Mineração de Dados **Nome:** _____
2 semestre de 2016
Relatório 4 - Pré-processamento
10/10/2016
André C. P. L. F. de Carvalho
Victor H. Barella

O objetivo desta prática é aplicar os principais conceitos sobre pré-processamento de dados vistos em aula. Espera-se que o aluno consiga limpar uma base de dados corretamente, fazer transformações nesta base e diminuir sua dimensionalidade.

Para esta prática você vai precisar:

- RapidMiner
- Editor de texto com suporte para figuras (ex: Microsoft Office Word, LibreOffice Writer, LaTex)
- Base de dados de Dyrskjøt et al sobre expressão gênica

A seguir são apresentados 4 tópicos: entendendo a base, limpeza de dados, transformação e redução de dimensionalidade. Escreva um relatório cumprindo os requisitos dos tópicos.

O relatório deve estar no formato pdf para submissão. Ao final da atividade o aluno terá uma base pré-processada. Esta base também deverá ser submetida e deverá estar no formato csv.

Em Mineração de Dados e Aprendizado de Máquina, é importante pré-processar o conjunto de dados antes de extrair qualquer conhecimento. Um esquema comum de pré-processamento é apresentado da Figura 1. Observe o esquema e a seguir aplique as etapas de pré-processamento utilizando o Rapid Miner.

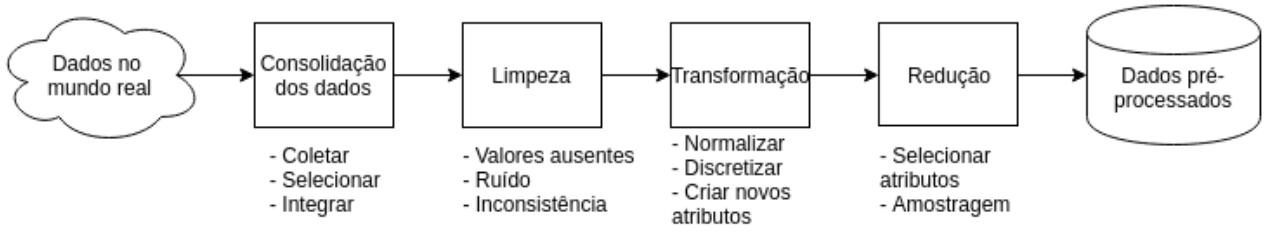

Figure 1: Esquema de pré-processamento de dados

1. **Entendendo a base**: A fim de entender melhor a base de dados, responda os seguintes questionamentos
   - Qual o contexto em que a base de dados se encontra?
   - O que cada instância/exemplo representa?

- O que os atributos representam? Quais o(s) atributo(s) alvo(s)? Quais os atributos preditivos?

2. **Limpeza de dados**:

   - Utilizando o operador *Impute Missing Values*, preencha os valores ausentes da base de dados. Mostre no relatório os parâmetros que utilizou.
     - Você deve escolher um modelo para predizer os valores ausentes. Você pode utilizar o operador *k-nn* para isso.
     - Lembre-se que se trata de uma base com alta dimensionalidade. Escolha cuidadosamente os parâmetros do operador *Impute Missing Values*.

3. **Transformação**:

   - Normalize os dados de forma que a média seja zero e o desvio padrão seja 1. Mostre no relatório os parâmetros utilizados.

4. **Redução**:

   - Remova os atributos irrelevantes. Mostre no relatório os parâmetros utilizados.
   - Remova os atributos correlacionados. Mostre no relatório os parâmetros utilizados.
   - Selecione os atributos que possuem maior correlação com o atributo alvo. Para isso, você pode utilizar os operadores *Weight by correlation* e *Select by weights*. Mostre no relatório os parâmetros utilizados.
   - Apresente no relatório a redução de dimensão alcançada.

   Utilize o operador *Write csv* para salvar a base de dados pré-processada. Ela deverá ser enviada junto com o relatório.