Especificação do Projeto

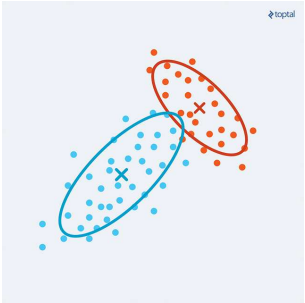
Teoria e Aplicação de Grafos - 2017/1

Prof. Dr. Vinicius Ruela Pereira Borges
viniciusrpb@unb.br

Objetivo

Dado um conjunto de dados, deve-se realizar um agrupamento utilizando uma técnica de agrupamento espectral. Especificamente, deve-se agrupar os registros (instâncias) de um conjunto de dados de acordo com os atributos que os descrevem. Espera-se que instâncias similares em relação aos seus atributos pertençam ao mesmo grupo, enquanto que instâncias bem diferentes entre si estejam em grupos distintos.

A imagem abaixo apresenta um gráfico de dispersão ilustrando um agrupamento de um conjunto de dados em dois grupos, representados pelas cores vermelho e azul. Os símbolos circulam representam cada instância do conjunto, enquanto que os símbolos "×" são os centróides de cada grupo, isto é, representantes dos grupos calculados como sendo a média dos valores dos atributos das instâncias de cada grupo.



Tarefas

As tarefas a serem seguidas neste projeto são basicamente descritas abaixo:

1. Escolher um conjunto de dados
2. Preparar o conjunto de dados
3. Agrupar as instâncias por meio de uma abordagem espectral
4. Apresentar os resultados da aplicação da técnica de agrupamento no conjunto de dados

1. Escolher um conjunto de dados

Escolha um conjunto de dados no site UCI Machile Learning Repository: `https://archive.ics.uci.edu/ml/datasets.html`. Cada conjunto de dados descreve vários registros considerando um determinado domínio do conhecimento. Por exemplo, o conjunto de dados Íris possui 150 registros de 3 espécies de flores (*Iris Setosa*, *Iris Versicolour*, *Iris Virginica*), sendo que 4 atributos descrevem cada um dos registros: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala, todos medidos em centímetros. O último atributo, também denominado `atributo classe`, apenas identifica a espécie da flor para cada instância.

2. Preparar o conjunto de dados

A coleta de dados muitas vezes não é feita de maneira apropriada. Por exemplo, quando cria-se um conjunto de dados em que existem os atributos "gênero" e "Está grávida?" e existe uma instância cujo gênero é masculino, o valor para o atributo "Está grávida?" pode ser inconsistente. Outro exemplo seria um atributo que descreve o peso de uma pessoa, que pode aparecer um valor negativo devido às falhas em sensores ou erros no processo de coleta.

Por isso, nesta etapa, caso o conjunto de dados apresente instâncias com valores inconsistentes em relação ao seu atributo, ou então valores ausentes, pode-se optar por duas soluções: ou remove-se a instância ou seu valor é estimado por meio da média dentre todos os valores do atributo.

Outro "problema" se refere à presença de atributos nominais no conjunto de dados. O uso de desses atributos juntamente com os atributos numéricos exigiriam a elaboração de medidas de dissimilaridade mais complexas para viabilizar a comparação de instâncias. Além disso, a construção do grafo na técnica de agrupamento espectral demanda que o conjunto de dados esteja descrito apenas por atributos numéricos ou binários. Por exemplo, como comparar pessoas de acordo com um atributo nominal que descreve a cor do cabelo, sendo seus valores: *branco*, *castanho*, *loiro*, *ruivo*? Como obter um valor que expresse a dissimilaridade entre eles? Por isso, é mais indicado converter este atributo para nominal, fazendo-se a seguinte associação: valor 0 é *branco*, valor 1 é *castanho*, valor 2 é *loiro* e valor 3 é *ruivo*. Mais informações podem ser encontradas no livro mencionado neste endereço: `http://www-users.cs.umn.edu/~kumar/dmbook/index.php`. No livro, as seções 2.1, 2.2 e 2.3 podem esclarecer caso seja necessário pré-processar os dados.

3. Agrupamento espectral

Para conhecer melhor sobre técnicas de agrupamento espectral, leia o artigo:

NG, Andrew Y.; JORDAN, Michael I.; WEISS, Yair, On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems, pp. 849-856, 2002

Neste artigo (página 2), existe um pseudo algoritmo que detalha as etapas para se realizar um agrupamento espectral.

Pode-se utilizar bibliotecas para calcular autovalores e autovetores de matrizes Laplacianas, como por exemplo a técnica Singular Value Decomposition[1].

4. Resultados experimentais

Seja K um número de grupos criados na etapa anterior e o conjunto de dados $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x_N}\}$ em que N é o número de instâncias. Utilizando-se os agrupamentos obtidos na etapa anterior, deve-se gerar como saída, a atribuição feita de um grupo para cada instância, conforme mostra o exemplo abaixo

- Instância $\mathbf{x_1} \rightarrow$ pertence ao agrupamento 2
- Instância $\mathbf{x_2} \rightarrow$ pertence ao agrupamento K
- Instância $\mathbf{x_3} \rightarrow$ pertence ao agrupamento 3
- Instância $\mathbf{x_4} \rightarrow$ pertence ao agrupamento 1

[1] `http://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm`

- Instância $\mathbf{x_5}$ → pertence ao agrupamento 1
- ...
- Instância $\mathbf{x_N}$ → pertence ao agrupamento 3

Sugere-se também gerar informações mais resumidas à respeito dos agrupamentos obtidos, como por exemplo, o número de instâncias

Como proceder

Realizar o trabalho em grupos de **no máximo 3 alunos**. Enviar email para viniciusrpb@unb.br informando o nome de todos os integrantes e o conjunto de dados escolhido. Não será permitido que dois grupos realizem o trabalho utilizando o mesmo conjunto de dados. Por isso, o grupo que informar ao professor o conjunto de dados escolhido com maior antecedência terá prioridade.

Apresentações

As apresentações ocorrerão nos dias 04 e 06 de julho de 2017, no horário e local das aulas de Teoria e Aplicação de Grafos. O professor se reserva ao direito de arguir o grupo caso julgue necessário. Todos deverão estar presentes na apresentação, pois 50% da nota é individual.

Observações

1. Trabalhos plagiados terão nota atribuída em zero.
2. O professor poderá atualizar a especificação deste trabalho caso seja necessário. Desta maneira, em caso de atualização, será inserido no cabeçalho desta especificação a última data de atualização.
3. A nota deste trabalho substituirá a nota da Prova 2.