Clustering Data Mining models to identify patterns in weaning patient failures

Sérgio Oliveira, Filipe Portela, Manuel Filipe Santos, José Machado, António Abelha, Álvaro Silva and Fernando Rua

***Abstract*—** The weaning process of ventilated patients need to be carefully performed. This type of procedure is very common in Intensive Medicine. The procedure is well-defined and it is executed according the patient condition. During the weaning process, the patient can be in vary stages. At the end the extubation tentative can be considered as successful or not. Before the extubation, the patient is submitted to a set of tests in order to validate the procedure. When this procedure is wrong executed, it can provoke long term injuries to the patient. This work arises in order to avoid weaning failures by early detecting the procedure result. This work has as main goal identify possible patient patterns associated to weaning failures. In this context Clustering data mining was used to select and identify the features and the patterns associated to failures. As result an Index-Davies Bouldin of 0.51 was achieved and the most significant variables associated to a failure were identified. The physicians has now new and useful knowledge able to help to take a decision about weaning before it be initiated.

***Keywords*—** Ventilation Weaning, Extubation, Mechanical Ventilation, Respiratory Diseases, Intensive Medicine, Intensive Care Unit, INTCare, Data Mining and Clustering.

I. Introduction

In Intensive Medicine (IM) around of 75% of the admitted patients requires mechanical ventilation. This type of patients are admitted in Intensive Care Units and in many cases they are suffering multiple organ failures as is the Respiratory System. Mechanical ventilation (MV) is used to support their breath function, however a long ventilation process can provoke lung injuries.

MV can have negative effects to the patients. Its mortality rate ranges is from 41% to 65% [1]. Prevent weaning failures is an important asset essentially due the high number of re-intubations. This number varies from 2% to 25%[2].

An automatic mechanical ventilation control can improve the quality of patient care and consequently the patient condition. It also can contribute to reduce healthcare costs and to help in reducing the morbidity and mortality rates associated with provision of inappropriate ventilator treatments.

Having conscience of this reality this study was executed in order to find patient variables associated to wrong weaning. Identify patient patterns associated to weaning failures is the main goal of this work. With this work a set of information are provided to the clinicians in order to early detect possible failures. The goal is not to predict if a patient can be or not extubated but identify patterns and consequently set of patients that should not be put in a weaning process.

In this study, Clustering Data Mining techniques (K-means and K-medoids) were used to create weaning failures patters. Clustering is used for grouping a set of data in such a way that the objects in the same cluster (group) are more similar.

INTCare research project [3, 4] is the base of this study. The data collected in real-time [5, 6] from the Intensive Care Unit (ICU) of Centro Hospitalar do Porto (CHP), Portugal were used to induce Data Mining models.

This work arises also in sequence of several studies performed in this area [7-9]

As result it was possible to make a feature selection and identify a set of patient characteristics associated to weaning failures. The Index-Davies Bouldin achieved was 0.51. The features with most impact and identified by the better cluster were: Compliance Dynamic (CDYN), Mean Air Pressure (MAP), Plateau Pressure and Support Pressure.

The paper is divided in six sections after introduce the work a set of related concepts are presented in the second section. In the third section is presented the material and methods used in this work. Section four present the work developed following CRISP-DM methodology. In section five the results achieved are discussed having in consideration the main target. In this section also it is presented an analysis of the most significant

This work was FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

Sérgio Oiveira is with Algoritmi Research Centre, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal and ESEIG, Porto Polytechnic, Portugal (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

José Machado and António Abelha is with Algoritmi Research Centre, University of Minho, Braga, Portugal. (e-mail: {jneves@di.uminho.pt).

Álvaro Silva and Fernando Rua are with Intensive Care Unit of Centro Hospitalar do Porto, Portugal (e-mail: {moreirasilva@me.com; fernandorua.sci@hgsa.min-saude.pt).

variables. Finally some conclusion are taken.

II. Background

A. Intensive Care Units

The patients who are admitted to Intensive Care Units (ICU) are constantly monitored. They have a set of sensors connected from the body to the bedside monitors. The most common monitoring process is patient vital signs and mechanical ventilation. In fact these patients need intensive care and typically they are in a risk-life condition, being their life condition supported by ventilators.

The main goal of intensive medicine it is use the artefacts available in ICU and the intensivists knowledge to diagnose and treat patients with serious illnesses, restoring them to their previous health condition [10].

Most recently became more difficult to make decision in ICU taking in attention all the data collected from the patient. The existence of vital signs monitors and ventilators allow a continuous data streaming. This situation difficult the data analysis in a short period of time due to a high number of data collected. ICU is a potential source of implicit knowledge. This knowledge can be used to improve the decision making process.

B. Mechanical Ventilation and Weaning

Respiratory failure is a syndrome in which the respiratory system fails in one or both of its gas exchange functions: oxygenation and carbon dioxide elimination [11].

The goal is to reduce lung injury due to over distention. However, the efficacy of this approach has not been established [12]. To overcome this problem the patients are ventilated using artificial ventilation.

Nowadays mechanical ventilator, are only used by the clinicians to consult the patient values. The data observed are not stored in a database. This situation results in a wasting of data that could be transformed in knowledge and it could be very useful to the decision-making process.

In addition, the process of ventilator weaning is based in a medical assumption [13] and in a tentative-error procedure, which sometimes seriously compromises the patient condition.

Mechanical ventilation is commonly used in ICU and it is very important to treat many different illnesses, however is relatively costly [2].

Weaning is a gradual process of liberation from, or discontinuation of, mechanical ventilator support resulting in an extubation. In Intensive Medicine an extubation process is considered successful when a patient can breathe from himself for a period upper than one hour.

The Intelligent Decision Support Systems (IDSS) for mechanical ventilation can be grouped in two types: an expert advisory systems or an automatic control of ventilation or weaning [14]. In the ICUs there is a set of IDSS system to ventilators, however, most of them are rule-based system. They are not adaptive and they do not use the results obtained to improve the models.

After an overview [14] it is also possible to verify that most of the existing systems is not using data mining to predict the results. The most far as they can go it is in the input data that can be based in clinical rules and guidelines. Many of its rules can be adaptive and can be derived on the basis of physiological models.

C. INTCare

In the ICU of CHP was deployed a Pervasive Intelligent Decision Support System (PIDSS). This PIDSS is in a continuous developing and test and it is a result of INTCare project [3].

INTCare is a multi-agent system [15-17] and it is able to monitoring the patient condition in real-time by collecting, processing and displaying the information collected [18, 19] from the bedside monitors and other hospital sources in an intuitive and easy way.

It also has a module to support the decision process through Data Mining models. This module can induce in real-time and using online learning several models able to predict clinical events, as is for example patient outcome [20], organ failure [20], length of stay [21], readmission [22, 23] and barotrauma [7, 9], among others.

INTCare uses intelligent agents [16, 17, 24] to perform their tasks automatically and without human intervention.

This work is inserted in the second phase of the project where the main concern is the respiratory system.

After make a first research to predict barotrauma [7, 9] now it is time to explore a new field: weaning and extubation.

D. Data Mining

Data Mining is the process of using artificial intelligence techniques and statistical and mathematical functions to extract knowledge from the data stored in the database. The achieved knowledge can be presented in multiple forms: business rules, similarities, patterns or correlations [25]. Clustering is inserted in the group of Data Mining problems.

Clustering has as main goal divides the data collected in datasets with similar values. The groups created by the clusters represents a natural catchment of the data and data aggregations. The groups created should make sense, be helpful or both. Clustering rules are not pre-defined. They are discovered along the clustering process. The clusters are characterized by a great internal homogeneity and external heterogeneity [26].

The use of cluster to identify groups of variables is an important asset in many areas like psychology and social sciences, biology, statistical, pattern recognition, information recovery, machine learning and data mining [27, 28].

Clustering offers a high number of algorithms. The choice of the best algorithm to use it is depending from the data collected and project goal. The majority of the clustering methods are grouped into five categories.

The hierarchical methods execute a hierarchical decomposition of the data. These methods can be divisive or agglomerative.

Divisive methods behave the other way. The density-based methods are useful to filter outliers or discovering data with arbitrary form.

The agglomerative methods start with singular objects to create an isolated group. Then the groups or objects are successively merged until a group is missing.

The Partition Methods build a set of partitions on the data, where each partition represents a cluster.

Grid-based methods restrict the space of objects to a finite number of cells forming a grid structure. The Model-based methods formulate a model hypothesis for each cluster and find the best fit the data to the model [29].

Clustering assessment can be done by laying on two factors: compactness and separability. The compactness expresses how much the cluster elements are near. How lesser the variance value it is, greater it will be the cluster compactness. The calculation of the intra cluster distance is very useful to assess this characteristic. The separability evaluates how diverse the clusters are. This can be evaluated by the inter-cluster distance. How higher the distance is better the clusters are [30].

The simplest and most fundamental method of cluster analysis is partitioning. This method aims to organize a set of objects in various exclusive groups namely clusters. In order to maintaining the problem a concise specification of the number of clusters should be identified. This is the parameter used as a starting point in the implementation partitioning methods. From a data set D of n objects, where K is the number of clusters to create, a partitioning algorithm organizes the objects in partitions ($K \leq n$) [29].

The agglomerates are created in order to optimize an objective partitioning criteria so that objects within a cluster are "similar" to another and "dissimilar" objects are in other groups in accordance with attribute data sets. Two of the methods most used classic partition are K-means and K-medoids [29].

The K-means algorithm sets the centroid of a cluster from the average value of the points that are within the cluster. First, are randomly selected K of D objects, and initially each of the K is the center of a cluster. Then the objects are assigned to cluster with greater similarity. The assignment of the object to a cluster is based on the Euclidean distance between the object and the mean cluster [29].

In order to better understand the K-means a representation was made in Algorithm 1.

Algorithm 1: k-means

Input: *K* (number cluster) and *D* (dataset).
Output: Group of *K* clusters.
1: choose *K* objects such as the centers of the initial group *D*;
2: **Do**
3: (re) assign each object to the cluster to which the object is more like having based the average value of the objects in the cluster;
4: update the cluster means;
5: **Until** no change

The k-medoids algorithm is also known as partitioning around medoids. It is a variant of k-means method. The k-medoids is based on the use of medoids rather than the mean values of observations belonging to each group. Aiming to decrease the sensitivity of partitions created with respect to external values from a data set. For each next iteration the algorithm tries to perform the replacement of each medoid with an observation that does not represent a medoid, but since the overall quality of the subdivisions is improved.

The quality of the divisions are evaluated by a measure of lack of homogeneity average, between each observation and the medoid associated to the cluster [31].

In order to better understand the K-medoids follows is the representation of Algorithm 2.

Algorithm 2: k-medoids

Input: *K* (number cluster) and *D* (dataset).
Output: Group of *K* clusters.
1: arbitrarily choose *k* objects from *D* as the initial cluster centers;
2: **Do**
3: place the missing objects in the cluster with the nearest representative object;
4: randomly select a non-representative object;
5: compute the total cost of swapping (*S*) representative object;
6: **If** $S < 0$ **Then**
7: swap to form the new set of k representative objects;
8: **End If**
9: **Until** no change

III. Material And Methods

This is a Design Science Research work where at the end an artefact (Cluster) is produced. This artefact are assessed using the clinical domain and the intensivists' knowledge.

Cross Industry Standard Process for Data Mining (CRISP-DM) was the methodology chosen to conduct this work. CRIS-DM is divided in six steps: Business Understanding, Data Understanding, Data preparation, Modelling, Evaluation and Deployment.

This is a methodology in cycle where each step provides a structured approach to planning a data mining project.

R tool were used to perform this work. R is an environment of statistical programming language for development [32]. The library "cluster" was used primarily to implement Partitioning Around Medoids (PAM) and the k-means algorithms. Then, R also were used for making the graphics. For finding the optimum number of cluster the library "fpc" were used. The library "clusterSim" and the Davies-Bouldin Index were used to evaluate the cluster.

IV. Knowledge Discovering Process

As stated CRISP-DM was the methodology chosen to develop this study. In the following sections the work made and the results achieved for each CRISP-DM phase are presented.

A. Business Understanding

The main goal of this study is identify the patient ventilation variables (patterns) which can interfere in a non-successful weaning / extubation. The goal of this study is not to predict the probability of a patient be successful extubated but it is to identify some patterns and clusters associated to weaning

failures. The clusters were designed using only data monitored and real-time collected by the ventilators.

In a clinical point of view it is expected creating new knowledge to the intensivists in order to helping them to take the better decision in the moment of a patient be weaned.

B. Data Understanding

In this study the data used were exclusively collected from the ventilators connected to the patients admitted in the ICU of CHP. These data corresponds to a period between 2014-09-19 and 2015-02-03 in a total of 15325 records and 50 patients. Each one of the records contains thirteen fields:

- CDYN – (F_1): Dynamic compliance in mL/ cmH_2O;
- CSTAT – (F_2): Static compliance from inspiratory pause measured in mL/ cmH_2O;
- FIO2 – (F_3): Fraction of inspired oxygen (%);
- Flow – (F_4): Peak flow setting in litters per minute;
- RR – (F_5): Respiratory rate setting in berths per minute;
- PEEP – (F_6): Positive End-Expiratory Pressure in cmH_2O;
- MAP – (F_7): Mean airway pressure in cmH_2O;
- Plateau pressure –(F_8): End inspiratory pressure in cmH_2O;
- Peak pressure – (F_9): Maximum circuit pressure in cmH_2O;
- RSTAT – (F_10): Static resistance from inspiratory pause measured in cmH_2O/L/s;
- Volume EXP – (F_11): Exhaled tidal volume in litters;
- Volume INS – (F_12): Tidal volume settings in litters;
- Support Pressure – (F_13): Exhaled minute volume litters;

Table 1 presents a statistical analysis of the variables used in this study. The minimum (MIN), maximum (MAX), average (AVG), standard deviation (SD) and coefficient of variation (CV) of each variable were calculated.

The values obtained with the CV showed that there is some dispersion in variables used. Only four variables had a coefficient of variation lower than 20%. These variables were: Plateau Pressure, MAP, Peak Pressure and FIO2. This measure of dispersion is calculated for each variable and it is the ratio between the standard deviation and the average.

Table 1 – Distribution of variables

	MAX	MIN	AVG	SD	CV
Plateau Pressure	36	6.2	19.98	3.83	19.18
CDYN	200	0	46.31	40.45	87.34
CSTAT	71	0	20.88	20.93	100.22
MAP	18	3.1	10.52	1.84	17.48
Peak Pressure	40	9	20.53	3.85	18.75
RSTAT	29	0	8.89	8.60	96.74
FIO2	100	35	49.68	7.43	14.96
FLOW	60	0	23.83	23.61	99.06
RR	24	0	1.69	5.50	324.73
PEEP	10	3	5.09	1.04	20.45
Volume EXP	2.46	0	0.53	0.17	31.83
Volume INS	0.56	0	0.26	0.25	98.06
Support Pressure	27	4	14.1	3.56	25.24

C. Data Preparation

The data used in this study did not have information related about the patient inability to be submitted to the extubation process (e.g. respiratory diseases, chronic diseases, other clinical situations, ventilation time, among others).

In order to identify these occurrences it was necessary to determinate the respective scenario. Patients who were not submitted to extubation process were patients who had mechanical ventilation variations or the support pressure level was continuous for more than one hour but they never were an attempt to extubation under this scenario. In this process five levels were identified. The development process of the various levels can be seen in Algorithm 3.

Algorithm 3: Patient's condition scenarios

Input: Ventilation data and identification of patients

Output: Scenarios by occurrence

```
Function Extubation scenarios
  N = number row of values;
  For weach patient Do
    If (time_row(N+1) – time_row(N)) < 60 Then
      If F_13(N) == F_13(N+1) Then
        If ventilation time >=3 And <30 Then
          patient_setting = -1; N++;
            ventilation_time = ventilation_time + 3;
        Else If ventilation_time >=30 And <60 Then
          patient_setting = 0; N++;
            ventilation_time = ventilation_time + 3;
        Else If ventilation_time >= 60 Then
          patient_setting = 1; N++;
          ventilation_time = ventilation_time + 3;
        End If
      Else
        patient_seting = -2; N++;
        ventilation_time = 0;
      End If
    Else
      patient_setting = 2; N++;
    End If
  End For
End Function
```

To this work only the patient with level equal to 1 were considered. Then all the records having at least a null value were excluded. This operation had to be performed because it was not feasible to correct fill these fields. Furthermore the data mining models cannot be induced containing null values. All clinical values were validated using clinical information. In this way it was possible considering only the correct values (validation performed with intelligent agent). Through the

various procedures and function executed on the dataset, their quality and consistency was ensured.

After all the changes are made in the dataset, it was identified that the number of records used by models would be 13135 registration, representing 28 patients.

D. Modeling

Two clustering algorithms were used to implement data mining models: K-means and K-medoids. The selection of these two algorithms was due to two characteristics: follow the principle of partition method and the ability to be sensitive to outliers.

The K-means algorithm is able to create simple iterative groups in which a dataset is divided into a number priori defined. It is an algorithm of simple implementation and enforcement [33]. The K-means algorithm is sensitive to outliers, because the objects are far from the majority which can significantly influence the average value of the set. Thus assigning other objects to the clusters will be affected. This is an effect which is significantly affected by the use of square error function [34].

The K-medoids algorithm in turn takes real objects and represents the clusters, it treats each object at a time, instead of using the value of an object in a cluster as a reference point. The remaining objects are assigned to the most similar cluster. The partitioning method is then performed based on the principle teaches the sum of the differences between each of p and its corresponding object representative object [34]. The two algorithms are in fact similar except that the centroid must belongs to the collected dataset [33].

To accomplish the implementation algorithm, two adjustments were made, particularly in the identification of the most appropriate number of *K*. The number of *K* was achieved through the calculation of the Sum Square Error (SSE). The SSE determines the squared distances between each cluster member and the cluster centroid.

$$SSE(o_i) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ji}^{p} dist(o_i, c_j)^2$$

When the SSE value decreases the number of clusters increases. The identification of the K model number is identified when there is not a continuous decreasing of SSE value (when the value is stabilized). So it is possible to identify the large number of K [34].

For each model was determined the number of execution: 10 times. To the configuration of the parameter relating to the calculation of dissimilarity, the "Euclidean" distance was used. The developed models can be represented by the following expression:

$$M_n = \{A_f;\ F_i;\ D_x;\ AG_y\}.$$

The model M_n belongs to an approach (A) and it is composed by a set of fields (F), a type of variable (TV) and an algorithm (AG):

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

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

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

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

E. Evaluation

In this phase, the Davies-Bouldin Index (DBI) was used to assess the models generated.

Among the algorithms used the one which presented the most satisfactory results was the K-means algorithm.

Some of the models developed had interesting results but only one appears to present satisfactory results (DBI). Table 2 shows the most relevant models.

Table 2 – Models for clustering

Model	Fields	Number Clusters	Algorithm	DBI
M_1	$F_{\{2,3,4,5,6,13\}}$	10	AG_2	0.82
M_2	$F_{\{1,2,4,13\}}$	4	AG_2	0.51
M_4	$F_{\{1,2,3,4,13\}}$	7	AG_1	0.72
M_5	$F_{\{1,2,3,4,11,13\}}$	7	AG_1	0.73

Analyzing Table 2 it is possible to identify that the model M_2 is best model generated. The Davies Bouldin Index tends to $+\infty$ however M_2 model has an index of 0.51. This is not the ideal case but it is the model closest to 0.

In order to present a representation of the data segments generated by the best model, the Figure 1 was drawn.

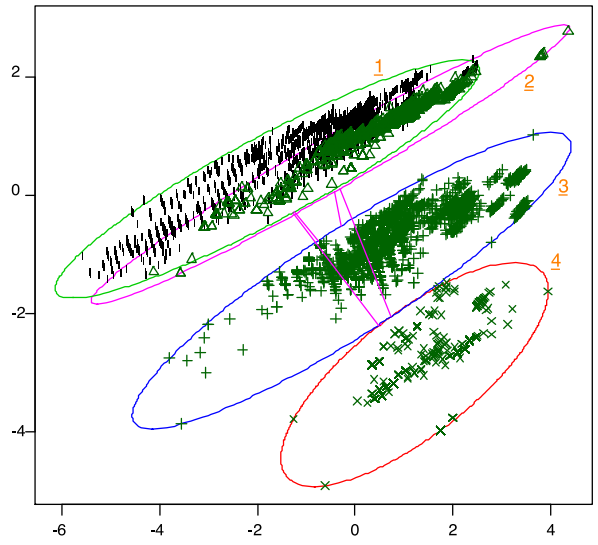

Figure 1: Graphical representation of the clusters presented in table 2

Figure 1 illustrates the four clusters created by the model M_2. It should be noted that in Cluster 1 and Cluster 2 there is a large number of intercepts having quite similar characteristics. There is only one record that could belong to both Clusters without having a big interference: CDYN (value 13). However there are 5242 records that could belong to any of the clusters, Cluster 1 or Cluster 2 if the variable is ignored.

It is also possible to verify that Cluster 3 and Cluster 4 have no interception, although they are quite near to each other the division was duly achieved. Another observation is the population number where Cluster 3 has a greater population than Cluster 4.

Since the goal is select and identify patient features which were not capable of being subjected to extubation process, it was necessary to identify the cluster that best characterizes these patients.

In order to identify the cluster that best identifies such patients, it was necessary to find the cluster which presented the better characteristics. The Table 3 presents information of the seven clusters created by the model.

Table 3 – Information of clusters

Cluster	Size	Max distance	Avg distance	Diameter	Separation
1	5565	23.69	9.40	41.50	1.00
2	2573	22.61	3.60	30.61	1.41
3	3832	44.40	9.95	62.03	1.00
4	1163	52.85	10.17	88.78	6.08

Based on the purpose of identifying the features that best determine which patients are unable to be extubated, by using clustering is possible to identify the characteristics of these same patients. As such it is intended to identify a set of more differentiated data, Cluster 3 has the second highest average distance and the second better distance between the cluster observations and cluster medoid.

Cluster 3 also presents the second largest diameter, since it has the highest dissimilarity between two points of the respective cluster.

The lowest dissimilarity achieved between an observation in Cluster 3 and another cluster was 1.00. This represents a very small value namely dissimilarity low. The Cluster 4 could also be an excellent alternative to Cluster 3, however the Cluster 4 contains only 1/3 of Cluster 3 records.

Table 4 presents the variables distribution used in Cluster 3. When a patient presents a value between the minimum and the maximum values for each variable of the cluster, he is associated to Cluster 3.

Table 4 – Distributions of Cluster 3

Cluster	Fields	MIN	MAX	MEAN	SD	CFV
Cluster 3	$F_{\{F\_1\}}$	49	108	68.19	11.33	16.62%
	$F_{\{F\_7\}}$	3.1	17	10.13	1.54	15.16%
	$F_{\{F\_8\}}$	12	36	18.58	2.91	15.68%
	$F_{\{F\_13\}}$	8	24	13.22	2.70	20.45%

Analyzing the achieved results it can conclude that although Cluster 3 had a greater dispersion, this cluster has a great homogeneity in the distribution for variable. To analyze the variables variation for each data groups, Figure 2 was designed.

In Figure 2, CDYN was identified as the variable with more importance in the creation of clusters. It is the better variable in creating the boundaries of separation between the clusters. CDYN presents an almost perfect distribution.

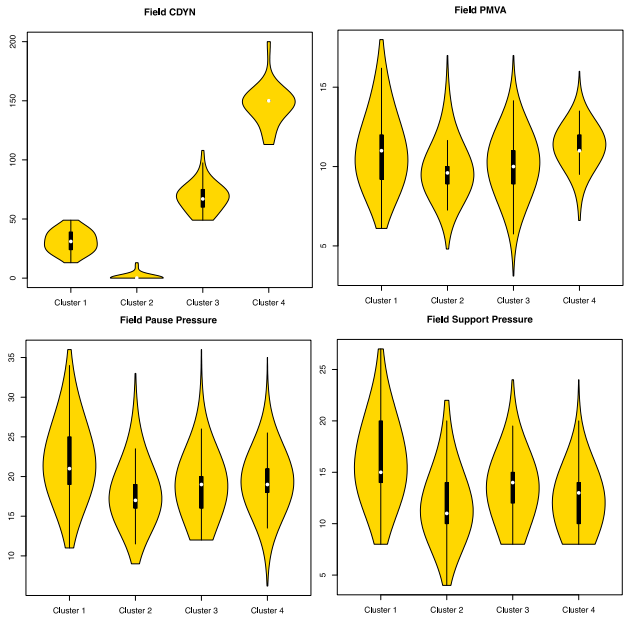

Figure 2: Diagram box

V. Discussion

The best model M_2 created 4 clusters. Cluster 3 is the only cluster able to properly identify the properties of the patients that should not be weaned / extubated, because presents a low level of intersections. 29.18% of the data used are in cluster 3.

Cluster 3 has a significant amount of registers. There is also a great homogeneity in the variables that make up the Cluster 3. Cluster 3 has not intersection with the remaining clusters. In this cluster there is a positive data separation. The MAP is the variable that demonstrates to have more homogeneous values. CDYN has a greater capacity to divide the data presented in the cluster. In general this cluster is constituted by variables with a little dispersion.

All of the patients records used in the study represent a non-possibility of extubation, however in some cases there are records that have some characteristics demonstrating the opposite. A Support Pressure value below than 7 shows that there is a full evidence of patient extubation. To this procedure be successful, it is necessary perform a clinical trial during 1 hour without the patient be re-intubated. 8 is the lower value achieved by Support Pressure in Cluster 3. This value represents an important threshold for the non-occurrence of extubation.

VI. Conclusion

After conclude this study was possible to identify a set of patterns, patient features and the variables which presents a great similarity on weaning and extubation failures.

The most satisfactory result was attained by the model M_2 with a Davies-Bouldin Index of 0.51. M_2 is the model which presented results closest to the optimum value: zero. The better Davies Bouldin Index achieved is acceptable, being this value

below to the maximum acceptable level (1). This result means that there are similarities in the clusters created. The variables used to provide the creation of the best model (cluster) were: CDYN, MAP, Pause Pressure and Support Pressure.

With this cluster there is a new possibility to improve the decision making process. The idea is presenting to the physicians the clusters created and to design dashboards containing the variables and values of the Best Cluster. This cluster will be updated in real-time always a new patient arrives to the ICU. In the dashboards will present the boundaries of each variable and some indicators associated.

The physician will not have information about the success probability but the patient variables and values associated to a not successful weaning.

With this information the physician has new information able to decide if the patient are or not prepared to make a weaning procedure. Combining these results with the clinical knowledge it is possible to provide best care to the patients.

This study is a viable work to analyze the weaning process and consequently contribute to avoid wrong extubation and avoid long injuries associated to respiratory system (e.g. lungs injuries).

In the future the dashboards containing these results will be designed and a set of other Data Mining models containing more variables will be explored.

ACKNOWLEDGMENT

The authors would like to FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II). This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

REFERENCES

[1] A. S. Fauci, "Harrison's Principles of Internal Medicine, 17e," ed: Silverchair Science: Minion, 2008.

[2] F. T. Tehrani, "Automatic control of mechanical ventilation. Part 2: the existing techniques and future trends," *Journal of clinical monitoring and computing*, vol. 22, pp. 417-424, 2008.

[3] 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*, ed: Springer, 2014, pp. 87-102.

[4] Filipe Portela, Jorge Aguiar, Manuel Filipe Santos, Álvaro Silva, and Fernado Rua, "Pervasive Intelligent Decision Support System - Technology Acceptance in Intensive Care Units," in *Advances in Intelligent Systems and Computing*, Springer, Ed., ed: Springer, 2013.

[5] S. Oliveira, C. F. Portela, and M. F. Santos, "Pervasive Universal Gateway for Medical Devices," *Recent Advances in Electrical Engineering and Education Technologies*, pp. 205-210, 2014.

[6] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelha, and J. Neves, "Information Architecture for Intelligent Decision Support in Intensive Medicine," *8th International Conference on Applied Computer and Applied Computational Science (ACACOS '09)*, vol. 8, pp. 810-819, 2009.

[7] S. Oliveira, F. Portela, M. F. Santos, J. Machado, A. Abelha, Á. Silva, *et al.*, "Predicting Plateau Pressure in Intensive Medicine for Ventilated Patients," in *New Contributions in Information Systems and Technologies*, ed: Springer, 2015, pp. 179-188.

[8] F. P. Sérgio Oliveira, Manuel Filipe Santos, José Neves, Álvaro Silva and Fernando Rua, "Feature selection for detecting patients with weaning failures in Intensive Medicine," in *Mathematical Methods and Computational Techniques II*, CPS, Ed., ed, 2015.

[9] F. P. Sérgio Oliveira, Manuel Filipe Santos, José Machado, António Abelha, Álvaro Silva and Fernando Rua, "Intelligent Decision Support to predict patient Barotrauma risk in Intensive Care Units," in *Procedia Technology - HCIST 2015 - Healthy and Secure People*, Elsevier, Ed., ed, 2015.

[10] F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado, "Adoption of Pervasive Intelligent Information Systems in Intensive Medicine," *Procedia Technology*, vol. 9, pp. 1022-1032, 2013.

[11] A. Kaynar and S. Sharma. (2010, Respiratory Failure. 39. Available: http://emedicine.medscape.com/article/167981-print

[12] T. E. Stewart, M. O. Meade, D. J. Cook, J. T. Granton, R. V. Hodder, S. E. Lapinsky, *et al.*, "Evaluation of a Ventilation Strategy to Prevent Barotrauma in Patients at High Risk for Acute Respiratory Distress Syndrome," *New England Journal of Medicine*, vol. 338, pp. 355-361, 1998.

[13] S. P. Stawicki, "Mechanical ventilation: weaning and extubation," 2007.

[14] F. T. Tehrani and J. H. Roum, "Intelligent decision support systems for mechanical ventilation," *Artificial Intelligence in Medicine*, vol. 44, pp. 171-182, 2008.

[15] L. Cardoso, F. Marins, F. Portela, A. Abelha, and J. Machado, "Healthcare interoperability through intelligent agent technology," *Procedia Technology*, vol. 16, pp. 1334-1341, 2014.

[16] L. Cardoso, F. Marins, F. Portela, M. Santos, A. Abelha, and J. Machado, "The Next Generation of Interoperability Agents in Healthcare," *International journal of environmental research and public health*, vol. 11, pp. 5349-5371, 2014.

[17] F. Marins, L. Cardoso, F. Portela, M. F. Santos, A. Abelha, and J. Machado, "Improving High Availability and Reliability of Health Interoperability Systems," in *New Perspectives in Information Systems and Technologies, Volume 2*, ed: Springer, 2014, pp. 207-216.

[18] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelha, J. Neves, *et al.*, "Information Modeling for Real-Time Decision Support in Intensive Medicine," in *Proceedings of the 8th Wseas International Conference on Applied Computer and Applied Computational Science - Applied Computer and Applied Computational Science*, S. Y. Chen and Q. Li, Eds., ed Athens: World Scientific and Engineering Acad and Soc, 2009, pp. 360-365.

[19] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelhaa, J. Neves, *et al.*, *Nursing Information Architecture for Situated Decision Support in Intensive Care Units*, 2009.

[20] F. Portela, M. F. Santos, J. Machado, A. Abelha, and Á. Silva, "Pervasive and Intelligent Decision Support in Critical Health Care Using Ensembles," in *Information Technology in Bio-and Medical Informatics*, ed: Springer Berlin Heidelberg, 2013, pp. 1-16.

[21] R. V. Filipe Portela, Sérgio Oliveira, Manuel Filipe Santos, António Abelha, José Machado, Álvaro Silva and Fernando Rua, "Predict hourly patient discharge probability in Intensive Care Units using Data Mining," *ScienceAsia Journal (ICCSCM 2014)*, 2014.

[22] Pedro Braga, F. Portela, and M. F. Santos, "Data Mining Models to Predict Patient's Readmission in Intensive Care Units," in *ICAART - International Conference on Agents and Artificial Intelligence*, Angers, France, 2014.

[23] R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, *et al.*, "A clustering approach for predicting readmissions in intensive medicine," *Procedia Technology*, vol. 16, pp. 1307-1316, 2014.

[24] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelha, and J. Neves, "INTCARE - Multi-agent approach for real-time Intelligent Decision Support in Intensive Medicine," in *3rd International Conference on Agents and Artificial Intelligence (ICAART)*, Rome, Italy, 2011.

[25] E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*, 9th Edition ed.: Prentice Hall, 2010.

[26] S. Tufféry, *Data mining and statistics for decision making*: John Wiley & Sons, 2011.

[27] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management—Vol*, vol. 19, p. 65, 2011.

[28] P.-N. Tan, Steinbach, M., & Kumar, V, *Introduction to Data Mining* 1ed.: Addison-Wesley Longman Publishing Co., Inc., 2005.

[29] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Morgan kaufmann, 2006.

[30] J. C. Krzysztof, W. S. Roman, and A. Lukasz, "Data Mining: A Knowledge Discovery Approach," ed: Springer, 2007.

[31] J. Ranjan, "Business intelligence: concepts, components, techniques and benefits," *Journal of Theoretical and Applied Information Technology*, vol. 9, pp. 60-70, 2009.
[32] L. Torgo, *Data mining with R: learning with case studies*: Chapman & Hall/CRC, 2010.
[33] X. Wu and V. Kumar, *The top ten algorithms in data mining*: CRC Press, 2009.
[34] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*: Elsevier, 2011.