Pré-processamento de Dados

Inteligência computacional

Gabriel de Souza Rosa
Design de gestão e informática
CEFET-MG - Campus V
Divinópolis, Brasil
gabrieldesouza2305@gmail.com

Gustavo Rodrigues Barcelos
Design de gestão e informática
CEFET-MG - Campus V
Divinópolis, Brasil
gustavo05rb@gmail.com

Sávio Rodrigues
Design de gestão e informática
CEFET-MG - Campus V
Divinópolis, Brasil
saviorodrigues012@gmail.com

***Resumo*—O presente trabalho visa analisar três diferentes bases de dados selecionadas a partir de atributos previamente estipulados. A partir dos datasets, foi realizada a aplicação de técnicas de pré-processamento de dados, além do treinamento e teste de algoritmos tanto de classificação quanto de regressão.**

***Palavras-chave*—Dataset, regressão, classificação, pre-processamento, análise de dados.**

I. Introdução

O pré-processamento de dados é um conjunto de atividades que são feitas sobre um dataset cujo objetivo é sua organização, preparação e estruturação. Nesse sentido, com o intuito de aplicação de técnicas de análise e predições e sabendo da importância da etapa de pré-processamento, o presente trabalho aplica essas atividades em três diferentes bases de dados, são elas : Adult Data Set (Classificação), Airfoil Self-Noise Data Set (Regressão) e um conjunto de dados de email classificados como spam e não spam.

A. Adult Dataset

A primeira base de dados, Adult Dataset [1], se refere a uma extração feita por Barry Becker do banco de dados do Census (Departamento do Censo dos Estados Unidos) no ano de 1994. A tarefa associada a esse dataset é uma classificação binária na qual deve-se prever se um indivíduo ganha mais que que cinquenta mil dólares por ano ou não.

Para realização dessa tarefa o dataset conta com dados pessoais anonimizados como: idade, informações de ativos, classes trabalhadora, estado civil etc... Com isso, o conjunto de dados pode ser classificado como multivariável, com quinze atributos, trinta e dois mil quinhentos e sessenta e um registros, com valores ausentes e referente a assuntos sociais.

B. Airfoil Self-Noise

A segunda base de dados, Airfoil Self-Noise [2] se trata de um conjunto de dados da NASA o qual compreende aerofólios modelo NACA 0012, desenvolvido na década de 1930 pelo Comitê Nacional para aconselhamento aeronáutica. Esses equipamentos são usados no setor automobilístico e principalmente na aeronáutica cujo objetivo é provocar uma variação na direção da velocidade do fluido (ar) gerando uma força. Em aeronaves, essa força proporciona a sustentação e permite que realize o voo, por exemplo.

O dataset possui um conjunto multivariável, seis atributos reais sem valores ausentes e a tarefa associada é a regressão. Os experimentos foram feitos com variantes do tamanho em diferentes velocidades de vento e ângulos. Por outro lado, atributos que não foram mudados durante a coleta de dados foram o vão do aerofólio e a posição do observador, e apenas uma saída - o nível de pressão sonora. Ademais, todas as etapas de pre-processamento foram aplicadas na base de dados de forma a conseguir resultados consistentes para analisar e predizer resultados.

C. Emails dataset - Spam x Ham

Por fim, a terceira base de dados consiste em dados brutos obtidos nos anos de 2004 e 2005 através do repositório assassin [3]. No total, foram coletados 3539 emails para realizar os experimentos. O objetivo principal desta análise é utilizar métodos de ciência de dados e inteligência artificial para classificar emails em dois grupos, spam ou ham (e-mail verídico).

Para realizar tal feito, foi-se utilizado a linguagem de programação python com o intuito de extrair desses dados brutos, informações relevantes para a classificação. Gerando um dataset multivariável com 41123 atributos inteiros sem valores ausentes.

Aplicando metodologias estatísticas disponíveis no software matlab, foi possível reduzir o dataset para 60 atributos mais relevantes. Além do mais, foram aplicadas as etapas de pré-processamento cabíveis a base de dados com o intuito de facilitar a analise e predição dos resultados.

II. Desenvolvimento

A. Adult Dataset

Para aplicação de modelos de classificação no dataset, é necessários que os dados passem antes por um tratamento e uma análise, etapa conhecida como pré-processamento de dados. A realização do pré-processamento é crucial para que as funções de machine learning receba os parâmetros corretos e relevantes para uma boa análise.

Identify applicable funding agency here. If none, delete this.

1) Eliminação de atributos irrelevantes e redundantes: Dentre os quinze atributos, três foram removidos com base em uma análise visual por não terem relevância ou serem redundantes. O atributo ”fnlwgt” corresponde há um peso para cada indivíduo que leva em conta sua escolaridade, raça, sexo, etc... Porém, não há uma padronização nos pesos e a mesma pessoa pode receber pesos diferentes em diferentes estados. Por isso, esse atributo foi considerado como irrelevante. Já os atributos ”educatioin-class” e ”relationship” são atributos redundantes que podem ser obtido pela combinação de outros atributos no conjunto de dados.

2) Tratamento de valores ausentes: O conjunto de dados possui quatro mil duzentos e sessenta e dois valores ausentes distribuído em três atributos categóricos que são: ”workclass”, ”ocupation” e ”native_country”. Devido ao fato de todos os atributos serem categóricos foi realizado uma imputação por valor e onde havia valores nulos, foi acrescentado uma nova classe denominada: ”nao_informado”.

3) Discretização dos dados: Por fim, foi realizado a discretização de todos os dados categóricos em dados inteiros a fim de adequar os dados de entradas para funções de machine learning. Para realizar esse processo cada categoria de atributo foi substituído por um índice, por exemplo: no atributo ”workclass” a categoria ”State-gov” foi substituída pelo índice um e ”Private” pelo índice dois.

B. Airfoil Self-noise

Os atributos dos aerofólios, seja em uma aeronave ou em um carro de corrida, são fundamentais estarem ajustados para que seja possível cumprir seu objetivo da melhor forma. Isso acontece, pois eles modificam seu funcionamento e consequentemente seu resultado. Toda a coleta de dados foi realizada com o intuito de relacionar todos esses atributos e observar a saída de nível de pressão sonora, sendo um total de 1503 instâncias.

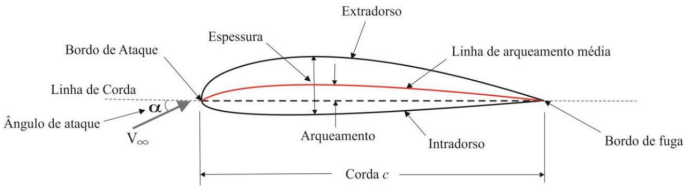

Fig. 1. Atributos de um aerofólio

A fig. 1 mostra alguns dos atributos levados em consideração na construção da base de dados. Todas as características analisadas são:

- Frequência.
- Ângulo de ataque.
- comprimento da corda.
- Velocidade de fluxo livre.
- Espessura do deslocamento lateral de sucção.
- Nível de pressão sonora.

Foi feita uma análise dos dados em linguagem python utilizando a ferramenta google colab. Nessa etapa constatou-se que não havia nenhum dado nulo e consequentemente, não seria necessário nenhum tipo de imputação. Além disso, através de uma análise gráfica, todos os atributos considerados tem relação direta com o resultado de saída, não possuindo nenhum sem significância. Portanto, concluiu-se que a base, por si só, já está ”pronta” para aplicação de algoritmos de regressão para realiza predições de dados.

C. Emails dataset - Spam x Ham

Antes de realizar a aplicação do dataset a algum algoritmo de classificação, é necessário realizar a extração e pré-processamento dos dados que se dá da seguinte forma:

1) Extração dos dados: Tendo sido realizada através da linguagem de programação python, está etapa consiste em coletar o conjunto de dados brutos e organiza-los de forma tabular através dos seguintes passos:

- Tokenização: leitura de todos os emails, verificando palavra por palavra a fim de realizar a contagem da frequência de cada uma em um determinado e-mail, gerando uma tabela de informação com 3539 instâncias e 41123 atributos de entradas
- Discretização: definição de valores inteiros para representar cada palavra encontrada nos emails. Para isso, foi-se utilizada uma estrutura chave valor, o qual suposto índice de cada palavra na estrutura passou a representa-la.
- Armazenamento: geração de duas tabelas no formato ’.csv’, sendo, a primeira com a relação entre índices e palavras, e a segunda, uma tabela de correlação entre as instâncias e os atributos.

2) Pré processamento: Sendo realizada em matlab, está etapa consiste em preparar e filtrar os dados para o algoritmo de classificação através dos seguintes passos:

- Amostragem: realizou-se uma amostragem de forma sistemática, considerando somente as 300 palavras que mais aparecem no conjunto de emails.
- Seleção: para selecionar os sessenta melhores atributos, utilizou-se um método estatístico (fscchi2) que compara os resultados observados com os resultados esperados e busca uma relação entre as variáveis estudadas.
- Remoção: os dados repetidos foram removidos do conjunto de treinamento a fim de evitar a super saturação do modelo. Excluindo um total de quinhentos e quatro amostras repetidas.
- Normalização: a normalização dos dados foi realizada na escala de 0 a 1 através do método min-max, que realiza a diferença do dado amostral pelo menor valor do conjunto dividido pela amplitude da classe de determinado atributo. Tal que as palavras que tiveram maior frequência se aproximaram do valor 1 e as com menores frequência se aproximaram do valor 0.

Através da figura [2] é possível observar a distribuição de frequência das palavras no conjunto de dados, havendo uma grande presença de artigos nas duas classes de e-mail. Outra observação que se torna possível, é realizada através da figura [3] que exibe a densidade de cada e-mail em relação aos atributos selecionados, indicando a concentração das amostras

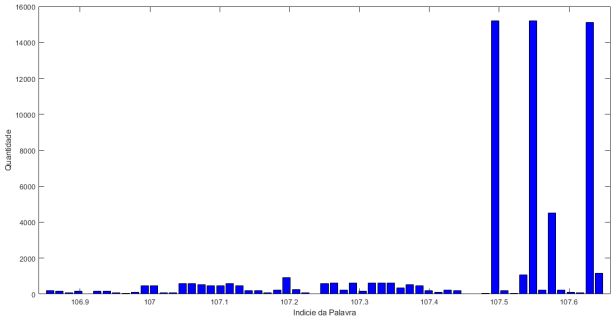

Fig. 2. Distribuição de frequência das palavras

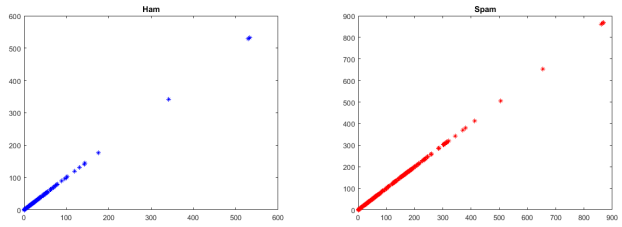

Fig. 3. Distribuição de densidade das palavras selecionadas em cada e-mail

em razão da quantidade de palavras que cada classe de e-mail contém.

REFERENCES

[1] Adult. 1996. UCI Machine Learning Repository.

[2] BROOKS THOMAS, P. D. . M. M. *Airfoil Self-Noise*. 2014. UCI Machine Learning Repository.

[3] SPAM Assassin Dataset.(2002) Apache SpamAssassin Disponivel em. Disponível em: <https://spamassassin.apache.org/old/publiccorpus/>.