Feature analysis to predict treatment outcome in rheumatoid arthritis

Cátia Sofia Tadeu Botas
catia.botas@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2017

Abstract

Rheumatic diseases are causing a major impact in the daily life of the patients and although not fatal, these diseases cause severe pain and can impair life quality. One of these diseases is rheumatoid arthritis (RA), a chronic condition, being the focus of this thesis. Conventional disease-modifying anti-rheumatic drugs (DMARDs) have long been the mainstream treatment for RA. Unfortunately, one third of the patients fail to respond and many are not able to sustain a good response. For these, biological treatments provide a clinical alternative but there is no biomarker to predict both response code and biological treatment. One uses the Reuma.pt database, developed by the Portuguese Society of Rheumatology (SPR), containing only biological treatment patients. Some modifications are applied to the database in order to present the data in a time-series format. Taking this into account, it is compared different classification methods with and without feature selection. Further, three cases studies are introduced, being one the data resulted from the database (time series data) and the other two presenting their characteristics through time (data summarization and data representation). One concludes that the case studies containing information about the features perform better than the remaining. Besides, it is possible to forecast the response code at the 24th month in the 6th-month appointment, with an average accuracy of 65%. This is an exploratory work, where the initial focus was on data treatment to enable future studies in this field.

Keywords: rheumatic diseases, biologic treatment, classification, time-series, data representation, data summarization, treatment outcome

1. Introduction

With the progress in medicine, the improvements of living conditions and medical assistance, longevity in Portugal has increased in the last years and had a tendency to increment. However, this is not always reflected in a good life quality in the last years of life. This can be explained by the fact that while severe infectious diseases are being increasingly fought, the number of chronic conditions increases. Having this is fundamental to study techniques to improve life quality especially concerning diseases that do not kill but deteriorate the human body causing pain and impair quality of life with the evolution of the disease. One of this chronic diseases is rheumatoid arthritis (RA), which is going to be the focus of this report.

RA is defined by pain and reduction in a scope of movements, and it can affect several areas of the musculoskeletal system. In some conditions exist signs of inflammation, e.g. redness, swelling and warmth which can affect the internal organs [5]. When not adequately treated, the majority of daily activities such as cooking, climbing stairs or walking are affected. This disease has a severe effect on work capability, being the most significant cause of sick leave and premature retirement [5].

Long-term outcome in patients with RA disease is highly dependent upon an aggressive pharmacological control of inflammation early in the disease course. Although the concerns of selecting the best biologic treatment, nowadays there is no reliable biomarker predictor of drug treatment response. One consequence to avoid is the irreversible joint destruction while a physician searches for an effective drug. Some studies are being done in the area, investigating the possibility of adapting the drugs to the patient on an early stage [12].

In Portugal, the Portuguese Society of Rheumatology (SPR) developed a database in 2008, the Rheumatic Diseases Portuguese Register (RNDR), available at Reuma.pt [20]. This database only has patients who started the biologic treatments, and these treatments are only started when the synthetic disease modifying anti-rheumatic drugs (DMARDs) stop working, however, although the

register is recent, contains enough data of lifestyle habits, disease activity and functional assessment scores, previous and current therapies and laboratory measurements registered at each visit. After some transformations, in order to represent the database in a time series format, some patterns and correlations were developed, making possible the forecast of essential features such as medication or health improvement.

The goal of this work is to investigate the correlation between the factors described before and the response code to a biologic treatment of the patient during follow-up. It is planned to compare different classification methods, in order to evaluate the adequate model to the database. It's also intended to analyse which features have the most impact in determining the response code of a patient to a biologic treatment.

This is a pioneer work in this field, opening new possibilities in the field of precision medicine. For example, the study of the best biologic treatment for each patient, according to its habits, age, gender, and other covariates related to the disease.

This thesis contributes to the development of an algorithm that pre-process the database and classify treatment outcome [3]. Besides, the ability to predict treatment outcome at 24 months from the 6th-month medical appointment with an average accuracy of 65%.

In addition, is being envisaged to be submitted, in an international journal or conference, an article to provide these results to the scientific community.

2. Data Mining

Data mining, which is the process of identifying patterns in large databases, is being widely used to understand, analyze, and be able to extract information. This chapter will analyze the data comprised int Reuma.pt with the tools to unveil information from it.

There are many areas in which TS (TS) can occur naturally such as finance, economics and medicine. A TS is a series of observations, $x_i(t); [i = 1, ..., n; t = 1, ..., m]$, made sequentially through time where i indexes the measurements made at each time point t [21]. When the value of n is equal to 1, the TS is called univariate, otherwise, is considered a multivariate TS (MTS).

2.1. Data Summarization

Summarization represents a TS through its global characteristics. When applied, it is common to use more than one summarization technique to represent it, resulting in the creation of a feature vector that contains several global characteristics of the TS [16].

Some of the characteristics used are related to statistical measures very recognized. The **mean**, also known as the first moment, represents the average of a TS. The **median** represents the central value in an ordered set of TS. The **mode** represents only the most common value in a TS. Finally, the **variance**, also known as the second moment, represents the amount of variation of the TS around the mean value.

The **fractal dimension** was first presented in 1982 [14] and there are several ways of calculating the fractal dimension of a curve and the one presented here utilizes the variance of the TS as a measure of self-similarity. However, the variance formula has some changes in order to be applied to the fractal dimension. Thus, the variance is given by

$$Var = \frac{\sum_{i=1}^{N(d)} (x(t_i) - x(t_i + d))^2}{2N(d)}, \quad (1)$$

where $x(t_i)$ is the value of a TS at time t_i, $x(t_i + d)$ is a point of the TS with distance d from t_i and $N(d)$ is the number of all points on the TS with distance d.

After that, a graph of the log of variance versus the log of the distance is plotted to obtain the TS semi-variogram. Therefore, the fractal dimension, FD, is estimated by

$$FD = \frac{4 - s}{2}, \quad (2)$$

where s is the slope of the semi-variogram.

The **run-length** can be defined as a sequence of consecutive values that have the same value. From this, a matrix can be constructed where each element, $r(i, j)$, indicates that level i has a run-length of length j. From the run-length matrix, some measures can lead to short-run length and long-run length parameters.

The **Short Run-Length** has a significant focus on short run-lengths and when it has high values, indicates the presence of short run-lengths. It can be defined as

$$SRE = \frac{\sum_{i=1}^{NL} \sum_{j=1}^{NR} \frac{r(i,j)}{j^2}}{\sum_{i=1}^{NL} \sum_{j=1}^{NR} r(i, j)}, \quad (3)$$

where NL is the number of values in the TS, NR is the number of run-lengths, and $r(i, j)$ is the run-length for value i and level j.

The **Long Run-Length** has focused on long run-lengths, and high values in this parameter indicate the presence of long run-lengths. Therefore, this

new metric can be defined as

$$LRE = \frac{\sum_{i=1}^{NL}\sum_{j=1}^{NR} j^2 \cdot r(i,j)}{\sum_{i=1}^{NL}\sum_{j=1}^{NR} r(i,j)}. \quad (4)$$

A **histogram** is a graph that shows the frequency, that is, the occurrence number for each value in the TS. Given this representation and considering that the histogram has M values, with μ as mean and $f(m)$ as the frequency of the m-th level, a TS can be characterized by the following statistical measures:

Skewness measures the asymmetry of the histogram form and can be defined by

$$sk = \frac{\sum_{m=1}^{M} (f(m) - \mu)^3}{M\sigma^3}, \quad (5)$$

Kurtosis measures the peakedness of the histogram and can be defined by

$$kur = \frac{\sum_{m=1}^{M} (f(m) - \mu)^4}{M\sigma^4}. \quad (6)$$

2.2. Data Representation

Data representation techniques describe the features in other formats [16]. Then, in this chapter are defined the data representation techniques utilized.

The **discrete Fourier transform** (DFT) has the ability to represent the TS in the frequency domain.

The wavelet transform uses functions known as wavelets that allow the location of the TS in frequency and space. The **discrete wavelet transform** (DWT) is a discretized version of a wavelet.

The **Piecewise Aggregate Composition** (PAA) can substantially reduce the dimensionality of the TS by dividing the series into segments of equal size where they are replaced by the mean value of the segment. Then, these mean values are grouped into a vector that becomes the segment signature.

Taking into consideration the PAA, it is possible after this result to discretize the same using an alphabet of strings. **Symbolic Aggregate Approximation** (SAX) produces symbols with equal probability because, knowing that a normalized TS has a Gaussian distribution, it is possible to divide into areas of the same size.

Shape Definition Language (SDL) represents the TS through the changes that exist in its shape over time. The SDL vocabulary is {Up, up, stable, zero, down, Down}.

Clipping transforms a TS into a series of bits, taking into account the average value of the series. If the value is higher than the average, this value is replaced by 1 and if it is lower, it is replaced by 0.

2.3. Feature Selection

Features can also be called attributes, properties or characteristics, where each of them describes an instance (case, example or record) [13]. Feature selection aims at reducing the subset of features. First, it may be desired to improve performance (learning speed, predictive accuracy). It may be desired to visualize the data in a different way to choose the model to be used and, finally, to reduce dimensionality and remove noise. In this way, feature selection can be defined as the process that chooses an optimal subset of features according to a specified criterion [13].

2.4. Predictive Models

Predictive models are fundamental in the data mining process because they are the ones that allow the application of classification tasks to the database. One will introduce the different classification techniques used.

The **Logistic** got its name because of the function that governs it. The Logistic function, also known as a sigmoid function, was developed by statisticians to describe the growth of a population. From this function, the Logistic regression was developed in 1958 [22] where the input values are combined with coefficients to predict the output value.

The **Support Vector Machine** (SVM) is a recent new statistical learning theory that has been receiving increasing attention for classification and forecasting[2]. One of the objectives of SVM is to find a rule which selects some particular subset of training data, also known as support vectors.

The goal is to find a function $f(x)$ that has at most an error, ε, from the obtained targets y_i from all the training data, and at the same time, is as flat as possible. In other words, errors are ignored as long as they are less than ε, but any deviation more massive than this should not be accepted [19].

A **Bayesian Network** (BN) assigns probability factors to various results based on an analysis of a set of input data. Like many other Machine Learning algorithms, a BN is taught using training data. Once trained, a BN can be queried to make predictions about new data that was not represented by the training set.

From BN is possible to obtain a **BN** classifier and one of the most efficient classifiers in practice, **Naive Bayes** (NB) classifier [8, 18]. The NB is a BN classifier were each attribute has only the class

attribute as a parent. NB is based on the conditional class independence, and due to its fixed graph structure, the computational process does not have expensive as when compared to the BN classifier.

Finally, **decisions trees** (DT) are data structures, in a tree form together with the rules to make predictions. The internal nodes compose these structures, reflecting the attribute, the leaf nodes that represent the classes [15]. J48 and Random Forest are some decisions tree vastly applied in the literature.

J48, an improved version of the C4.5 decision tree, generates a pruned or un-pruned C4.5 decision tree [15]. **Random Forest**, by another hand, is an ensemble learning algorithm [9]. The ensemble classifiers are based on the principle that a set of classifiers perform better than an individual one. Random Forest has the advantage of running efficiently on large databases and can handle thousands of inputs variables without deleting them. This method applied a combination of classifiers, and each classifier infer a vote for the selection of the most voted class [9].

3. Reuma.pt database

The object of study of this thesis is the Reuma.pt database [20]. The patient's record come from two health centres in Portugal. There are a total of 424 patients in the database, and all of them were diagnosed with RA. However, this database only has patients who started the biologic treatments, and these treatments are only started when conventional treatments stop working. Thus, it can exist a gap between the detection of the disease and the beginning of the biologic treatment.

3.1. Dataset description

The data consists of lifestyle habits, disease activity, functional assessment scores, previous and current therapies, laboratory measurements and response to biologic treatment, registered at each visit. This data is known as **panel data**, where each patient has more than one record during the follow-up period. In addition, a patient can switch between available biologic therapies during this follow-up period. Having this, the database is composed of the number of instances and attributes present in Table 3.1. It is clear that even though only 424 patients are present in the database, the number of instances exceeds highly this value due to the data format.

Considering static data (does not change during follow-up) and dynamic data (may change during follow-up), one presents in Table 2 the number of nominal and numerical attributes for both static and dynamic data. One concludes that the number of static attributes is much lower than the number of dynamic attributes and that mainly the static data is composed of nominal attributes as opposed to the dynamic data, mostly composed of numerical data.

Table 1: Number of instances and attributes in the database.

Instances	Attributes
9305	433

Table 2: Number of nominal and numerical attributes.

	Static	Dynamic
Nominal	24	4
Numerical	16	289
Total	433	

More than 400 features are recorded in the database. Some of these records are taken by applying some questions to the patient or are the result of laboratory measures, but others are the result of some mathematical equations based on the patient data. These calculations are relevant to the study of the response to the biologic treatment since its result can define the continuation or abandonment of treatment.

The Disease Activity Score 28 (DAS28) [1] is present in the dataset and is a system developed and validated by the European League Against Rheumatism (EULAR) to measure the progress and improvement of RA. Number 28 is a subset of a total of 75 joints that have been shown to be most relevant in RA assessment. DAS28 is calculated as:

$$DAS28 = 0.56 \cdot \sqrt{t28} + 0.28 \cdot \sqrt{s28} + 0.70 \cdot \ln(ESR) + 0.014 \cdot SA \quad (7)$$

where $t28$ is the number of tender joints, $s28$ the number of swollen joints, the ESR the measured erythrocyte sedimentation rate (ESR) and SA the subjective assessment of disease activity by the patient.

The values resulted from this score range from 2 to 10, where higher values mean a higher disease activity and where a value below 2.6 means a disease remission. From the comparison with the value of DAS28 from the first appointment, a response criterion was developed by EULAR (Table 3).

To consider that a patient has a good response to the treatment, the DAS28 score at a certain appointment has to be less or equal than the value of 3.2, and the difference between this value and the DAS28 score from the beginning of the treatment has to be higher than 1.2.

Table 3: EULAR Response Criteria.

	Improvement		
Current	> 1.2	$]0.6,\ 1.2]$	≤ 0.6
≤ 3.2	yes	moderate	no
$]3.2,\ 5.1]$	moderate	moderate	no
> 5.1	moderate	no	no

3.2. Dataset processing

In order to make possible the forecasting study, some changes needed to be made to the dataset. One represents in Figure 1 the process of modifying the database, in order to originate new files.

Figure 1: Processing of data.

First, static and dynamic data were separated into different sheets in order to preprocess the features which change over time. This new dataset, with split data, is now represented as **ReumaA**. After that, since this file was still not enough to perform classification tasks, the transformation from panel data to one line per patient per biologic were made that gave rise to the **ReumaB** dataset. However, there were many missing values, creating the need to apply imputation methods, originating the dataset **ReumaBImp**. Finally, with these two datasets, the files for classification were created.

Concerning the transformation process, one presents in Table 4 the number of instances and attributes in the **ReumaB** dataset.

Table 4: Number of instances and attributes in the **ReumaB** dataset.

	Static	**Dynamic**
Instances	719	
Attributes	40	293

One draws from Table 4 that there was a 92% reduction in the number of instances, these files having only 719 instances. This value is not 424, the total number of patients, because there are patients who have a record of taking several biologic and, in these files, each line represents the biologic and patient combination.

However, not all sheets contained information because the longer the time series, the smaller the number of series present. In order to simplify the dataset, only lengths up to 24 months were considered, due to the small number of time series from this value. Thus, the **ReumaB** data resulted in the configuration present in Table 5.

Table 5: Sheets present in dataset **ReumaB**.

Static	**0**	**3**	**6**	**12**	**18**	**24**

Having this, the **ReumaB** contains seven different sheets, wherein the first it is possible to access all the static data of the patients. The remaining six sheets present the information related to the appointments segmented by the different months up to a maximum of 24 months. Each of these sheets represents a subset of the data and are named T_i, where $i = 0, 3, 6, 12, 18, 24$ is the corresponding time point.

Note that all these sheets have the same number of rows, where each row displays the information per patient and biologic. When there is no information for the respective month about the patient, these lines are blank.

However, this file is not yet capable of performing any classification in order to forecast the response to the biologic treatment. The EULAR response criteria is the base of classification and the class to be forecast but first a transformation from EULAR response code to class code needed to be made. One can find the transformation on Table 6. So, in the last column of file **ReumaB** will be present a class of response code for each time series.

Table 6: Transformation of Eular response code in **ReumaB**.

no	moderate	yes
C0	**C1**	**C2**

Afterward, the **ReumaBImp** dataset was created after performing longitudinal imputation to **ReumaB** dataset.

The question that follows is how many and what data to put on this new files in order to get the best classification results. For this, several types of files were made, with different compositions in order to perceive later what would be the best case.

The first files to be created were composed of all the data of month 0 and all the data of another month, having several possibilities, and containing in the last column the response code to the biologic treatment. The first time point (t.p.) will always be 0 due to the EULAR response criteria need of DAS28 value at the initial time point (t.p.). One represents in Table 7 all the possible combinations used to create the TS data files.

One draws from Table 7 that 20 different files were created to perform classification. Each file has the following composition, **0-A:B**, where 0 is the

Table 7: Combinations of data used to create time series data files.

1st t.p.	2nd t.p.	Response code
0	0	3, 6, 12, 18, 24
	3	3, 6, 12, 18, 24
	6	6, 12, 18, 24
	12	12, 18, 24
	18	18, 24
	24	24

first time-point, A the second, and B the class to predict. For instance, 0-3:6 means that all features from T_0 and T_3 were used to predict the EULAR response 6 months after the beginning of the treatment.

Then, the data were transformed to obtain other datasets with information on the time series. The methods used for these transformations were data summarization (DS) and data representation (DR) and their compositions are present in Table 8.

Table 8: Combinations of data used to data summarization and representation calculations.

TS considered	Response code
{ T_0, T_3 }	3, 6, 12, 18, 24
{ T_0, T_3, T_6 }	6, 12, 18, 24
{ T_0, T_3, T_6, T_{12}}	12, 18, 24
{ T_0, T_3, T_6, T_{12}, T_{18} }	18, 24
{ T_0, T_3, T_6, T_{12}, T_{18}, T_{24} }	24

From Table 8 it is possible to extract that for each method (summarization and representation) 15 different files are created. Another factor to note is that to calculate these parameters for each feature, a minimum of two points is required to be considered as a time series. However, the notation to represent each file maintains, for instance, **0:12-18** means that the time series $T = \{T_0, T_3, T_6, T_{12}\}$ was used to predict the EULAR response 18 months after the beginning of the treatment.

Note that Tables 7 and 8 represent the creation of files for classification tasks. In this way, knowing that for the study are being used two databases, **ReumaB** and **ReumaBImp**, we will have in total to carry out our study the number of files present in Table 9.

Table 9: Total number of files created.

	TS	DS	DR
ReumaB	20	15	15
ReumaBImp	20	15	15
Total per composition	40	30	30
Total		**100**	

Thus, 100 different files were obtained to perform classification and, with this, to find out which one can more accurately predict the response code of the patient to the biologic treatment.

3.3. Data treatment on WEKA

Having the files ready, the next step is to use the WEKA software [10] to perform the comparison of various classification methods. However, before doing so, some tools inherent in the software will be considered and will improve performance at the time of implementation.

One of the tools used was the Synthetic Minority Over-sampling Technique (SMOTE) [4]. This allows balancing classes so that they have all the same weight within classifiers.

In WEKA this process is performed taking into account the members of each class, that is, if elements are added to a class, they will be coherent with those that already belong to the class.

Another crucial task before classification is discretization. Discretization is an instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. To do so, one uses supervised discretization based on the Fayyad and Irani method [6]; in some classification methods, this task is mandatory.

Concerning the feature selection filter used from WEKA, the parameters considered for the filter were the Correlation-based Feature Subset Selection Evaluator (CfsSubsetEval) [11] as a attribute evaluator and as the search method used, the best-first search strategy to navigate attribute subsets [17] (c.f. Section 2.3).

Having this, six methods will be applied to the files created from **ReumaB** and **ReumaBImp** datasets. One presents in Table 10 the classifiers used to perform classification tasks.

Table 10: Classifiers used in the experiments.

Classifier	Nomenclature	WEKA
BN	BayesNet	BayesNet
NBs	NaiveBayes	NaiveBayes
Logistic	Logistic	Logistic
SVM	SMO	SMO with PolyKernel (E=1)
DT	J48	J48
	RandomForest	RandomForest

Afterward, and with the tools presented previously, the schema present in Figure 2 will be used to perform classification tasks. Note that SMOTE, discretization and feature selection are optional so one can create several distinct schema combinations.

Having this, the different combinations are using: SMOTE and discretization, discretization or none. After that, the resulted dataset can only undergo

Figure 2: Steps to perform classification tasks.

classification tasks to discover their accuracy or apply a feature selection filter before performing these tasks.

4. Results

This chapter will present the main results for the three case studies: TS, DS and DR. Afterward, the conclusions of the study will be presented.

4.1. Time Series Data

First, starting with TS data files, for the combination of files per dataset, the average accuracy obtained for the classification methods is present in Table 11.

Table 11: Analysis of average accuracy per method in **ReumaB** and **ReumaBImp** datasets for TS files.

	ReumaB	ReumaBImp
NaiveBayes	65.39	64.68
BayesNet	64.52	64.60
Logistic	61.50	62.15
SMO	67.56	66.93
J48	65.30	65.54
RandomForest	65.24	66.53

One draws from Table 11 that the method with best value in average of accuracy is for the SMO. However, the **ReumaB** dataset is the best dataset because it has the highest value of performance, considering the average accuracy as measure. Afterward, feature selection were applied to the TS files for both datasets. One presents in Table 12 the classification results.

Table 12: Analysis of average accuracy per method in **ReumaB** and **ReumaBImp** datasets after feature selection for TS files.

	ReumaB	ReumaBImp
NaiveBayes	68.96	70.04
BayesNet	68.14	69.85
Logistic	67.83	69.54
SMO	67.78	70.13
J48	66.56	67.28
RandomForest	65.88	68.08

It is notable that the feature selection increases the classification results in all methods. Considering the same choice measure, the best dataset is **ReumaBImp** with SMO classifier.

4.2. Data Summarization

Concerning DS files, the same results were considered. One presents in Table 13 the classification results for the DS files in both datasets.

Table 13: Analysis of average accuracy per method in **ReumaB** and **ReumaBImp** datasets for DS files.

	ReumaB	ReumaBImp
NaiveBayes	66.53	66.92
BayesNet	65.91	66.84
Logistic	59.35	61.02
SMO	67.35	68.12
J48	65.25	67.13
RandomForest	68.63	70.12

Comparing with the previous section, the results are similar; however, values below 60% are reached and the method with the best average accuracy is RandomForest, for both datasets.

One presents in Table 14 the same results after applying feature selection. All methods improve its performance; for RandomForest its the opposite, for both datasets the performance is lower with feature selection. Another factor to note is the majority of values higher than 70% average accuracy reached by the **ReumaBImp** dataset.

Table 14: Analysis of average accuracy per method in **ReumaB** and **ReumaBImp** datasets after feature selection for DS files.

	ReumaB	ReumaBImp
NaiveBayes	70.02	70.40
BayesNet	69.88	70.45
Logistic	69.86	71.53
SMO	69.69	70.99
J48	66.78	69.67
RandomForest	67.66	69.81

4.3. Data Representation

Finally, considering the DR files, in Table 15 the values of average accuracy for both datasets, **ReumaB** and **ReumaBImp**, are presented. The Logistic classifier was not considered due to processing limitations.

Table 15: Analysis of average accuracy per method in **ReumaB** and **ReumaBImp** datasets for DR files.

	ReumaB	ReumaBImp
NaiveBayes	62.89	67.60
BayesNet	61.96	67.72
SMO	65.71	73.39
J48	57.46	68.27
RandomForest	61.97	68.53

One extract from Table 15 that as in previous section, some values are below 60% accuracy. Con-

cerning the best method, with the best average accuracy, the SMO classifier is the one with highest values for both datasets. Afterward, feature selection was applied to the datasets, presenting in Table 16 the classification results.

Table 16: Analysis of average accuracy per method in **ReumaB** and **ReumaBImp** datasets after feature selection for DR files.

	ReumaB	**ReumaBImp**
NaiveBayes	70.66	77.58
BayesNet	68.83	77.60
SMO	69.18	77.88
J48	57.47	71.58
RandomForest	64.30	74.15

From Table 16, one can conclude that for the **ReumaBImp** the values obtained are always higher than 70% for all classification methods.

4.4. Conclusions

Due to the amount of information, a summary of the obtained results is developed in this section. First, the methods that obtained the best accuracy had to be collected. However, only **ReumaBImp** dataset was considered, since this was the one that obtained the best results in the majority of the case studies.

One presents in Table 17 the best methods for the case study without feature selection. It is notorious that the best method, with the best average accuracy, was the SMO classifier with DR files. In Table 18 is present the same comparison but in the case of usage of feature selection.

Table 17: Comparison between the best results of the case studies used with the **ReumaBImp** dataset without feature selection.

	TS **SMO**	DS **RandomForest**	DR **SMO**
Average [%]	66.93	70.12	73.39

Table 18: Comparison between the best results of the case studies used with the **ReumaBImp** dataset with feature selection.

	TS **SMO**	DS **NaiveBayes**	DR **SMO**
Average [%]	70.13	71.53	77.88

It is clear that, as was shown earlier, the worst case study is with the TS files without feature selection. In addition, it is unquestionable that the best case study is the one that has as processing the DR methods followed by feature selection due to a 77.88% average accuracy. However, it is important to verify with another evaluation measure. For that purpose, an analysis of the values of F-measure was developed.

One concluded that the best classification results are always present when one tries to predict the response code of month B with the data from that month (A=B). However, when A is further away from B in time results become less accurate an this can also be proven true in the results of F-measure. However, given some focus to the non-responder class, one concluded that this class had almost always higher valuer than the other classes.

Whereas the improvement in the results it is very significant to study the selection of features. For the three case studies, the features related with the result of DAS28 and EULAR criteria are present, demonstrating its importance. Furthermore, in two case studies, are present the features representing the values of some components of DAS28, as well as the feature that indicates if biologic treatment, Tocilizumab, is prescribed. Having this, afterward, was developed a study about Tocilizumab therapy.

One concluded after an analysis of the effect of Tocilizumab in the response code that at 3 months of follow-up if the patient is being treated with Tocilizumab, the percentage of responder patients is lower. However, when comparing to the cases in which the patient is taking the medicine, the number of C0 response code is even lower. In fact, this behaviour is more notorious in the next months, where when it is prescribed Tocilizumab there is a large number of responder patients, a fact also proved in [7]. Combining these conclusions with the fact that the algorithms consider this an important attribute, it may indicate that this drug plays an essential role in the positive evolution of the disease. This should be further investigated from a clinical point of view.

However, from the analysis of feature selection can be taken other conclusions. First, it takes very few features, compared to the initial dataset, to get good results. This fact is elicited in the previous subchapters where after a reduction of 96% in the dimensionality, the performance increases.

Since this is a study with a strong biomedical presence, there are some characteristics that, even though they are not linked to the performance of the algorithm, their study can be significant. One presents in Table 19 and 20 the study developed about the best month to predict the response code. All options are concerning the response code after 24 months of follow-up.

It is noticeable that for all case studies the month 6 is the best month in all of them, with an accuracy between 60 and 70%, with the exception of DR after feature selection where after 3 months is as accurate as later on to predict. However, it is important to

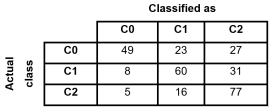
Table 19: Best month for all case studies without feature selection.

	Month	Accuracy [%]
TS	6	65
DS	6	65
DR	6	60

Table 20: Best month for all case studies with feature selection.

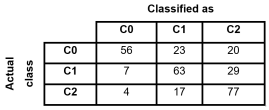
	Month	Accuracy [%]
TS	6	65
DS	6	70
DR	3	65

verify, as before, with another evaluation measure how well behaves the classification methods. One presents in Figure 3 the confusion matrixes for the best-performing methods at the month considered in Table 19 for each case study: NaiveBayes for TS, RandomForest for DS and NaiveBayes again for DR.



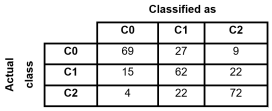
Actual class \ Classified as	C0	C1	C2
C0	49	23	27
C1	8	60	31
C2	5	16	77

(a) Confusion Matrix for **T0:6-24** file applying NaiveBayes at TS dataset.



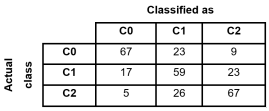
Actual class \ Classified as	C0	C1	C2
C0	56	23	20
C1	7	63	29
C2	4	17	77

(b) Confusion Matrix for **T0:6-24** file applying NaiveBayes at TS dataset with feature selection.



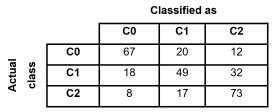
Actual class \ Classified as	C0	C1	C2
C0	69	27	9
C1	15	62	22
C2	4	22	72

(c) Confusion Matrix for **T0:6-24** file applying RandomForest at DS dataset.



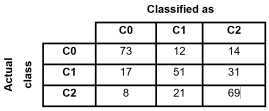
Actual class \ Classified as	C0	C1	C2
C0	67	23	9
C1	17	59	23
C2	5	26	67

(d) Confusion Matrix for **T0:6-24** file applying RandomForest at DS dataset with feature selection.



Actual class \ Classified as	C0	C1	C2
C0	67	20	12
C1	18	49	32
C2	8	17	73

(e) Confusion Matrix for **T0:6-24** file applying NaiveBayes at DR dataset.



Actual class \ Classified as	C0	C1	C2
C0	73	12	14
C1	17	51	31
C2	8	21	69

(f) Confusion Matrix for **T0:3-24** file applying NaiveBayes at DR dataset with feature selection.

Figure 3: Comparison between confusion matrixes with feature selection.

First, one draws from Figure 3 that for all cases the higher values are present in the diagonal of the confusion matrix. This indicates that the models correctly classify the majority of patients response. However, there is in all cases a number of erroneously classified patients. The best method is the method with lower values outside the diagonal.

Having this, one concludes from Figure 3 that from the methods presented, considering the confusion matrix, the best method is RandomForest when applied to DS files achieving the highest values in the diagonal, well classifying the three classes. Nevertheless, given importance and focus to the C0 class, the best method classifying it is NaiveBayes applied to the TS files. It is also important to note that only 3 months of follow-up with NaiveBayes method applied to DR dataset the confusion matrix has values that, are not as good as the previous method, but are also very good.

5. Conclusions

The main goal of this master thesis is the analysis of the Reuma.pt database [20] in order to predict the response code of a patient to a biologic treatment. This is a pioneer study on this database, and, regarding that fact, the database was in need of plenty of processing. After this work, an understanding of the database is richer leading to new possibilities and approaches to studies in the area. Nonetheless, this thesis relies on the importance of some features in the forecasting process. One concludes that the most important features are always related to the calculation of DAS28 and DAS28 itself, with the exception of the feature related to the Tocilizumab therapy indicator and some painful joints. Another relevant factor is the use of data representation and data summarization, in a different perspective from the literature, and achieving good results.

As a global conclusion of this thesis, one can affirm that the results are better with **ReumaBImp** dataset, after applying the feature selection. Another factor that can be extracted from the study of the F-measure is that the lowest value is always in the class C1, corresponding to be a moderate responder. This result is intended since one wants the higher values of F-measure in the extremes, C0 and C2. Concerning the more relevant features, besides the ones related to the EULAR response criteria, and highlighting the indicator of the Tocilizumab therapy, one concludes that when this feature has the value 1, a large number of the patients belongs to the C2 class. This shows that this feature can be highly related to the response code, as seen in literature [7]. Concerning the best month study to predict, all methods achieved an accuracy higher than 60%, and one concludes that after analyzing the confusion matrix, the method with the best performance is the RandomForest applied to the data summarization dataset without feature selection. Finally, some graphs of the J48 classification method are presented, and it is visible the appearance of some painful and tumefact joints in the obtaining of the response code.

Despite the fact that these conclusions are not

supported with statistical measures, one can affirm that the use of SMOTE and discretization improved the results, and their consideration is essential. Besides the use of data representation and data summarization with a different approach lead to better classification results in average. This fact needs to be considered future work, due to its innovative approach and good conclusions. Lastly, reinforcing the idea that this thesis is an exploratory study, the conclusions obtained can lead to different ranges of scientific work.

Having this, some improvements could be made to the study, such as the use of other parameters like a different kernel in SVM. Another improvement is the reduction of the number of files in order to obtain a more concise study. Furthermore, due to the meritorious results obtained for DS and DR, the combination of these two methods could be developed. Finally, a statistical analysis should have been applied to prove which method performs better in all parameters.

Regarding the future work, considering this thesis as a pioneer work, many different paths of investigation could be applied. However, the work developed in the processing of the database and the selection of some features relevant to the forecasting of the response code it is a helpful start for any future work. This is explained by the enormous reduction in the number of missing values, which is the biggest problem of data mining processes. Further, in this study was already possible to find classification methods that at the 6th-month of follow-up of the patient, can forecast with a 70% of accuracy the response code of the same patient at the 24th-month, giving to the community an tool to help in the course of biologic treatment.

Acknowledgements

The author would like to thank the supervisors, Prof. Alexandra Carvalho and Prof. Susana Vinga.

References

[1] DAS28 - Home of the Disease Activity Score and DAS28. `https://www.das-score.nl/das28/en/`. Accessed: August 2017.

[2] R. Agrawal and R. Adhikari. *An Introductory Study on Time Series Modeling and Forecasting*. Lambert Academic Publishing, Germany, 2013.

[3] C. Botas. Master thesis software. `https://github.com/catiabotas/MasterThesis.git`, 2017.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[5] EULAR - European League Against Rheumatism. 10 things you should know about rheumatic disease. `http://www.eular.org/myUploadData/files/10\%20things\%20on\%20RD.pdf`.

[6] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuousvalued attributes for classification learning. In *Thirteenth International Joint Conference on Articial Intelligence*, volume 2, pages 1022–1027. Morgan Kaufmann Publishers, 1993.

[7] J. Freitas. Analysis of electronic medical records of rheumatoid arthritis patients on biological therapies. Master's thesis, Instituto Superior Técnico, 2015.

[8] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 1997.

[9] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. Random forests for land cover classification. *Pattern Recogn. Lett.*, 2006.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[11] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[12] C. Lin, E. W. Karlson, H. Canhão, T. A. Miller, D. Dligach, P. J. Chen, R. N. G. Perez, Y. Shen, M. E. Weinblatt, N. A. Shadick, R. M. Plenge, and G. K. Savova. Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. *PLoS ONE*, 8(8), 2013.

[13] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[14] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Company, 1977.

[15] M. Mayilvaganan and D. Kalpanadevi. Comparison of classification techniques for predicting the cognitive skill of students in education environment. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–4, 2014.

[16] T. Mitsa. *Temporal Data Mining*. Chapman & Hall/CRC, 1st edition, 2010.

[17] J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984.

[18] I. Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[19] A. J. Smola and B. Schölkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, 2004.

[20] E. Sousa, J. Silva, M. Santos, J. Silva, J. Gomes, C. Duarte, J. Silva, L. Cunha-Miranda, A. Teixeira, W. Castelão, J. Branco, J. Costa, D. Araujo, T. Nóvoa, G. Figueiredo, H. Jesus, A. Quintal, A. Cravo, G. Sequeira, P. Pinto, R. André, M. Bernardes, F. Ventura, I. Cunha, A. Barcelos, P. Nero, and M. Cruz. Reuma.pt - The rheumatic diseases portuguese register. *Acta Reumatologica Portuguesa*, 36(1):45–56, 2011.

[21] A. Tucker, S. Swift, and X. Liu. Variable grouping in multivariate time series via correlation. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 31(2):235–245, 2001.

[22] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–179, 1967.