ANÁLISE DE FILTROS DE PRÉ-PROCESSAMENTO DE DADOS: *NORMALIZE* E *DISCRETIZE*

Júlio César Machado Álvares [(1)], Marcus Vinícius Rodrigues Campos [(1)], Marcos Roberto Ribeiro [(2)]

RESUMO

O processo de extração de conhecimento de bases de dados é um processo um tanto quanto complexo e é uma área de estudo crescente da computação atual. O processo é conhecido como KDD (*Knowledge discovery database*) e é composto por várias fases, sendo o seu conjunto, o procedimento de extração de conhecimento. A fase que será abordada no presente artigo trata-se do pré-processamento dos dados, estágio onde várias técnicas são submetidas na base de dados para preparar a mesma para os processos de mineração. Serão abordados dois filtros, *Normalize* e *Discretize*, respectivamente não-supervisionado e supervisionado.

Palavras-chave: *Data Mining*. Pré-processamento. *Normalize*. *Discretize*.

1 INTRODUÇÃO

Com diversos problemas surgindo a partir do avanço desenfreado da computação nos dias de hoje, várias técnicas são criadas todos os dias para contornar tais problemas.

Um desses é o crescimento absurdo da quantidade de dados que são gerados todos os dias. O problema criado com tal crescimento é o fato de que nem sempre as bases de dados trazem conhecimentos explícitos consigo. Assim, é necessário aplicar técnicas para extrair o conhecimento das bases *in natura*.

Para a solução de tal problema, foi proposta uma metodologia, chamada de KDD (*Knowledge discovery database*), onde a mesma trata-se de um conjunto de técnicas e ações para extrair da melhor forma possível uma quantidade de conhecimento das bases de dados.

O KDD trata-se de um processo não trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir de dados armazenados em um banco de dados (PRASS, 2016). A Figura 1 demonstra o processo completo. Pode-se notar que o início é dado a partir de uma base sem nenhum tratamento (*in natura*) e o seu resultado é o conhecimento implícito na mesma.

A área de interesse do presente trabalho está na segunda parte do processo de KDD, o pré-processamento dos dados.

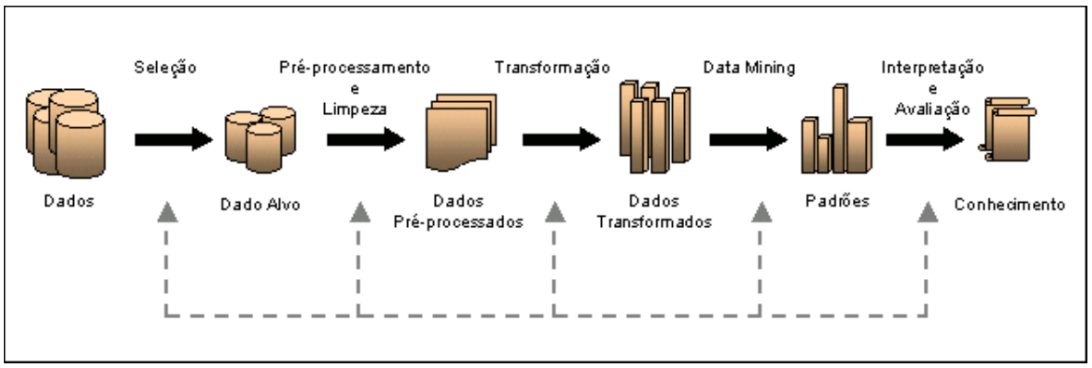

Figura 1 – Processo de *KDD* (PRASS, 2016).

1.1 Pré-Processamento

Essa parte é crucial no processo de KDD, sendo que a mesma pode impactar diretamente na qualidade dos resultados obtidos no final.

Trata-se de um conjunto de técnicas para tratar certas características presentes nos dados *in natura*, como por exemplo, deixar todos os dados do conjunto no mesmo intervalo.

Nesta etapa deverão ser realizadas algumas tarefas para eliminar dados redundantes e ou inconsistentes. Também devem ser tratados possíveis dados que sejam muito discrepantes em relação ao conjunto, tais dados são chamados de *outliers* e, além desses, pode ser aplicado algumas técnicas de redução de dimensionalidade para que a base de dados diminua de tamanho e não perca representatividade (TAN et al., 2007).

Para os testes com os filtros que serão apresentados, foi utilizado a ferramenta *Weka* (WAIKATO, 2018).

1.2 *Discretize*

Discretizar dados trata-se de transformar um atributo puramente numérico em um atributo puramente categórico. Um exemplo, transformar uma quantidade de valores aferidos de temperatura em uma cidade em valores do conjunto {frio, quente, muito quente}. A forma como cada valor será incluindo em um dos elementos do conjunto varia conforme o filtro aplicado, podendo ele ser supervisionado ou não.

Uma distinção básica entre métodos de discretização para classificação é se informações sobre classes são usadas ou não (TAN et al., 2007). Quando não é levado em conta informações sobre as classes, faz-se o uso de abordagens relativamente simples, como a igualdade de largura ou frequência. Todavia, quando a classe é levada em conta, a tarefa torna-se não trivial. Abordagens estatísticas se mostram mais relevantes, e, entre elas, as que utilizam do cálculo de entropia, as mais promissoras.

1.3 *Normalize*

Quando trabalhamos com dois ou mais atributos torna-se necessário a normalização dos valores de tais atributos, especialmente em casos onde os valores tem escalas muito diferentes (ZAKI; JR; MEIRA, 2014).

O tipo mais comum é a *Range normalization*, que leva em consideração os intervalos da amostra para calcular cada um dos valores das *features*. Tal algoritmo deixa os valores dentro de um mesmo intervalo, mesmo que tais valores apresentem intervalos iniciais muito divergentes. No caso do *Range normalization*, o intervalo admitido é $[-1, 1]$.

O modelo matemático do algoritmo é apresentado a seguir.

$$x_i^{'} = \frac{x_i - min_i\{x_i\}}{max_i\{x_i\} - min_i\{x_i\}}$$

2 DESENVOLVIMENTO

Para os testes desenvolvidos no presente trabalho, primeiramente foram escolhidas duas bases de dados que fossem cabíveis de aplicação dos filtros. As bases foram retiradas do site *UCI Machine Learning* (LEARNING; SYSTEMS, 2018b).

Primeiramente, a base que será aplicada a Discretização chama-se *Vertebral Column Data Set* e se trata de um conjunto de dados sobre doenças da coluna vertebral dos pacientes do Centro Médico Cirurgico de Massues, na França (LEARNING; SYSTEMS, 2018c).

Para a aplicação do filtro de normalização, foi escolhida uma base de dados do mesmo repositório. A base de dados chama-se *Hill-Valley Data Set* e trata-se de um conjunto de pontos de um plano cartesiano de duas dimensões, onde cada um dos indivíduos apresenta um vale ou um pico quando plotados em um gráfico (LEARNING; SYSTEMS, 2018a).

Todos os testes foram desenvolvidos no software *Weka* e os mesmos tratam-se de uma análise dos atributos antes e depois da aplicação dos filtros.

3 RESULTADOS E DISCUSSÃO DOS RESULTADOS

Ao observar a base de dados *Vertebral Column*, notamos que suas *features* tem uma quantidade significativa de classes, estando elas dentro de um intervalo determinado.

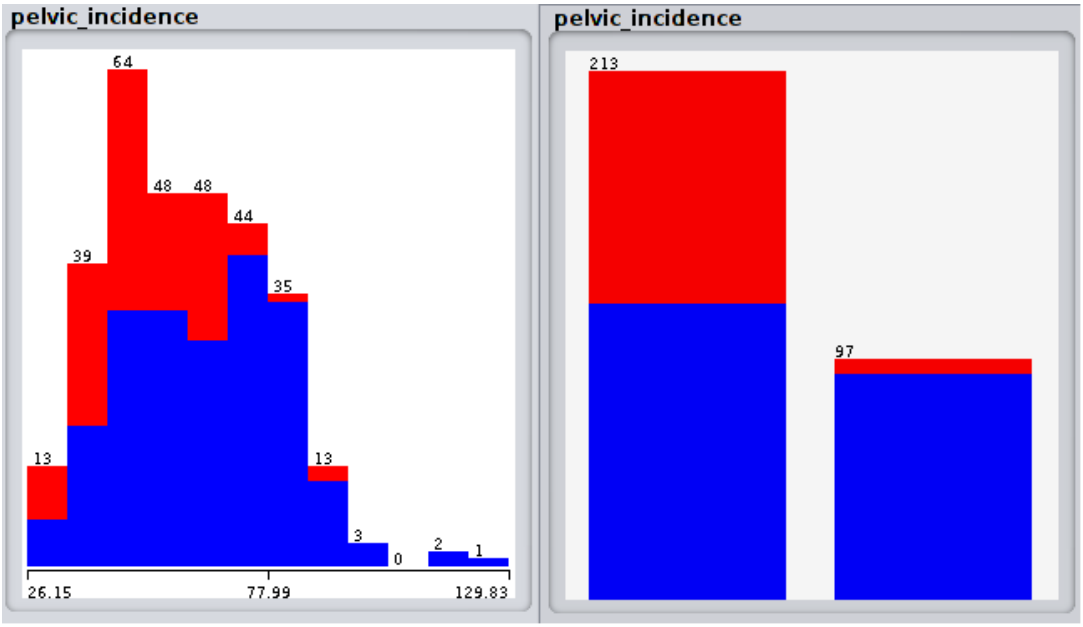

Figura 2 – Atributo *pelvic incidence* da base de dados *in natura* e após discretização.

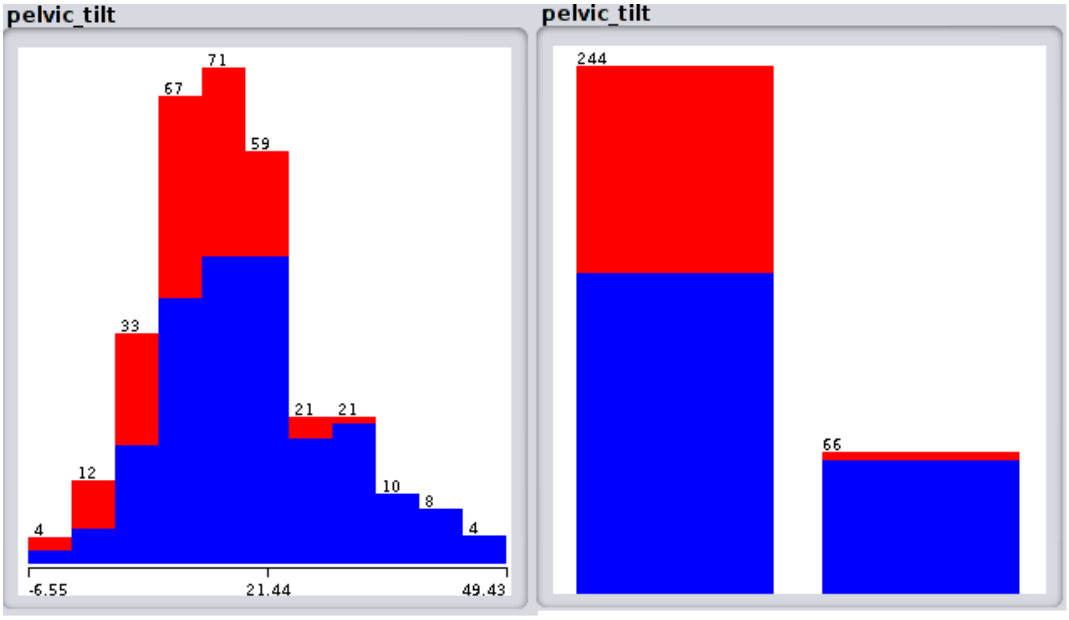

Figura 3 – Atributo *pelvic tilt* da base de dados *in natura* e após discretização.

Como é possível observar na figura 2, o atributo *pelvic incidence* apresenta 12 intervalos para o atributo e, após a aplicação do filtro, tais intervalos foram reduzidos em apenas 2.

Já a figura 3 apresenta 11 intervalos inicialmente e, após a aplicação do filtro, apenas 2. Tal comportamento é apresentado por todas as *features* do conjunto.

Portanto, após a discretização do conjunto, os dados estão mais aptos a serem submetidos a um classificador de extração de regras, como uma *Decision Tree*. Tal afirmação deve-se ao fato de que o filtro reduziu a dimensão do intervalo do conjunto (transformando-

os em atributos categóricos), assim, reduzindo a quantidade de regras que serão necessárias para modelar um classificador a partir dos dados. Também espera-se que não seja perdido a representatividade dos dados no conjunto.

Ao observar a base de dados *Hill-Valley* sem a normalização, notamos que os intervalos dos valores são absurdos. Dentre todos os atributos do conjunto, o intervalo encontrado é $[0, 164627]$.

Após a aplicação do filtro *Normalize* com o algoritmo *Range Normalization*, o intervalo encontrado está entre $[0, 1]$, devido a ajustes do algoritmo no *Weka*.

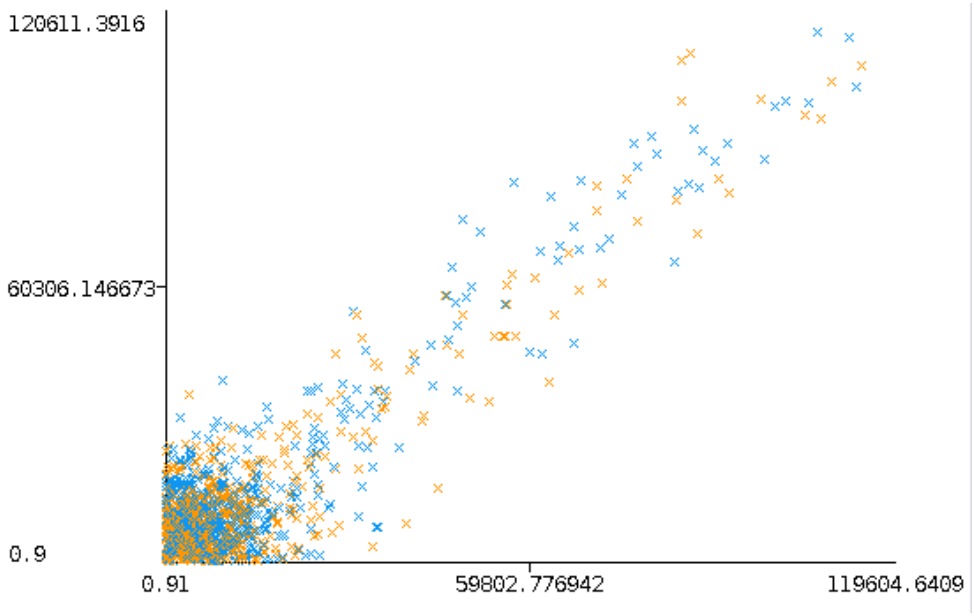

Figura 4 – Atributo x_1/x_{100} da base de dados sem aplicação de filtros.

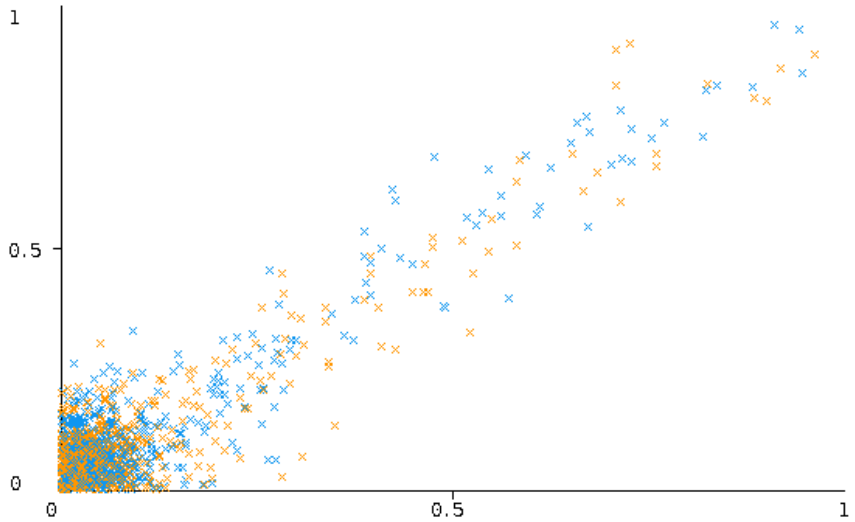

Figura 5 – Atributo x_1/x_{100} da base de dados após aplicação normalização.

Como pode-se observar na figura 4, que se trata de um gráfico de disperção de x_1

por x_{100}, os dados apresentam densidade muito alta perto da origem e também, intervalos muito grandes.

Já na figura 5, após a aplicação do filtro, nota-se que a configuração de densidade dos atributos se manteve e, os intervalos agora estão constantes entre $[0, 1]$. Também é possível notar leves diferenças entre as distâncias dos pontos entre as figuras. Tal problema é ocasionado porque os limites na figura 4 não são iguais como os da figura 5, assim, a distribuição é levemente diferente, mas mantém a mesma representatividade.

Com tais afirmações, é possível concluir que a base de dados está mais apta para ser submetida a modelos de classificação baseados em distância entre pontos ou classificadores como a *Multilayer Perceptron* e a *Support Vector Machine*, que traçam funções para separar as classes dos dados.

4 CONCLUSÃO

Primeiramente podemos concluir que os objetivos do trabalho foram alcançados, trazendo a nós um maior entendimento sobre o processo de KDD, com ênfase no pré-processamento, utilizando os filtros *Discretize* e *Normalize*.

Pudemos também, perceber a importância de trade-offs tais como: dado um *database* qualquer, diminuir sua quantidade de instâncias, sem perder a representatividade, para usar algoritmos computacionalmente mais custosos. Essa visão mais ampla, de saber o que é mais adequado usar, é a chave da etapa de pré-processamento de dados.

REFERÊNCIAS

LEARNING, C. for M.; SYSTEMS, U. o. C. I. **Hill-Valley Data Set**. 2018. <https://archive.ics.uci.edu/ml/datasets/Hill-Valley>. Accessed: 2018-09-15.

_____. **UCI Machine Learning Repository**. 2018. <http://archive.ics.uci.edu/ml/index.php>. Accessed: 2018-09-15.

_____. **Vertebral Column Data Set**. 2018. <http://archive.ics.uci.edu/ml/datasets/vertebral+column>. Accessed: 2018-09-15.

PRASS, F. S. Kdd–uma visão geral do processo. **Recuperado em**, v. 15, 2016.

TAN, P.-N. et al. **Introduction to data mining**. [S.l.]: Pearson Education India, 2007.

WAIKATO, U. of. **Weka 3: Data Mining Software in Java**. 2018. <https://www.cs.waikato.ac.nz/ml/weka/index.html>. Accessed: 2018-09-15.

ZAKI, M. J.; JR, W. M.; MEIRA, W. **Data mining and analysis: fundamental concepts and algorithms**. [S.l.]: Cambridge University Press, 2014.

ANALYSIS OF PREPROCESSING FILTERS: NORMALIZE AND DISCRETIZE

ABSTRACT

The process of extracting knowledge from databases is a rather complex process and is a growing area of study of current computing. The process is known as KDD (Knowledge Discovery database) and is composed of several phases, the whole being the procedure of knowledge extraction. The phase that will be addressed in this article is the pre-processing of the data, stage where several techniques are submitted in the database to prepare the same for the mining processes. Two filters, Normalize and Discretize, respectively unsupervised and supervised, will be addressed.

Keywords: Data Mining. Pré-Processing. Discretize. Normalize.