Trabalho de Estatística Aplicada II

Matheus Machado de Almeida, Leandro Dias Gomes de Carvalho, Maqueise de Medeiros Pinheiro, Caio Macedo Alves

08/04/2021

Análise de Cluster

são técnicas estatísticas utilizada com objetivo de agrupar elementos, de forma que haja o mínimo de diferença entre os elementos de um mesmo grupo e o máximo de diferença entre elementos de grupos diferentes. O grau de diferença entre cada elemento é medido pela distância entre os mesmos, onde as noções de distância variam de situalçao para situação.

Quando usar?

- Quando a preocupação principal é dividir os elementos em grupos, de forma que os elementos de um mesmo grupo sejam homogêneos e os elementos em grupos diferentes sejam heterogêneos.
- Considerações teóricas, conceituais e práticas devem ser observadas ao selecionar variáveis para a Análise de Agrupamentos.
- Como medir similaridades entre indivíduos?
- Como agrupar indivíduos semelhantes?

Objetivos:

essa técnica pode ser usada para dividir e classificar os elementos entre dois ou mais grupos levando em consideração sua similiridade a paritr das variáveis disponíveis, nesse caso, o algoritmo encontra diretamente uma partição de n itens em k clusters usando dois critérios, semelhança interna e isolamento dos clusters formados. Esse tipo de técnica também pode ser usado para simplificar os dados coletados e identificar relações entre os elementos e variáveis.

Desvantagens:

a análise é descritiva, não-teórica, e não inferencial. Divide os elementos em grupos mesmo que não haja qualquer direfença significativa entre os elementos dados. Não é generalizável, pois depende das variáveis usadas como base na medição de similaridade.

Técnicas não-hierárquicas de agrupamento

Existem dois tipos diferentes de técnicas de agrupamento, as hierárquicas e as não hierárquicas, e é importante notar suas diferenças. Os algoritmos das técnicas não-hierárquicas possuem maior capacidade de análise de conjunto de dados além de serem iterativos. A quantidade de número de clusters precisa ser pré-definida. Em cada etapa novos clusters podem ser formados por divisão ou junção de clusters inicialmente difinidos.

K-Médias

Essa técnica consiste em agrupar cada elemento para o cluster cuja média estiver mais próxima.

Etapas: 1. escolher k médias (centróides) para iniciar o processo de partição; 2. comparar cada item com o centroide inicial por uma distância (exemplo: euclidiana). Os itens são alocados ao cluster mais próximo; 3. após a alocação dos n itens, recalcular os centroides para cada novo cluster formado, repetindo o passo (2) com estes novos centróides. 4. repetir os passos (2) e (3) até que todos os elementos estejam bem alocados em seus grupos.

Em k-médias, a escolha das sementes iniciais influencia na partição final. Assim, seguem algumas sugestões para essa escolha:

- Sugestão 1: Use alguma técnica hierárquica para obter os k clusters iniciais. Calcule o vetor de médias de cada grupo, as sementes iniciais.
- Sugestão 2: escolha aletoriamente os k centroides iniciais. Selecione m amostras aleatórias com k centroides e repetir a amostragem m vezes e, no final, calcula-se os m centroides para cada grupo.
- Sugestão 3: Escolha a variável de maior variância. Em seguida, divida o domínio da variável em k intervalos. A semente inicial será o centroide de cada intervalo.
- Sugestão 4: Escolha os k outliers identificados, que serão as sementes iniciais.
- Sugestão 5: Escolha prefixada (ad hoc) – não muito recomendável.
- Sugestão 6: selecione os k primeiros valores do banco de dados. Grande parte dos softwares usa como padrão esta sugestão para atribuir as sementes iniciais. Fornece bons resultados quando os itens são bem discrepantes entre si. Logo, não é recomendável quando os elementos são bem semelhantes.

Mingoti (2005, p.194) aponta que a solução da k—médias, utilizando como sementes iniciais a técnica de Ward, gera melhores resultados que a solução e k-médias, usando os quatro primeiros valores

Fuzzy C-Médias

- Técnica iterativa e exige a definição inicial de k clusters. Sendo n itens e p variáveis aleatórias, busca-se a partição que minimiza:
- A função é minimizada quando as probabilidades:
- Para ter solução final, deve-se ter os centroides e probabilidades iniciais, _____, geradas de uma distribuição uniforme [0,1].
- Os centroides se modificam a cada iteração e o processo cessa quando a distância entre os centroides dos 2 últimos passos é: _____ .
- Nessa técnica, a partição final alocará os itens nos clusters conforme a sua maior probabilidade, o que torna possível identificar os itens que se assemelham a mais de um cluster.
- Em oposição, a técnica de k-Médias gera uma partição na qual cada elemento pertence a um único cluster.

Comentários

- Tais técnicas são também sensíveis às escalas e aos outliers. As variáveis de maior dispersão dominam na distância euclidiana.
- Pode-se padronizar as variáveis ou usar distâncias ponderadas.
- Há razões para não fixar o n° de clusters, como nessas técnicas:
- Se 2 ou mais sementes estão dentro de um cluster, os clusters resultantes serão pobremente diferenciados;
- Outliers => pelo menos 1 cluster com itens muito dispersos.
- Mesmo que saiba os itens nos k clusters, perde-se grupos raros e latentes na amostra. Os k grupos iniciais => partição sem sentido.
- Comparando às técnicas, pode-se afirmar:
- Quando os grupos estão bem separados, qualquer técnica leva a resultados satisfatórios;

- Quando há interseção inicial entre os grupos, Fuzzy é melhor por gerar a probabilidade dos itens;
- Para definir o n° final de grupos, pode aplicar bootstrap a fim de delinear um intervalo de confiança.