Research Article

Rita Reis, Hugo Peixoto*, José Machado, and António Abelha

Machine Learning in Nutritional Follow-up Research

https://doi.org/10.1515/comp-2017-0008
Received Nov 17, 2017; accepted Nov 29, 2017

Abstract: Healthcare is one of the world's fastest growing industries, having large volumes of data collected on a daily basis. It is generally perceived as being 'information rich' yet 'knowledge poor'. Hidden relationships and valuable knowledge can be discovered in the collected data from the application of data mining techniques. These techniques are being increasingly implemented in healthcare organizations in order to respond to the needs of doctors in their daily decision-making activities. To help the decision-makers to take the best decision it is fundamental to develop a solution able to predict events before their occurrence. The aim of this project was to predict if a patient would need to be followed by a nutrition specialist, by combining a nutritional dataset with data mining classification techniques, using WEKA machine learning tools. The achieved results showed to be very promising, presenting accuracy around 91%, specificity around 97% and precision about 95%.

Keywords: Health Information Systems; Data Mining; Classification Techniques; Decision Support Systems; Nutrition Evaluation

1 Background

Every day, millions of patients go to health institutions to doctor appointments. Usually, physicians perform a general check-up to understand patients' clinical status. When patients are in a nutritional state that is not appropriate, a request is sent to the nutrition service to decide whether or not the patient should have nutritional monitoring. If the process time between the request and the answer from the nutritionist is substantial, the malnourished patient will not have the needed follow-up. Information technologies are being increasingly implemented in healthcare organizations in order to respond to the needs of doctors in their daily decision-making activities. In critical environments, for example when a patient is malnourished and needs immediate monitoring, decisions need to be performed quickly [1].

As health organizations generate and store large volumes of data every day, clinical decisions could be made not only based on doctor's intuition and experience but also based on hidden knowledge stored over time in healthcare databases [2]. Thus, and to satisfy the urgent need for extraction of useful information from the large amount of data collected, it is fundamental to develop a solution able to predict events before their occurrence. In this sense, the aim of this project was to predict the nutritionist's response to a request from a physician, by applying data mining techniques, to reduce the process time between the request and the answer, therefore provide an immediate and adequate treatment to the patient. Wu, *et al.* proposed that integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data mining has the potential to generate a knowledge-rich environment which can help to improve the quality of clinical decisions and consequently improve the quality of service provided to patients [3].

Identifying the risk of malnutrition in patients from predictive variables is the first step towards an adequate nutritional control. Given its prevalence, the traceability and monitoring of nutritional status should be available in the hospital environment to prevent, treat and improve its prognosis. With this, morbidity, mortality, as well as hospitalization time and hospital costs will be reduced, enhancing the quality of patients' life. Given this reality, the nutritionist plays a crucial role, since it's able to identify early cases of nutritional risk and, consequently, prevent and control malnutrition. Each health institution should

***Corresponding Author: Hugo Peixoto:** Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal; Email: hpeixoto@di.uminho.pt
Rita Reis: University of Minho, Campus Gualtar, Braga 4710, Portugal
José Machado, António Abelha: Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal

Table 1: Dataset discrete attributes

Discrete Attribute	*Description*	*Variable*	*Class*	*Cases (%)*
Nutri Follow-up (NF)	Indicates if a patient need to be followed by a nutrition specialist or not.	Yes	1	41,64
		No	0	58,36
Nutrition Classification (NC)	Categories that indicates the nutritional status of the patient, according to BMI.	Underweight	1	58,25
		Normal Weight	2	32,16
		Pre-obesity	3	7,35
		Obesity Class I	4	1,70
		Obesity Class II	5	0,21
		Obesity Class III	6	0,32

Table 2: Dataset continuous attributes

Continuous Attribute	*Description*	*Min*	*Max*
BMI	Ppatient's weight divided by the square of the person's height (kg/m^2)	13,38	46,30
Weight (W)	Patient's weight (kg)	26	136
Height (H)	Patient's height (m)	1,20	1,92
Age (A)	Patient's age	13	102

be responsible to identify nutritional risk factors that affect the population and implement an instrument of nutritional tracking [4].

Machine learning is the study of computer algorithms that improve automatically through experience and can be used to develop systems resulting in increased efficiency and effectiveness [5]. Machine learning provides the technical basis of data mining. It is used to extract information from the raw data in databases. Data mining is the application of specific algorithms in extracting patterns from data [6]. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage [7]. In J. Han and M. Kamber's book, data mining was defined as "the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories" [8].

2 Materials and Methods

The data used for this project was extracted from a hospital in Portugal and allowed the prediction of nutritionist's answers to physician's requests regarding a malnourished patient. Five different data mining models were included in this study: Decision Trees, Support Vector Machines, Bayesian Networks, Decision Rules and Nearest Neighbours.

The CRISP-DM (CRoss Industry Standard Process for Data Mining) provides a framework for carrying out data mining activities. In this project, CRISP-DM was used. It consists of six phases which are well structured and defined. The first phase is the understanding of the business activities while the data for carrying out business activities are collected and analyzed in the second phase. Data preprocessing and modelling is done in the third and fourth phase respectively. Fifth phase evaluates the model and last phase is responsible for the deployment of the constructed model [9]. The data mining goal was to create useful models able to predict the probability of a patient to be followed and monitored by a nutrition specialist, by applying classification techniques. The solution should be able to support medical decisions and make more information available to the intensivists, providing new knowledge in this field. The raw data of this experiment consisted of 2892 patients' medical records, recording a period between August 1st, 2011 to January 4th, 2017 (1984 days). There were 15 attributes extracted, such as doctor's request date, nutritionist's answer date, the service that demands the request, patient's birthdate, patient's height and weight and if the patient was followed or not by a nutritionist.

After careful analyses, 6 attributes were selected to be applied data mining techniques in the modelling phase. Nutrition Follow-Up as the decision attribute, meaning the

patient not have been or have been followed by a nutrition specialist, respectively. The dataset was composed by discrete attributes, presented in Table 1, and continuous attributes, presented in Table 2. Databases are highly susceptible to noisy, missing, and inconsistent data due to huge size, complexity and their likely origin from multiple heterogeneous sources [2]. Data that are not considered reliable may cause confusion during the mining process, which may lead to inaccurate results [10]. To solve this problem and guaranty data quality, the next step was data cleansing.

It was conducted a search for errors, data omission and data integrity and then several solutions to correct the errors were proposed. The errors encountered were blank spaces, where data that were not filled by doctors and nutritionists, or was information with writing errors and symbols. The next step was to convert "yes" and "no" values to Boolean values (0 and 1).

Another key component was data transformation. Therefore, and also in this step, data was transformed into appropriate formats for the mining process, using WEKA, through several operations. Namely: Attribute Construction where new attributes were built from the given set of attributes. Smoothing, in which there was a search for the occurrence of values out of the acceptable range (noise values). These noise values were removed from the data. Normalization, where data was sized to be inserted at a short reference interval: 0-1. Finally, discretization where it was divided the range of continuous attributes into intervals.

Several algorithms of automatic learning consider that the values that a class can assume will present equal probabilities [11]. This fact does not always occur, as in this case study, as the number of patients that needed to be followed by a nutrition specialist was considerably lower than the number of patients that didn't need to be followed. Since this imbalance would affect the hit rate for the lowest occurrence class, it was used an approach to balance the dataset that involved the removal of tuples from the dominant class. Despite the existence of the problem of potentially useful data being eliminated, this approach significantly increased the performance of classifiers. The modelling phase began by choosing the best algorithms to be applied to the dataset. Several data mining methods are implemented in WEKA software. Some are rule-based like ZeroR, or Bayesian learning algorithms, like NaïveBayes. The algorithms that implement SVMs use the SMO method (for classification tasks). For a K-nearest neighbor's classifier it is used IBk (Instance-based learning with parameter k). As far as decision trees are concerned, one of the algorithms used for classification is J48, which is a simple implementation of the C4.5 algorithm. A set of data mining models were induced using five data mining techniques, Bayesian Networks (T1 - BN); Support Vector Machines (T2 - SVM); Decision Trees (T3 - DT); Decision Rules (T4 - DR); Nearest Neighbor (T5 - NN), and two sampling methods: holdout sampling (SM1) and cross validation (SM2). The final dataset presented 1877 rows. To the holdout sampling method, the subset of 1314 (70%) was used as a training set and the remaining 563 cases (30%) was used as a testing set, in order to evaluate the performance of the classifiers. In the cross validation method, the data were divided into 10 folds. The last step of the modelling phase was to group attributes into varied scenarios to generate different models. The first set of attributes created was Case Mix, which contained all the attributes from Tables 1 and 2. Scenario 2, 3 and 4 were created to understand if any given attribute had positive or negative impact in the final results. The scenarios are:

S1: {NF; BMI; A; NC; W; H};
S2: {NF; BMI; NC; W; H};
S3: {NF; A; NC};
S4: {NF; BMI; A; W; H}.

3 Results

The application of different methods of automatic learning requires a process that ensures that the results are reliable and statistically significant. Thus, performance metrics were applied to assure the evaluation of the quality and characteristics of the models, guaranteeing the reliability of the results. These metrics are numerical measures that quantify the performance of a given classifier. The metrics used in this project were accuracy (1), sensitivity (2), specificity (3) and precision (4). To calculate these metrics, the confusion matrix obtained for each model with WEKA software was used. This matrix presents the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) of a given model. Table 3 and Table 4 present the scenarios that achieved the best results, by technique, and for both sampling methods, holdout sampling and cross validation, respectively.

$$\text{Accuracy} = TP + TN/(TP + TN + FP + FN) \quad (1)$$

$$\text{Sensitivity} = TP/(TP + FN) \quad (2)$$

$$\text{Specificity} = TN/(TN + FP) \quad (3)$$

$$\text{Precision} = TP/(TP + FP) \quad (4)$$

Table 3: Best results obtained, by technique, with holdout sampling method

Holdout Sampling								
Tech.	***Accuracy***		***Sensitivity***		***Specificity***		***Precision***	
BN	**S1**	0,842	**S4**	0,700	**S1**	0,950	**S1**	0,913
SVM	**S4**	0,851	**S1; S3**	0,713	**S4**	0,963	**S4**	0,933
DT	**S4**	0,908	**S1; S4**	0,829	**S4**	0,972	**S4**	0,954
DR	**S1; S2; S3**	0,838	**S3**	0,713	**S1; S2; S4**	0,957	**S1; S2; S4**	0,921
NN	**S1; S2**	0,888	**S1**	0,825	**S4**	0,947	**S4**	0,919

Table 4: Best results obtained, by technique, with cross validation method

Cross Validation								
Tech.	***Accuracy***		***Sensitivity***		***Specificity***		***Precision***	
BN	**S1**	0,833	**S1**	0,694	**S2**	0,946	**S2**	0,896
SVM	**S1**	0,839	**S1**	0,745	**S2**	0,921	**S2**	0,868
DT	**S1**	0,887	**S1**	0,808	**S1**	0,955	**S4**	0,918
DR	**S1; S2; S4**	0,815	**S3**	0,677	**S1;S2;S4**	0,959	**S1; S2; S4**	0,897
NN	**S1**	0,895	**S4**	0,847	**S1**	0,929	**S1**	0,893

Table 5: Best models achieving the defined threshold

Model	***Accuracy***	***Sensitivity***	***Specificity***	***Precision***
Scenario 1 DT SM1	0,904	0,829	0,960	0,939
Scenario 4 DT SM1	0,908	0,829	0,966	0,948

To choose the best model, a threshold was introduced. The threshold combines four metrics to find the most suitable model to predict the probability of having a patient following nutrition (sensitivity) with a high specificity, an acceptable accuracy and a high precision in order to avoid a high number of false positives. It is important to note that, in clinical context, losing accuracy (in practice, incorrectly classifying negative instances) can be tolerated as long as it does not involve a high cost. The threshold defined was: Accuracy >=88%, Sensitivity >=80%, Specificity >=95% and Precision >=90%. Table 5 presents the two models that achieved the defined threshold. In machine learning, current research has shifted away from simply presenting accuracy results when evaluating algorithms that output probabilities of class values [12]. For this reason, there were constructed two curves: Receiver Operator Characteristic (ROC) and Precision-Recall (PR). The first one is created by plotting the true positive rate (sensitivity) against the false positive rate (specificity). The second one shows how precision varies with sensitivity (recall). In Figure 1. it is possible to observe both curves for the model that achieved the best results«, which is Scenario 4; Technique 3 and Sampling Method 1.

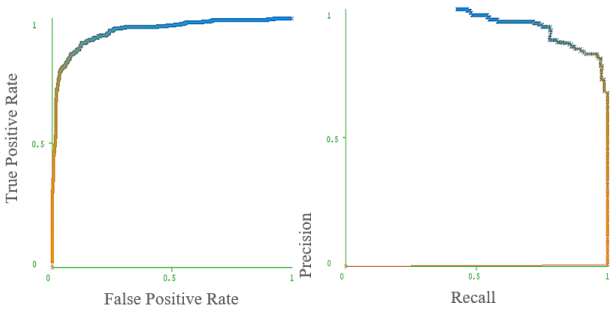

Figure 1: ROC curve (left) and PR curve (right).

4 Discussion and Conclusions

By analyzing Tables 3 and 4, it is possible to observe that the achieved results were generally better with holdout sampling method. In S3 scenario there was a significant negative impact on results because BMI, Weight and Height attributes, that are directly related, were removed. In this scenario, the results for the great majority of the techniques were lower than the other scenarios, as can be observed by the almost inexistence of S3 in Tables 3 and 4. On the other hand, there was a positive impact

on results when the attribute NutritionClassification was withdrawn, in scenario S4. Scenario S2 had a neutral impact on results, which means that the Age attribute does not substantially influence results. To be able to choose the best model, a threshold which combining four metrics that were introduced in the previous phase. The induced models that achieved the defined threshold combined hold-out sampling method, Random Forest algorithm from Decision Tree data mining technique and scenarios 1 and 4. From these two, and by observing the obtained results in Table 5, it can be concluded that model S4T3SM1 was the best constructed model with about 91% of accuracy, 83% of sensitivity, 97% of specificity and 95% of precision, all high metric values. Furthermore, Figure 1 shows ROC and PR curves for the classification of the dataset instances. The goal in ROC space is to be in the upper-left-hand corner, having an area under the ROC curve on a range from 0 to 1, where 1 represents the perfect classifier, and 0 is a classifier that is always wrong [12]. The area under ROC curve, in Figure 1, was 0,947. In PR space, the goal is to be in the upper-right-hand corner [12]. When observing both curves in Figure 1 they appear to be close to optimal. Data mining has great importance for the healthcare industry, and it represents comprehensive process that demands the understanding of needs of healthcare organizations. Knowledge gained with the use of data mining techniques can be used to make successful decisions that will improve the quality of services provided to patients. In this paper, it's shown the role of data mining for the use of evidence-based medicine in the context of nutrition evaluation. There were constructed useful models able to predict the probability of a patient to be followed and monitored by a nutrition specialist, through data mining classification techniques. The best constructed model was certified by different metrics to assure the quality of the results. It presented a level of accuracy of 91%, a level of specificity of 97%, sensitivity rounded 83% and precision 95%. In this sense, and as a quality model, it can be used for the construction of clinical scenarios which will support the healthcare providers to predict patients' outcomes in the context of nutrition evaluation. The solution provided is able to support medical decision-making and could contribute to the improvement of the nutritional condition of the population.

For future work, it could be suggested to include datasets from different hospital facilities from various regions to identify patterns in data at national level. Additionally, more experiments could be done on using different parameters and data mining techniques.

Acknowledgement: This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

[1] Portela, F., Santos, M. F., Machado, J., Abelha, A., Rua, F., & Silva, Á. (2015). Real-time decision support using data mining to predict blood pressure critical events in intensive medicine patients. In Ambient Intelligence for Health (pp. 77-90). Springer International Publishing.

[2] Abirami, N., Kamalakannan, T., & Muthukumaravel, A. (2013). A Study on Analysis of Various Data Mining Classification Techniques on Healthcare Data. International Journal of Emerging Technology and Advanced Engineering, 3(7), 604-607.

[3] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

[4] Reis, R., Mendonça, A., Ferreira, D. L. A., Peixoto, H., & Machado, J. (2017). Business Intelligence for Nutrition Therapy. In Next-Generation Mobile and Pervasive Healthcare Solutions (pp. 203-218). IGI Global.

[5] Eapen, A. G. (2004). Application of Data mining in Medical Applications.

[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[7] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[8] Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Fco., CA., USA.

[9] Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). Int. J. Innov. Sci. Res, 12(1), 217-222.

[10] Milovic, B., & Milovic, M. (2012). Prediction and decision making in health care using data mining. Kuwait Chapter of the Arabian Journal of Business and Management Review, 1(12), 126.

[11] Ferreira, P. M. S. (2010). Aplicação de Algoritmos de Aprendizagem Automática para a Previsão de Cancro de Mama (Master Thesis). Faculdade de Ciências da Universidade do Porto, Porto, Portugal.

[12] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM