Seminários II

Metaheurísticas para o Problema da Seleção de Atributos em Bases de Dados, Modelado como Empacotamento de Conjuntos

Aluno: Bruno C. do Nascimento
Orientador: Marcos Henrique Fonseca Ribeiro

Sumário

- O Problema
- Por que reduzir a dimensão dos dados?
- Como reduzir a dimensão dos dados?
- Modelagem do Problema
- Métrica da Silhueta
- Aplicação do trabalho
- O que já foi feito e o que está sendo feito
- O que será feito
- Cronograma
- Dúvidas e sugestões

Problema

- Utilizar busca heurística para a seleção dos melhores atributos de uma base de dados com a finalidade de reduzir a sua dimensão.

Por que reduzir a dimensão dos dados?

- A maldição da dimensionalidade

A maldição da dimensionalidade

- Alto custo computacional
- Atributos redundantes (possuem alta correlação)
- Problemas com a eficácia das métricas de distância

Ex: Atrapalha o agrupamento realizado por algoritmos de clusterização

Como reduzir a dimensão dos dados?

- PCA (Principal Component Analysis)
- Busca Heurística (Presente trabalho)

Modelagem do problema

- O problema foi modelado como Set Packing Problem.
- Set Packing Problem é NP-Completo
- Encontrar $c \subset S$ que maximize a silhueta, onde S é o conjunto com todos atributos da base.
- É informado os valores tam_min e tam_max
  - Onde $tam\_min \leq |c| \leq tam\_max$

Restrição do problema

- Se dois atributos a_1 e $a_2 \in S$ possuem módulo da correlação acima de de um limiar th, eles não podem aparecer juntos em c. Por terem uma alta correlação, eles representam informações muito parecidas, e isso causa redundância.

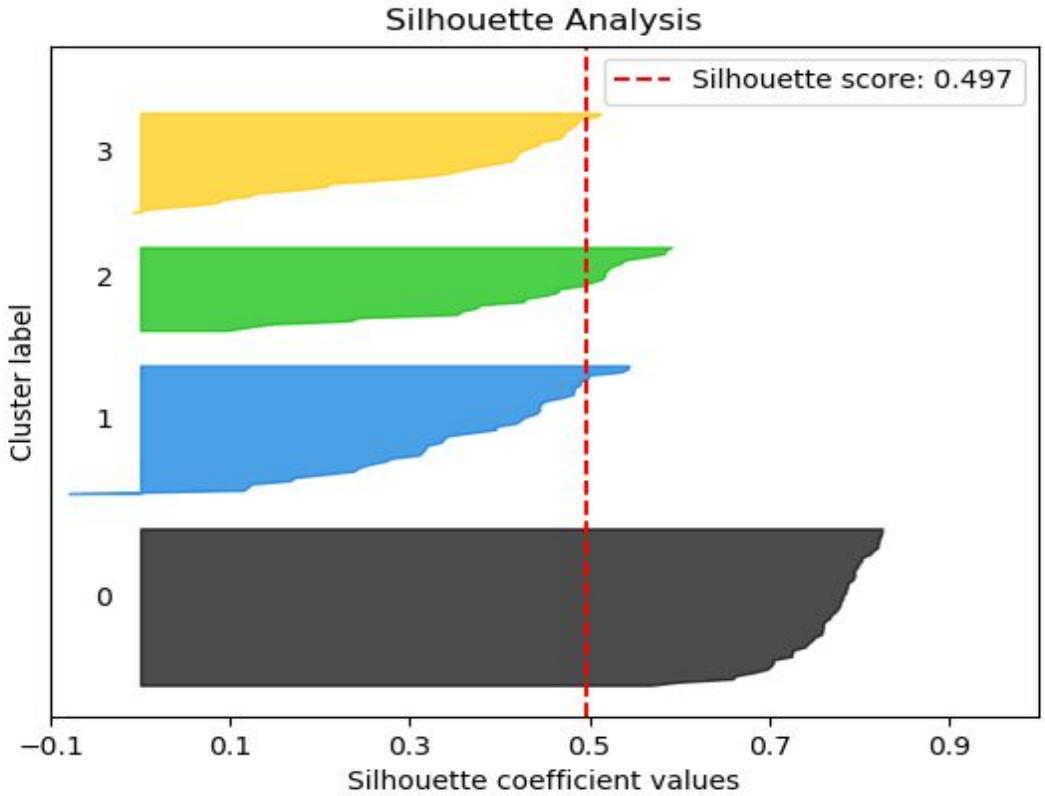
Avaliação do subconjunto de atributos

- K-Means

Como avaliar um bom agrupamento?

- Coeficiente de silhueta
  - Esse coeficiente varia entre [−1.0, 1.0], e valores mais próximos de 1.0 indicam um bom agrupamento.

Coeficiente de silhueta



https://scikit-plot.readthedocs.io/en/stable/metrics.html acessado em 04/09/2018

Função de avaliação

- C_{sol} (c) = silhueta(k-means(c)) − penalidade(c))
  - penalidade(c) = log(1 + NV(c))
    - NV(c): restrições violadas

Aplicação do trabalho

- Pré-processamento dos dados
  - Algoritmos de clusterização, classificação e etc.

O que já foi e o que está sendo feito

- GRASP (Greedy Randomized Adaptive Search Procedure)
- ILS (Iterated Local Search)
- Comparação do resultado gerado pelo GRASP com método exato

O que será feito

- VNS (Variable Neighbourhood Search)
- GA (Genetic Algorithm)

Cronograma

Mês / Metas	A	B	C	D	E
Agosto	x				
Setembro	x	x	x	x	
Outubro				x	x
Novembro					x

A - GRASP e primeiras comparações
B - Implementação do ILS
C - Implementação do VNS
D - Implementação do GA
E - Comparação dos resultados e redigir artigo

Dúvidas

Contato:

Bruno Conceição do Nascimento

Email: bcnbruno17@gmail.com / b_cnbruno@hotmail.com

Github: github.com/bcnbruno/brunotcc

Possíveis dúvidas

- Custo de inserção
  - $C_{ins}(a) = 10 \cdot var(a) / 1 + NR(a)$

Possíveis dúvidas

- Coeficiente de silhueta
  - (b - a) / max(a, b)