hence their sensitivity to different doses. Finally a Markov model to capture the dynamics of the patient's response over the years is proposed.

Keywords: Oral Anticoagulation Therapy, Clustering, Classification, Markov Models

1. Introduction

Warfarin and its companion anticoagulant drugs are the first line treatment to mitigate the risks related to atrial fibrillation, ventricular dysfunction, deep vein thrombosis and aortic valve replacement. However their therapeutic range is very narrow: therefore frequent sampling (at least once in 2-3 weeks) of the INR index (International Rationalized Ratio), i.e. the time required for the blood to clot, and careful dosage adjustments are needed for the INR to stay within its assigned range.

The "trial &error" basis of the methods currently in use to fine tune the dosage for a given patient along with the response's variability due to genetic and behavioral factors can result in out of range periods and therefore in a non negligible risk of thromboembolic and bleeding events. Warfarin initiation is associated with one of the highest adverse events for any single drug due to high inter-individual variability. About 50% of patients fail to stabilize within the therapeutic range: for this reason most of these patients even with no contraindication to warfarin therapy are not receiving it because physicians are reluctant to initiate it in patient's elderly or with risk of bleeding.

Genotyping of patients has been recently suggested in order to understand inter-individual variability and control its dose-INR relationship, particularly in the induction phase. This fact as been recognized by FDA whose labeling for Warfarin 2007 reads: "It cannot be emphasized too strongly that the treatment of each patient is a highly individualized matter". A notable contribution to patient genotyping is: [14] where it is shown that genetics variants of the enzyme that metabolized Warfarin cytochrome P450 CYP2C9 and VKORC1 contribute to differences in patients' response. Basically the same results have been obtained in a wide range of genetics investigations. More recently [3][12][15] variance of a new gene CYP4F2 have been shown to alter warfarin requirements.

While there is a relative large agreement of the value of genotypes for the induction phase, the debate is still open on its effectiveness in the long term therapeutic management [7]. Indeed in [1] it is shown that pharmacogenomics guided dosing failed to achieve a reduction of the patient average percentage INR outside the therapeutic range.

This paper has two objectives:

1. Use data mining techniques in order to characterize patients according to their warfarin metabolism and hence their sensitivity to different doses.
2. Develop a Markov model to capture the dynamics of the patients response over the years

Section 2 explains in detail the data sources and its preparation. The data mining algorithms are considered

in Sect. 3 along with their results. Sect.4 is devoted to the Markov model, while sect.5 presents a comprehensive assessment of the results along with further research direction.

2. Data sources

We tested our approach on a sample of 1013 elderly (65+) patients. We imported data collected from the computerized databases in a database with three entities: patients, treatments and visits. For each patient we have information about date of birth, sex, medical evidence leading to OAT (Atrial Fibrillation, Deep Venous Thrombosis, other), patient's INR range (2-3, 2.5-3.5, 2.5-3) and target INR.

Furthermore, for each patient, we memorize the concurrent medications in the treatment entity. In particular, we classified all treatment in different categories: digitalis, amiodarone, furosemide, nitrates, beta blockers, calcium channel blockers, ACE Inhibitors, diuretic tiazidic, sartanic, farmaco lipidi and other. So for each patient and for a particular category, we have a value "yes" if patient assumes a drug belong to this category and value "no" otherwise.

Finally, for each visit we collected the date of visit, the result of the INR measurement, the weekly dose and the dose calendar, the drug used for OAT therapy (Coumadin 5 mg, Sintrom 1 or 4 mg).

For a subset of patients we collected in the patient entity genomic data. In particular the polymorphism of CYP2C9 and VKRC01 are collected. For each patient CYP2C9 gene feature can have the following values: WT (wild Type), CT (*2 allele: the SNP in exon 3 (CGT->TGT; Arg144Cys)), AC (*3 allele: the SNP in exon 7 (ATT->CTT; Ile359Leu)). The possible variants for gene VKORC1 are: WT (wild type), CT and TT.

Thus each patient is characterized by 23 features.

Entry characteristics for both 1013 patients and the subset of 135 patients with genomic data are summarized in Table 1:

Tab. 1:principal characteristic of studied population

Characteristics	Patients without genomic data	Patients with genomic data
Patients number	1013	138
Age, y, mean (dev.std)	76 (10)	76 (11)
Gender:		
Women N(%)	502 (49.5%)	59 (43.70%)
Men N(%)	511 (50.44%)	76 (56.30%)
Primary reason for anticoagulation, N(%)		
Atrial fibrillation	771 (76.11 %)	135 (98.0%)
Deep vein thrombosis	80 (7.9%)	1 (0.75 %)
Other diagnosis	162 (15.99 %)	2 (1.35%)
Clinical Variables:		
Target INR, mean (SD)	2,56 (0.2)	2,50 (0.3)
Takes amiodarone, N (%)	175 (17.20%)	26 (19,25%)
Takes ASA (acetylsalicylic acid), N(%)	110 (10.85%)	15 (11.11%)
Takes Farmaco Lipidi, N(%)	213 (21.02%)	29 (21.48%)

The sample shows a prevalence of atrial fibrillation (76.11%). The genotyped sub-sample mirrors in a balanced way the relative weight of the features in the large one. In our study we extract from the 138 patients, only those with atrial fibrillation and so we work on a dataset of 135 patients.

The allelic variant frequencies for the subset of 135 patients are summarized in Table 2.

Tab. 2: Allelic variant frequencies

Allelic variant frequencies		
CYP2C9	*WT*	66.67%
	CT	20.74%
	AC	12.59%
VKORC1	*WT*	33.33%
	CT	40.74%
	TT	25.93%

The overall allelic frequency distribution is similar to what is reported in the literature [1][5].

3. The data mining algorithms

In this work, we introduce a particular index, called *drug sensitivity* (D_{sens}) to capture the dose-INR relationship which better characterizes the patient behavior. This index is represented by the ratio between dose and INR's variations, as follows:

$$D_{sens} = \frac{\sum \Delta d_i}{\sum \Delta INR_i}$$

Where:

$$\Delta d_i = \frac{(d_{i+1} - d_i)}{7} \qquad \Delta INR_i = \frac{(INR_{i+1} - INR_i)}{Nd}$$

As INR measurements are not taken at regular intervals the dose values are replaced by their daily variations (Δd_i), and the INR values by ΔINR_i (computed with the above formula) where N_d is the number of days between successive measurements. Note that a negative value of D_{sens} means that patient is not responding to the therapy because increasing (decreasing) doses are likely to correspond to decreasing (increasing) INR values. In this case a high absolute value of D_{sens} correspond to patients whose response in highly

unpredictable. Positive values of D_{sens} indicate that patient is responding to the therapy, in this case the absolute value indicate the response sensitivity with respect to the dosage, patients falling in this class have a more predictable drug response behavior.

In our study we compute the drug sensitivity index (D_{sens}) by using 6 dose-INR measurement time courses. Fig. 1 shows the empirical distribution of this variable. Three principal drug sensitivity's classes (Negative [-125 $\leq D_{sens} \leq$ -3,738], Medium [-3.738 $\leq D_{sens} <$ 7.38], Positive [7.38 $< D_{sens} \leq$ 107]) are obtained by discretizing the variable by using the minimum description length approach.

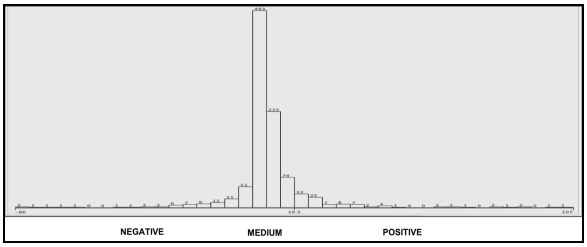

Fig. 1: Drug Sensitivity distribution

In particular, 123 patients belong to Negative D_{sens} class, 572 patients belong to Medium D_{sens} class and finally, 318 patients belong to Positive D_{sens} class.

3.1 Clustering

In particular, since data are both categorical and numerical, we use a modified k-prototypes algorithm, proposed by Bushel et al. [2], following an earlier paper of Huang [9], for handling mixed data. In Fig. 2 is represented the components of the modified k-prototypes algorithm for our mixed numeric and categorical dataset. The approach follows the k-means paradigm with randomization of initialization of the algorithm. The strategy involves constructing an objective function from the sum of the squared Euclidean distances for numeric data with simple matching for categorical values in order to measure dissimilarity of the samples. Furthermore, separate weighting terms are used to control the influence of each data domain on the clustering of the patients.

A cluster's prototype is formed from the mean of the values for numeric features and the mode of the categorical values of all the samples in the group.

Finally, the dynamic validity index (DVI) for numeric data was modified with a category utility (CU) measure, obtaining a DVI_CU index. With this index we can determine the optimal number of clusters in the mixed type data, like in Bushel et al. [2]. The DVI_CU index is computed as:

$$DVI\_CU = DVI + 1/CU$$

The *DVI* index, proposed by Shen et al. [13], is based on an intra/inter ratio validity index that also includes scaling of the intra- and the inter-cluster distances. Furthermore, *CU* measure [8] defines the probability of matching a categorical feature value given a cluster versus the probability of the categorical feature value given the entire data set.

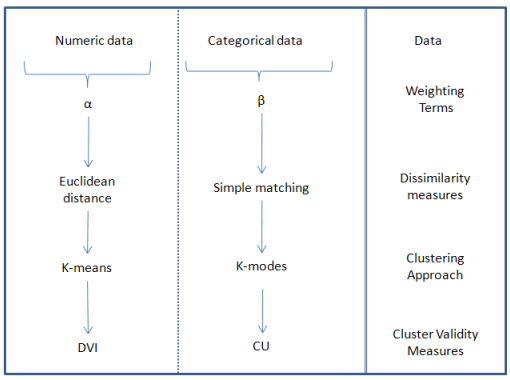

Fig. 2: schema of the implemented algorithm

In particular, *DVI_CU* is minimized over all *k* sets for each run of the mod*k*-prototypes clustering algorithm.

In our work, both the complete dataset and the genetic dataset are clustered using the modified k-prototypes clustering algorithm at values of *k* increasing from 2 to N, number of clusters.

Values of *DVI_CU* index obtained for each value of *k* are reported in Fig. 3 and in Fig. 4. We have the minimum value of this index, for complete dataset, with *k=3* and *DVI_CU* equals to 1.11, as represented in Fig. 3. Fig. 4 shown the results obtained on the dataset composed by patients with genetic data. In this case, minimum *DVI_CU* index is obtained with *k=3* and *DVI_CU* equals to 1.08.

Results of clustering algorithms confirm the division of patients into three different classes, as proposed in the last subsection.

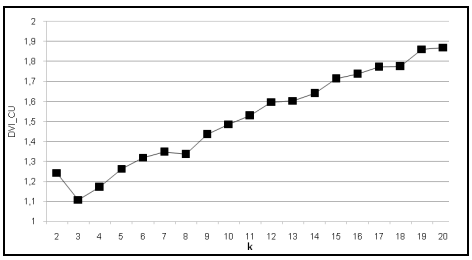

Fig. 3: DVI_CU index variation for *k* from 0 to 20 for complete dataset

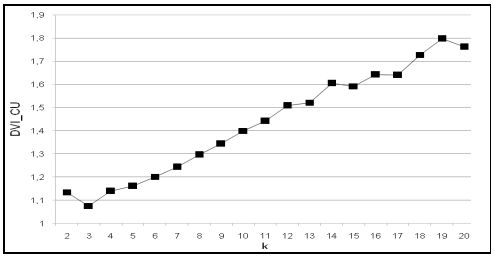

Fig. 4: DVI_CU index variation for *k* from 0 to 20 for genomic dataset

3.2 Drug sensitivity based classification

Patient classification models, based on personal and clinical data, have been proposed in [4][10].However, the traditional machine learning applications on Oral Anticoagulation Therapy (OAT) problems, classifies patients on their average INR value (below, in and over patient range) but do not consider their drug sensitivity. In this paper we propose classification models using drug sensitivity index, explained above, as class variable.

In order to build a classification model we considered the following features: personal data (age and gender), OAT therapeutic data (drug used for OAT therapy and medical evidence leading to OAT) and concomitant medications.

We train and test, using 10-fold cross validation, four different machine learning classification algorithms (Multi Layer Perceptron (MLP), Support Vector Machines (SVM), K-Nearest Neighborhoods (kNN) and Bayesian Networks (BN)). For our experiments we use the Weka [16] implementation of the used classification algorithms.

In this first stage we do not use INR average and variance. The model thus can be applied in the induction phase.

Results in terms of correctly classified instances (CCI) and F-measure (the weighted harmonic mean of precision and recall) are shown in Tab. 3.

Tab. 3: D_{sens} based classification results

	DRUG SENSITIVITY based Classification			
	MLP	SVM	kNN	BN
% CCI	60.61%	64.06%	59.32%	62.29%
F-measure	0.581	0.595	0.578	0.589

To characterize better the behavior of a patient we compute INR average and variance of a time course of 6 INR measurements and include both these data in the feature set. So, we built new classification models with this new feature set and the obtained results, reported in Tab. 4, are better both in term of CCI and F-measure.

Tab. 4: D_{sens} based classification results with new features

	DRUG SENSITIVITY based Classification			
	MLP	SVM	kNN	BN
% CCI	63.71%	68.70%	64.30%	64.46%
F-measure	0.6214	0.613	0.598	0.611

The model thus learned i.e. with the full set of features has been applied also in the induction phase i.e. without considering INR values. Obtained results are reported in Tab. 5.

Comparing these results with those reported in Tab. 3, we can see that models learned using the two additional features about INR are better in term of CCI and f-measure than those learned without these two features.

Tab. 5: D_{sens} based classification results

	DRUG SENSITIVITY based Classification			
	MLP	SVM	kNN	BN
% CCI	61.61%	65.26%	60.88%	63.8%
F-measure	0.5964	0.61	0.584	0.597

3.3 Classification with Genetic data

In this study we know genomic data of a subset (135 patients) of the original 1013. We now present classification results obtained including two genomic features:

- *CYP2C9 polymorphism*: this feature can assume value [WT,CT,AC]
- *VKORC1 polymorphism*: this feature can assume value [WT,CT,TT]

Also with this dataset three different tests are performed, the same presented in subsection 4.2. In the induction phase first stage we do not use INR average and variance. Results obtained at this step are reported in Tab. 6.

Tab. 6: D_{sens} based classification with genomic data results. In this phase INR average and variance are not considered.

	DRUG SENSITIVITY based Classification			
	MLP	SVM	kNN	BN
% CCI	61.66%	65.6%	62.3%	63.5%
F-measure	0.591	0.62	0.58	0.60

Results obtained using the complete set of features (including INR average and variance) are reported inTab. 7. A new improvement is visible compared to the results inTab. 4.

Tab. 7: D_{sens} based classification with complete genomic data results

	DRUG SENSITIVITY based Classification			
	MLP	SVM	kNN	BN
% CCI	68.61%	74.41%	79.07%	75.58%
F-measure	0.675	0.747	0.665	0.645

Also in this case, the model thus learned i.e. with the full set of features has been applied also in the induction phase i.e. without considering INR values. Obtained results are reported in Tab. 8.

Tab. 8: D_{sens} based classification with genomic data results. In this phase INR average and variance are not considered.

	DRUG SENSITIVITY based Classification			
	MLP	SVM	kNN	BN
% CCI	62.2%	66.8%	63.1%	64.1%
F-measure	0.62	0.64	0.60	0.625

In this way we can see that genomic data allow a better characterization of patient's behavior.

Tab. 9: Genomic variant distribution in the three D_{sens} classes

	GENES			DRUG SENSITIVITY		
	CYP2C9	VKORC1	TOT. PATIENT NUMBER	POSITIVE	MEDIUM	NEGATIVE
WILD TYPE	WT	WT	29	51,72%	44,83%	3,45%
ONE POLYMORPHISM	WT	CT	38	34,21%	50,00%	15,79%
	WT	TT	23	34,78%	60,87%	4,35%
	CT	WT	12	8,33%	75,00%	16,67%
	AC	WT	4	0,00%	25,00%	75,00%
TWO POLYMORPHISMS	CT	CT	10	10,00%	30,00%	60,00%
	CT	TT	6	0,00%	83,33%	16,67%
	AC	CT	7	14,29%	14,29%	71,43%
	AC	TT	6	0,00%	16,67%	83,33%

"Wild type" patients are predominantly in positive and medium drug sensitivity classes, as can be seen in Tab. 9. Patients with only one polymorphism are distributed principally in medium class and, finally, patients with variants on both genes are in negative class. To understand better the behavior of patients belonging to the three different classes we plot INR values of three different patients. In Fig. 5 are plotted INR values of a wild type patient belonging to the positive drug sensitivity class. Comparing this plot with that reported in figure Fig. 6, is possible to see that the hemorrhagic or thrombotic risk of a patient in negative D_{sens} class is higher then that of a positive D_{sens} patient.

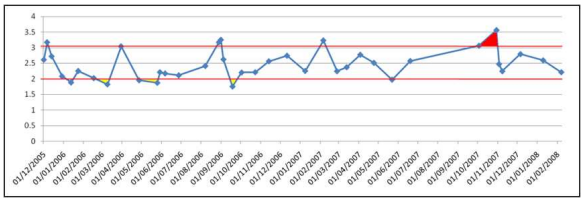

Fig. 5: Wild type patient (gene CYP2C9: WT; gene VK0RC1: WT), positive Drug Sensitivity class

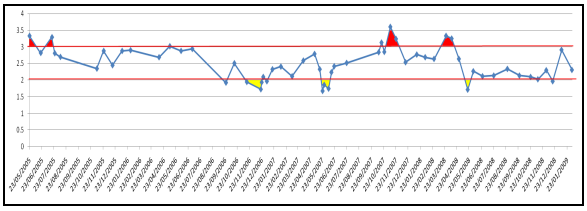

Fig. 6: Patient with two polymorphisms (gene CYP2C9:CT; gene VK0RC1:TT), medium Drug Sensitivity class

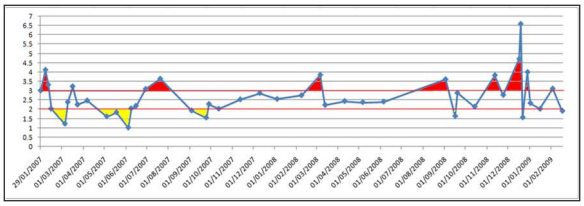

Fig. 7: Patient with two polymorphisms (gene CYP2C9: AC; gene VK0RC1:CT), negative Drug Sensitivity class

4. The Markov model

A Markov Chain (MC) is a discrete time stochastic process $\{X_k\}$, with a finite number of states and transition probability matrix. The simplest possible model of our problem considers: three states (*HIGH* (over range), *IN* (in range), *LOW* (under range)); three dosing actions (dose decrease, increase and constant) and a different transition probability matrix for each drug sensitivity class. Assessment of results is in Fig. 8, Fig. 9 and Fig. 10.

Dose decrease

	LOW	IN	HIGH
LOW	0,4545	0,4545	0,0909
IN	0,2353	0,4706	0,2941
HIGH	0,2830	0,4528	0,2642

Dose constant

	LOW	IN	HIGH
LOW	0,5849	0,3774	0,0377
IN	0,3750	0,4063	0,2188
HIGH	0,1212	0,4242	0,4545

Dose increase

	LOW	IN	HIGH
LOW	0,5248	0,3762	0,0990
IN	0,5849	0,2642	0,1509
HIGH	0,3158	0,4211	0,2632

Fig. 8: Transition matrices for negative D_{sens} class

Dose decrease

	LOW	IN	HIGH
LOW	0,4762	0,4127	0,1111
IN	0,4902	0,3529	0,1569
HIGH	0,3220	0,4878	0,1902

Dose constant

	LOW	IN	HIGH
LOW	0,7395	0,2140	0,0465
IN	0,1302	0,7359	0,1339
HIGH	0,0000	0,3442	0,6558

Dose increase

	LOW	IN	HIGH
LOW	0,2500	0,5212	0,2288
IN	0,1818	0,5289	0,2893
HIGH	0,0667	0,3556	0,5778

Fig. 9: Transition matrices for medium D_{sens} class

Dose decrease

	LOW	IN	HIGH
LOW	0,7377	0,2459	0,0163
IN	0,7173	0,2608	0,0217
HIGH	0,5238	0,3714	0,1047

Dose constant

	LOW	IN	HIGH
LOW	0,8314	0,1685	0
IN	0,0638	0,8776	0,0585
HIGH	0	0,0294	0,9705

Dose increase

	LOW	IN	HIGH
LOW	0,0636	0,5682	0,3682
IN	0,0000	0,3113	0,6887
HIGH	0,0000	0,2857	0,7143

Fig. 10: Transition matrices for positive D_{sens} class

5. Assessments of results and future directions

In this paper authors use different classification methods to characterize patients according to their warfarin metabolism and hence their sensitivity to different doses. Finally, through a Markov model they try to capture the dynamics of the patient's response over the years. These promising results should be prospectively validated.

Since all clinical studies so far failed to demonstrate a beneficial impact of pharmacogenetic guided warfarin dosing and to achieve the primary end point reduction of out of range INR [7], we feel that an accurate patient characterization and subset analysis is required to capture how the dynamics of INR is impacted by the genomic patients' features. The authors believe that a reduction in out of range time can be obtained through dosing approaches that capture this dynamic.

6. References

[1] J.L. Anderson, B.D. Horne, S.M. Stevens, A.S. Grove, S. Barton, Z.P. Nicholas, S.F. Kahn, H.T. May, K.M. Samuelson, J.B. Muhlestein, J.F. Carlquist, Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation, *Circulation*, 2007, Vol. 116, No.22, pp.2563-70,

[2] P.R. Bushel, R.D. Wolfinger, G. Gibson, Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes, *BMC Systems Biology*, 2007, pp. 1-15

[3] M.D. Caldwell, T. Awad, J.A. Johnson, B.F. Gage, M. Falkowski, P. Gardina, J. Hubbard, Y. Turpaz, T.Y. Langaee, C. Eby, C.R. King, A. Brower, J.R. Schmelzer, I. Glurich, H.J. Vidaillet, S.H. Yale, K. Qi Zhang, R.L. Berg, J.K. Burmester, CYP4F2 genetic variant alters required warfarin dose, *Blood*, 2008, Vol. 111, No.8, pp.4106-12

[4] M. Carney, P. Cunningham, The Benefits of Using a Complete Probability Distribution when Decision Making: An Example in Anticoagulant Drug Therapy, Trinity College Dublin, Department of Computer Science, *Technical Report*, 2005, pp.22.

[5] B.F. Gage, C. Eby, J.A. Johnson, E. Deych, M.J. Rieser, Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin, *Clin Pharmacol Ther*, 2008, Vol. 84, No. 3, pp. 326-31

[6] B.F. Gage, P.E. Milligan, Pharmacology and pharmacogenetics of warfarin and other coumarins when used with supplements. *Thromb Res.*, 2005, Vol.117, pp.55–59.

[7] D.A. Garcia, E. Hylek, Warfarin Pharmacogenetics, Correspondence, *NEJM*, 2009, Vol. 360, No. 23, pp. 2474-2475

[8] M. Gluck M, J. Corter: Information, uncertainty, and the utility of categories. *Proc 7th Ann Conf Cog Soc*, 1985, pp.283-287.

[9] Z. Huang, Clustering large data sets with mixed numeric and categorical values. *Proceedings of the 14th International Joint Conference on Knowledge Discovery and Data Mining*, 1997

[10] S. McDonald, C. Xydeas, P. Angelov, A Retrospective Comparative Study of three Data Modelling Techniques in Anticoagulation Therapy, *Proceedings of the International Conference on BioMedical Engineering and Informatics BMEI2008*, China, 2008, pp. 219-225.

[11] L. Lesko, The critical path of warfarin dosing: finding an optimal dosing strategy using pharmacogenetics, *Clin Pharmacol Ther*, 2008, Vol. 84, No.3, pp. 301-303

[12] V. Pérez-Andreu, V. Roldán, A.I. Antón, N. García-Barberá, J. Corral, V. Vicente, R. González-Conejero, Pharmacogenetic relevance of CYP4F2 V433M polymorphism on acenocoumarol therapy, *Blood*, 2009, Vol. 113, No.20, pp.4977-9.

[13] J. Shen, Y. Deng, E.S. Lee, S.I. Chang, Determination of cluster number in clustering microarray data. *Applied Math and Computation*, 2005, Vol.169, pp.1172-1185.

[14] U.I. Schwarz, M.D. Ritchie, Y. Bradford, C. Li, S.M. Dudek, A. Frye-Anderson, R.B. Kim, D.M. Roden, C.M. Stein, Genetic determinants of response to warfarin during initial anticoagulation, *N Engl J Med.*, 2008, Vol. 358, No. 10, pp.999-1008

[15] F. Takeuchi, R. McGinnis, S. Bourgeois, C. Barnes, N. Eriksson, N. Soranzo, P. Whittaker, V. Ranganath, V. Kumanduri, W. McLaren, L. Holm, J. Lindh, A. Rane, M. Wadelius, P. Deloukas, A Genome-Wide Association Study Confirms VKORC1, CYP2C9, and CYP4F2 as Principal Genetic Determinants of Warfarin Dose. *PLoS Genet*, 2009, Vol. 5, No.3.

[16] Weka. A multi-task machine learning software developed by Waikato University, 2006. www.cs.waikato.ac.nz/ml/weka.