

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?
- A maldição da dimensionalidade
- Diferencial do Trabalho
- O que já foi feito
- Resultados obtidos com a base de vinhos
- 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

Diferencial do Trabalho

- O problema foi modelado como Empacotamento de Conjuntos.
- Restrições criadas a partir da análise da correlação dos atributos.
- Atributos com correlação alta não devem estar juntos na solução.

Diferencial do Trabalho

- Técnica específica para seleção de atributos para clusterização, onde não se tem uma classe associada a cada entrada da base de dados.

O que já foi feito

- GRASP (Greedy Randomized Adaptive Search Procedure)
- ILS (Iterated Local Search)
- VNS (Variable Neighborhood Search)
- GA (Genetic Algorithm)
- Teste na base de dados de vinhos

Base de vinhos

- Obtida em UCI Machine Learning Repository
<https://archive.ics.uci.edu/ml/datasets/wine>
- 12 Atributos
- 179 instâncias
- 161 instâncias (com remoção de outliers)

Atributos da base de vinhos

0 - Malic acid

1 - Ash

2 - Alcalinity of ash

3 - Magnesium

4 - Total phenols

5 - Flavanoids

6 - Nonflavanoid phenols

7 - Proanthocyanins

8 - Color intensity

9 - Hue

10 - OD280/OD315 of diluted wines

11 - Proline

Resultados com a base de vinhos

Exato	GRASP	ILS	VNS	GA
[4, 7, 11] 0.687259	[0, 3, 11] 0.698179	[0, 3, 11] 0.698179	[0, 3, 11] 0.698179	[0, 3, 6] 0.671707
[0, 3, 11] 0.662934	[0, 3, 6] 0.671707	[0, 3, 6] 0.671707	[0, 3, 10] 0.631337	[1, 7, 10] 0.615981
[1, 9, 10] 0.663934	[1, 9, 10] 0.649208	[4, 7, 11] 0.666886	[0, 3, 9] 0.618954	[0, 3, 6, 9, 10] 0.601749
[6, 7, 9] 0.652378	[3, 6, 7] 0.638472	[1, 9, 10] 0.649208	[0, 3, 9, 10] 0.604974	[7, 9, 10] 0.597212
Tempo médio (s)	61	106	33	12

O que será feito

- Comparação dos resultados em bases maiores
- Implementar outra metaheurística
- Escrever artigo

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 escrever artigo

Dúvidas

Contato:

Bruno Conceição do Nascimento

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

Github: github.com/bcnbruno/brunotcc