Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Descoberta do Conhecimento - 2017/2018

Predicting the Number of Recovery Days for Breast Tumour Patients

Hugo Carvalho[a], João Nuno Almeida[b], Marcos Luís[c]

[a]*MIEI - A74219, Departamento de Informática, University of Minho*
[b]*MIEI - A75209, Departamento de Informática, University of Minho*
[c]*MIEI - A70676, Departamento de Informática, University of Minho*

Abstract

Treatments for breast tumours are delicate and complicated procedures. Each day a patient spends in post-surgery recovery, translates into an extra expense for the hospital in question. On top of that, the life quality of each patient also decreases. Given these premises, the goal behind this paper, is to predict and to understand which aspects influence the number of days each patient, who has been submitted to a specific kind of treatment, will remain in post-surgery care. This way we may hopefully, increase the life quality of each patient by decreasing the number of days they will have to spend in this stage, and we may improve the fund management of the medical centres giving these treatments. To reach this goal, we used previous patients medical records combined with Data Mining techniques following the CRISP-DM methodology and software tools like WEKA, Excel and RapidMiner to generate the DM models. Some of our models were able to achieve accuracy values of –%, specificity values of –% and sensitivity values of –%.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining; CRISP-DM; WEKA; RapidMiner; Excel; Breast Tumour; Post-Surgery;

1. Introduction

Currently, organisations are relying more and more in decision support tools.

2. Background and Related Work

2.1. Breast Tumours

A breast tumour, just like any other tumour, is a mass of abnormal tissue, created by an unnatural accelerated cell reproduction within the tissue in question. There are two kinds of breast tumours, benign tumours or malign tumours, most commonly known as cancer.

When not treated, benign breast tumours, may cause pain and discomfort, even though they are not as aggressive as malign tumours, benign tumours may occasionally grow and start affecting surrounding tissues and organs. Malign tumours or cancers, are extremely aggressive towards surrounding tissues, for their only goal is to attack and destroy them, creating metastasis, i.e., secondary tumours in other body parts.

Peer-review under responsibility of the Conference Program Chairs.

The most common breast tumour symptom is a lump, that feels different from the rest of the breast tissue. Breast tumours are usually gender biased with 99% of the total number of cases being diagnosed in women. The most affected age group is woman with 65 years or more.

Like any other evasive surgery, there are risks associated with the treatment of breast tumours. Studies have concluded that age is not a risk factor for post-operative complications, neither is something that influences the days spent in post-operative care, and that the rate of POCs (post-operative complications) is currently 15.2%. However there is still no conclusion on what is the predicted number of days needed to recover from such treatments, because no one can really put their finger on what factors affect that variable, even though the current average time is 7 days. Having the knowledge beforehand of what motivates the need to remain in the post-operative condition or not, would allow doctors and health professionals to, if possible, avoid certain types of treatments or approaches that they knew would increase the number of days admitted in the hospital in a certain variety of patients. This would improve the patients life quality by reducing the total time spent in a hospital environment. Having a predicted value of days for each patient in advance, would also allow the hospital to manage their funds more accurately, distributing the remaining funds more accordingly.

2.2. Related Work

In this section we will presented some works that have applied Data Mining techniques, in an effort to support the existence of behavioural patterns in the breast cancer research area.

Delen used three Data Mining techniques to predicted the survivability of patients diagnosed with breast cancer. Real data from more than 200000 cases was compiled in a dataset, used, and applied in two of the most popular Data Mining algorithms: artificial neural networks and decision trees. Along with these, a statistical method (logistic regression) was used, to develop the prediction models. A 10-fold cross validation, was also used to measure the unbiased estimate, of the predictions obtained using the algorithms referenced above. The results achieved in terms of accuracy, with the use of the decision tree algorithm, were the best ever recorded in literature with a value of 93.6%. In second came the values achieved by the artificial neural network algorithm, with an accuracy level measured at 91.2%. Lastly came the results achieved by the logistic regression models with an accuracy of 89.2%.

Asri used machine learning algorithms to predict the risk of breast cancer and to diagnose it. The data used was real, and extract from the Wisconsin Breast Cancer datasets. For this study, four machine learning algorithms were considered: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbours (k-NN). In the end, the most accurate machine learning algorithm proved to be the SVM algorithm, with and accuracy measured at 97.13% and the lowest error rate of the four at 0.02%.

3. Methodology, Data and Data Mining

In this section we will explore the subjects of the methods used to reach conclusions, the data mining process itself and techniques used and the data sources available.

3.1. Dataset

The data sample used to represent the population of breast tumour patients, was extracted from medical records of 176 real anonymous patients, from ages 14 to 85. It comprises the months of January and February of the year 2017, and there is no information within the data sample that specifies from which hospital(s) were the records taken, therefore we can not make any conclusions about the nationalities of the people involved. Because the patients are anonymous we can not take any elation about their sex either. The data sample contains, for each patient, information in the form of 23 attributes, about their personal health history, the treatment they were submitted to, their behaviour in the post-surgery period and any complication that might have occurred.

3.2. CRISP-DM

We followed the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, during the Data Mining Process. Essentially it divides the Data Mining process into six phases: *Business Understanding, Data Un-*

derstanding, Data Preparation, Modelling, Evaluation, Discussion. This *framework* was followed incrementally, by the order specified in order to increase the success of the Data Mining process as a whole and its final conclusions.

3.3. Data Mining

To generate the Data Mining models needed, three different techniques were used: *Linear Regression* (LR), *Support Vector Regression* (SVR) and *Artificial Neural Networks* (NN). We used Weka to implement these machine learning algorithms.

LR works by estimating coefficients for a line or hyperplane that best fits the training data. It is a very simple regression algorithm, fast to train and can have great performance if the output variable for your data is a linear combination of your inputs. LR only supports regression type problems.

SVR was developed for binary classification problems, although extensions to the technique have been made to support multi-class classification and regression problems. Unlike SVM that finds a line that best separates the training data into classes, SVR works by finding a line of best fit that minimizes the error of a cost function.

NN are a complex algorithm to use for predictive modelling because there are so many configuration parameters that can only be tuned effectively through intuition and a lot of trial and error.

4. Data Mining Process

4.1. Business Understanding

The goal of the work this paper is presenting, is to predict the number of days a breast tumour patient will remain in post-surgery care. The aspects to be considered consist of, the personal health history of each patient, the treatment he was submitted to and the complications that may have occurred in the meantime, given the way each patient deals with the disease and the treatment. Knowing beforehand what patterns, may affect the number of days spent in post-surgery care, not only increases the life quality of future breast tumour patients by avoiding, if possible, certain treatments that may extend that period, but it also allows hospitals and medical centres to predict and manage the costs involved in keeping the patients admitted with a lot more care and knowledge ahead of time.

4.2. Data Understanding

The dataset used consists of 176 entries. Each entry represents, the records and information of an anonymous real former breast tumour patient. Unfortunately we can not specify from which hospital(s) were the records taken. Each entry is described by a set of 23 attributes, divided in 4 groups: Personal History, Surgery, Post-Operative, Complications.

Table 1 and Table 2 show the statistical distribution of the attributes in Personal History, and presents the attributes themselves to the reader. Table 3 shows the statistical distribution of the attributes in Surgery, and presents the attributes themselves to the reader. Table 4 shows the statistical distribution of the attributes in Post-Operative, and presents the attributes themselves to the reader. Table 5 shows the statistical distribution of the attributes in Complication, and presents the attributes themselves to the reader.

Table 1: Numerical attributes in group: Personal History.

Name	Minimum	Maximum	Mean	Standard Deviation
Idade	14	85	53.023	15.560

Within the dataset there is an unnamed target attribute. This unnamed attribute, draws a conclusion on whether there was a post-treatment complication or not, for each patient. The attribute specifies which complication occurred if that is the case. Because that is not the issue of this work, we will normalise the information in a binary way: yes or no, for the emergence of a complication. As we can see in the following graph, a great percentage of the

Table 2: Nominal attributes in group: Personal History.

Name	Range	Percentage (%)
Tabaco	Sim	15.9
	Não	84.1
Diabetes	Sim	8.0
	Não	92.0
Imunosupressores	Sim	3.4
	Não	96.6
Hipocoagulação	Sim	2.3
	Não	97.7
QTx NA	Sim	8.8
	Não	91.2

Table 3: Nominal attributes in group: Surgery.

Name	Range	Percentage (%)
Data Cx	02/01/2017 a 27/02/2017	100.0
Cx/Ambulatório	Cirurgia	58.0
	Ambulatório	42.0
Benígno/Malígno	Benígno	19.9
	Malígno	80.1
Diagnóstico	CI NST	51.7
	Fibroadenoma	9.7
	CDIS Alto Grau	4.5
	CL Invasor	4.5
	Others (37)	29.6
Lateralidade	D	49.7
	E	46.9
	B	3.4
Intervenção Mama	TA	38.9
	BE	17.7
	MRT	8.8
	MRM	7.1
	Others (15)	27.5
Intervenção Axila	GS	81.6
	EA	17.1
	Exérese de gânglio	1.3
Outras Intervenções	CVC TI	54.5
	Remoção CVC TI	12.5
	Simetrização	5.7
	RMI	2.3
	Others (21)	25.0

Table 4: Numerical attributes in group: Post-Surgery.

Name	Minimum	Maximum	Mean	Standard Deviation
Dias	0	26	1.710	2.972

patients, *80.2%*, did not show signs of any complication that required them to be admitted for longer than predicted, nevertheless, there is a percentage of patients, *19.8%*, that suffered from these complications. However, even though this is the only explicit target attribute, this is not the target attribute we will use considering our goals with this work. Our goal is to predict the number of recovery days for breast tumour patients, and therefore our target attribute will be the attribute *Dias*.

Table 5: Nominal attributes in group: Post-Surgery.

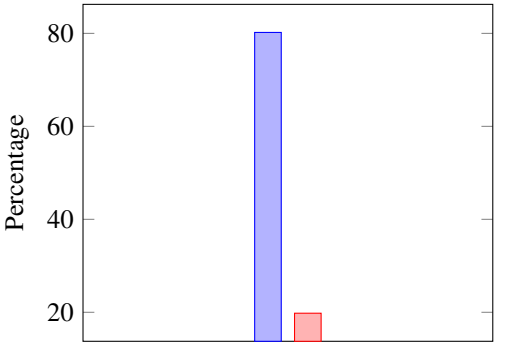
Name	Range	Percentage (%)
Antibióticos	Sim	10.2
	Não	89.8
Hipocoagulação	Sim	54.3
	Não	45.7

Table 6: Numerical attributes in group: Complications.

Name	Minimum	Maximum	Mean	Standard Deviation
Dias Pós	1	85	20.714	24.724
Dias Tx	—	—	—	—

Table 7: Nominal attributes in group: Complications.

Name	Range	Percentage (%)
Data Dx	02/01/2017 a 27/02/2017	2.3
Complicação	Hematoma	14.8
	Seroma	11.1
	Atraso Cicatrização	7.4
	Deiscência	7.4
	Others (15)	59.2
Tratamento	Conservador	45.5
	Médico	22.7
	Drenagem Aspirativa	18.2
	Drenagem Cirúrgica	4.5
	Penso	4.5
	Pico	4.5



Graph 1. Target Attribute distribution

4.3. Data Preparation

In order to create accurate DM models, the data had to be prepared. The first step in the data preparation process was to eliminate all the null entries of the dataset, that were causing nothing but inconsistency. First of all, we started by eliminating the attributes ”Nome”, ”Data Cx” and ”Dias Tx”, because they weren’t in any way relevant considering the goals of this study. In total, at the end of this phase, we took out 169 entries of the original 176. We handled these missing values, by deleting every entry that corresponded to a missing value of less than 5% in a certain attribute. By replacing the missing values of a certain attribute, by its mean or mode, depending on whether it was a numerical or

nominal type, if the percentage of missing values was greater than 5% and less than 80%. And by deleting the attribute from the dataset in its whole if the percentage of missing values was bigger than 80%. We also felt the need to change the instances of some attributes. For nominal attributes where the range of values consisted in two possibilities, for example, "Yes" or "No", we normalized the data in order to have a range of values of 0 and 1. In cases where the range of values was more extensive, for example in the attribute "Intervenção Mama", we used the 4 most common occurrences and flagged the rest with the value "Outras". Next we handled the outliers within the remaining data to prevent them from influencing our final conclusions. We excluded from the dataset every entry that had a numerical attribute, out of the range [Mean-2SD, Mean+2SD]. Ultimately the dataset we were left to work with, had 156 entries, each one with 16 attributes. Table 8 presents the decisions taken for each attribute.

Table 8: Data Preparation decisions.

Name	Missing Values	Decision	Name	Missing Values	Decision
Nome	0%	Remove	Idade	0%	Keep
Diabetes	0%	Normalize	Imunosupressores	0%	Normalize
QTx NA	3.41% (Remove)	Normalize	Cx/Ambulatório	0%	Normalize
Diagnóstico	0%	Normalize	Lateralidade	0.57% (Remove)	Keep
Intervenção Axila	56.2% (Mode)	Keep	Outras Intervenções	50.0% (Mode)	Keep
Name	Missing Values	Decision	Name	Missing Values	Decision
Tabaco	0%	Normalize	Hipocoagulação	0%	Normalize
Benígno/Malígno	0%	Normalize	Intervenção Mama	35.8% (Mode)	Keep
Dias	0%	Keep	Dias Pós	92.1%	Remove
Antibióticos	0%	Normalize	Hipocoagulação	0.57% (Remove)	Normalize
Complicação	84.1%	Remove	Tratamento	87.5%	Remove

4.4. Modeling

This phase consisted of inducing the Data Mining Models (DMM) in Weka using the prepared data. As the described approach corresponds to a linear regression problem, we used 3 different DM techniques: Linear Regression (LR), Support Vector Regression (SVR) and Multi-Layer Perceptron (MLP). These algorithms were used with the default settings in Weka.

Two different data approaches were made, one of them using oversampling and the other without it, both testing on 1/3 of the data (Holdout Sampling) and on all the data (Cross Validation with 10 folds). The different scenarios presented below were also created combining different variables in order to identify which factors have more impact on predict the number days of hospitalization for a patient:

S1: All variables

S2: Idade, Diabetes, Imunossupressores, Hipocoagulação, QTx NA

S3: Cx/Ambulatório,Ben./Malig., Diagnostico, Lateralidade, Intervencão Mama, Intervencão axila, Outras intervencões

S4: Antibióticos, Hipocoagulação

S5: Diabetes, Hipocoagulação, Intervenção Mama, Intervenção Axila, Outras Intervenções, Antibióticos, Hipocoagulação

Each DMM can be described as belonging to an approach (A), being composed by a scenario (S), a data mining technique (DMT), a sampling method (SM), a data approach (DA) and a target (T):

$$DMM_s = \{A_f, S_i, DMT_y, SM_c, c, TG_t\}$$

$A_f = LinearRegression$
$S_i = S1, S2, S3, S4, S5$
$DMT_y = LR, SVR, MLP$
$SM_c = HouldoutSampling, CrossValidation$
$DA_b = WithoutOversampling, WithOversampling$
$TG_t = Days$

4.5. Evaluation

With the help of the confusion matrix (CMX) we were able to determine the quality of each one of our DMM. This matrix allows us to determine the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Having these values lets us to calculate the following measures:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FT + TN + FN}$$

The following tables 9, 10 and 11 show for each DM algorithm, which models were able to reach the best values of Sensitivity, Specificity and Accuracy.

Table 9: DM models sensitivity values for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
LR				
SVR				
NN				

Table 10: DM models specificity values for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
LR				
SVR				
NN				

Table 11: DM models accuracy values for each DM technique

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
LR				
SVR				
NN				

As we you see on the tables above the best overall sensitivity value achieved was ——, the best overall specificity value achieved was ——- and the best overall accuracy value achieved was ——-.

To rank the models accordingly, we decided to define, for each measure, a threshold value beneath which the models would be excluded. This threshold was defined as ¿–% for sensitivity, ¿–% for specificity and ¿–% for accuracy.

Table 12 shows the models that reached the established default threshold values for all the three measures.

Table 12: Top DM models with the best accuracy results

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity	Specificity	Accuracy

5. Discussion

6. Conclusion and Future Work

7. Acknowledgements

We thank Professor José Manuel Ferreira Machado, and Professor Hugo Peixoto for sharing with us their knowledge and experience in Data Mining studies, allowing us to conduct our own research.

References

1. Van der Geer J, Hanraads JAJ, Lupton RA. The art of writing a scientific article. *J Sci Commun* 2000;**163**:51-9.
2. Strunk Jr W, White EB. *The elements of style*. 3rd ed. New York: Macmillan; 1979.
3. Mettam GR, Adams LB. How to prepare an electronic version of your article. In: Jones BS, Smith RZ, editors. *Introduction to the electronic age*. New York: E-Publishing Inc; 1999. p. 281-304.