PUC-MG - Pontifícia Universidade Católica de Minas Gerais
Curso : Ciência da Computação
Disciplina : Inteligência Artificial
Professora : Cristiane Neri Nobre
Alunos : Filipe Arthur Ferreira Silva, Henrique Augusto Rodrigues

Lista de exercícios 06

Questão 1 -

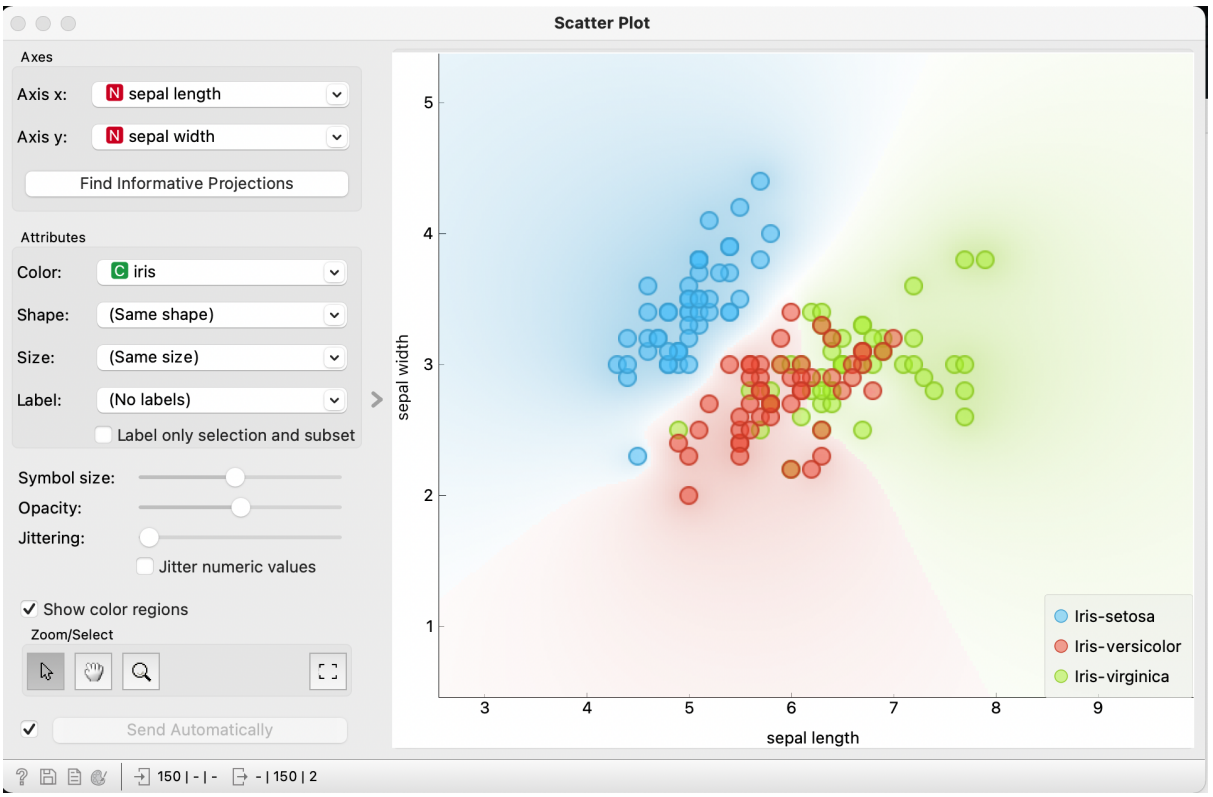
Distância Ex3 = 2.0
Distância Ex4 = 2.0

Questão 2 -

- **Índice de silhueta (Silhouette Index) :** Avalia a separação dos clusters a partir da diferença entre a distância média entre pontos de clusters diferentes (tanto entre pontos do mesmo cluster quanto do outro cluster), mostrando assim quais pontos estão agrupados de forma equivocada.
- **Índice de Dunn (Dunn Index) :** Mostra quão compacto determinado cluster é a partir da divisão da distância mínima entre clusters pelo tamanho do cluster observado. Quanto maior o resultado, melhor agrupado o cluster está.

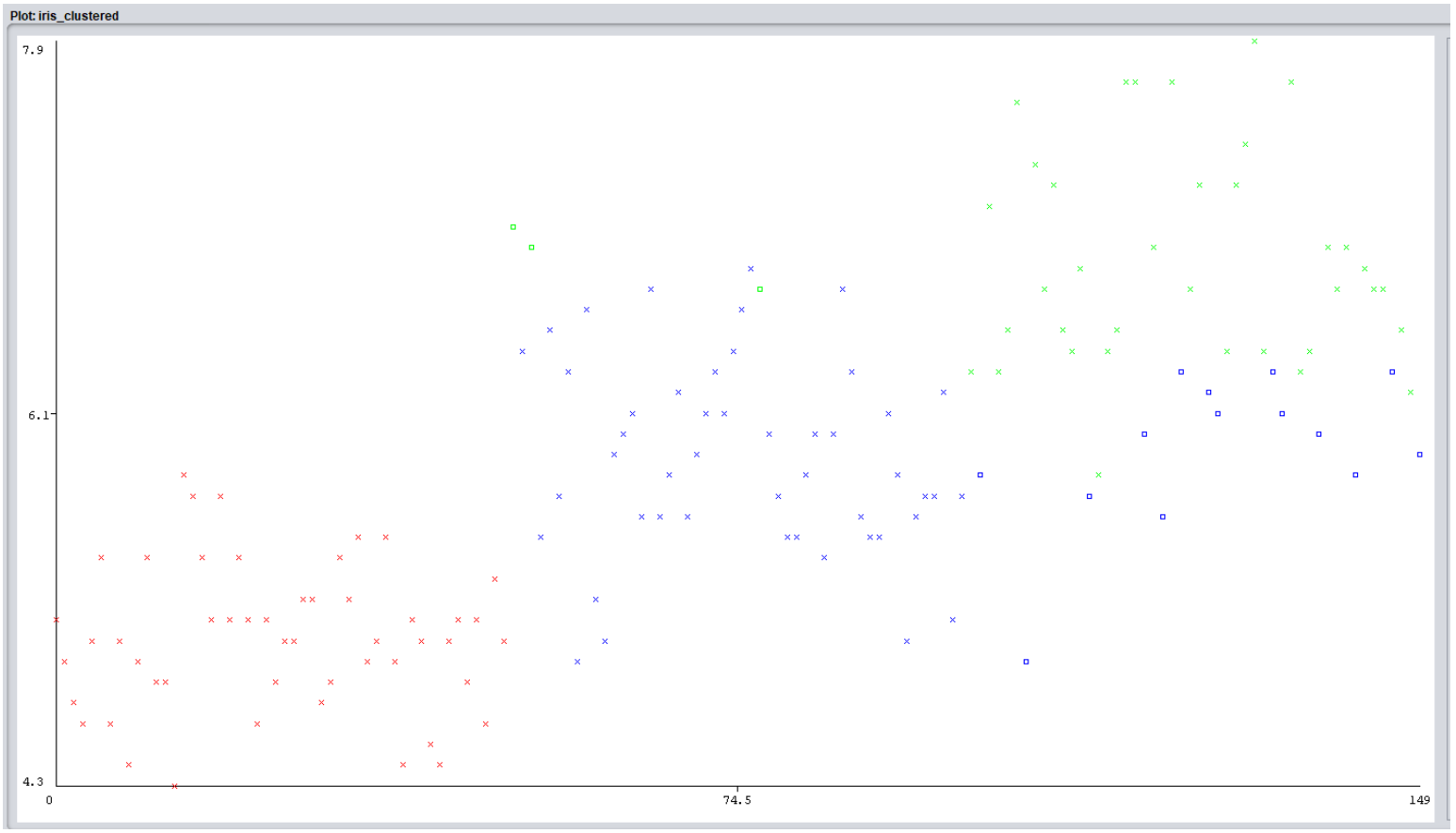
Questão 3 -

a)



Iris-Versicolor ocupando a mesma região que a maioria da Iris-virgínica. O conjunto (K) não está bem segmentado.

b)



Vale ressaltar que os clusters não possuem o mesmo número de instâncias: um é consideravelmente maior que os outros. Possivelmente pela falta de algum processo de balanceamento.

Questão 4 -

O k-means se aplica a tuplas de dados completamente numéricos, funciona bem compactos, clusters esféricos, mas, falha em outras formas.

Quando há K clusters, a probabilidade de selecionar um centroide é pequena, exemplo, K = n, n!/nˆn

Questão 5 -

O clustering, considerado a questão mais importante do aprendizado não supervisionado, trata da partição da estrutura de dados em área desconhecida e é a base para o aprendizado posterior.

A definição completa para clustering, no entanto, não chega a um acordo, e um clássico é descrito a seguir: As instâncias, no mesmo cluster, devem ser semelhantes tanto quanto possível;

Processo padrão de armazenamento em cluster pode ser dividido nas seguintes etapas: Extração e seleção de recursos: extraia e selecione os recursos mais representativos do conjunto de dados original;

Os indicadores de avaliação podem ser divididos em duas categorias, os indicadores de avaliação interna e os indicadores de avaliação externa, em termos dos dados de teste, seja no processo de construção do algoritmo de agrupamento.

A ideia central do K-means é atualizar o centro do cluster que é representado pelo centro dos pontos de dados, por computação iterativa e o processo iterativo irá continuar até que alguns critérios de convergência sejam atendidos.

K- mediods é uma melhoria do K-means para lidar com dados discretos, o que leva o ponto de dados, mais próximo ao centro dos pontos de dados, como o representante do cluster correspondente.

Suponha que cada ponto de dados represente um cluster individual no início e, em seguida, os dois clusters mais vizinhos são mesclados em um novo cluster até que haja apenas um cluster restante.

O Camaleão, princípio, dividir os dados originais em clusters de menor tamanho com base no gráfico do vizinho mais próximo e, em seguida, os clusters de menor tamanho são mesclados em um cluster de maior tamanho, com base no algoritmo de aglomeração, até que sejam apresentados.

O FCS, diferente dos algoritmos tradicionais de agrupamento fuzzy, toma a hiper-esfera multidimensional como protótipo de cada agrupamento, de modo a se agrupar com a função de distância baseada na hiper-esfera.

A distância entre um cluster e seu ponto de dados mais próximo satisfaz a distribuição da distância esperada que é gerada a partir dos pontos de dados existentes desse cluster, o ponto de dados mais próximo deve pertencem a este cluster.

A ideia central do GMM é que o GMM consiste em várias distribuições Gaussianas a partir das quais os dados originais são gerados e os dados, obedecendo à mesma distribuição Gaussiana independente, são considerados pertencentes ao mesmo cluster.

A ideia básica deste tipo de algoritmos de agrupamento é que os dados que estão na região com alta densidade do espaço de dados sejam considerados pertencem ao mesmo cluster.

No processo de deslocamento médio, a média de deslocamento do ponto de dados atual é calculada primeiro, o próximo ponto de dados é calculado com base no ponto de dados atual e o deslocamento e, por último, a iteração será continuada até alguns critérios são atendidas.

Desvantagens: resultando em um resultado de clustering com baixa qualidade quando a densidade do espaço de dados não é uniforme, uma memória com grande tamanho necessária quando o volume de dados é grande e o resultado de clustering altamente sensível aos parâmetros;

De acordo com esse tipo de algoritmo de agrupamento, o agrupamento é realizado no gráfico onde o nó é considerado o ponto de dados e a borda é considerada a relação entre os pontos de dados.

A ideia central do STING, que pode ser usado para processamento paralelo é que o espaço de dados é dividido em muitas unidades retangulares pela construção da estrutura hierárquica e os dados em diferentes níveis de estrutura são agrupados, respectivamente.

Desvantagens: o resultado do clustering sensível à granularidade (o tamanho da malha), a alta eficiência de cálculo ao custo de reduzir a qualidade dos clusters e diminuir a precisão do clustering;

O algoritmo típico desse tipo de agrupamento é o FC, cuja ideia central é que a alteração de quaisquer dados internos de um agrupamento não tem nenhuma influência na qualidade intrínseca da dimensão fractal.A complexidade de tempo de FC é O (n);

A ideia central do COBWEB é construir uma classificação árvore, com base em alguns critérios heurísticos, a fim de realizar agrupamento hierárquico no pressuposto de que a distribuição de probabilidade de cada atributo é independente.

A ideia central do SOM é construir um mapeamento de redução de dimensão do espaço de entrada de alta dimensão para o espaço de saída de baixa dimensão na suposição que existe topologia nos dados de entrada.

A ideia básica deste tipo de algoritmos de agrupamento é que os dados no espaço de entrada são transformados no espaço de características de alta dimensão pelo mapeamento não linear para a análise de agrupamento.

A ideia básica do kernel K-means, kernel SOM e kernel FCM é tirar vantagem do método do kernel e do algoritmo de agrupamento original, transformando os dados originais em um espaço de recursos de alta dimensão por função de kernel não linear para realizar o algoritmo de agrupamento original.

A ideia central do SVC é encontrar a esfera com o raio mínimo que pode cobrir todos os pontos de dados no espaço de características de alta dimensão e,

em seguida, mapear a esfera de volta aos dados originais espaço para formar a isolinha, ou seja, a borda dos clusters, cobrindo os dados, e os dados na isolinha fechada devem pertencer ao mesmo cluster.

Vantagens: mais fácil de agrupar no espaço de recursos de alta dimensão, adequado para dados com forma arbitrária, capaz de analisar o ruído e separar os clusters sobrepostos, e não é necessário ter o conhecimento preliminar sobre a topologia dos dados;

O algoritmo de clustering baseado em ensemble também é chamado de ensemble clustering, cuja ideia central é gerar um conjunto de resultados de clustering inicial por um método particular e o resultado final do clustering é obtido integrando os resultados do clustering inicial.

Existem principalmente 4 tipos de métodos para obter o conjunto de resultados de agrupamento inicial da seguinte forma: Para o mesmo conjunto de dados, empregue o mesmo algoritmo com os diferentes parâmetros ou as diferentes condições iniciais;

A ideia central do LF, o algoritmo típico do ACO_based, é que os dados são distribuídos aleatoriamente na grelha de duas dimensões primeiro, depois os dados são seleccionados ou não para operação posterior com base na decisão de uma formiga e este processo é iterado até se obter um resultado de agregação satisfatório.

Os clusters iniciais de partículas são obtidos primeiro pelo outro algoritmo de clustering, depois os clusters de partículas são actualizados continuamente com base no centro dos clusters e na localização e velocidade de cada partícula, até se obter um resultado satisfatório de clustering.

A ideia central dos algoritmos baseados em SFLA_ é simular a interacção de informação dos sapos e tirar partido da pesquisa local e da interacção de informação global.

A ideia central dos algoritmos baseados em ABC_ é simular o comportamento forrageiro de três tipos de abelhas, dos quais o dever é determinar a fonte alimentar, numa população de abelhas e fazer uso da troca de informação local e de informação global para o agrupamento.

O algoritmo de agrupamento baseado na teoria quântica chama-se agrupamento quântico, do qual a ideia básica é estudar a lei de distribuição de dados de amostras no espaço de escala através do estudo da lei de distribuição de partículas no campo energético.

A ideia central de QC (aglomeração quântica), adequada para dados de alta dimensão, é obter a energia potencial de cada objecto pela Equação de Schrodinger utilizando o algoritmo de descida de gradiente iterativo, considerar o objecto com baixa energia

potencial como o centro do aglomerado, e colocar os objectos em diferentes aglomerados através da função de distância definida.

DQC, uma melhoria do QC, adopta a Equação de Schrodinger baseada no tempo para estudar a mudança do conjunto de dados original e a estrutura da função de energia potencial quântica de forma dinâmica.

A ideia básica deste tipo de algoritmos de agrupamento é considerar o objecto como o vértice e a semelhança entre objectos como a borda ponderada, a fim de transformar o problema de agrupamento num problema de partição gráfica.

E a chave é encontrar um método de partição gráfica que torne o peso da ligação entre diferentes grupos tão pequeno quanto possível e o peso total da ligação entre as arestas dentro do mesmo grupo tão alto quanto possível.

Os algoritmos típicos deste tipo de agrupamento podem ser divididos principalmente em duas categorias, espectral recursivo e espectral multidireccional e os algoritmos típicos destas duas categorias são SM e NJW respectivamente.

E a NJW efectua a análise de agrupamento no espaço de características construído pelos vectores próprios correspondentes aos k maiores valores próprios da matriz Laplaciana.

A ideia central da AP é considerar todos os pontos de dados como potenciais centros de aglomeração e o valor negativo da distância euclidiana entre dois pontos de dados como a afinidade.

Os dados espaciais referem-se aos dados com as duas dimensões, tempo e espaço, ao mesmo tempo, partilhando as características de grande escala, alta velocidade e complexo em informação.

DenStream, que toma a ideia central do algoritmo de clustering baseado na densidade, é adequado para o conjunto de dados não convexo e pode lidar com os outliers de forma eficiente, em comparação com os algoritmos mencionados acima nesta secção.

As definições básicas de agrupamento e o procedimento típico, enumeram as funções de distância (dissimilaridade) comumente utilizadas, funções de semelhança, e indicadores de avaliação que estabelecem as bases do agrupamento, e analisa os algoritmos de agrupamento a partir de duas perspectivas, os tradicionais que contêm 9 categorias incluindo 26 algoritmos e os modernos que contêm 10 categorias incluindo 45 algoritmos.

É difícil apresentar uma lista completa de todos os algoritmos de agrupamento devido à diversidade da informação, à intersecção de campos de investigação e ao desenvolvimento da moderna tecnologia informática.

Assim, são seleccionadas 19 categorias dos algoritmos de agregação comummente utilizados, com elevado valor prático e bem estudados, e um ou vários algoritmos típicos de cada categoria é(são) discutido(s) em detalhe de modo a dar aos leitores uma visão sistemática e clara do importante método de análise de dados, a agregação.