Leonardo Santos Miranda - Mestrado

1. A independência entre os atributos de uma base de dados é um requisito fundamental para a aplicação de várias técnicas de mineração de dados. Entretanto, nem sempre é possível alcançar essa independência, por exemplo usando métodos de redução de dimensionalidade como PCA e SVD, sem que haja uma perda significativa de informação. Métodos kernel, por outro lado, promovem um aumento na dimensionalidade, frequentemente recorrendo a mapeamentos não lineares. Discuta quais critérios e princípios utilizar para construir estratégias que considerem ambas as direções, nominalmente métodos kernel e estratégias de redução de dimensionalidade, na preparação de dados previamente à aplicação de técnicas. Apresente exemplos, mesmo que hipotéticos, considerando características dos dados, das técnicas e das tarefas a serem realizadas.

R: O método kernel é comumente utilizado em dados não linearmente separáveis e promove uma transformação (aumento de dimensionalidade) nesses dados de tal maneira que técnicas da álgebra linear possam ser aplicadas para as tarefas em cima desses dados. O método kernel trabalha com a informação que expressa as relações entre os pontos em um espaço de característica e representa os objetos de entrada em um mapeamento do espaço em uma matriz de similaridade n por n. Por outro lado, técnicas de redução de dimensionalidade como o PCA procura uma r-dimensão base que melhor captura a variância dos dados, sendo a direção que maximiza a variância é também aquela que minimiza a média quadrática do erro. Dito isso, o principal critério antes de utilizar qualquer uma das duas técnicas é a análise dos seus dados, além das tarefas que desejam realizar em cima deles, por exemplo, caso queira representar dados em altas dimensionalidades (tipos de diabetes, por exemplo) a utilização da técnica de mapas não lineares pode ser a melhor estratégia. De outra forma, caso os dados de entrada estejam em altas dimensionalidade (vetores de palavras, por exemplo), a técnica de redução de dimensionalidade pode ser a melhor opção a ser tomada.

2. Você é cientista de dados de uma empresa de auditoria e tem por tarefa identificar comportamentos anômalos de funcionários de uma instituição financeira ao longo do tempo. Você tem informações sobre a natureza e magnitude das atividades, assim como o momento que elas ocorreram, os funcionários envolvidos e, se for o caso, entidades externas. Discuta como você poderia utilizar técnicas de mineração de padrões frequentes, sejam eles conjuntos, sequências ou grafos. Explicite os padrões de interesse à luz do seu entendimento dos dados mencionados como disponíveis e métricas de interesse a serem utilizadas na análise demandada para priorizar os padrões mais relevantes.

R: Para a identificação de anormalidades eu primeiramente utilizaria a técnica de mineração de conjuntos frequentes, ajustando o suporte mínimo e analisando os conjuntos mais frequentes até que eu identifique todos os conjuntos possíveis. Após isso, eu analisaria os conjuntos obtidos e iria verificar quais são os conjuntos infrequentes (aqueles que não atingiram o suporte mínimo), analisando cada funcionários destes conjuntos não frequentes e obtendo a resposta de qual funcionário possui comportamentos anômalos aos demais.

3. Algoritmos de mineração de grafos são um desafio tanto do ponto de vista de modelagem, que pode ser muito variada, quanto do ponto de vista computacional, tendo em vista a explosão combinatória em termos de padrões minerados. Discuta como você estenderia o algoritmo GSpan para minerar subgrafos maximais, ou seja, subgrafos que sejam os maiores possíveis entre os subgrafos frequentes. Quais operações do gSpan continuariam a ser realizadas e quais teriam que ser alteradas tendo em vista o novo padrão a ser minerado? Com relação à complexidade computacional do algoritmo proposto, ela seria maior, menor ou não se alteraria? Por que?

R: O algoritmo GSpan pode ser estendido de tal forma que os subgrafos sejam os mais possíveis dentre os subgrafos frequentes modificando a função que verifica a canonização mas isso pode afetar a eficiência da canonização, além de ter um custo computacional mais alto por realizar um pós processamento na enumeração dos múltiplos vértices de padrões hierárquicos.

4. Algoritmos baseados em representantes (k-means e EM) e baseados em densidade (DBSCAN e DenClue) têm dificuldades em lidar com agrupamentos de densidade variável. Algoritmos hierárquicos lidam melhor com essa variação de densidade e o algoritmo MST pode ser visto como uma extensão hierárquica do k-means. Considerando algoritmos espectrais e baseados em grafos, você considera que eles são capazes de lidar com variação de densidade nos agrupamentos inerentes aos dados? Se sim, justifique a sua resposta, se não, discuta como os algoritmos espectrais e baseados em grafos podem ser estendidos para lidar melhor com densidade variável nos agrupamentos nativos.

R: Acredito que não. Os algoritmos baseados em grafos e espectrais podem ser estendidos ajustando a similaridade entre os pontos de amostras e aumentando a força do local de correlação entre os pontos dos dados.

5. A avaliação da qualidade de agrupamentos, em geral, se baseia nas noções de similaridade intra e inter agrupamento. Antes de realizar o agrupamento, entretanto, é importante saber o quanto os dados são agrupáveis, o que é conhecido como tendência de agrupamento. Estratégias tradicionais de avaliação de tendência de agrupamento se baseiam em contrastar os dados com distribuições aleatórias, o que é necessário, mas não suficiente para garantir que haja agrupamentos com diferenças significativas entre as similaridades intra e inter agrupamento. Proponha uma estratégia de avaliação de tendência de agrupamento baseada em execuções sobre amostras da base de dados e sementes aleatórias para técnicas como k-means e EM. Em particular descreva como você pode utilizar os resultados das várias execuções para avaliar a tendência de agrupamentos.

R: O principal quesito a ser avaliado são os próprios dados a serem agrupados. O que é similar é realmente similar? O que é diferente é realmente diferente? Existem técnicas de agrupamentos de dados que só funcionam em um tipo de dado específico, além disso, é necessário ver se os dados de entrada podem ser de fatos agrupados. Para isso, como nas

técnicas de agrupamento é geralmente utilizado uma lógica randomizada (semente aleatória), é necessário executar os algoritmos diversas vezes e analisar a estabilidade dos agrupamentos nas diversas execuções, ou seja, verificar se os agrupamentos são consistentes mesmo na mudança de alguns parâmetros.

6. Padrões frequentes e agrupamentos são modelos ditos descritivos, pois buscam explicitar propriedades entre as entidades representadas nos dados de entrada. Considerando os algoritmos ECLAT e EM, é possível usar mineração de padrões frequentes para subsidiar a criação de métricas de similaridade para fins de agrupamento? Se sim, descreva como, se não, explique porque não acredita ser possível.

R: Sim**.** Algoritmos de mineração de dados utilizam técnicas que usam métricas de similaridade e dissimilaridade para tomada de decisões, em particular para decidir se determinado dado é similar ou não a um protótipo ou se a distância entre um dado ou para verificar, em um grupo, qual é o par de dados mais próximos em um espaço qualquer (no caso de técnicas de agrupamento).