# Constrained Clustering

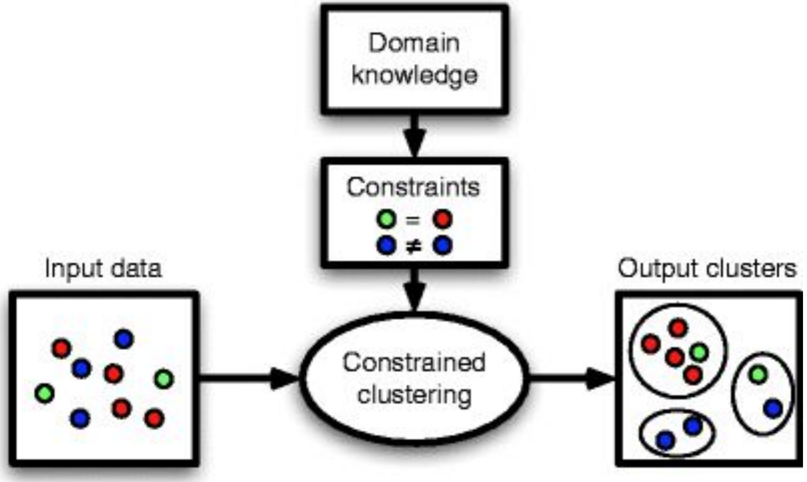
2021

# Resumo

1. Definição
2. Agrupamento de dados
3. Artigos Relacionados
4. Proposta

# 1.Definição

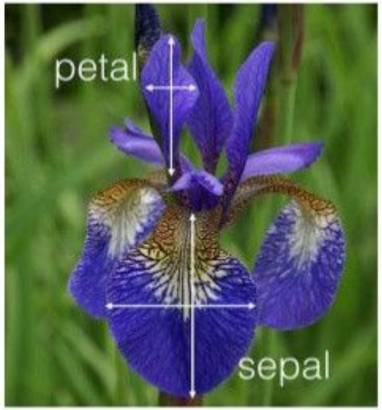
É uma abordagem **semi-supervisionada** para agrupar dados enquanto incorpora **conhecimento de domínio** na forma de **restrições**. As restrições são geralmente expressas como declarações em pares, indicando que dois itens **devem** ou **não podem** ser colocados no mesmo cluster.




Fonte: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_163

# 2. Agrupamento de dados

# Exemplo agrupamento conjunto de dados iris




| Features | | | | Labels |
|---|---|---|---|---|
| Sepal length | Sepal width | Petal length | Petal width | Species |
| 5.1 | 3.5 | 1.4 | 0.2 | Iris setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | Iris setosa |
| 7.0 | 3.2 | 4.7 | 1.4 | Iris versicolor |
| 6.4 | 3.2 | 4.5 | 1.5 | Iris versicolor |
| 6.3 | 3.3 | 6.0 | 2.5 | Iris virginica |
| 5.8 | 3.3 | 6.0 | 2.5 | Iris virginica |

[ 1 , 1 , 2 , 2 , 3 , 3 ]

Algoritmo de Clustering

[ 1 , 2 , 2 , 2 , 3 , 3 ]

ARI

0.44

[-1,1]

# 3. Artigos relacionados

# Propôs BRKGA e Geração de Colunas

OBS: Fez contagem ML e CL satisfeitos, não utilizou ARI pois se baseou em artigo anterior que só contava ML e CL satisfeitos

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

## A comparison of two hybrid methods for constrained clustering problems

Rudinei Martins de Oliveira[a,*], Antonio Augusto Chaves[a], Luiz Antonio Nogueira Lorena[b]

Link: https://www.sciencedirect.com/science/article/abs/pii/S1568494617300388

# Comparou o BRKGA do artigo anterior com heurísticas (COPKM,LCVQE,RDPM,TVClust,CECM) e o algoritmo proposto por ele (DILS)

OBS: Não fez contagem de ML e CL satisfeitos. Utilizou ARI

ELSEVIER

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

## DILS: Constrained clustering through dual iterative local search

Germán González-Almagro[a,*], Julián Luengo[a], José-Ramón Cano[b], Salvador García[a]

[a] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Spain
[b] Dept. of Computer Science, EPS of Linares, University of Jaén, Campus Científico Tecnológico de Linares, Cinturón Sur S/N, Linares 23700, Jaén, Spain

Link: https://www.sciencedirect.com/science/article/abs/pii/S0305054820300964

On the k-Medoids Model for Semi-supervised Clustering

Rodrigo Randel[1(✉)], Daniel Aloise[1], Nenad Mladenović[2], and Pierre Hansen[3]

OBS:

1. Utilizou apenas ARI
2. dij = distância euclidiana

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} d_{ij}$$

subject to

$$\sum_{j=1}^{n} x_{ij} = 1, \quad \forall i = 1, ..., n$$

$$x_{ij} - x_{wj} = 0 \quad \forall (p_i, p_w) \in \mathcal{ML}, \quad \forall j = 1, ..., n$$

$$x_{ij} + x_{wj} \leq 1 \quad \forall (p_i, p_w) \in \mathcal{CL}, \quad \forall j = 1, ..., n$$

$$x_{ij} \leq y_j \quad \forall i = 1, \ldots, n, \forall j = 1, ..., n$$

$$\sum_{j=1}^{n} y_j = k$$

$$x_{ij} \in \{0, 1\} \quad \forall i = 1, \ldots, n, \forall j = 1, \ldots, n,$$

$$y_j \in \{0, 1\} \quad \forall j = 1, \ldots, n,$$

OBS: O modelo do paper anterior pode falhar com o tipo de dados abaixo

# Constrained Overlapping Clusters: Minimizing the Negative Effects of Bridge-Nodes

Jerry Scripps, *Member, IEEE* and Pang-Ning Tan, *Member, IEEE*

***Abstract*—This paper presents a new approach to forming overlapping clusters of objects by balancing the effects of incompleteness, impurity and overlap. Incompleteness results from similar objects separated into different clusters while impurity arises when a cluster contains dissimilar objects. Overlap is caused by nodes that appear in more than one cluster. The key to balancing these effects is the identification of bridge-nodes. We show the limitations of traditional clustering algorithms in handling bridge nodes and demonstrate the intractability of minimizing all three effects. Approximation algorithms based on graph mincut and genetic algorithm are proposed to minimize these effects. Our results with real data sets show significant improvement over traditional methods with regard to incompleteness, impurity and overlap.**

***Index Terms*—Constrained Clustering and Overlapping Clustering**

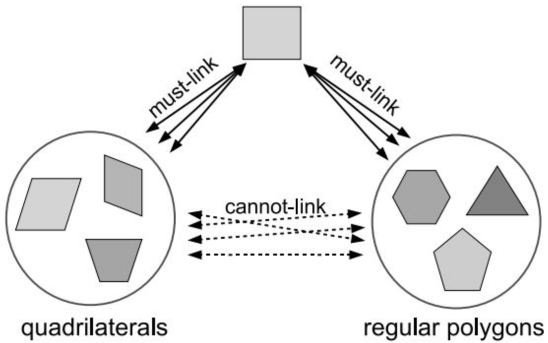



Fig. 1. Bridge-node example

OBS: O modelo do paper anterior pode falhar com o tipo de dados criados da forma abaixo

# Clustering in the Presence of Bridge-Nodes

Jerry Scripps
*Computer Science and Engineering*
*Michigan State University*
*scripps@msu.edu*

Pang-Ning Tan
*Computer Science and Engineering*
*Michigan State University*
*ptan@msu.edu*

Our experiments were performed using k-means and three agglomerative hierarchical clustering algorithms (complete-link, single-link and group-average) [10]. The number of clusters was varied from 20 to 500. To identify the ML and CL edges, we used thresholds based on the top 1% and the bottom 1% of the similarity values.

Modelo matemático baseado no k-means (força ML/CL nas restrições)

# A Binary Linear Programming-Based K-Means Algorithm For Clustering with Must-Link and Cannot-Link Constraints

Publisher: IEEE



Philipp Baumann All Authors



Abstract



**Abstract:**
Clustering is probably the most extensively studied problem in unsupervised learning. Traditional clustering algorithms assign objects to clusters exclusively based on features of the objects. Constrained clustering is a generalization of traditional clustering where additional information about a dataset is given in the form of constraints. It has been shown that the clustering accuracy can be improved substantially by accounting for these constraints. We consider the constrained clustering problem where additional information is given in the form of must-link and cannot-link constraints for some pairs of objects. Various algorithms have been developed for this specific clustering problem. We propose a binary linear programming-based k-means approach that can consider must-link and cannot-link constraints. In a computational experiment, we compare the proposed algorithm to the DILS $_{CC}$ algorithm, which represents the state-of-the-art. Our results on 75 problem instances indicate that the proposed algorithm delivers better clusterings than the DILS $_{CC}$ algorithm in much shorter running time.

**Published in:** 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)

Link: https://ieeexplore.ieee.org/document/9309775

OBS: Comparou com o DILScc

TABLE III: Average ARI values

| Dataset | $CS_{10}$ BLPKM$_{CC}$ | $CS_{10}$ DILS$_{CC}$ | $CS_{15}$ BLPKM$_{CC}$ | $CS_{15}$ DILS$_{CC}$ | $CS_{20}$ BLPKM$_{CC}$ | $CS_{20}$ DILS$_{CC}$ |
|---|---|---|---|---|---|---|
| Appendicitis | 0.573 | **0.611** | **1.000** | 0.957 | **1.000** | **1.000** |
| Breast Cancer | **0.979** | 0.755 | **1.000** | 0.792 | **1.000** | 0.796 |
| Bupa | **0.931** | 0.889 | **1.000** | 0.993 | **1.000** | 0.988 |
| Circles | **0.850** | 0.781 | **1.000** | **1.000** | **1.000** | **1.000** |
| Ecoli | **0.686** | 0.039 | **0.912** | 0.091 | **0.974** | 0.264 |
| Glass | **0.286** | 0.008 | **0.763** | 0.076 | **0.942** | 0.258 |
| Haberman | **0.929** | 0.802 | **1.000** | **1.000** | **1.000** | **1.000** |
| Hayesroth | **0.173** | 0.057 | **0.978** | 0.478 | **0.923** | 0.816 |
| Heart | **0.885** | 0.846 | **1.000** | **1.000** | **1.000** | **1.000** |
| Ionosphere | **0.943** | 0.809 | **1.000** | 0.973 | **1.000** | 0.984 |
| Iris | **0.584** | 0.550 | 0.598 | **0.832** | 0.574 | **0.953** |
| Led7Digit | **0.611** | 0.013 | **0.877** | 0.012 | **0.988** | 0.017 |
| Monk2 | **0.963** | 0.823 | **1.000** | 0.899 | **1.000** | 0.899 |
| Moons | **0.987** | 0.963 | **1.000** | **1.000** | **1.000** | **1.000** |
| Movement Libras | **0.312** | 0.019 | **0.348** | 0.018 | **0.503** | 0.020 |
| Newthyroid | **0.865** | 0.040 | **0.984** | 0.390 | **0.984** | 0.845 |
| Saheart | **0.983** | 0.788 | **1.000** | 0.870 | **1.000** | 0.867 |
| Sonar | **0.743** | 0.710 | 0.981 | **0.981** | **1.000** | **1.000** |
| Soybean | **0.607** | 0.289 | **0.607** | 0.468 | **0.805** | 0.629 |
| Spectfheart | 0.871 | **0.895** | **1.000** | **1.000** | **1.000** | **1.000** |
| Spiral | **0.857** | 0.849 | **1.000** | **1.000** | **1.000** | **1.000** |
| Tae | **0.046** | 0.028 | **0.547** | 0.386 | **0.982** | 0.846 |
| Vehicle | **0.956** | 0.023 | **1.000** | 0.066 | **1.000** | 0.171 |
| Wine | **0.397** | 0.326 | 0.536 | **0.740** | 0.583 | **0.898** |
| Zoo | **0.629** | 0.221 | **0.788** | 0.193 | **0.819** | 0.250 |
| Mean | **0.706** | 0.485 | **0.877** | 0.649 | **0.923** | 0.740 |

$$
\text{BLP}\begin{cases}
\text{Min.} \ \sum_{i=1}^{n}\sum_{j=1}^{k} d_{ij}y_{ij} & & (1)\\
\text{s.t.} \ \sum_{j=1}^{k} y_{ij} = 1 & (i = 1,\dots,n) & (2)\\
\sum_{i=1}^{n} y_{ij} \geq 1 & (j = 1,\dots k) & (3)\\
y_{ij} = y_{i'j} & ((i,i') \in ML;\ j = 1,\dots,k) & (4)\\
y_{ij} + y_{i'j} \leq 1 & ((i,i') \in CL;\ j = 1,\dots,k) & (5)\\
y_{ij} \in \{0,1\} & (i = 1,\dots,n;\ j = 1,\dots,k) & (6)
\end{cases}
$$

# 4. Proposta

# Constrained clustering baseado em consenso

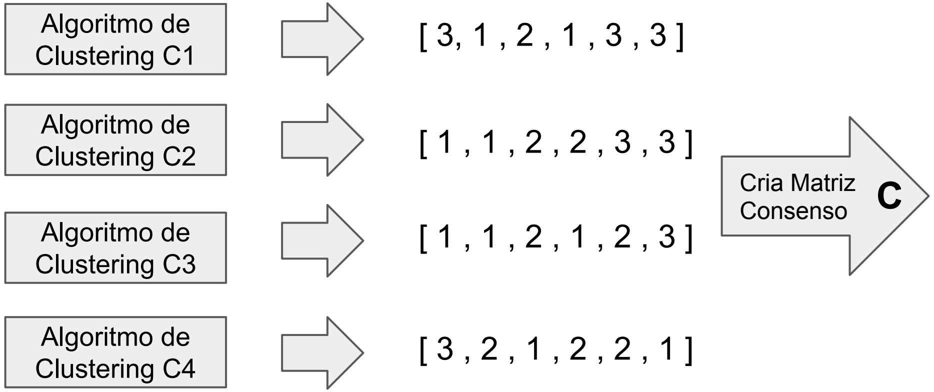
Utilizará o modelo de p-medianas (abordagem **não-supervisionada**)

A idéia é colocar a semi-supervisão no modelo.

Diferente de Randel et al., que insere restrições explícitas para o ML e CL no modelo matemático, será utilizada a matriz de consenso entre o agrupamento de uma série de algoritmos

Para montar a matriz conta-se o número de vezes que o par (i,j) aparece junto em cada uma das soluções obtidas pelos algoritmos de agrupamento.

| Features | | | | Labels |
|---|---|---|---|---|
| Sepal length | Sepal width | Petal length | Petal width | Species |
| 5.1 | 3.5 | 1.4 | 0.2 | Iris setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | Iris setosa |
| 7.0 | 3.2 | 4.7 | 1.4 | Iris versicolor |
| 6.4 | 3.2 | 4.5 | 1.5 | Iris versicolor |
| 6.3 | 3.3 | 6.0 | 2.5 | Iris virginica |
| 5.8 | 3.3 | 6.0 | 2.5 | Iris virginica |




$$
\begin{bmatrix}
0.0 & 0.5 & 0.0 & 0.25 & 0.25 & 0.25 \\
0.5 & 0.0 & 0.0 & 0.75 & 0.25 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.75 & 0.25 & 0.0 & 0.25 & 0.0 \\
0.25 & 0.25 & 0.25 & 0.25 & 0.0 & 0.5 \\
0.25 & 0.0 & 0.25 & 0.0 & 0.5 & 0.0
\end{bmatrix}
$$

# Modelo de p-medianas com matriz de consenso

$$
\begin{aligned}
\text{Min} \quad & \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}\, x_{ij} & \\
\text{Sub.} \quad & \textstyle\sum_{i=1}^{n} x_{ii} = p & , i = 1 \ldots n \\
& x_{ii} \geq x_{ij} & , i = 1, \ldots, n; j = 1, \ldots, n \\
& \textstyle\sum_{i=1}^{n} x_{ij} = 1 & , j = 1, \ldots, n \\
& x_{ij} \in \{0, 1\} &
\end{aligned}
$$

$$\Downarrow$$

$$
\begin{aligned}
\text{Min} \quad & \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} [(1-\alpha)\, d_{ij} - \alpha\, C_{ij}]\, x_{ij} & \\
\text{Sub.} \quad & \textstyle\sum_{i=1}^{n} x_{ii} = p & , i = 1 \ldots n \\
& x_{ii} \geq x_{ij} & , i = 1, \ldots, n; j = 1, \ldots, n \\
& \textstyle\sum_{i=1}^{n} x_{ij} = 1 & , j = 1, \ldots, n \\
& x_{ij} \in \{0, 1\} &
\end{aligned}
$$

Testar os valores de alfa entre 0.0 e 1.0