PUCRS
Pontifícia Universidade Católica do Rio Grande do Sul

Aprendizado Não Supervisionado

Prof. Thomas da Silva Paula

Trabalho II - Comparação de Algoritmos de Agrupamento de Dados

Descrição

É muito comum realizarmos comparação de algoritmos durante o desenvolvimento de soluções que utilizam Machine Learning. No caso de Aprendizado Supervisionado, existem muitas formas de avaliar os resultados, tais como Matriz de Confusão, Medida F1 e Curva ROC. No entanto, a avaliação de algoritmos de Aprendizado Não Supervisionado tem desafios adicionais, pois não temos um objetivo explícito sendo otimizado. Isso torna a avaliação mais complexa e um tanto quanto subjetiva, principalmente no caso de Agrupamento de Dados.

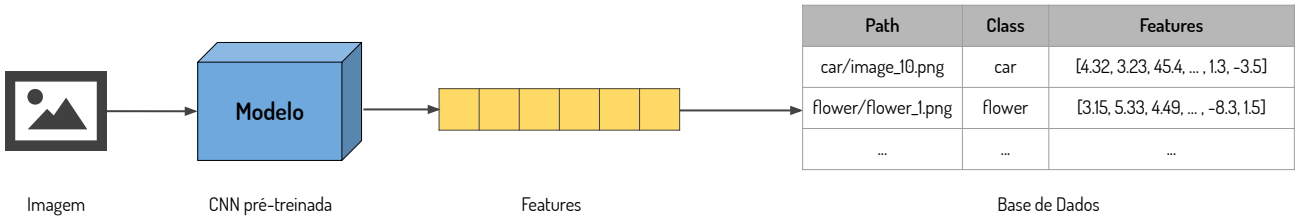
Neste trabalho, iremos exercitar esses conceitos, realizando tarefas muito similares às que Cientistas de Dados fazem no seu dia a dia. Este trabalho consiste em duas partes principais:

1. Aplicação de algoritmos de agrupamento de dados
2. Avaliação do resultado dos agrupamentos de dados

1 Aplicação de Algoritmos de Agrupamento de Dados

1.1 Base de Dados

Iremos realizar o agrupamento de dados em um conjunto de imagens. Para isso, é necessário realizar um processo intermediário para extração das *features* das imagens, com o objetivo de termos uma representação mais compacta (e "significativa") dos dados antes de aplicarmos o agrupamento de dados. A Figura 1 mostra um fluxo de exemplo.



Path	Class	Features
car/image_10.png	car	[4.32, 3.23, 45.4, ... , 1.3, -3.5]
flower/flower_1.png	flower	[3.15, 5.33, 4.49, ... , -8.3, 1.5]
...	...	...

Figura 1: Fluxo de exemplo para extração de features e construção da base de dados.

Como ainda não vimos como fazer este processo, o professor está disponibilizando Numpy arrays com features extraídas com duas redes neurais convolucionais diferentes: *VGG16* e *ResNet18*. Tais features devem ser utilizadas como entrada para os diferentes algoritmos de agrupamento de dados escolhidos.

1.2 Requisitos

a) Utilizar as imagens da base de dados *Natural Images 100*

- Google Drive - Base de Dados

b) Utilizar Numpy arrays disponibilizados (features da VGG16, da ResNet18, caminhos das imagens e rótulos das imagens)

- Google Drive - Numpy Arrays

c) Escolher três algoritmos diferentes de agrupamento de dados (utilizando as bibliotecas *scikit-learn* e *scipy*), para realizar agrupamento de diferentes imagens de uma base de dados.

d) Idealmente (não é obrigatório) os algoritmos devem ser de tipos diferentes (por exemplo, um hierárquico, um particional e um por densidade). Podem ser utilizados algoritmos não vistos em aula, como Mean-Shift, Spectral Clustering, entre outros.

e) Realizar diferentes experimentos com os algoritmos e as features disponibilizadas. Idealmente, os experimentos com as features da VGG16 e da ResNet18 devem ficar em dois notebooks separados.

2 Avaliação do resultado dos agrupamentos de dados

2.1 Requisitos

a) Comparação dos agrupamentos de dados

(a) Quantitativa: Para comparação dos algoritmos, devem ser escolhidos **três** métodos de avaliação diferentes, sendo um *interno*, um *externo* e um *relativo*. O objetivo é avaliar a diferença dos índices para diferentes hiperparâmetros de um mesmo algoritmo e para comparar os resultados de diferentes algoritmos. Lembre-se das diferenças de cada um deles e quando eles se aplicam.

(b) Subjetiva: Para a avaliação subjetiva, deve-se criar uma visualização das imagens dos grupos. Um exemplo pode ser visto na Figura 2.

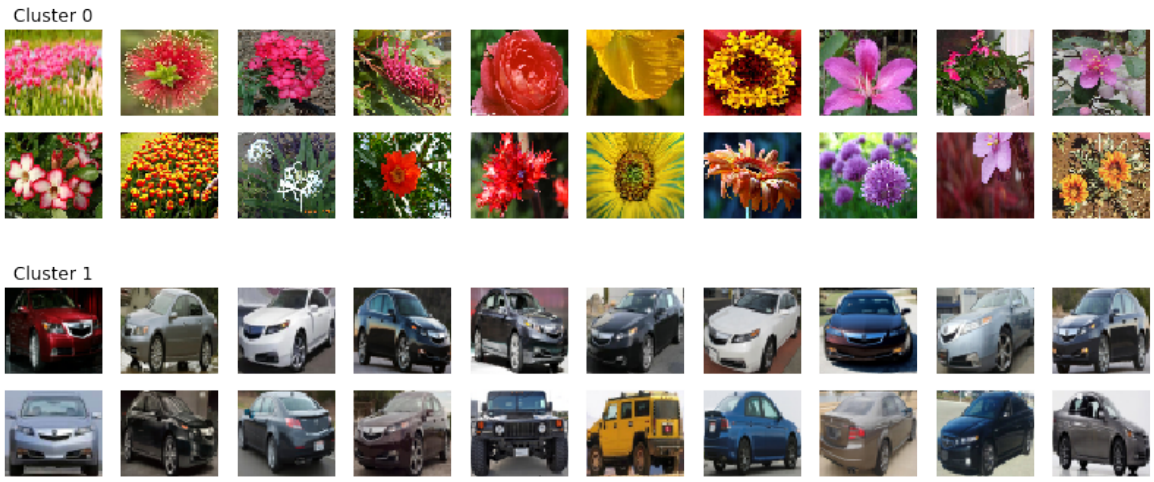


Figura 2: Exemplo de avaliação subjetiva.

b) A partir dos experimentos realizados, deve-se escrever um relatório detalhando as escolhas feitas, os experimentos realizados e, principalmente, o aprendizado. Os diferentes hiperparâmetros utilizados (e.g. ϵ do DBSCAN) em diferentes experimentos não precisam ser todos colocados, porém o racional para ter chegado em tais valores deve ficar claro (o que implica em explicar por onde começou). É importante colocar os desafios e dificuldades enfrentados.

c) **Atenção!** Existe um número "correto"de grupos para essa base de dados, porém a análise não deve se concentrar apenas neste número. Use a criatividade e explore diferentes formas em que os dados podem ser agrupados, fugindo do que seriam os agrupamentos mais "naturais"!

Extra

Caso você se sinta encorajado a explorar um pouco mais, segue uma ideia de experimentos extras que você pode fazer para aprender e exercitar ainda mais seus conhecimentos.

- **Extra 1**: Vamos aprender nas próximas aulas sobre redução de dimensionalidade. Faça experimentos com redução de dimensionalidade e compare com os agrupamentos anteriores.
- **Extra 2**: Foram disponibilizadas também features da ResNet101, uma versão mais poderosa da ResNet. Faça experimentos com as features das ResNet101 e compare com os agrupamentos anteriores da ResNet18 e VGG16.

Entrega Final

- Jupyter notebook com os experimentos realizados (.ipynb e um .html ou .pdf).
  - Fique à vontade caso queira dividir as análises em mais de um notebook.
- Relatório em formato .pdf com os resultados e conclusões. Caso prefira, os resultados e conclusões podem ser colocados diretamente no Jupyter notebook, desde que fique bem explicado e claro.