Predicting death and morbidity in perforated peptic ulcer

Hugo Peixoto[1], Lara Correia e Silva[2], Soraia Pereira[2], Tiago Jesus[2], Vítor Lopes[3], António Abelha[1] and José Machado[1]

[1] Algoritmi Research Center, University of Minho, 4710, Braga, Portugal.
[2] University of Minho, Campus de Gualtar. 4710, Braga, Portugal.
[3] Centro Hospitalar do Tâmega e Sousa, 4564 Penafiel, Portugal
{hpeixoto, abelha, jmac}@di.uminho.pt
{a74923, a74713, a73787}@alunos.uminho.pt; 72777@chts.min-saude.pt

Abstract. Peptic ulcers are defined as defects in the gastrointestinal mucosa that extend through the muscularis mucosae. Although not being the most common complication, perforations stand out as being the complication with the highest mortality rate. To predict the probability of mortality, several scoring systems based on clinical and biochemical parameters, such as the Boey and PULP scoring system have been developed. This article explores, using data mining in the medical data available, how the scoring systems perform when trying to predict mortality and patients' state complication. We also try to conclude, from the two scoring systems presented, which predicts better the situations described. Regarding the results, we concluded that the PULP scoring allows a better mortality prediction achieving, in this case, above 90\% accuracy, however, the results may be inconclusive due to the lack of patients who died in the dataset used. Regarding the complications, we concluded that, on the other hand, the Boey system achieves better results leading to a better prediction when it comes to predicting patients' state complication.

Keywords: Data Mining, dataset, scoring systems, PULP, Boey, death, health complications.

1 Introduction

Peptic ulcers are defined as defects in the gastrointestinal mucosa that extend through the muscularis mucosae [1]. Complications of peptic ulcer disease include bleeding, perforation, penetration, and gastric outlet obstruction. Its incidence varies geographically – in developed countries, hemorrhage is the most common cause (up to 73%), followed by perforation (9%) and obstruction (3%). Although not being the most common complication, perforations stand out as being the complication with the highest mortality rate [2]. Different numbers present in countries categorized as developing countries, as a review from Nigeria demonstrates, with obstruction being the most common cause of complication (56%), followed by perforation (30%) and bleeding (10%) [3]. Pieces of evidence implicates that the primary causes of

complicated peptic ulcer disease are H. pylori infection and nonsteroidal inflammatory drugs. This work is based on a dataset that evaluates several clinical and biochemical parameters in patients with the diagnosis of peptic ulcer perforation, in order to classificate them in two scoring systems – Peptic Ulcer Perforation (PULP) [5] and Boey [6] – previously established to determine a patient's prognosis with this pathology.

The main objective of this work is to establish a relation between the scoring systems (PULP and Boey) in patients with perforated peptic ulcer and the outcome (mortality and morbidity) [4]. In order to achieve all the goals proposed was necessary to follow a detailed process that allows you to analyze large amounts of data. This process is entitled "Data Mining". To easily understand the method used, first will be explored, superficially, the concept of Data Mining and the main features of all the process. Then, we will explore, briefly, some of the work already done in this area. Only after that, the Data Mining process will effectively start, passing through all the five phases that constitute the process. In the end, it will be done an extensive analysis of the results and drawn some conclusions relevant to the propose.

2 Methodologies, Materials and Methods - Data Mining

Data mining is the process of extracting knowledge that matters from massive amounts of data. The idea is to build computer programs that sift through databases automatically, seeking patterns. If found, these regularities will likely generalize to make accurate predictions or descriptions on future data. It is an interdisciplinary field with contributions from many areas, such as statistics, machine learning, information retrieval, pattern recognition and bioinformatics. The main techniques for data mining include classification and prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and text mining, and some new techniques such as social network analysis and sentiment analysis. It should also be noted that this process can be divided into two major categories, different from each other, the supervised and the unsupervised methods. The major difference between them is the final goal sought with of the application of the method. While the supervised search the achievement of a value, requiring a target attribute, the unsupervised search for relations between the data, such as affinities and intrinsic structures [7][8][9].

Six Data Mining techniques were used to induce the data mining models: JRip, OneR, NB, DT, J48 and K-Means. JRip is one of the basic and most popular algorithms. It implements a propositional rule learner, repeated incremental pruning to produce error reduction (RIPPER). It as a building state (formed by a growing phase and a pruning phase) and an optimization stage [10][11]. In its turn, OneR (short for One Rule) is a class for building a OneR classifier: it is a simple classification algorithm that uses a one-level decision tree. In other words, it uses the minimum-error attribute for prediction, discretizing numeric attributes [12]. Then, the NB, stands for Naive Bayes, it is a class for a Naive Bayes classifier that uses estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. The NB classifier learns very quickly and is easy to apply [13][14].

Another model used was DT, that stands for Decision Table, and it is a class for building and using a simple decision table majority classifier. Decision tables are very easy to understand and one of the simplest hypothesis spaces possible [11]. J48, the fifth of the six models used, is a class for generating a pruned or unpruned C4.5 decision tree. At least, it was used K-Means model which is a cluster data using the k means algorithm. It can use either the Euclidean distance (default) or the Manhattan distance. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean [15]. In order to implement those data mining models, it was necessary to follow a specific process, that delimits all the steps and decisions that must be taken. The decision fell on the CRISP-DM process, a Cross-Industry Process for Data Mining. The CRISP-DM methodology provides a structured approach to planning a data mining project and consists on a hierarchical process model that divides the process of data mining into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [16][17][18]. This methodology was followed throughout this work. Therefore, in order to correctly perform each of the steps of CRISP-DM, it was necessary to resort to different tools and try to combine their results. The main software's used were WEKA and R. WEKA is a machine learning software that allows a specific analyzation of the data, based on a collection of machine learning algorithms for data mining tasks [12]. On the other hand, R is a free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing and based on R language, widely used for developing statistical software and data analysis [16].

3   Related Work

Data mining, as already mentioned, is a powerful tool that allows us to find certain patterns in a dataset for instance. As described in the paper "Principles of Data Mining" it allows discovering interesting aspects of the dataset, some patterns, that can be very useful [19]. The process, however, when applied in the healthcare area has some peculiarities, because not all the data is available to be treated since medical information can sometimes be sensitive and for ethical and legal aspects cannot be revealed. The paper "Uniqueness of medical data mining" addresses these differences in data mining when applied in medical data [20]. Despite this uniqueness, data mining is very useful, even in medical data. The paper "Data Mining in Healthcare and Biomedicine: A Survey of the Literature" describes how data mining can reveal knowledge hidden within all the medical data and how this knowledge can be very useful [21]. Already in the paper "Application of Data Mining Techniques to Healthcare Data" is described how data mining is useful in medical data and shows its advantages. It also shows the steps involved in data mining and successful applications of the process [22]. When it comes to the comparison of the results obtained by the scoring systems, the paper "Scoring systems for outcome prediction in patients with perforated peptic ulcer" is possible to visualize an evaluation of several score systems, in which is concluded that the most used system is the Boey

system and that the PULP system, although promising, need to be validated externally before the implementation [23].

4 Data Mining Process

As told, previously, all the process of classification and analysis of the data provided was based on the CRISP-DM methodology. The process followed, and all the results obtained, are presented in this section.

4.1 Business Understanding

The death and the existence of complications in patients with ulcers is a reality, unfortunately. This leads to a necessity, the attempt to predict these adverse situations and try to protect the patient both of death and of the existence of complications in their clinical reports. This study aims to achieve a prediction of the two parameters in the study in order to focus all the resources available to the situations that need them the most.

4.2 Data Understanding

In order to achieve all the results to which we propose the study base has to be a well-achieved dataset that contains good and adequate data. This allows achieving reliable results that can be applied to real situations. The ulcer dataset, on which the whole study is based, has 69 instances and multiple attributes that allows achieving PULP and Boey scoring punctuation

Table 1. Analysis of the statistical distribution of PULP/Boey parameters.

Punctuation System	Nominal Variable	Positive cases (%)
PULP	Hepatic cirrhosis	1.45
	Use of steroids	4.35
	Shock on Admission	17.39
	Drilling time up to diagnosis >24 hours	43.48
	Serum Creatnine	17.39
	ASA 2	62.32
	ASA 3	23.19
	ASA 4	10.10
	ASA 5	1.45
Boey	Severe co-morbidity	14.49
	Preoperative Shock	17.39
	Drilling time	44.93

Punctuation System	Numerical Variable	Mean	Standard Deviation	Minimun	Maximun
PULP/Boey	Age	55.24	17.65	21	90

The dataset contains, also, the PULP and Boey result as well as the class death and complicate in which can be found the result of the patient, if he died and got complicated. Through the analyze of the data is possible to visualize that 31.88% of them suffer of complications derived from the disease but only 5.8% of the patients died from it. In regard of gender distribution, it is possible to see that 63.66% of the patients are males, and 36.36% are females.

As for the classes in the study, is possible to analyze the distribution of the values for each attribute available. Since the dataset is based on "yes" or "no" information, almost all the attributes are nominal. Only one of the attributes is numerical, the age of the patient. Therefore, is possible to analyze the number of positive cases regarding each attribute, as shown in Table 1.

Once analyzed the attributes that allow to calculate the PULP score and Boey score and, since this is the main parameters in the prediction of death in real situations. So, is possible to verify the distribution of these parameters' values by all the dataset, as shown in Figure 1.
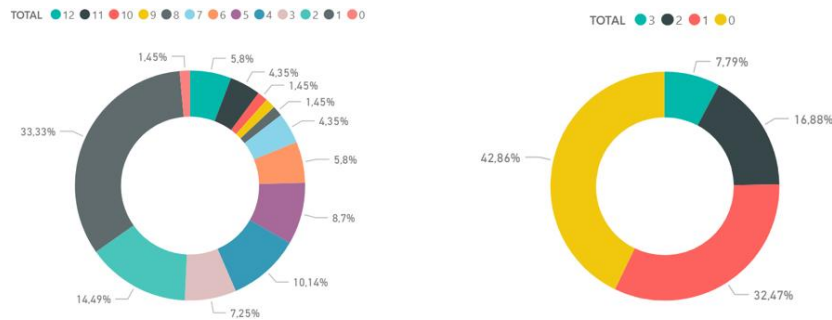

Fig. 1. Distribution of the PULP score (left) and the Boey score (right) by the dataset.

4.3 Data Preparation

One fundamental step of the CRISP-DM is the preparation of the data and the selection of all the important attributes in order to determine which are the most important ones. In the dataset used, it made no sense to eliminate any of the attributes, essentially because the most important ones are the PULP score and the Boey score. However, a dataset with this score and the class was too small, and in order to counteract this, all the attributes used to estimate this score are used. The dataset had, however, to be divided in order to study each of the systems, independently, so that, in the end, was possible to compare the accuracy results and conclude which of the punctuation systems is more assertive.

Focus our further study in the dataset that analysis the influence of PULP score on the prediction of death is possible to highlight some procedures that were followed. In addition to the several tests made using the original data, it was important to analyze the influence of some procedures in the final results. One of this was the oversampling. For such, a dataset with the triple of death cases, this is, instead of the 4 death cases, it has become 12 death cases, in a total of 75 cases. The other approach

was proceeding to the age normalization. In the original dataset, the age has values between 21 and 90. So that, it can be normalized and tested in another different way, to be possible to conclude which is the best way to study this particular data, in this case.

Number of Clusters.
In addition to the normalized dataset and the oversampling, the influence of clustering technique in the final results was also studied. This study begins with the determination of the ideal number of clusters that feat the dataset. For that is necessary to analyze the sum of squared errors returned in Clutterer output. For the case of PULP versus death, the ideal number is 6, since it is the minimum value achieved after a jump in them. However, analyzing the information, it was concluded that the expected number of clusters should be 2, because of the dimension of the target attribute, this is, because the target attribute is "Yes" or "No", it is more logical to split the results into two groups, one for the "Yes" and other for the "No". In this way, both possibilities will be studied and compared, in order to obtain the best results.

4.4 Modelling

The next step of the study is to induce the prepared data in WEKA, an open-source machine learning software tool. For these, it was necessary to choose an approach and identify a number of Data Mining Techniques, in order to obtain results.

The combination of different algorithms, scenarios and other parameters allows to obtain different models and set the best one. Each model depends on six parameters that can elaborate a huge number of different models and can be described as represented in Equation 1.

$$DMMn = \{Af + Si + DMTy + SMc + DAb + TGj\} \tag{1}$$

Af = {Classification} Si = {Scenarios}={S1}
DMTy = {Data Mining Techniques} = {JRip,OneR,NB,DT,J48,K-Means}
SMc = {Sampling Method}={Cross-Validation, Percentage Split}
DAb = {Data Approach}= {With Oversampling, Without Oversampling, With Cluster, Without Cluster, With Prunning, \\Without Prunning, With Normalization, Without Normalization}
TGj = {Target}= {Death}

Analyzing all the information specified is possible to conclude that the data mining model will be:
DMM = {1 classification, 1 scenario, 6 techniques, 2 methods, 8 approaches', 1 target} which leads to a total of 96 models induced.

4.5 Evaluation

The evaluation of the models constructed was made through three different statistic metrics based on the confusion matrix resulting from the application of the model. Based on the value of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN), it is possible to calculate Accuracy, according to eq. 2:

$$Accuracy = (TP+TN)/(TP+FN+TN+FN) \qquad (2)$$

This way is possible to analyze all the results obtained and conclude which of the models and data approaches can achieve the best results. As referred there are several approaches to the data information that were studied in order to find the best way to predict death through PULP score, trying to understand which combination of factor can achieve a higher accuracy. Therefore, will be studied factors such as Number of Clusters and the influence that this can lead in accuracy results, as well as the influence that oversampling and normalization can represent in the result of accuracy.

Table 2. Evaluation results obtained to death prediction through PULP score.

| Technique	Method	Data Approach	Accuracy (%)
NB	Cross-Validation	-	92.75
NB	Cross-Validation	Cluster (k=2)	94.20
JRip	Percentage Split	Prunning and Cluster (k=6)	86.96
JRip	Percentage Split	Prunning and Oversampling	96.15
OneR	Cross-Validation	Oversampling and Cluster (k=2)	97.40
JRip	Cross-validation	Oversampling, Prunning and Cluster (k=6)	94.81
NB	Cross-Validation	Normalized	92.75
NB	Cross-Validation	Normalized and Cluster (k=6)	94.20
J48	Cross-Validation	Normalized, Prunning and Cluster (k=7)	91.30

Analyzing the results is possible to withdraw the results to the three principal data approaches (oversampling, normalized and original) and compare their results, as we can see in Figure 2.
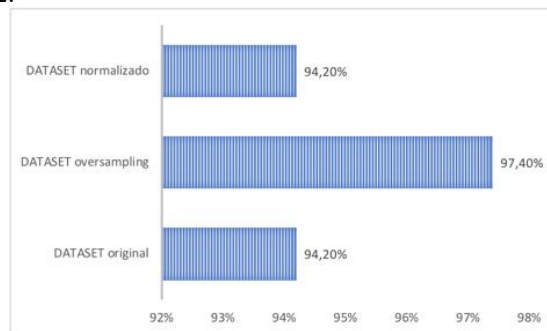
Fig. 2. Results obtained for the three principal data approaches.

Similarly, to the procedure made for obtaining the accuracy for death prediction using PULP score, the prediction of death using Boey score, and the prediction of

complications using both, PULP and Boey score, were studied and the results compared. The results obtained can be found in Figure 3.
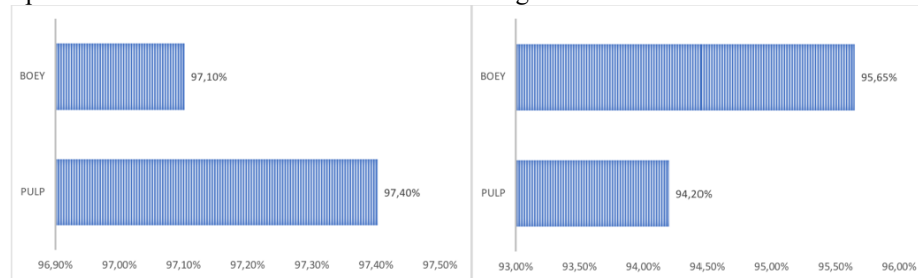
Fig. 3. Results obtained for predicting death (left) and complications (right) in patients with ulcers.

The final step of this study is the analysis of the correlation existing between all the attributes and the two targets studied. This will allow finding the most relevant attributes among all, this is, the ones that have a major influence in the final state of the patient with regard to death and the complications of the state of health. The results obtained can be found in Figure 4.
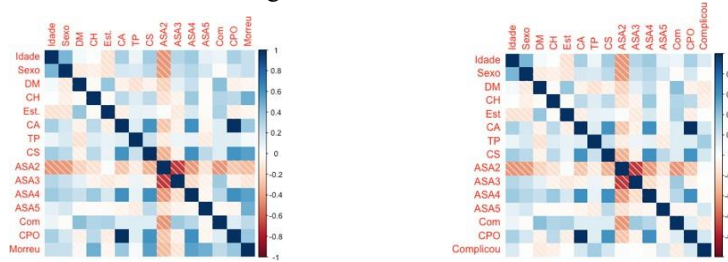
Fig. 4. Results obtained for correlation with death (left) and with complications (right) in patients with ulcers.

5   Discussion

With the analysis of the results obtained is easy to understand that, in the vast majority, the values of accuracy are greater than 90%, with respect to the prediction of death using the PULP score. This fact can be explained by the number of deaths in the whole dataset. There are only 4 cases of death in 69 patients. This fact can influence, negatively, the results obtained, making them less relevant due to the universe of study. However, the results can be improved, by the reformulation of the dataset, especially with regard to case numbers.

Focusing the analysis on the results obtained, more specifically the accuracy values, is easy to conclude that the best value achieved is 97.40%, using a normalized dataset, and with the application of clustering techniques. Concerning the method used, the one that leads to a better result is OneR. The number of clusters was one of the parameters in the study. Despite the fact that the determination of this number by the squared errors leads to a different number, the rational analysis makes us conclude

that the ideal cluster number is 2 because the classification according to the target attribute is "Yes" or "No". Therefore, makes sense try to fit the results in only two groups. Another parameter in the study is the influence of oversampling in the results. How can be checked by the analysis of the results, the oversampling produces an increasing on the accuracy result, which can be explained by the simple fact that this approach leads to the replication, in this case, triplication of data, that can lead to the addiction of the model. Crossing the remaining results is possible to infer that the PULP score allows better predictions of death, as it is mentioned in the literature, and as expected because takes in account a much larger number of factors. However, with regard to complications, the results are precisely the opposite. In this situation, Boey can achieve better results, leading to better predictions of complications in the health status of the patients. The last one is a totally new approach, which does not yet have a comparison term with the results obtained.

At last, with regard to the established correlations, as we can see in Figure 5, it was found that the four main parameters, with a strong positive correlation with the target attribute, that should be incorporated in a new system to predict death effectively, were ASA 4, ASA 5, serum creatinine and liver cirrhosis. In the other hand, the three parameters that most influence the occurrence, or not, of complications are perforation time greater than 24 hours, the shock in the patient's admission, and creatinine.

6 Conclusion and Future Work

With this work is possible to conclude that Data Mining can be a very important step in trying to solve some problems that can be found in health systems nowadays and, in trying to create new approaches and protect patients from some situations like death and health complications. In this specific case, is possible to conclude that PULP is a better scoring system to predict that than Boey. However, Boey shows up to be superior when it comes to complications. Regarding the correlations, there are 4 major attributes, in the prediction of death, ASA 4, ASA 5, serum creatinine and liver cirrhosis, and 3 in the prediction of complications, perforation time greater than 24 hours, the shock in the patient's admission, and creatinine. The results obtained, however, are not conclusive due to the dimension of the dataset. Once it has few arguments, the conclusions are not totally reliable. This fact, and overwhelming the possibility of future work, can be bypassed by the increase of the dataset, increasing the number of patients in it, and balancing the number of patients in each possible value of the target.

Acknowledgments

References

1. Sandler RS, Everhart JE, Donowitz M et al. The burden of selected digestive diseases in the United States. Gastroenterology. 2002 May;122(5):1500-11.
2. Wang YR, Richter JE, Dempsey DT, Trends and outcomes of hospitalizations for peptic ulcer disease in the United States, 1993 to 2006. Ann Surg. 2010;251(1):51.
3. Irabor DO. An audit of peptic ulcer surgery in Ibadan, Nigeria. West Afr J Med. 2005;24(3):242.
4. Sánchez-Delgado J, GenéE, Suárez D, García-Iglesias P., et al. Has H. pylori prevalence in bleeding peptic ulcer been underestimated? A meta-regression. Am J Gastroenterol. 2011;106(3):398.
5. Lohsiriwat V, Prapasrivorakul S, Lohsiriwat D. Perforated peptic ulcer: clinical presentation, surgical outcomes, and the accuracy of the Boey scoring system in predicting postoperative morbidity and mortality. World J Surg. 2009;33:80–85.
6. M. H. Møller, M. C. Engebjerg, S. Adamsen, J. Bendix4, R. W. Thomsen. The Peptic Ulcer Perforation (PULP) score: a predictor of mortality following peptic ulcer perforation. A cohort study. Acta Anaesthesiol Scand. 2012 May;56(5):655-62.
7. Neto, C., Peixoto, H., Abelha, V., Abelha, A., & Machado, J. (2017). Knowledge Discovery from Surgical Waiting lists. Procedia Computer Science, 121, 1104-1111.
8. Zhao, Y. (2012). R and data mining: Examples and case studies. Academic Press.
9. Pereira, J., Peixoto, H., Machado, J., & Abelha, A. (2017). A Data Mining Approach for Cardiovascular Diagnosis. Open Computer Science, 7(1), 36-40.
10. Cohen, W. W. (1995). Fast effective rule induction. In Machine Learning Proceedings 1995 (pp. 115-123).
11. Kohavi, R. (1995, April). The power of decision tables. In European conference on machine learning (pp. 174-189). Springer, Berlin, Heidelberg.
12. Najm, W. I. (2011). Peptic ulcer disease. Primary Care: Clinics in Office Practice, 38(3), 383-394.
13. Parsania, V., Bhalodiya, N., & Jani, N. N. (2014). Applying Naïve Bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis.
14. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.
15. Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier..
16. Arthur, D., & Vassilvitskii, S. (2007, January). k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete Algorithms (pp. 1027-1035). Society for Industrial and Applied Mathematics.
17. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.
18. Reis, R., Peixoto, H., Machado, J., & Abelha, A. (2017). Machine Learning in Nutritional Follow-up Research. Open Computer Science, 7(1), 41-45.
19. Morais, A., Peixoto, H., Coimbra, C., Abelha, A., & Machado, J. (2017). Predicting the need of Neonatal Resuscitation using Data Mining. Procedia Computer Science, 113, 571- 576.
20. Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622.
21. Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. Artificial intelligence in medicine, 26(1-2), 1-24.
22. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua,
a. L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.
23. Thorsen, K., Søreide, J. A., & Søreide, K. (2013). Scoring systems for outcome prediction in patients with perforated peptic ulcer. Scandinavian journal of trauma, resuscitation and emergency medicine, 21(1), 25.