Métodos de Agrupamento
Aprendizagem de Máquina
2024

André Luiz Brun[1]

[1]Colegiado de Ciência da Computação
Campus de Cascavel - UNIOESTE

Resumo. *Este documento consiste na especificação formal do segundo trabalho da disciplina de Aprendizagem de Máquina (Csc3040) para o ano letivo de 2024. Aqui são apresentadas as atividades a serem desenvolvidas e como cada processo deverá ser realizado. Além disso, o documento contém as informações sobre a data de entrega dos relatórios.*

1. Introdução

O objetivo do segundo trabalho da disciplina consiste em comparar o comportamento, em termos de competência, de métodos de agrupamento baseados em diferentes conceitos sobre uma mesma base de dados. Como critério de avaliação serão computadas as medidas intrínsecas de coesão, separação e coeficiente de silhueta. Além disso, serão avaliadas as métricas (externas) de homogeneidade, completude, entropia e índice randômico para o comportamento de cada método de agrupamento.

Espera-se, através da execução dos experimentos, que cada equipe possa identificar a abordagem que foi mais adequada ao seu conjunto de dados.

2. Implementação

Nesta seção é descrito como cada etapa do desenvolvimento deve ser realizada segundo os conceitos vistos durante a disciplina. Deverão ser implementadas uma estratégia hierárquica (Aglomerativa), uma abordagem baseada em densidade (DBScan) e outra com foco em centralidades (K-means).

2.1. Análise descritiva dos dados

Cada equipe ficará encarregada de uma base de dados distinta. A definição do conjunto alvo é de responsabilidade do próprio grupo. No entanto, conforme explicitado, é necessário que a base contenha uma coluna com os rótulos, identificando quais instâncias pertencem a um mesmo grupo.

2.2. Treinamento e Calibração dos Modelos

Como pretendemos agrupar todas as instâncias do conjunto de entrada em um determinado número de clusters, o processo de ajuste dos parâmetros se dará sobre todo o conjunto de dados, sem a necessidade de separação entre treino, teste e validação. Assim, os valores mais adequados aos parâmetros serão aqueles que possibilitarem a obtenção das melhores métricas.

Uma vez que cada estratégia possui seus próprios parâmetros a serem definidos, na Tabela 1 estão detalhados quais deles devem ser refinados para cada método de agrupamento.

Tabela 1. Conjunto de parâmetros a serem calibrados

Método	Parâmetros
K-means	n_clusters max_iter
DBScan	eps min_samples
AGNES (Aglomerative)	n_clusters linkage

2.3. Avaliação dos Modelos

Definidos os melhores parâmetros para cada método de clustering, o passo seguinte consiste em estimar as medidas de avaliação especificadas. Para cada uma das três abordagens deverão ser obtidas as métricas descritas a seguir:

- Intrínsecas
  - Coesão
  - Separação
  - Coeficiente de Silhueta Médio
- Extrínsecas
  - Homogeneidade
  - Índice Randômico
  - Completude
  - Entropia

2.4. Análise Comparativa

Esta etapa consiste na comparação das medidas de desempenho dos métodos para descobrir qual deles se sobressaiu.

De acordo com o comportamento dos dados de entrada ou dos parâmetros dos modelos, quais seriam as métricas mais interessantes para seu trabalho?

Dentre as métricas escolhidas, comparando-se os modelos de agrupamento, qual seria o mais indicado? Por que razão?

2.5. Análise dos Grupos Formados

A última etapa do trabalho envolve uma análise aprofundada dos grupos formados. É preciso compreender quem são as instâncias que compõe cada cluster identificando o que elas têm em comum e o que as diferem dos outros grupos.

Além disso, é necessário entender quais atributos foram mais interessantes para a separação dos clusters e quais praticamente não contribuíram para a identificação dos grupos.

2.6. Como fazer?

A linguagem adotada é de escolha da dupla. Entretanto, é fortemente indicado o uso de Python ou Java.

Não é necessário implementar os métodos de agrupamento. Neste caso, pode-se e é indicado, que sejam utilizadas implementações prontas dos métodos, ficando a carga da equipe apenas a implementação do framework e análise dos parâmetros e resultados.

A seguir constam algumas sugestões de repositórios em que estão disponíveis conjuntos de dados para serem empregados neste trabalho.

Não devem ser usadas bases já empregadas no primeiro trabalho. Todas as equipes devem selecionar conjuntos que ainda não foram usados.

- UCI Machine Learning Repository
- Kaggle
- Awesome Public Datasets Collection
- Harvard Dataverse
- Microsoft Datasets
- Amazon Datasets

3. O que deve ser entregue

3.1. Relatório

Deve ser elaborado um relatório técnico em formato pdf contendo:

- Detalhamento de quais foram os parâmetros empregados em cada método de clustering e em qual faixa de valores cada parâmetro foi variado. Por exemplo, no K-means, seria possível variar o valor de k entre 1 e 5.
- Análise detalhada das métricas de desempenho (internas e externas) obtidas para cada modelo.
- Análise pertinente indicando quais métricas melhor representam o desempenho dos algoritmos perante o conjunto de entrada.
- Comparação adequada e embasada das três estratégias de agrupamento testadas.
- Análise dos clusters formados e *insights* sobre observações pertinentes.

O formato do relatório deve ser a formatação presente neste texto. As regras para tal podem ser obtidas no link download. No arquivo disponível pode-se utilizar a formatação em arquivo .doc ou em latex.

3.2. Código-fonte

Além do relatório citado, cada equipe deverá enviar os códigos-fontes construídos para a execução dos experimentos e o conjunto de dados empregado. Os arquivos podem ser compactados e enviados como arquivo único.

4. Para quando?

O trabalho deverá ser submetido no link disponibilizado na turma de disciplina dentro do ambiente Microsoft Teams até as **23:59 do dia 30/09/2024**.

Para este trabalho as equipes não apresentarão os resultados alcançados.