Universidade: presente!

UFRGS PROPESQ

XXXI SIC

Salão UFRGS 2019 – Conhecimento o Formação o Inovação

21 . 25 . OUTUBRO . CAMPUS DO VALE

Estudo de algoritmos de agrupamento para aplicação em dados petrográficos

Júlia Eidelwein *{jeidelwein@inf.ufrgs.br}*

Introdução

O agrupamento de dados (*clustering*) é uma técnica que busca separar um conjunto de dados em grupos com características semelhantes. Diferentemente das técnicas de classificação, o agrupamento não possui exemplos associados aos dados de treinamento e, por isso, é denominado como um método de aprendizado de máquina não supervisionado.

No campo da Geologia, a separação de *petrofácies* é uma tarefa que exige muita experiência desses profissionais. *Petrofácies* é um padrão de textura, estrutura e composição que se apresenta em amostras de rocha em nível microscópico e que controla a porosidade dessas rochas. Nesse contexto, a aplicação de algoritmos de agrupamento pode ser uma alternativa para automatizar essa tarefa.

Objetivo

Identificar, dentre os algoritmos de *clustering* analisados, qual é o mais apropriado para realizar a separação de *petrofácies*.

Descrição dos dados

O conjunto de dados explorado neste trabalho é constituído por 651 descrições de lâminas de rocha-reservatórios de petróleo coletadas em 6 diferentes bacias sedimentares brasileiras. As instâncias possuem atributos relativos ao percentual de componentes primários, diagenéticos, texturais e de localização observados em 300 pontos de cada lâmina.

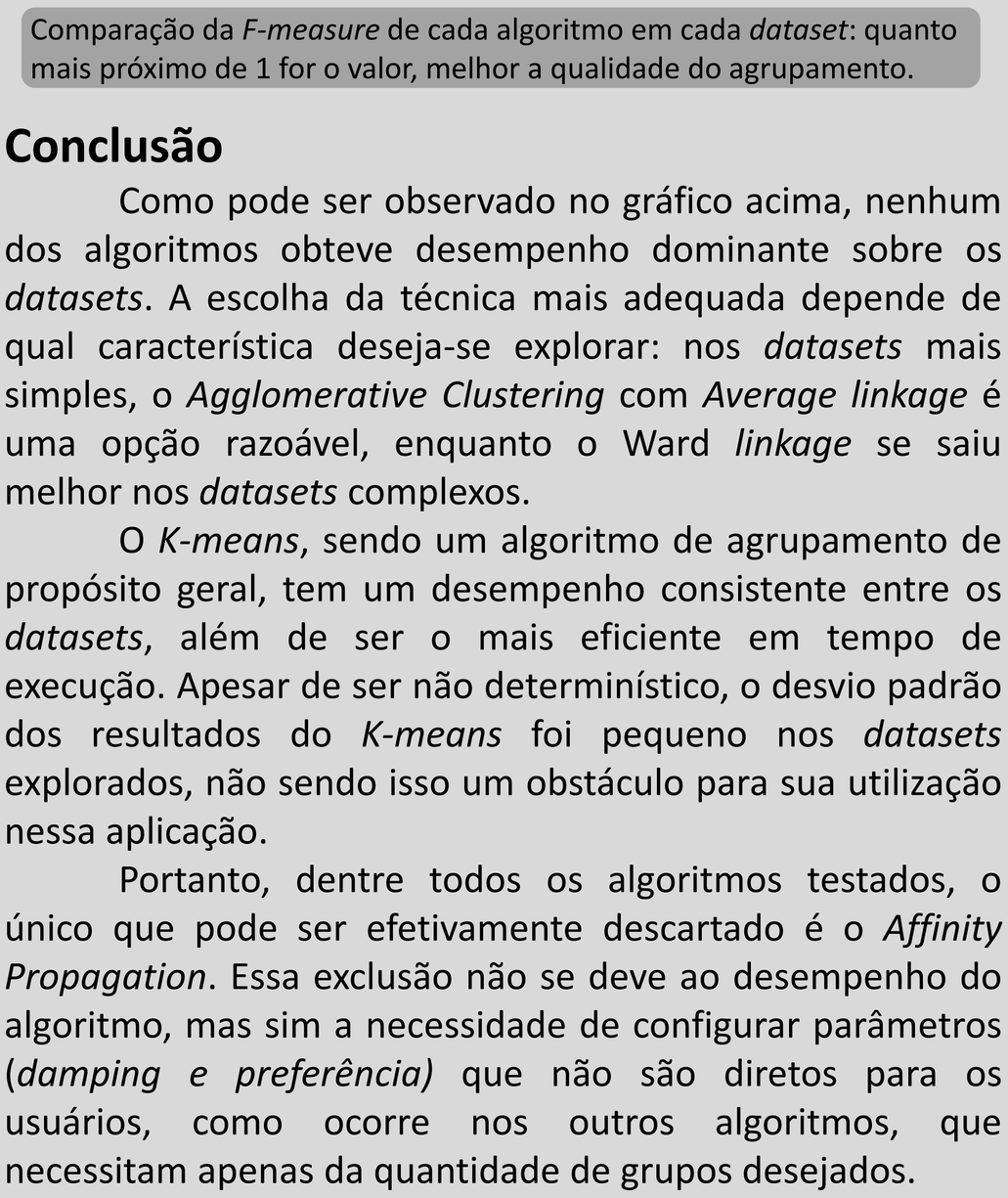
Algoritmos analisados

Foram analisados, no total, três algoritmos de regressão: *K-means*, *Affinity Propagation* e *Agglomerative Clustering*. Foram testadas variações nos parâmetros e os modelos apresentados são as configurações com os melhores resultados.

A comparação entre os algoritmos foi feita por meio da métrica *pairwise F1-measure* e *Adjusted Rand Score*. Esta computa a similaridade entre dois clusters, aquela é uma média harmônica entre a precisão (quantos elementos agrupados no mesmo grupo pertencem, de fato, ao grupo) e o *recall* (quantos elementos de um mesmo grupo foram de fato agrupados juntos).

Dataset	Número de instâncias	Número de clusters*	Número de atributos
A	53	10	45
B	120	17	45
C	66	14	33
D	143	20	49
E	264	14	52
F	61	12	41

Descrição dos dados
*Clusters determinados por um especialista



Comparação da *F-measure* de cada algoritmo em cada *dataset*: quanto mais próximo de 1 for o valor, melhor a qualidade do agrupamento.

Conclusão

Como pode ser observado no gráfico acima, nenhum dos algoritmos obteve desempenho dominante sobre os *datasets*. A escolha da técnica mais adequada depende de qual característica deseja-se explorar: nos *datasets* mais simples, o *Agglomerative Clustering* com *Average linkage* é uma opção razoável, enquanto o Ward *linkage* se saiu melhor nos *datasets* complexos.

O *K-means*, sendo um algoritmo de agrupamento de propósito geral, tem um desempenho consistente entre os *datasets*, além de ser o mais eficiente em tempo de execução. Apesar de ser não determinístico, o desvio padrão dos resultados do *K-means* foi pequeno nos *datasets* explorados, não sendo isso um obstáculo para sua utilização nessa aplicação.

Portanto, dentre todos os algoritmos testados, o único que pode ser efetivamente descartado é o *Affinity Propagation*. Essa exclusão não se deve ao desempenho do algoritmo, mas sim a necessidade de configurar parâmetros (*damping e preferência)* que não são diretos para os usuários, como ocorre nos outros algoritmos, que necessitam apenas da quantidade de grupos desejados.

Instituto de Informática - UFRGS
Av. Bento Gonçalves, 9500.
CEP:91501-97 Porto Alegre - RS

Grupo Banco de Dados Inteligentes
http://www.inf.ufrgs.br/bdi

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico