Universidade Federal do Ceará
Centro de Ciências/Departamento de Computação
Código da Disciplina: CKP8466
Professor: Ismayle de Sousa Santos

Aulas 22 e 23

Lógica da Pesquisa Científica

Mineração de Dados (de Software)

nctt3tj ismaylesantos@great.ufc.br @IsmayleSantos

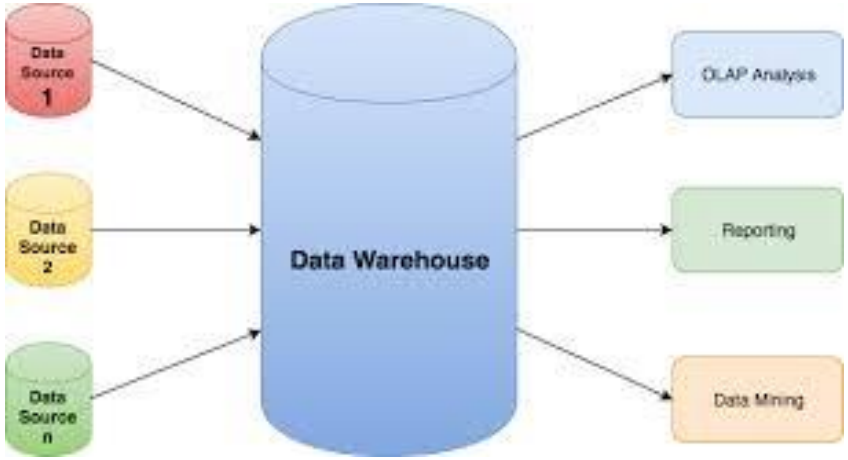
História da Mineração de Dados

- Na década de 70, muitos especialistas foram instruídos a armazenar seus dados em discos rígidos, fitas magnéticas, banco de dados, etc..., para fornecer segurança
- Com o surgimento de novos métodos de armazenamento de dados e da popularização dos Sistemas de Gerenciamento de Banco de Dados (SGBD) houve à proliferação da informação
- Os sistemas de informações construídos para apoiar o processo decisório geralmente armazenam seus dados em sistemas de banco de dados ou até mesmo em grandes repositórios de dados, data warehouse

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

História da Mineração de Dados

- Segundo Inmon apud Coradine (2011), "um data warehouse consiste de um banco de dados especializado capaz de manipular um grande volume de informações obtidas a partir de bancos de dados operacionais e de fontes de dados externas à organização"
- A tecnologia data warehouse permite atender sistemas de informação capazes de produzir transações de alto desempenho com objetivo de armazenar e cruzar grande volume de dados



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Mineração de Dados de Repositório de Software

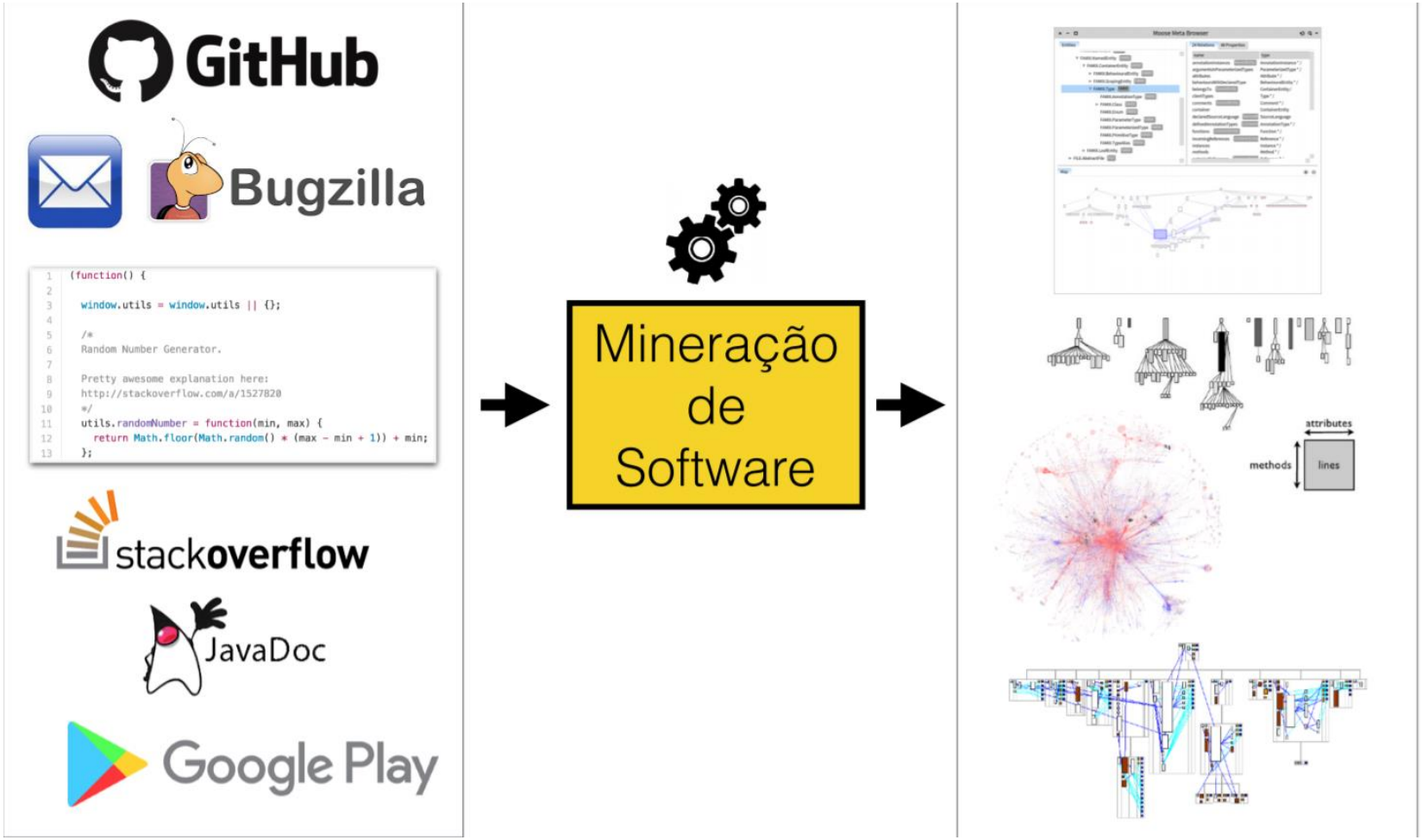
- Utilizando técnicas voltadas para mineração de repositórios de software:
  - Desenvolvedores
    - Adquirem maior conhecimento sobre como manter e evoluir sistemas de software
    - Exemplo: Detecção de códigos candidatos a refatoração
  - Pesquisadores
    - Exploraram esses dados para entender e melhorar as práticas de desenvolvimento
    - Exemplo: Investigando refatorações no histórico de versões

Mineração de Dados de Repositório de Software

- O que podemos analisar de repositórios de software?

Fonte: https://homepages.dcc.ufmg.br/~andrehora/teaching/msr/2-introducao-msr.pdf

Mineração de Dados de Repositório de Software



Fonte: https://homepages.dcc.ufmg.br/~andrehora/teaching/msr/2-introducao-msr.pdf

Exemplo de Mineração de Dados

Ver todas as 2 imagens

Leia com nossos **apps gratuitos** 11 Novo(s) a partir de R$ 156,95

Em até 5x R$ 31,39 sem juros Calculadora de prestações

Entrega para o CEP 01319-900 na Sexta-feira, 1 de Setembro, se você finalizar o pedido em **23 horas e 20 minutos**

Leia Enquanto Enviamos

Leia Enquanto Enviamos
Compre e comece a ler a amostra digital deste livro enquanto espera ele chegar. Saiba mais aqui.

No intuito de atender às necessidades de alunos e professores dos cursos de ciência da computação, engenharia de computação e sistema de informação, esta nona edição de Engenharia de software teve seu conteúdo reestruturado e totalmente atualizado. A obra conta agora com novos capítulos que focalizam o desenvolvimento ágil de software e os sistemas embarcados, além de trazer novas abordagens sobre engenharia dirigida a modelos, desenvolvimento open source, modelo Swiss Cheese de Reason,

Leia mais

Frequentemente comprados juntos

+

Preço total: R$ 308,93

Adicionar ambos ao carrinho

Este item: Engenharia de Software por Ian Sommerville Capa comum R$ 156,95

Engenharia de Software. Uma Abordagem Profissional por Roger S. Pressman Capa comum R$ 151,98

Data mining

Sommerville → Pressman

Fonte: https://homepages.dcc.ufmg.br/~andrehora/teaching/msr/2-introducao-msr.pdf

Introdução a Mineração de Dados

- Para que os dados sejam devidamente manipulados para serem transformadas em conhecimento, faz-se necessário a utilização de técnicas que propiciem a automação desse processo a partir de estruturas artificialmente inteligentes

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Introdução a Mineração de Dados

- O que é mineração de dados?
  - Mineração de dados pode ser visto como um conjunto de métodos para fazer inferências a partir de dados (CORADINE, 2011)
    - provém da análise inteligente e automática de dados para descobrir padrões ou regularidades em grandes conjuntos de dados, através de técnicas que envolvam métodos matemáticos, algoritmos baseados em conceitos biológicos, processos linguísticos e heurísticos, os quais fazem parte do processo KDD responsável pela busca de conhecimentos em banco de dados

Fonte: KITCHENHAM, Barbara Ann et.al.. Evidence-based Software Engineering and Systematic Reviews. Boca Raton, FL: CRC Press, 2016.

Introdução a Mineração de Dados

- No contexto da engenharia de software geralmente é usado para desenvolver modelos de previsão que visam identificar alguma característica importante de um projeto ou de um componente de software desconhecido
  - Exemplos:
    - técnicas de mineração podem ser utilizadas para tentar predizer se há probabilidade da geração de um erro, dado um estado S de um programa e um evento E

Fonte: KITCHENHAM, Barbara Ann et.al.. Evidence-based Software Engineering and Systematic Reviews. Boca Raton, FL: CRC Press, 2016.

Introdução a Mineração de Dados

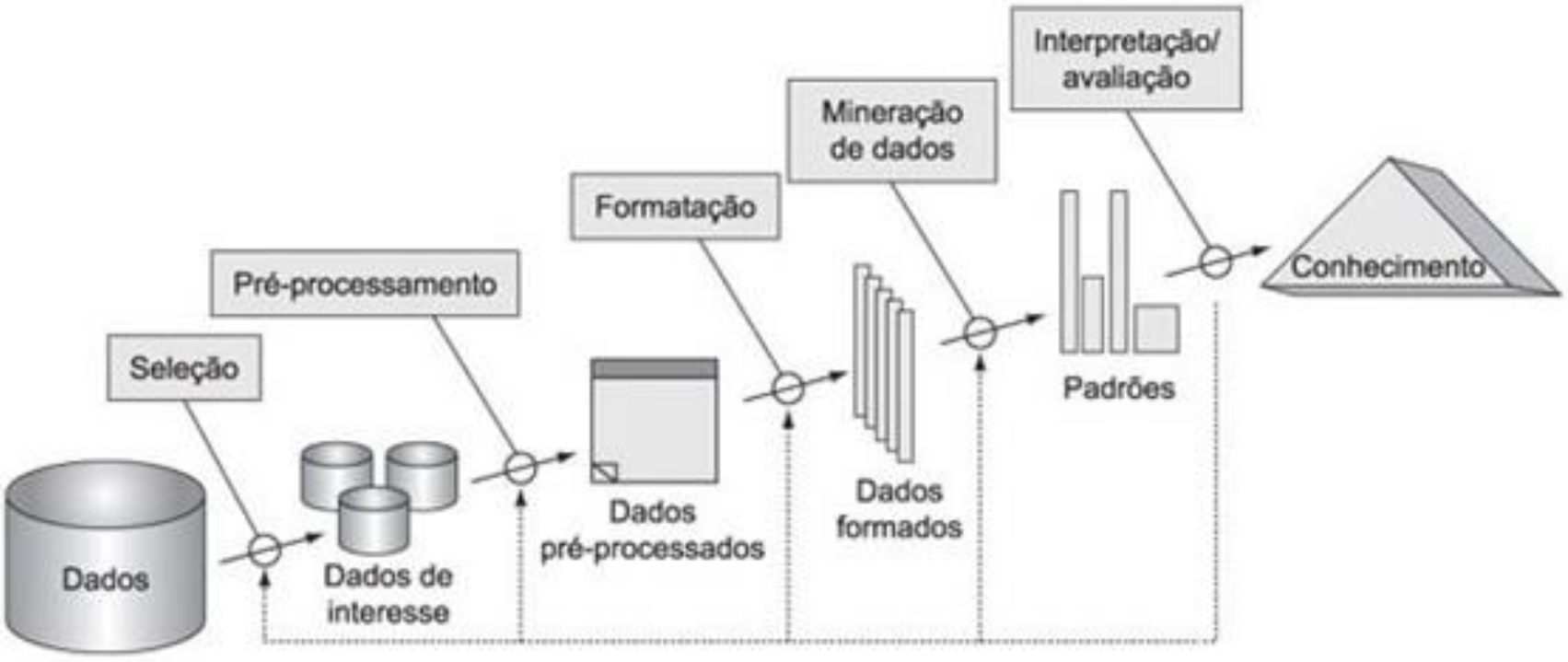
- knowledge discovery in databases (KDD)
  - O objetivo é a descoberta dos conhecimentos engendrados no banco de dados
  - É o processo de extração da informação relevante ou de padrões nos dados contidos em grandes BD e que sejam não triviais, implícitos, previamente desconhecidos e potencialmente úteis, objetivando a tomada de decisão

Data mining = etapa que transforma dados em informação

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Mineração de Dados: Processo do KDD

- Principais tarefas do processo do KDD



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Mineração de Dados: Processo do KDD

- Seleção de dados
  - Busca identificar conjunto de dados relevantes e seus subconjuntos de variáveis objetivando a criação de um conjunto de restrito de dados para a descoberta de conhecimento
- Pré-processamento
  - Envolve a limpeza dos dados, com operações de remoção dos ruídos, elaboração de esquemas e mapeamentos de valores desconhecidos
- Transformação(Formatação)
  - Onde se busca características úteis nos dados, utilizando métodos de redução ou transformação da dimensionalidade dos dados para um o melhor desempenho

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Mineração de Dados: Processo do KDD

- Mineração de dados
  - Onde se aplica técnicas específicas em dados pré-processados com objetivo de buscar modelos de interesse numa representação incluindo regras de classificação, árvores de decisão, regressão ou agrupamento
- Interpretação dos dados
  - Análise dos resultados obtidos, a qual permite avaliar padrões com objetivo de determinar quais as melhores maneiras de usar as informações na tomada de decisão

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

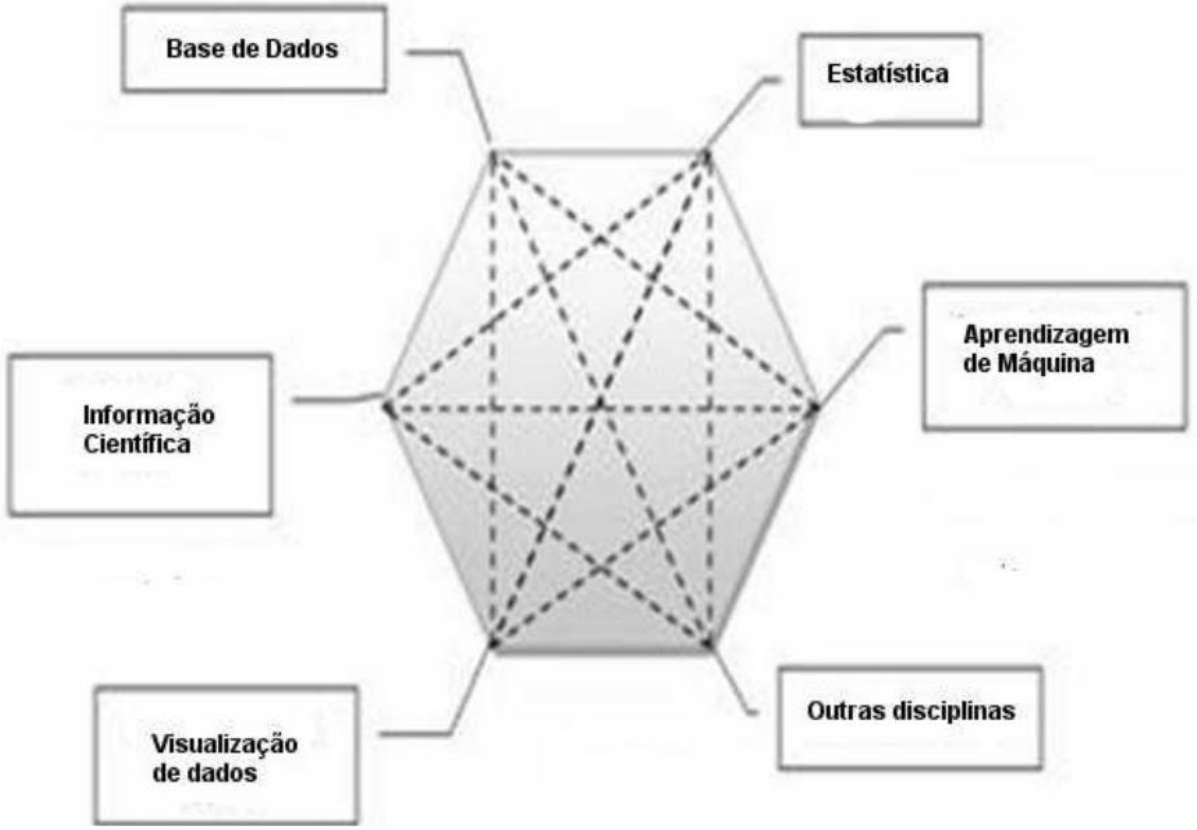
Mineração de Dados: Processo do KDD

- O KDD evoluiu, e continua a evoluir, a partir da interseção de campos de pesquisa, como
  - Aprendizagem de máquina
  - Reconhecimento de padrões
  - Bases de dados
  - Estatísticas
  - Visualização de dados
  - Computação de alto desempenho
- O objetivo é unificar a extração de alto nível de conhecimento a partir de dados de baixo nível, no contexto de grandes conjuntos de dados

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Multidisciplinaridade da Mineração de Dados

- Integração de diversas áreas de conhecimento no processo de análi



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Características da Mineração de Dados

- As análises de mineração de dados podem ser categorizadas em dois tipos amplos:
  - Supervisionado
    - O objetivo é **prever** o valor de alguma medida de resultado (por exemplo, o esforço esperado necessário para desenvolver um projeto de software) dadas uma série de variáveis de entrada (por exemplo, a estimativa tamanho do projeto, a experiência dos desenvolvedores)
    - Para aprendizado supervisionado, é comum separar o conjunto de dados em um conjunto de treinamento (ou seja, um subconjunto aleatório de casos) e um conjunto de dados de validação (ou seja, os outros casos)

Fonte: KITCHENHAM, Barbara Ann et.al.. Evidence-based Software Engineering and Systematic Reviews. Boca Raton, FL: CRC Press, 2016.

Características da Mineração de Dados

- Não supervisionado
  - Não há medida de resultado
  - O objetivo é encontrar padrões nos dados
    - Por exemplo, grupos de itens que compartilham propriedades semelhantes

Fonte: KITCHENHAM, Barbara Ann et.al.. Evidence-based Software Engineering and Systematic Reviews. Boca Raton, FL: CRC Press, 2016.

Características da Mineração de Dados

- **Exemplo Supervisionado**
  - Ensinar alguém que nunca viu maçãs ou laranjas a reconhecê-las, mostrando-lho várias frutas e identificação de cada uma (treinamento). Depois, esse alguém será capaz de classificar novas amostras.
- **Exemplo Não Supervisionado**
  - As frutas seriam mostradas sem identificação, devendo a pessoa descobrir os dois tipos de frutas (classes)

Fonte: KITCHENHAM, Barbara Ann et.al.. Evidence-based Software Engineering and Systematic Reviews. Boca Raton, FL: CRC Press, 2016.

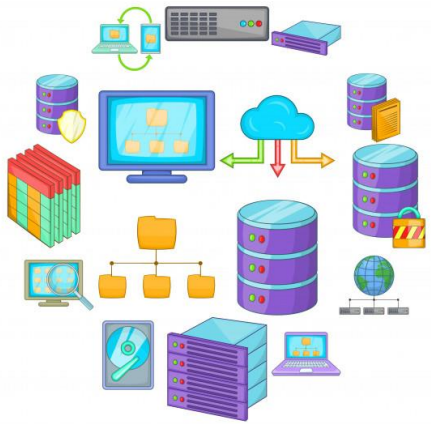
Funcionalidades da Mineração de Dados

- O processo de mineração de dados pode ser dividido em componentes capazes de favorecer a identificação mais adequada dos algoritmos de mineração quando se leva em consideração algumas informações relevantes tais como:
    - Função do modelo
    - Representação do modelo

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo

- As funções do modelo são utilizadas para especificar o tipo de aplicação do algoritmo minerador
- As funções mais comumente usadas são:
  - Classificação
  - Regressão
  - Análise de associação
  - Análise de seqüência
  - Sumarização
  - Visualização



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Classificação

- A função de classificação tem por premissa reconhecer, em um conjunto de dados, as observações que tenham as mesmas características
- A tarefa é descobrir se um item vindo do banco de dados pertence a uma das algumas classes, previamente definidas
- Na prática, as classes são muitas vezes definidas usando-se valores específicos de determinados campos nos registros de dados ou alguns derivados desses valores
    - Por exemplo, se um registro de dados contém o campo Região, então algum dos valores típicos do campo, por exemplo, Norte, Sul, Leste ou Oeste, pode definir a classe

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Classificação

- A classificação de dados pode ser vista um processo em duas etapas
  - Na primeira etapa, um modelo é construído descrevendo um conjunto pré-determinado de classes de dados ou conceitos
    - As listas de dados analisadas para construir o modelo formam coletivamente um conjunto de treinamento de dados
    - As listas individuais que compõem o conjunto de treinamento são referidas como amostras de treinamento e são selecionados aleatoriamente da população de amostra
    - Como o rótulo de classe de cada amostra de treinamento é fornecido, esta etapa é vista como um aprendizado supervisionado

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Classificação

- Na segunda etapa, o modelo é usado para a classificação
  - A precisão da previsão do modelo (classificador) é estimada usando um conjunto de testes das amostras de treinamento
  - Uma vez que o algoritmo classificador tenha sido desenvolvido de forma eficiente, ele poderá ser usado de forma preditiva para classificar novos registros naquelas mesmas classes pré-definidas

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Regressão

- A função de regressão é, basicamente, um conjunto de métodos que permite a interpretação da relação funcional entre variáveis com boa aproximação, considerando a existência de uma relação entre essas as variáveis, de modo que a medida possa estabelecer modelos utilizados para fins de predição
- No caso mais simples, a regressão é uma função de aprendizado que mapeia um dado item a uma variável de predição de valor real
- No caso geral tem-se a predição de uma ou mais variáveis dependentes, considerada como resposta, a partir de conjunto de variáveis independentes, os preditores

O objetivo é prever os valores de uma variável dependente com base em resultados da variável independente

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Regressão

- Os métodos de regressão podem ser utilizados em diversas áreas de conhecimento como, por exemplo:
  - Para previsão da economia nacional com base em certas informações (níveis de renda, investimentos, etc...)
  - Para a verificação de quais fatores ajudam a manter a qualidade dos serviços oferecidos
  - Na medida de viabilidade de um novo produto
  - Ou na construção de séries temporais onde as variáveis de entrada são versões atrasadas da variável de predição

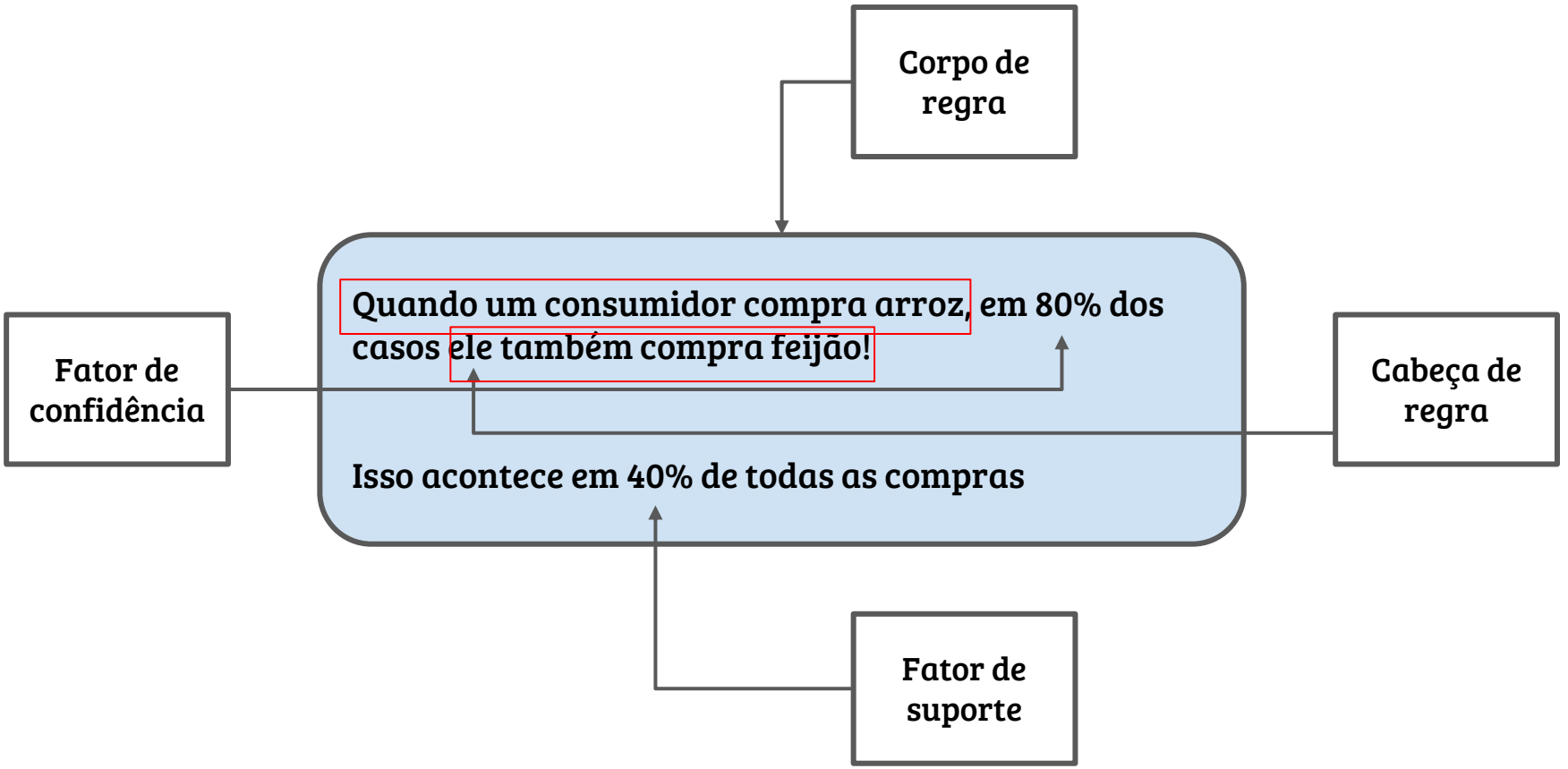
Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Análise de Associação

- Análise de associação tem como objetivo elaborar uma representação explícita entre os objetos, visando determinar relacionamentos entre conjuntos de itens de associação
- Gera redes de interações e conexões presentes nos conjuntos de dados usando as associações item a item
- A presença de um item implica necessariamente na presença do outro item na mesma transação
- Em geral, uma regra de associação pode ser representada formalmente através do tipo, se X então Y, considerados corpo e cabeça da regra, respectivamente

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Análise de Associação



A figura exemplifica uma regra de associação voltada a identificar afinidades entre itens de um subconjunto de dados, dentro de um conjunto de valores (produtos comprados por um cliente, sintomas apresentados por um paciente, etc.), destacando, ainda, dois fatores importantes, o de confidência e o de suporte.

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Análise de Seqüência

- Análise de seqüência constitui-se de uma variação da análise associativa objetivando extrair e registrar desvios e tendências no tempo
- As regras identificadas são usadas para reconhecer seqüências relevantes que possam ser utilizadas para prever comportamentos, modelar processos gerando uma seqüência ou relatar tendências de um processo ao longo do tempo

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Análise de Seqüência

- Exemplo
  - Seja um conjunto de dados ordenado pelo sobrenome do consumidor e pelo período de transação de compras
  - A tabela mostra as seqüências de transações de consumidores organizadas segundo o tempo

Tabela 4.1: Seqüências de transações dos consumidores data (Diniz e Louzada, 2000)

Consumidor*	**Seqüência diária de compras de bebidas**			
Oliveira	(Cerveja)	(Vodka)		
Soares	(Guaraná, Suco)	(Cerveja)	(Água, Licor, Vinho)	(Gin, Licor)
Tenório	(Cerveja)	(Água, Gin, Vinho)	(Vodka, Soda)	
Zacaria	(Vodka)			

***sobrenomes fictícios**

d1 **d2** **d3** **d4**

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Análise de Seqüência

- A característica seqüencial "*cerveja é comprada em uma transação anterior a que a vodka é comprada*" ocorre em dois dos quadros

Tabela 4.1: Seqüências de transações dos consumidores data (Diniz e Louzada, 2000)

Consumidor*	Seqüência diária de compras de bebidas			
Oliveira	(Cerveja)	(Vodka)		
Soares	(Guaraná, Suco)	(Cerveja)	(Água, Licor, Vinho)	(Gin, Licor)
Tenório	(Cerveja)	(Água, Gin, Vinho)	(Vodka, Soda)	
Zacaria	(Vodka)			

***sobrenomes fictícios**

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Sumarização

- A sumarização visa obter uma descrição compacta de um conjunto de dados, bastante usada em análise exploratória de dados
- A sumarização não é usada para a resolução de problemas, mas possibilita identificar características no conjunto de dados que possa estar contaminadas por ruídos, que interfiram no processo de análise, ou redundantes, gerando uma tendência errônea à análise

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

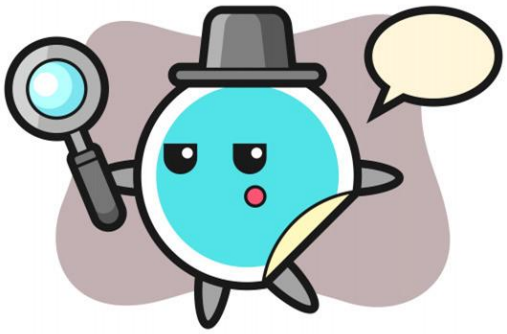
Funções do Modelo: Sumarização

- A sumarização é usada, principalmente, no pré-processamento dos dados
  - Valores inválidos, no caso de variáveis quantitativas, são determinados através do cálculo de medidas estatísticas e, no caso de variáveis categóricas, através da distribuição de freqüência dos valores
- O objetivo da sumarização em mineração de dados é propiciar a limpeza dos dados facilitando a análise e a geração automatizada de relatórios

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Visualização

- As técnicas de visualização podem ser consideradas ferramentas eficientes para se analisar grandes quantidades de dados
- Em muitas situações, elas são suficientes para a extração das respostas de interesse, descobrindo padrões, tendências, estruturas e relações, dentro de um conjunto de dados
- O método de visualização escolhido para análise dependerá basicamente do tipo de conjunto de dados disponível e como esses dados podem ser modelados



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Visualização

- Por exemplo, se o conjunto de dados envolve chamadas telefônicas feitas em um intervalo de tempo específico, então uma representação visual desta informação poderia ser sumarizada através de um simples diagrama de associação, disponibilizando todas as relações entre as chamadas, como na tabela 4.3

Tabela 4.3 Representação tabular das chamadas telefônicas (Diniz e Louzada-Neto, 2000)

De	1	1	2	4	4	8	7	8
Para	2	3	6	6	7	6	5	6
Horário	07:45	08:00	08:36	09:16	09:48	11:22	11:51	12:03
De	7	6	3	2	8	6	2	6
Para	4	2	2	6	6	2	6	7
Horário	14:03	14:18	14:53	15:34	16:19	16:38	17:05	17:28

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Visualização

- A figura 4.2, apresenta a visualização de várias camadas entre certos pares de telefones
- As linhas mais grossas no diagrama representam os números maiores de chamadas

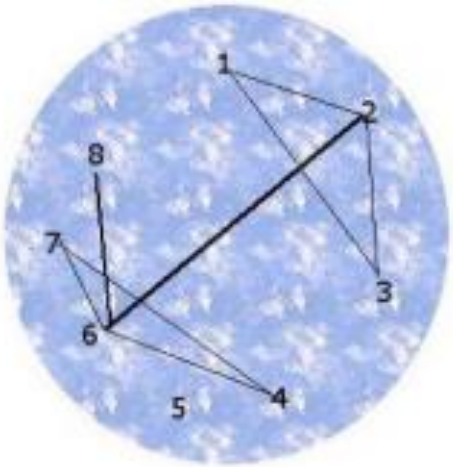


Figura 4.2: Diagrama de associação das chamadas telefônicas (Diniz e Louzada-Neto, 2000)

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Funções do Modelo: Visualização

- Representações de métodos de visualização bem comum são, por exemplo, os métodos de visualização simples de dados, os quais se baseiam em gráficos ou resumos rápidos que, de alguma forma, representam ou resumem características dos conjuntos de dados
- Outros métodos de visualização de dados incluem:
  - Diagramas baseados em proporções
  - Diagramas de dispersão
  - Histogramas
  - Box plots

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Representação do Modelo

- As funções do modelo têm um papel importante na análise e modelagem do problema
- Os modelos representados a partir de algoritmos de mineração de dados, podem determinar a flexibilidade do mesmo em representar o conjunto de dados e a sua interpretação
- Representações mais tradicionais incluem
  - árvore de decisão
  - conjunto de regras
  - métodos de agrupamento,
  - modelos lineares e não lineares

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Representação do Modelo: Árvores de Decisão e Regras de Decisão

- Quando o processo de mineração de dados é direcionado à classificação, a árvore de decisão pode ser conveniente quando o objetivo se relaciona à categorização dos dados
- As árvores de decisão são ferramentas eficientes e populares para classificação e diagnóstico
- A árvore é formada por nós e o primeiro, nó raiz, envolve todo o conjunto de dados, onde o processo de classificação se inicia

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Representação do Modelo: Árvores de Decisão e Regras de Decisão

- Cada nó interno identifica um dos atributos de previsão
- Cada linha que sai desse nó identifica um valor assumido por tal nó
- Cada nó terminal (folha) identifica o resultado da previsão ou objetivo

Classificação por árvore de decisão do formato de pinos dados comprimento e diâmetro

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Representação do Modelo: Árvores de Decisão e Regras de Decisão

- As regras de decisão podem ser consideradas um processo para analisar uma série de dados e a partir dela gerar padrões
- Também podem ser vistas como a expressão verbal das árvores de decisão
- A integração dos métodos de árvore de decisão e regra de decisão pode ser considerada ferramenta fundamental em previsão

Uma regra pode ser construída através da formação de um conjunto de testes que ocorre nos caminhos entre nó raiz e os nós terminais da árvore

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Representação do Modelo: Árvores de Decisão e Regras de Decisão

- Gerada uma solução utilizando árvore de decisão ou regra de decisão, esta pode ser usada para estimar ou predizer a resposta ou classe variável para um novo caso
- Exemplo de regras para árvore anterior

Se (comprimento ≤ 075)

Então Quadrado

Se (não (comprimento ≤ 075)) & (diâmetro ≤ 3,00)

Então Estrela

Se (não (comprimento ≤ 075)) & (não (diâmetro ≤ 3,00))

Então Losango

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento

- O objetivo da análise de agrupamento (cluster) está relacionado ao processo de agrupar elementos de dados mediante o particionamento de uma população heterogênea em subgrupos mais homogêneos
- A análise de agrupamento (*cluster*) é o estudo formal dos algoritmos e dos métodos para agrupar ou classificar objetos
- No agrupamento, não há classes pré-definidas, os elementos são agrupados de acordo com a semelhança, o que a diferencia da tarefa de classificação, buscando reunir indivíduos ou objetos em grupos

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

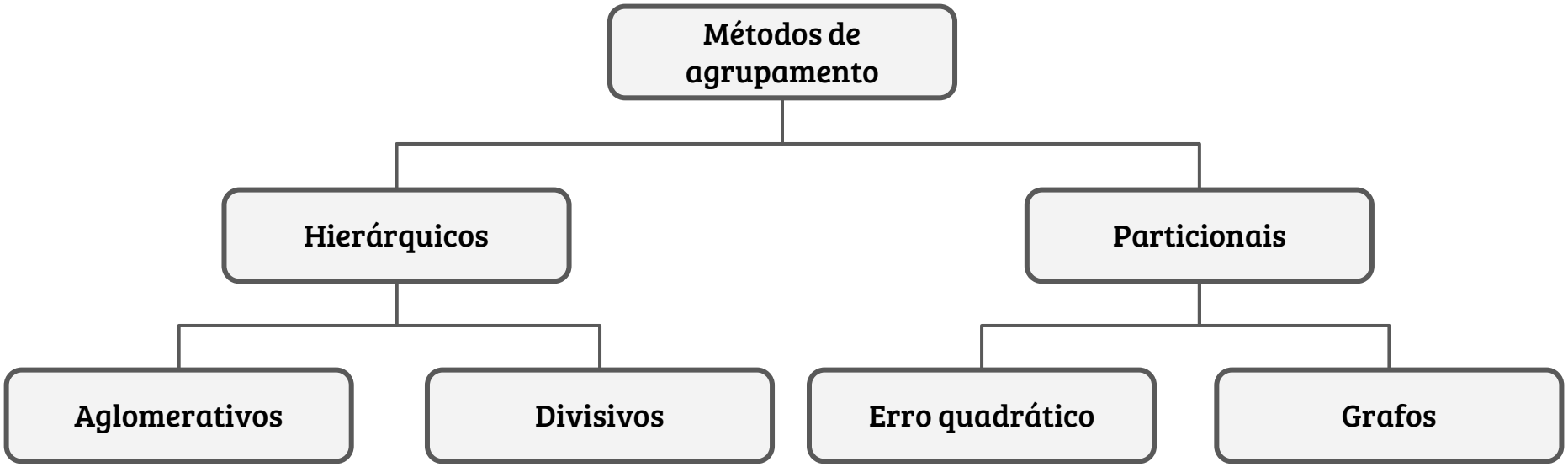
Análise de Agrupamento: Métodos de Agrupamento

- O método de agrupamentos é uma técnica analítica para desenvolver subgrupos significativos de indivíduos ou objetos
- Têm como objetivo classificar uma amostra de entidades em um pequeno número de grupos mutuamente exclusivos, com base nas similaridades entre eles
- Essa técnica pode ser dividida em três etapas:
  - A primeira relaciona-se a medida de similaridade ou associação entre as entidades para determinar quantos grupos realmente existem na amostra
  - A segunda refere-se ao processo de busca do agrupamento, no qual entidades são particionadas
  - E a terceira busca estabelecer o perfil das variáveis para determinar sua composição

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos de Agrupamento

- Existem várias adaptações ao modelo simplificado aos métodos de agrupamentos, representados graficamente a partir da figura abaixo



Classificação simplificada dos métodos de agrupamentos adaptados (Jain e Dubes, 1988)

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Agrupamento Hierárquico

- O procedimento hierárquico opera para formar um intervalo inteiro de soluções de agrupamento
  - Também pode operar como um método aglomerativo, objetivando fundir agrupamentos individuais (inicialmente, cada grupo contém um único objeto) em partições maiores até a obtenção de uma única partição contendo todos os objetos do conjunto
- O procedimento hierárquico trata o conjunto de dados como uma estrutura de partições, cada uma correspondendo a um agrupamento, hierarquicamente organizadas segundo a similaridade entre seus objetos

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

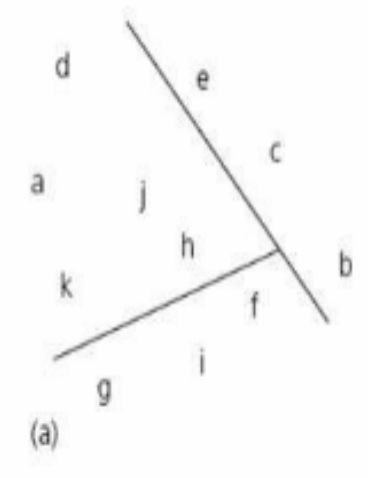
Análise de Agrupamento: Métodos Hierárquicos Aglomerativos

- Dentre os métodos hierárquicos que envolvem a construção de uma hierarquia numa estrutura em árvore, encontram-se as técnicas aglomerativas, utilizadas para descobrir agregados
- Exemplos:
  - Ligação Individual ou Simples: é baseado na distância mínima
    - Ele encontra os dois objetos separados pela menor distância e os coloca primeiro no agrupamento
    - A próxima distância mais curta é determinada, e um terceiro objeto se junta aos dois primeiros para formar um agregado, ou um novo agrupamento de dois membros é formado
  - Ligação Completa: é baseado na distância máxima

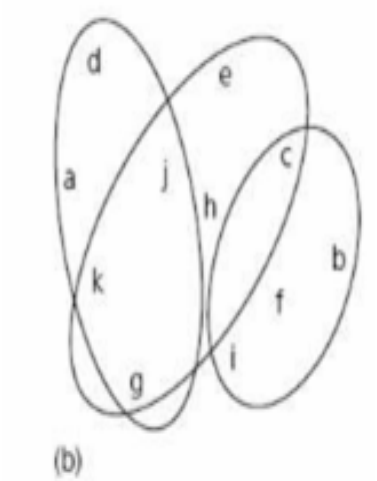
Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos Hierárquicos Aglomerativos

- Existem diferentes tipos de representação de agrupamentos
  - As representações mais comuns são do modelo dendrograma, do grego *dendro* (árvore), ou seja, diagrama em árvore, que representa as junções sucessivas de partições e que pode gerar agrupamentos diferentes conforme o nível em que é seccionada



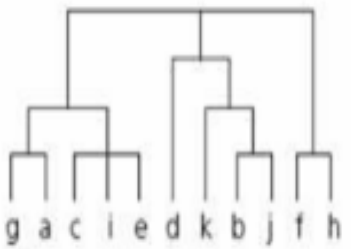
(a)



(b)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

(c)

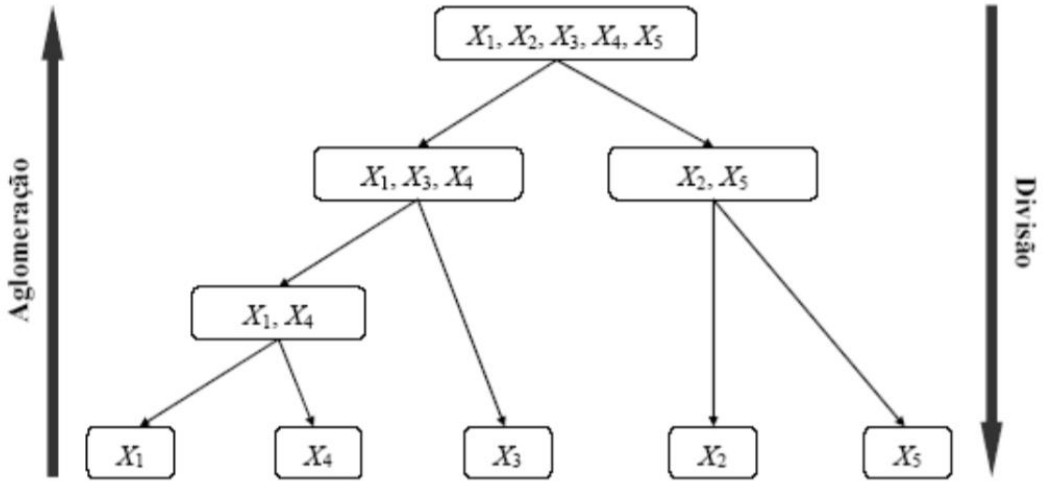


(d)

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos Hierárquicos Divisivos

- Quando o processo de classificação utiliza métodos divisivos ocorre uma inversão com relação ao método hierárquico aglomerativo
- Pode-se observar que um conjunto contendo todos os dados é particionado a partir de um aglomerado unificado
- Os métodos divisivos começam o processamento a partir de um grande agregado que contém todas as observações (objetos)



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos Particionais

- Os métodos particionais têm como objetivo dividir um conjunto de objetos em um número pré-estabelecido de clusteres
- Esse algoritmo divide o conjunto de dados em partes disjuntas, satisfazendo as seguintes recomendações:

  a) objetos de uma mesma parte estão próximos, de acordo com um critério de dado

  b) objetos de partes distintas estão longe, de acordo com este mesmo critério

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos Particionais

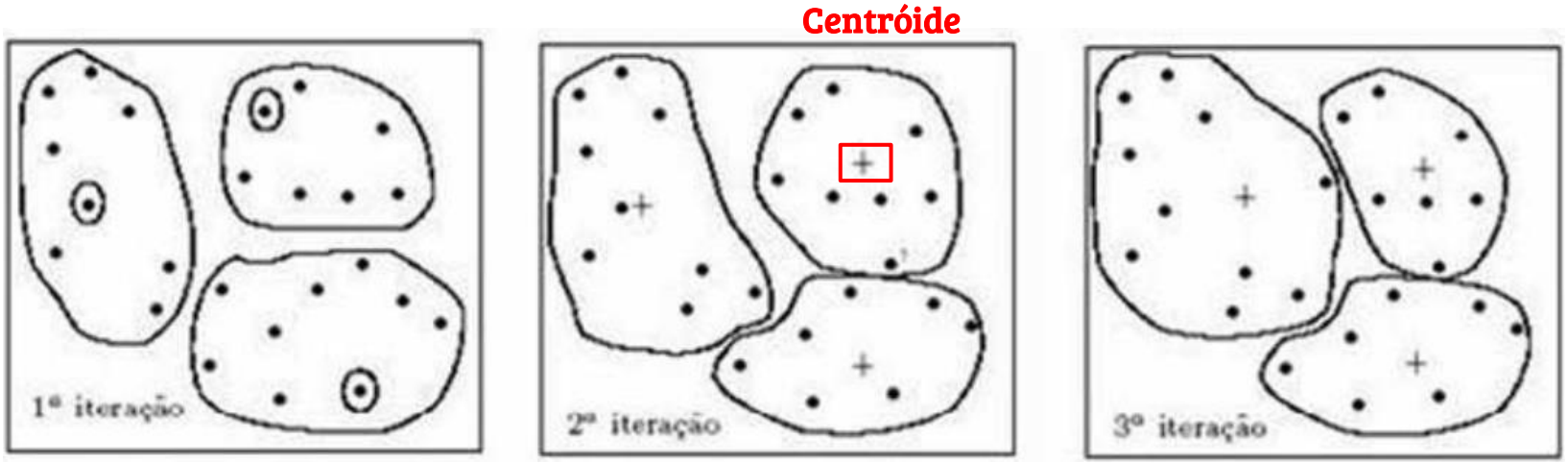
- O método mais popular é conhecido como k-means
- A subdivisão realizada pelo algoritmo k-means, como método particional, é feita da seguinte maneira:
  - Cria-se uma partição inicial aleatória de K partes e posteriormente, em um processo iterativo
  - Os elementos das partes vão sendo realocados para outras partes, de modo a melhorar o particionamento a cada iteração

Dividem o conjunto dos N objetos em K agrupamentos sem relacioná-los hierarquicamente entre si

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos Particionais

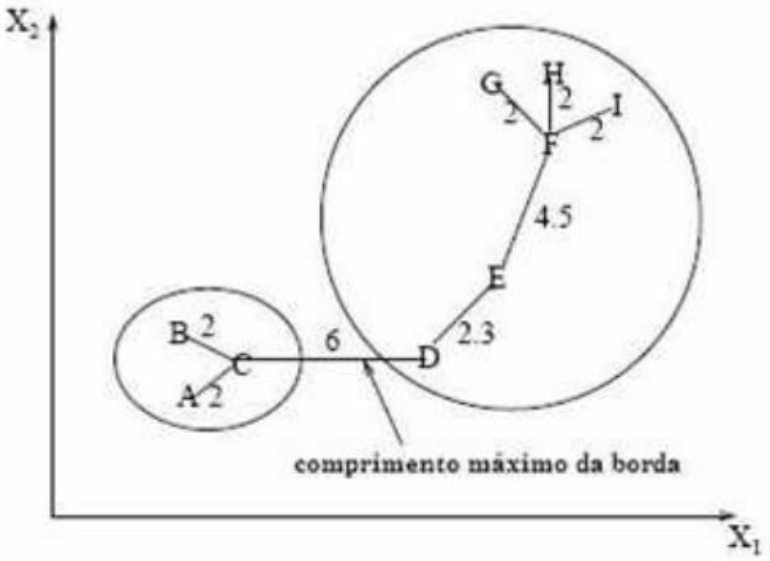
- O k-means recebe como entrada um número K de agrupamentos e atribui aleatoriamente um objeto como sendo o centróide inicial de cada agrupamento
  - Sucessivamente, cada objeto é associado ao agrupamento mais próximo e o centróide de cada agrupamento é então recalculado levando em conta o novo conjunto de objetos pertencentes a ele



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Análise de Agrupamento: Métodos Particionais - Grafos

- Os agrupamentos de dados baseado em grafos utilizam algoritmo baseado na construção de uma árvore geradora mínima (Minimum Spanning Tree - MST)
- Têm como objetivo à geração de um grafo, de modo que os objetos não possuam ciclos e sejam conectados por um arco, ou seja, uma árvore



Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Ferramentas

MAchine Learning for LanguagE Toolkit

WEKA

The workbench for machine learning

Ferramentas de Análise de Dados de Software

extração de dependências e cálculo de métricas para C, C++ e Java

Ferramentas para extração e visualização de dados de diferentes tipos de repositório

Para visualização estática de grafos

Plataforma agrega diversas ferramentas de análise de qualidade de código

Conferências

- MSR: International Conference on Mining Software Repositories
- ICSE: ACM/IEEE International Conference on Software Engineering
- ICSME: IEEE International Conference on Software Maintenance and Evolution
- SANER: IEEE International Conference on Software Analysis, Evolution and Reengineering
- VEM: Workshop on Software Visualization, Evolution and Maintenance

Sugestão de Leitura

- Outras técnicas podem ser vista na literatura, com detalhes, como forma complementar os estudos dos métodos de mineração de dados
  - Ferramentas supervisionadas e não supervisionadas com o uso da teoria de redes neurais
    - Principe, J. C., Euliano, N. R. e Lefebvre, W. C. (2000). Neural Adaptive Systems: Fundamentals Through Simulations, John Willey & Sons, New York, NY
  - Algoritmos bio-inspirados
    - Jain, A.K., Murty, M.N. e Flynn, P.J. (1999), Data clustering: A review, ACM Computing Surveys, 31:264–323.
  - Conjuntos nebulosos
    - Zadeh. L.A., Klir, G.J., Yuan. B. (1996), Fuzzy Sets, Fuzzy Logic, e Fuzzy Systems, World Scientific Publishing, New Jersey.

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Sugestão de Leitura

- Análise de componentes independentes
  - Hivärinen, A., Karhunen, J. e Oja, E. (2001), Independent Component Analysis, Jonh Wiley, New York
- No contexto da classificação
  - Metha, M.; Agraval R. e Rissanen, J. SLIQ: A Fast Scalable Classifier for Data Mining, IBM Almaden Research Center, 1996.
  - Shafer, J.; Agraval, R.; Mehta, M. SPRINT: A scalable parallel classifier for data mining. In Proc. Of the 22nd VLDB Conference, 1996.
- Novos métodos de agrupamento fuzzy cluster
  - Silvanandam, S.N., Sumatri, S. e Deepa, S.N., (2007). Introduction to Fuzzy Logic Using Matlab. Berlin.

Fonte: CORADINE, Luis C.. Mineração de Dados: Uma Introdução. – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Trabalho Prático TP-6

- Escolher um artigo de Mineração de Dados (de Software)
  - Única Restrição: Entre qualis A1 ~ A4
    - https://ppgcc.github.io/discentesPPGCC/pt-BR/qualis/
- Fazer uma resenha crítica
  - Foco: Metodologia utilizada
  - Tamanho: 3 à 5 páginas

Obrigado!

Por hoje é só pessoal...

Dúvidas?

 nctt3tj  ismaylesantos@great.ufc.br  @IsmayleSantos