Mineração de dados é o método de investigação de grandes quantidades de dados com o objetivo de encontrar irregularidade, padrões e correlações para arcar a tomada de decisões e possibilitar vantagens estratégicas.

A quantidade de dados produzido está dobrando a cada dois anos. Dados não-estruturados compõem sozinhos 90% do nosso universo digital. Entretanto, mais informação não significa necessariamente mais conhecimento. A mineração de dados nos permite filtrar todo o ruído caótico e repetitivo, entender o que é relevante e, então, fazer bom uso dessa informação para avaliar os prováveis resultados.

Os resultados obtidos com a mineração de dados podem ser usados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processo e muitas outras aplicações. A mineração de dados pode ser aplicada de duas formas: como um processo de verificação e como um processo de descoberta . No processo de verificação, o usuário sugere uma hipótese acerca da relação entre os dados e tenta prová-la aplicando técnicas como análises estatística e multidimensional sobre um banco de dados contendo informações passadas. No processo de descoberta não é feita nenhuma suposição antecipada. Esse processo usa técnicas, tais como, árvores de decisão, algoritmos genéticos e redes neurais.

Função executada por técnicas de Mineração de dados

As técnicas de mineração de dados podem ser aplicadas a tarefas como classificação, estimativa, associação, sumarização e segmentação. Essas tarefas são descritas a seguir

* **Classificação:** Reconhece modelos que descrevem o grupo ao qual o item pertence por meio do exame dos itens já classificados e pela inferência de um conjunto de regras. Equivale a construir um modelo de algum tipo que possa ser aplicado a dados não classificados visando categorizá-los em classes.

empresas de operadoras de cartões de crédito e companhias telefônicas preocupam-se com a perda de clientes regulares, a classificação pode ajudar a descobrir as características de clientes que provavelmente virão abandona-las e oferecer um modelo para ajudar os gerentes a prever quem são, de modo que se elabore antecipadamente campanhas especiais para reter esses clientes.

Estimativa (ou Regressão)

A estimativa é utilizada para determinar valores para alguma variável contínua desconhecida como, por exemplo, lucro, distância ou saldo na poupança . Ela lida com resultados contínuos, enquanto que a classificação lida com resultados discretos. Ela pode ser usada para executar uma tarefa de classificação, convencionando-se que diferentes intervalos de valores contínuos correspondem a diferentes classes. Como exemplos de tarefas de estimativa tem-se : estimar o número de alunos de uma escola; estimar o lucro total de uma empresa; prever a demanda de consumidores para um novo produto.

Associação

 a detecção de relações entre os registros, ocorrências ligadas a um único evento. O exemplo clássico é determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado outro exemplo seria um estudo de modelos de compra em lojas de carros pode revelar que, na compra de um carro automático, 85% das pessoas, querem ele com bancada de couro.

a tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo desta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados.

Sumarização

a tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo desta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados.

Segmentação (ou Clustering)

Funciona de maneira semelhante a classificação quando ainda não foram definidos grupos. uma técnica em que através de métodos numéricos e a partir somente das informações das variáveis de cada caso, tem por objetivo agrupar automaticamente por aprendizado não supervisionado os n casos da base de dados em k grupos, geralmente disjuntos denominados clusters ou agrupamentos.

Um dos métodos de clusterização é o algoritmo de Análise de Agrupamento k-means  um dos mais conhecidos e utilizados, além de ser o que possui o maior número de variações. O algoritmo inicia com a escolha dos  k elementos que formaram as sementes iniciais. Esta escolha pode ser feita de muitas formas, entre elas:

* selecionando as k primeiras observações;
* selecionando k observações aleatoriamente; e
* escolhendo k observações de modo que seus valores sejam bastante diferentes. Por exemplo, ao se agrupar uma população em três grupos de acordo com a altura dos indivíduos, poderia se escolher um indivíduo de baixa estatura, um de estatura mediana e um alto.

Escolhidas as sementes iniciais, é calculada a distância de cada elemento em relação às sementes, agrupando o elemento ao grupo que possuir a menor distância (mais similar) e recalculando o centróide do mesmo. O processo é repetido até que todos os elementos façam parte de um dos clusters.

Após agrupar todos os elementos, procura-se encontrar uma partição melhor do que a gerada arbitrariamente. Para isto, calcula-se o grau de homogeneidade interna dos grupos através da Soma de Quadrados Residual (SQRes), que é a medida usada para avaliar o quão boa é uma partição.

Após o cálculo, move-se o primeiro objeto para os demais grupos e verifica-se se existe ganho na Soma de Quadrados Residual, ou seja, se ocorre uma diminuição no valor da SQRes. Existindo, o objeto é movido para o grupo que produzir o maior ganho, a SQRes dos grupos é recalculada e passa-se ao objeto seguinte. Depois de um certo número de iterações ou não havendo mais mudanças, o processo é interrompido.

Técnicas de Mineração de dados

Não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens. A seguir são descritas as técnicas de mineração de dados normalmente usadas.

Árvores de Decisão

Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas. Permite que um indivíduo ou organização compare possíveis ações com base em seus custos, probabilidades e benefícios. Podem pode ser usadas tanto para conduzir diálogos informais quanto para mapear um algoritmo que prevê a melhor escolha, matematicamente.

Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore.

Algoritmos Genéticos

Os algoritmos genéticos são uma família de modelos computacionais inspirados na evolução, que incorporam uma solução potencial para um problema específico numa estrutura semelhante a de um cromossomo e aplicam operadores de seleção e "cross-over" a essas estruturas de forma a preservar informações críticas relativas à solução do problema.

Um algoritmo genético é um procedimento iterativo para evoluir uma população de organismos e é usado em mineração de dados para formular hipóteses sobre dependências entre variáveis, na forma de algum formalismo interno

Redes Neurais Artificiais

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência.

<https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html>

<http://www.din.uem.br/~intersul/intersul_arquivos/documentos/mineracao.pdf>