Pré-processamento em Data Mining

A etapa de pré-processamento, no processo de descoberta de conhecimento - KDD (Knowledge Discovery in databases), compreende a aplicação de várias técnicas para captação, organização, tratamento e a preparação dos dados. É uma etapa que possui fundamental relevância no processo de KDD. Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados.

Algumas das principais funções da etapa de pré-processamento dos dados:

Seleção de atributos – A Seleção de Atributos é uma etapa da fase de pré-processamento do processo de Descoberta de Conhecimento em Banco de Dados. Como o próprio nome já diz, o objetivo é escolher um subconjunto de atributos (também conhecidos como variáveis) ou criar outros atributos que substituam um conjunto deles a fim de reduzir a dimensão do banco de dados. Com essa redução de dimensão, reduz-se a complexidade do banco de dados e assim o tempo de processamento para extrair dele algum conhecimento. Além disso, atributos desnecessários podem causar ruído no resultado final e isto pode ser evitado com a aplicação de técnicas de Seleção de Atributos. Saiba mais.

Limpeza dos dados – abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos por eles representados. Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD.

Discretização - alguns algoritmos de mineração de dados, especialmente os algoritmos de classificação, requerem que os dados estejam na forma de atributos categorizados. Assim, muitas vezes é necessário transformar um atributo contínuo em categórico.

Binarização - Algoritmos utilizados para descoberta de padrões de associação requerem que os dados estejam na forma de atributos binários. Assim, muitas vezes tanto os atributos contínuos quanto os discretos necessitam ser transformados em um ou mais atributos binários.

Construção de atributos – essa operação consiste em gerar novos atributos a partir dos atributos existentes. A importância desse tipo de operação é justificada pois novos atributos, além de expressarem relacionamentos conhecidos entre atributos existentes, podem reduzir o conjunto de dados simplificando o processamento dos algoritmos de Mineração de Dados.

Transformação de variáveis - se refere a uma transformação que seja aplicada a todos os valores de um atributo. Em outras palavras, para cada objeto, a transformação é aplicada ao valor do atributo para aquele objeto. Uma transformação que podemos citar é a normalização dos dados, que consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como de -1 a 1 ou de 0 a 1. Tal ajuste se faz necessário para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa determinados métodos de Mineração de Dados.

Acesse aqui o servidor contendo a base de dados DIRETORIO do CNPq.

www.din.uem.br