Data Science Project

Grupo	**11**
Daniel Correia	80967
Pedro Orvalho	81151
Stéphane Duarte	81186

MEIC-A SAD 2017

INDEX

1 Introduction

The goal of the proposed project is to apply **unsupervised learning techniques** (association rules and clustering) over two distinct datasets: data on a population of crabs (Problem 1) and data on no-shows for medical appointments (Problem 2).

The first dataset represents data on a population of crabs. This dataset has 200 instances, the class attribute "sp" identifies the species of the crabs and the remaining attributes are of type character ("Gender") and numerical ("index" and physical characteristics "FL","CL","CW","BD","RW").

The second dataset represents data on attendance of medical appointments. This dataset has over 100.000 instances, the class attribute is "no-shows" and the remaining attributes are of type date ("AppointmentDay" and "ScheduledDay"), character ("Neighbourhood" and "Gender") and numerical ("Age", "Diabetes", etc…).

2 Pre processing

2.1 Problem 1 - Crabs dataset (crabs data)

The first step was **removing the class attribute "sp"** from the dataset since this attribute should only be used for classification (second delivery). Secondly, we **converted the binary character attribute "sex"** from "M" and "F" to numericals 0 and 1.

Finally, we **generated 3 different types of datasets** to be used in the exploration stage:

1. Base dataset where we **truncated the floats to integers**
2. Discretized dataset by **4 bins of equal-width** (intervals)
3. Discretized dataset by **5 bins of equal-height** (frequency)

We decided to discretize the dataset by bins because most of the attributes were of type float, which meant that there would be a low amount of frequent itemsets once we tried to apply association rules techniques, **and they had a similar order of magnitude**. It also allowed us to keep the resulting dataset as close as possible to its original form, when compared to the "truncate floats to integers" approach.

2.2 Problem 2 - No-shows dataset (noshows data)

The first step was **removing the class attribute "No-show"** from the dataset since this attribute should only be used for classification (second delivery).

Secondly, we **converted the date attributes** "AppointmentDay" and "ScheduledDay" to a numerical attribute: we **used taxonomy to replace the date content with the corresponding week number** (from 1 to 52). We made some experiments with replacing the date with the day of the month but overall the results were worse than with week numbers in both association rules and clustering exploration.

Finally, we **removed the attributes "AppointmentID" and "PatientID" because they provide no valuable insight into the dataset** (they are arbitrarily assigned to a person). We also **removed the "Neighbourhood" attribute because there was no appropriate normalization technique** for this attribute (e.g converting to numericals would ignore the geographical distances between different neighbourhoods).

From this point on, we had a base dataset that could be used for association rules and clustering algorithms.

To apply the clustering algorithms, we had to **split the dataset into 10 smaller samples** (10% of the original dataset) **using random sampling to avoid memory problems when running the k-means** algorithm or when calculating the distance matrix needed to obtain the silhouette coefficient of the clusters.

We also **generated a discretized version of the base dataset by 4 bins of equal-height** (frequency)

3 Exploration

3.1 Association rules exploration

3.1.1 Methods and Parametrization

Regarding the **crabs dataset**, we **applied the apriori algorithm over the 3 types of datasets** that we created (truncated, discretized by bins of intervals and discretized by bins of frequency). For the **no-shows dataset**, we **applied the apriori algorithm over the base and the discretized datasets**.

In terms of parametrization, we chose a **minimum support of 5% and minimum confidence of 75%** and, from these values, we did s**everal iterations where we increased support by 15% and/or confidence by 10%** to evaluate the evolution of the number and quality of the rules found by the Apriori algorithm.

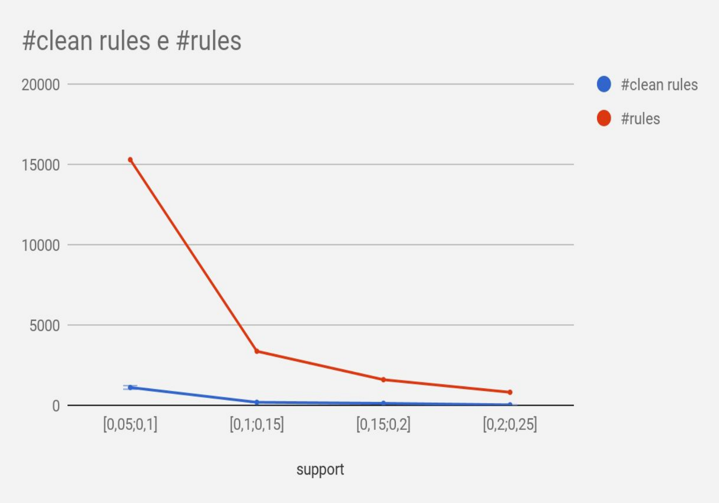
3.1.2 Results

Crabs dataset used	**# rules**	**# rules with supp > 20% and conf > 75%**	**Mean(lift)**
Truncated	3	0 (only 3 with supp < 10% and conf > 90%)	1.7
Discretized by equal-height	1635	78	3.6
Discretized by equal-width	1742	35	2.9

Based on observation of the top 5 rules sorted by lift of these attempts, we found that there's a **high correlation between 2 sets of attributes**:

- high values of CL and CW (around 37 to 42)
- medium values of FL and BD (around 10 to 25).

Regarding the no-shows dataset, we took into consideration the **rules with an empty set on the left**, since they have a high degree of support and confidence but **provide no interesting insight**. They are **dangerous because you can derive any rule from them**, so we **decided to remove them from the generated set**.



No-shows used	**# rules**	**# rules after cleaning**	**Mean(supp)**	**Mean(conf)**
Base dataset	6083	634	8.01%	90.29%
Discretized	20000+	1450 (34 with supp > 20% and conf > 75%)	20%	75%

On our second attempt, despite the low average lift of 1.37, there was an i**nteresting rule with a high support and confidence: "ScheduledDay[18,20] => {SMS_received=0}"**. Most of these 34 rules are a derivation of this rule, which **may indicate that there's an underlying event** that caused SMS's to not be received during weeks 18 to 20.

3.2 Clustering exploration

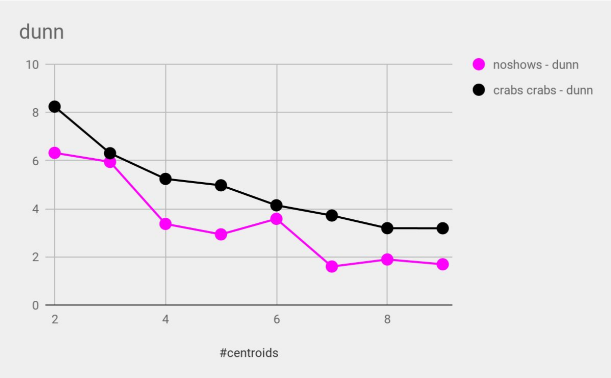
3.2.1 Methods and Parametrization

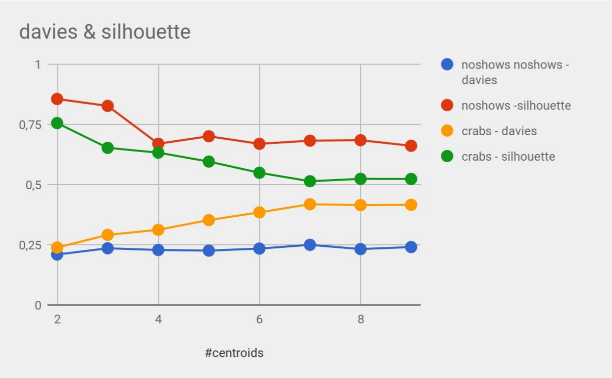
Regarding the **crabs dataset**, we **used the k-means algorithm over the base dataset with a default number of centroids of 2** and **iteratively increased the number of centroids until 9**.

Regarding the **no-shows dataset**, since we were **forced to split the dataset into smaller pieces to avoid memory issues,** we **applied the k-means algorithm over all the samples** and used the same strategy for parametrization from the previous dataset **(iteratively increased the number of centroids from 2 to 10)**.

3.2.2 Results

Regarding the clustering, we obtained the following results for the crabs dataset, and these results for the 10 samples of the noshows dataset. These values are based on the centroid diameter (intracluster distance) reflects the double average distance between all of the samples and the cluster's center, and on the centroid linkage distance (intercluster distance) the distance between the centres of two clusters.





Regarding the **dunn index for the crabs and the no-shows datasets,** the better value (higher) is when the number of centroids is 2, which indicates the presence of compact and well-separated clusters.

Looking at the **davies values for the crabs and the no-shows datasets,** the better (lower) value is when the number of centroids is 2 which means the clusters are compact and their centers are far away.

Analysing the **values of silhouette coefficient for both datasets**, when the number of centroids is 2 the silhouette coefficient is **0,7555535 for crabs** and **0,8556663 for no-shows**, which indicates an excellent separation between clusters, unlike the other numbers of centroids.

Based on these metrics, we concluded that **the optimal configuration for both datasets is 2 clusters**.

We also used **PCA** to obtain a different and better optimal solution, the values of the metrics (silhouette coefficient, davies and dunn indexes) **were better when the number of centroids was 3**, even though these values were still worse than without PCA.

4 Critical Analysis

Regarding the association rules exploration, we obtained **better results when we used a discretized dataset** by bins of equal-height instead of the base dataset in both problems. This may be due to the **sparseness of the attribute values in the original dataset**, which would result in **less frequent itemsets.** The discretization allowed for **more frequent itemsets** to be found **by concentrating the values in bins**.

In terms of the association rules obtained for the crabs dataset, **most of the rules indicate a relation between a set of attributes in certain intervals**. This may be caused by the existence of 4 different population of crabs within the crabs dataset, with each population having its specific set of values for each attribute.

Looking at the association rules exploration in the no-shows dataset, the 5 empty rules that we found corresponded with the absence of these attributes: Diabetes, Hipertension, Alcoholism, Handcap and Scholarship. These rules had a high degree of support and confidence. **We think that this happened because only a small percentage of the population has any of these, which means that given a random individual from the population there's a high chance that they don't have diabetes** (for example).

In terms of the sampling technique used in the no-shows dataset, **we need to take into account that we never analysed 100% of the dataset**, which means that we may have missed some insights that can only be obtained by using the entire dataset in unsupervised learning techniques. We tried to **minimize this by using 10 random samples of 10% of the dataset** and calculating an average of the results obtained.

Regarding the clustering exploration of the **crabs dataset, we concluded that the best number of centroids is 2 based on the metrics obtained**. This may imply that the k-means algorithm divided the crabs dataset based on the gender (a binary attribute) or based on the populations of crabs represented in the dataset (e.g. 2 populations per cluster).

Regarding the clustering exploration of the **no-shows dataset, we also concluded that the best configuration was 2 centroids based on the metrics obtained**. We also tried to use **PCA** to obtain a different and better result in terms of the metrics and, even though we obtained a different optimal configuration for the number of centroids with PCA, the values of the metrics (silhouette coefficient, davies and dunn indexes) were worse than without PCA.

5 Conclusions

Comparing the two methods used (association rules vs clustering), we think that the **association rules exploration was more useful than the clustering techniques** because the clustering results only indicated that **the dataset could be divided into 2 clusters**, which is basically the same as **separating the instances of the dataset based on the class** attribute (yes/no or by species).

The association rules exploration allowed us to find frequent itemsets from which we could try to extract interesting insights, for example, we were able to **find an anomaly event in the weeks 18 to 20 that may have caused the patients to not receive SMS**.