Clustering Barotrauma Patients in ICU – A Data Mining based approach using ventilator variables

Sérgio Oliveira[1], Filipe Portela[1], Manuel F. Santos[1], José Machado[1], António Abelha[1]
Álvaro Silva[2], Fernando Rua[2]
[1]Algoritmi Centre, University of Minho, Portugal
sergiomdcoliveira@gmail.com; {cfp, mfs}@dsi.uminho.pt; {jmac, abelha}@di.uminho.pt;
moreirasilva@me.com; fernandorua.sci@chporto.min-saude.pt

Abstract: Predicting barotrauma occurrence in intensive care patients is a difficult task. Data Mining modelling can contribute significantly to the identification of patients who will suffer barotrauma. This can be achieved by grouping patient data, considering a set of variables collected from ventilators directly related with barotrauma, and identifying similarities among them. For clustering have been considered k-means and k-medoids algortihms (Partitioning Around Medoids). The best model induced presented a Davies-Bouldin Index of 0.64. This model identifies the variables that have more similarity among the variables monitored by the ventilators and the occurrence of barotrauma.

Keywords: Barotrauma, Plateau Pressure, Intensive Medicine, Data Mining, Clustering, Similarity, Correlation.

1 Introduction

Data Minng (DM) process provides not only the methodology but also the technology to transform the data collected into useful knowledge for the decision-making process [1]. In critical areas of medicine some studies reveal that one of the respiratory diseases with higher incidence in the patients is Barotrauma [2]. Health professionals have identified high levels of Plateau pressure as having a significantly contribute to the Barotrauma occurrence [3]. This study is part of the major project INTCare. In this work a clustering process was addressed in order to characterize patients with barotrauma and analyze the similarity among ventilator variables. The best models achieved a Davies-Bouldin Index of 0.64. The work was tested using data provided by the Intensive Care Unit (ICU) of the Centro Hospitalar do Porto (CHP).

This paper consists of four sections. The first section corresponds to the introduction of the problem and related work. Aspects directly related to this study and supporting technologies for knowledge discovering from databases are then addressed in the second section. The third section formalizes the problem and presents the results in terms of DM models following the methodology Cross Industry Standard Process for Data Mining (CRISP-DM). In the fourth section some relevant conclusions are taken.

2 Background

2.1 Plateau Pressure, Acute Respiratory Distress Syndrome and Barotrauma

The occurrence of barotrauma happens when a patient has complications in mechanical ventilation. Patients with Acute Respiratory Distress Syndrome (ARDS) have shown that the incidence of pneumothorax and barotrauma varies between 0% and 76% [4]. The occurrence of barotrauma is one of the most dreaded complications when a patient is mechanically ventilated. This occurrence is associated with an increased morbidity and mortality. Several researchers argue that the Positive end-expiratory pressure (PEEP) is related to the occurrence of barotrauma, however there are other researchers not supporting this relationship and enforcing that there was not identified any relationship between PEEP and barotrauma [2]. The Plateau Pressure (PPR) values shall be continuously monitored providing important information for patient diagnosis. It is important to maintain the value of PPR <= 30 cmH_2O in order to protect the patient's lungs. An increased in PPR is associated to an increasing elasticity of the respiratory system and decreased compliance of the respiratory system [3].

2.2 Related work

Predicting barotrauma occurrence is important for the patient wellbeing. So it is fundamental to explore the prediction accuracy of the variables and their correlation with barotrauma – PPR values >= 30 cmH_2O. In a first stage of this project it was predicted the probability of occurring barotrauma considering only the data provided by the ventilator. This study [5] shown that it is possible to predict PPR class <30 cmH_2O and >= 30 cmH_2O, with an accuracy between 95.52% and 98.71%. The best model was achieved using Support Vector Machines and all the variables considered in the study. However, another good model was obtained (95.52% of accuracy) using only three variables. This model showed a strong correlation among: Compliance Dynamic (CDYN), Means Airway Pressure and Pressure Peak.

2.3 INTCare

This work was carried out under the research project INTCare. INTCare is an Intelligent Decision Support System (IDSS) [6] for Intensive Care which is in constantly developing and testing. This intelligent system was deployed in the Intensive Care Unit (ICU) of Centro Hospitalar do Porto (CHP). INTCare allows a continuous patient condition monitoring and a prediction of clinical events using DM. One of the most recent goals addressed is the identification of patients who may have barotrauma.

2.4 Data Mining

DM corresponds to the process of using technical features of artificial intelligence, statistical calculations and mathematical metrics able to extract information and

useful knowledge. The knowledge discovery may represent various forms, business rules, similarities, patterns or correlations [7].

This work is mainly focused on the development and analysis of clusters. This is a grouping process based on observing the similarity or interconnection density. This process aims to discover data groups according to the distributions of the attributes that make up the dataset [8]. To develop and assess the application of clustering algorithms in the barotrauma dataset, the statistical system R was choosen.

3 Knowledge Discovering Process

3.1 Business Understanding

The main goal is to use ventilation data in order to identify groups of objects that belong to the same class, i.e. group sets of similar objects in a single set and dissimilar in different sets. The data used to conduct this study were collected in the ICU of CHP. The clusters were supported only for data monitored by ventilators; the values used were numeric and were from discrete quantitative type.

3.2 Data Understanding

The initial data sample contained several records without patient identification (PID). This happens because sometimes the patients are admitted for a few hours in the ICU but are not assigned to an Electronic Health Record (EHR). These records were discarded for this study. The sample used was collected from the ventilators and comprises a period between 01.09.2014 and 10.12.2014 and a total of 33023 records. Each record contains fourteen fields: CDYN – (F_1); CSTAT – (F_2); FIO2 – (F_3); Flow – (F_4); RR – (F_5); PEEP – (F_6); PMVA – (F_7); Plauteau pressure – (F_8); Peak pressure – (F_9); RDYN – (F_10); RSTAT – (F_11); Volume EXP – (F_12); Volume INS – (F_13); Volume Minute – (F_14).

The coefficient of variation shows that the distributions are heterogeneous for all the attributes since the results obtained are higher than 20%. This measure corresponds to the dispersion ratio between the standard deviation and the average.

3.3 Data Preparation

Data transformations were necessary to perform data segmentation using clustering techniques based in resource partition methods. Because these techniques do not handle null values and qualitative data, two operations have been performed: Firstly, records having at least one null value were eliminated and then the records containing qualitative values were eliminated.

3.4 Modelling

The algorithms k-means and k-medoids were used to create the cluster. The choice is justified by the principle of partition method and the difference in their sensitivity to find outliers.

K-means algorithm is sensitive to outliers, because the objects are far from the majority, which can significantly influence the average value of the cluster. This effect is particularly exacerbated by the use of the squared error function [9].

On the other hand, K-medoids instead of using the value of a cluster object as a referencing point takes on real objects and represents the clusters, creates an object for each cluster. The partitioning method is then performed based on the principle of adding the differences between each p (intra-clusters distance). It is representing an object (dataset partition) [9] where the p is always >= 0. The K-medoids algorithm is similar to K-means, except that the centroids must belong to a set of grouped data [10]. Some configurations were atempted for each one of the algorithms. In the k-means algorithm the value K (cluster number) varies between 2 and 10. In order to obtain the appropriate number of K it was used the sum of squared error (SSE). Each dataset was executed 10 times.

The model M_n belongs to an approach A and it is composed by the fields F, a type of variable TV and an Algorithm AG:

$A_f = \{Discription\ (Clustering)_1\}$

$F_i = \{F\_1_1, F\_2_2, F\_3_3, F\_4_4, F\_5_5, F\_6_6, V\_7_7, F\_8_8, F\_9_9, F\_10_{10}, F\_11_{11}$
$, F\_12_{12}, F\_13_{13}, F\_14_{14}\}$

$TV_x = \{Qualitative\ variables\ ordinal_1\}$

$AG_y = \{K - means_1, K - medoids(PAM)_2\}$

Being this study related with Barotrauma and Plateau Pressure, all the models included the variable $F_{\{8\}}$. Some of the clusters induced are composed by the group of variables defined in the first approach.

3.5 Evaluation

This is the last phase of the study. It focuses mainly on the analysis of the results presented through the implementation of clustering algorithms (K-means and PAM). The evaluation of the induced models was made by using the Davies-Bouldin Index. The models which presented most satisfactory results were those obtained by means of the K-means algorithm. In general, some models presented good results, however the models did not achieve optimal results (index near 0). Table 2 presents the best models and the correspondent results.

Table 1 – Models for clustering

Model	Fields	Number of Clusters	Algorithm	Davies-Boldin Index
M_1	$F_{\{1,2,7,8,10\}}$	2	AG_1	0.82
M_2	$F_{\{1,2,7,8,10\}}$	5	AG_2	0.86
M_3	$F_{\{1,7,8,10\}}$	2	AG_1	0.64
M_4	$F_{\{1,7,8,10\}}$	6	AG_2	1.17

The M_3 model shown to be the most capable in designing a clusters with better distances. Davies-Bouldin Index tends to $+\infty$ however M_3 model has an index of

0.64. This is not the optimal value, but it is the most satisfactory because it is closest to 0. Figure 1 presents M_3 results.

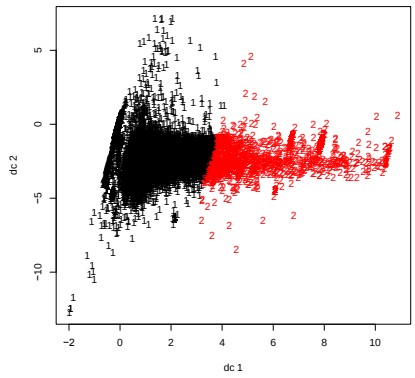

Figure 1 – Clusters of model M_3

Table 3 presents the minimum, maximum, average, standard deviation and coefficient of variation of each variable used to host the clusters in M_3.

Table 2 – Distributions for each cluster

Clusters	Fields	Min	Max	Average	StDev	CoeVariation
Cluster 1 (30017 rows)	$F_{\{1\}}$	0	73	31.45	16.55	52.63%
	$F_{\{7\}}$	0	38	11.94	2.71	22.73%
	$F_{\{10\}}$	0	100	14.18	7.26	51.2%
	$F_{\{8\}}$	0	99	22.04	6.57	29.82%
Cluster 2 (3006 rows)	$F_{\{1\}}$	73	200	114.72	32.62	28.43%
	$F_{\{7\}}$	3.9	22	11.00	1.68	15.3%
	$F_{\{10\}}$	1.4	46	13.22	5.27	39.88%
	$F_{\{8\}}$	0.2	84	17.88	6.24	34.87%

4 Conclusion

This study identified a set of variables that have a great similarity. These variables are related with Plateau Pressure - variable with greater influence in the occurrence of barotrauma. The better result was achieved with the model M_3 obtaining a Davies-Bouldin Index of 0.64, a value near to the optimum value (0).

It should be noted that most of the variables used presented some dispersion however in one of the clusters the higher dispersion value is quite acceptable: 19.42. This result was obtained with the implementation of the K-means algorithm. The CDYN is one of the variables that most influences the clustering, demonstrating a strong relationship with the PPR and Barotrauma. From the results shown in Table 3 it can be noted that the best corresponding field is CDYN, presenting only a few intersecting values (minimum and maximum). This means

that Cluster 1 has CDYN values ranging between [0; 73] and Cluster 2 has CDYN values between]73;200]. The remaining fields used have only few interceptions. Finally, this study demonstrated the feasibility of creating clusters using only data monitored by ventilators and analyzing similar populations. These results motivate further studies in order to induce more adjusted models reliable for classification and clustering at the same time.

Acknowledgements

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013 and the contract PTDC/EEI-SII/1302/2012 (INTCare II).

References

[1] H. Koh and G. Tan, "Data mining applications in healthcare," *J Healthc Inf Manag*, vol. 19, no. 2, pp. 64–72, 2005.

[2] A. Anzueto, F. Frutos-Vivar, A. Esteban, I. Alía, L. Brochard, T. Stewart, S. Benito, M. J. Tobin, J. Elizalde, F. Palizas, C. M. David, J. Pimentel, M. González, L. Soto, G. D'Empaire, and P. Pelosi, "Incidence, risk factors and outcome of barotrauma in mechanically ventilated patients," *Intensive Care Med*, vol. 30, no. 4, pp. 612–619, Apr. 2004.

[3] N. Al-Rawas, M. J. Banner, N. R. Euliano, C. G. Tams, J. Brown, A. D. Martin, and A. Gabrielli, "Expiratory time constant for determinations of plateau pressure, respiratory system compliance, and total resistance," *Crit Care*, vol. 17, no. 1, p. R23, 2013.

[4] M. Boussarsar, G. Thierry, S. Jaber, F. Roudot-Thoraval, F. Lemaire, and L. Brochard, "Relationship between ventilatory settings and barotrauma in the acute respiratory distress syndrome," *Intensive Care Med*, vol. 28, no. 4, pp. 406–413, Apr. 2002.

[5] S. Oliveira, F. Portela, M. Santos, J. Machado, A. Abelha, and F. Rua, "Predicting Plateau Pressure in Intensive Medicine for Ventilated patients," *WorldCist 2015 - Healthcare Information Systems: Interoperability, Security and Efficiency Workshop*, 2015.

[6] F. Portela, M. F. Santos, J. Machado, A. Abelha, Á. Silva, and F. Rua, "Pervasive and Intelligent Decision Support in Intensive Medicine – The Complete Picture," in *Information Technology in Bio- and Medical Informatics*, M. Bursa, S. Khuri, and M. E. Renda, Eds. Springer International Publishing, 2014, pp. 87–102.

[7] E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*, 9ª Edição. Prentice Hall, 2011.

[8] R. K. Anderson, *Visual Data Mining: The VisMiner Approach*, 1 edition. Chichester, West Sussex, U.K. ; Hoboken, N.J: Wiley, 2012.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3ª Edição. Morgan Kaufmann, 2012.

[10] W. Xindong and K. Vipin, *The Top Ten Algorithms in Data Mining*. CRC Press - Taylor & Francis Group, 2009.