TÉCNICO LISBOA

MSc Engenharia e Ciência de Dados

Métodos Estatísticos em Data Mining

Relatório 2

Authors:

Diogo Vilela (96193) diogo.pimpao.vilela@tecnico.ulisboa.pt
Pedro Rodrigues (96301) pedro.maria.rodrigues@tecnico.ulisboa.pt
José Pedro Antunes (96260) jose.pedro.m.c.a@tecnico.ulisboa.pt
Sebastião Caldas (96321) sebastiao.caldas@tecnico.ulisboa.pt
Nuno Marques (95758) nuno.figueiredo.marques@tecnico.ulisboa.pt

Grupo 04

2022/2023 – 1º Semestre, P1

CONTENTS

I. Introdução

O projeto realizado consiste na análise de um conjunto de dados com o objetivo de resolver um problema importante: prever algumas complicações do Enfarte do Miocárdio (EM) com base nas informações de vários pacientes, em diferentes momentos.

II. Objetivo

Nesta segunda parte, o objetivo do projeto é aplicar vários métodos de *Clustering*, que fazem parte do grupo de aprendizagem não supervisionada.

Para a resolução deste problema serão utilizados os seguintes métodos: Hierárquicos, com base na densidade (DBSCAN), com base na distribuição, de partição e com base em grafos.

Por fim, pretende-se resolver o problema de classificação utilizando os *clusters* e o melhor classificador do primeiro projeto (no nosso caso, o *Naive Bayes com Bernoulli*).

III. Métodos de Clustering

A. Métodos Aglomerativos

Os métodos de *clustering* aglomerativos baseiam-se na fusão recursiva de *clusters* em cada nível hierárquico de acordo com uma dada métrica ou medida. Inicialmente, as observações são os seus próprios *clusters*, que se vão unindo, sucessivamente, em cada iteração do algoritmo, fazendo com que cada nível tenha menos um *cluster* do que aquele que lhe precede. O algoritmo termina, naturalmente, com um único *cluster*, ao qual todas as observações originais pertencem.

Evidentemente, estes métodos dependem da escolha da relação de *dissemelhança* entre objetos e do critério de fusão de *clusters*; prendendo-se, sobretudo, com a natureza dos dados em questão. Visto que os nossos dados são maioritariamente categóricos binários, e após terem sido retiradas as variáveis contínuas, foram abordadas duas estratégias: aplicar *one-hot encodding* aos dados, considerando como métrica a distância de *hamming* e; não aplicar nenhuma transformação e considerar a distância de *gower*.

Apesar de se terem formulado duas estratégias distintas, a metodologia permaneceu igual. Efetivamente, consideram-se os métodos de agrupamento *complete-linkage*, *single-linkage* e, *average-linkage*, escolhendo-se o número de *clusters* de modo a maximizar o *silhouette coefficient*. Seguidamente, em função do método que apresenta ter *clusters* mais distintos, comparou-se os valores da partição com os valores da variável resposta, obtendo-se as métricas adequadas.

1) Distância de Hamming: Em seguida apresentam-se os resultados para a primeira estratégia de *clustering* aglomerativo

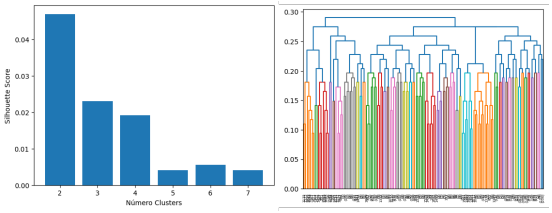

Fig. 1. *Silhoutte Scores* e dendrograma (truncado) para *complete-linkage*.

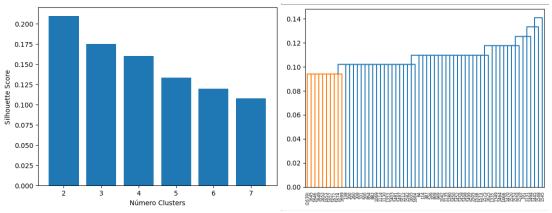

Fig. 2. *Silhoutte Scores* e dendrograma (truncado) para *single-linkage*.

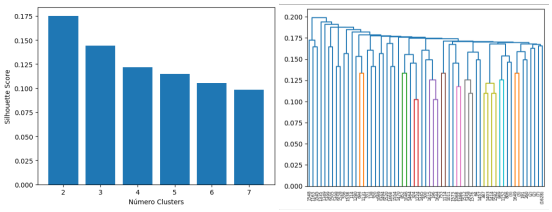

Fig. 3. *Silhoutte Scores* e dendrograma (truncado) para *average-linkage*.

Como se pode constatar, o número de *clusters* ideal é

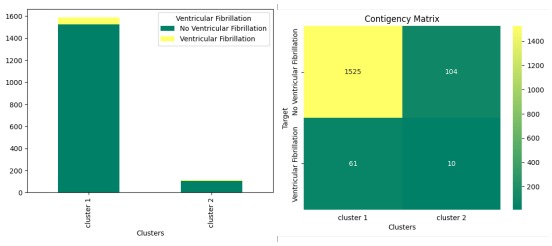

Fig. 4. Matriz de confusão para as *true labels* vs *clusters*.

	Random Index	**Adjusted Random Index**	**Adjusted Mutual Index**
Score	0.921	0.026	0.020

TABLE I
Métricas para distância de Hamming

2) Distância de Gower: Em seguida apresentam-se os resultados para a primeira estratégia de *clustering* aglomerativo

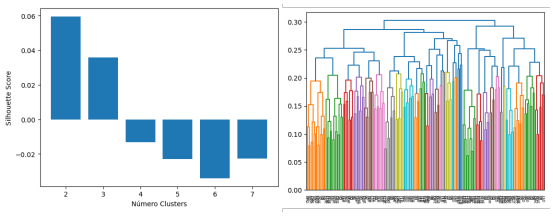

Fig. 5. *Silhoutte Scores* e dendrograma (truncado) para *complete-linkage*.

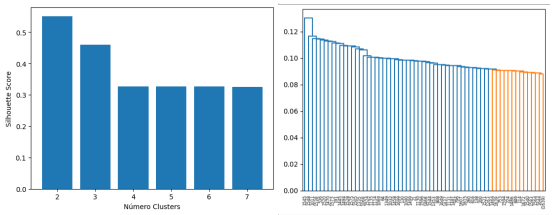

Fig. 6. *Silhoutte Scores* e dendrograma (truncado) para *single-linkage*.

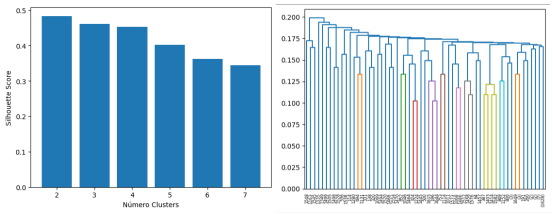

Fig. 7. *Silhoutte Scores* e dendrograma (truncado) para *average-linkage*.

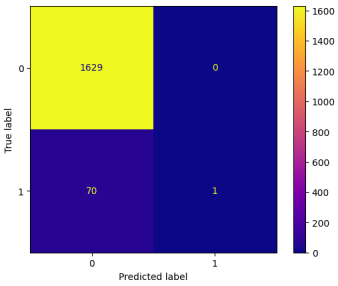

Fig. 8. Matriz de confusão para as *true labels* vs *clusters*.

	Random Index	**Adjusted Random Index**	**Adjusted Mutual Index**
Score	0.921	0.026	0.020

TABLE II
MÉTRICAS PARA DISTÂNCIA DE GOWER

B. Métodos de Partição

1) K-means:

C. DBSCAN

O DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um algoritmo de agrupamento baseado em densidade. Este método descobre *clusters* com forma arbitrária e é bastante eficiente para conjuntos de dados multi dimensionais.

O DBSCAN procura *clusters* através da vizinhança de cada ponto do *dataset* e verifica se contém mais do que um determinado número de observações.

Desta forma, é necessário definir dois parâmetros de entrada *à priori*: o raio de vizinhança (Eps) e o número mínimo de pontos para definir um *cluster* (minPts).

Pode-se definir três tipos de pontos neste algoritmo. Se uma determinada observação tiver mais pontos do que o número mínimo, definido num raio *Eps*, é considerado um *core point*. Se um ponto tiver menos pontos que número mínimo mas estiver na vizinhança de um *core point* é então definido como um *border point*. Um ponto que não seja nenhum dos anteriores é considerado um *noise point*.

O DBSCAN é capaz de detetar *clusters* de diferentes formas e densidades, e é menos afetado pelo ruído e *outliers* do que outros algoritmos de agrupamento. No entanto, é preciso ser muito rigoroso quando se define valores para o raio e para o número mínimo de pontos, uma vez que uma pequena alteração pode conduzir a resultados muito diferentes.

Para definir os parâmetros de entrada começou-se por calcular o número mínimo de pontos para um grupo ser considerado um *Cluster*. Este valor, o minPts, foi definido como 218, o dobro do número de variáveis do conjunto de dados.

Para o cálculo do raio Eps, recorreu-se à seguinte figura, que corresponde à representação gráfica das distâncias médias de cada ponto aos seus k vizinhos mais próximos. Assim, o "cotovelo" vai indicar o valor ótimo para este raio. Neste caso, por observação considerou-se Eps = 0.125.

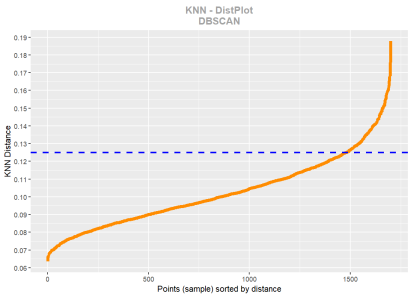

Fig. 9. KNN Distance Plot.

Como se pode ver nas figuras seguintes, este método sugere apenas um *Cluster*. Isto pode-se explicar devido ao facto dos pontos não apresentarem grandes diferenças entre si. Assim, o algoritmo apenas sugere um agrupamento, o que não facilita o nosso problema de classificação.

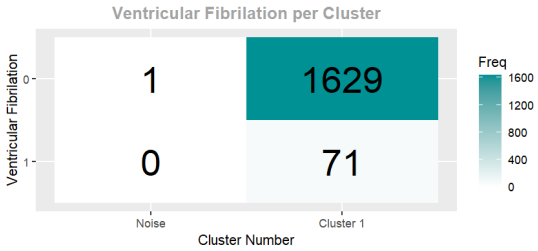

Fig. 10. Distribuição do target por cluster.

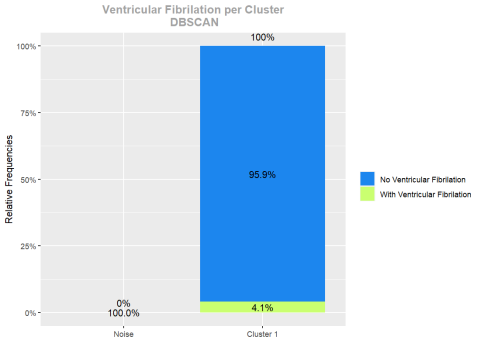

Fig. 11. Distribuição percentual do target por cluster..

D. Distribution-based

O método de *Cluster* com base na distribuição é o que está mais relacionado com estatística. Este modelo assume que as observações que pertencem a cada agrupamento têm uma distribuiçao de probabilidade específica. Assume-se que a distribuição global dos dados seja uma mistura de várias distribuições.

Neste algoritmo, o objetivo é identificar os vários aglomerados e os seus parâmetros de distribuição.

IV. Conclusão

References