Análise e redução de preditores do Glass Identification Dataset

*P*aulo Victor Felício Lage
Dept. Pós-Graduação em Engenharia De Teleinformática
Universidade Federal do Ceará
Campus do Pici - Fortaleza - Brasil
victorlage7@gmail.com

Talles Felix Gomes
Depto. de Engenharia de Teleinformática
Universidade Federal do Ceará
Campus do Pici - Fortaleza - Brasil
tallesfelixg95@gmail.com

***Resumo*—Este artigo aborda técnicas de pré-processamento aplicadas ao conjunto de dados Glass Identification[5] e análise do método *Principal Component Analysis* (PCA).**
Após a aquisição dos dados, um dos primeiros passos é a utilização de tecnicas de pré-processamento para se ter um melhor entendimento do conjunto de dados que está sendo trabalhado. Analises estatisticas como o calculo das medias, desvio padrão e *skewness*(assimetria), podem revelar diversos padrões sobre os dados. Além disso, a correlação e a covariancia entre os preditores, também podem nos guiar para outros métodos de pré-processamento, como o *Principal Component Analyses*(PCA), que foi utilizado nesse estudo para reduzir o conjunto de dados, retendo a maior quantidade possivel de informações.

***Index Terms*—pré-processamento, PCA, dados**

I. Introdução

As técnicas de pré-processamento de dados geralmente se referem à adição, exclusão ou transformação de um conjunto de dados. Essas técnicas podem ser necessárias por vários motivos como reduzir o impacto da distorção ou discrepância de dados bem como podem levar a melhorias significativas no desempenho. Alguns tipos de modelagem podem ter requisitos rígidos, como os preditores devem ter uma escala comum. A criação de um bom modelo pode ser difícil devido a características específicas dos dados como os outliers. Modelos diferentes têm diferentes sensibilidades ao tipo de predição.

Quanto maior o número de preditores é necessário um melhor nível de entendimento das características desses preditores bem como eles se relacionam. Essas características podem sugerir etapas importantes e necessárias de pré-processamento que devem ser tomadas antes da construção de um modelo. A necessidade de pré-processamento de dados é determinada pelo tipo de modelo que está sendo usado. Alguns procedimentos, como modelos baseados em árvore, são notavelmente insensíveis às características dos dados do preditor. Outros, como a regressão linear, não são.

Hill et al. (2007) [1] descrevem um projeto de pesquisa que utilizou triagem de alto conteúdo para medir vários aspectos das células. Observou-se que o software utilizado para determinar a localização e o formato da célula teve dificuldade em segmentar as células, um total de 2019 amostras foram utilizadas onde 1.300 foram consideradas mal segmentadas e 719 foram bem segmentadas. Para todas as células, 116 preditores foram medidos e utilizados para prever a qualidade de segmentação das células.

Dentre as técnicas de pré-processamento dos dados temos as técnicas de Centralizar e Escalonar os dados, Transformação dos dados para resolver *Skewness* (Assimetria), Transformação para resolver *Outliers* e Redução de dados e extração de *Feature* como por exemplo a tecnica *Principal Component Analysis* (PCA).

II. Métodos

O conjunto de dados utilizado possui 214 amostras de Tipos de Vidro, classificados em sete classes, porém a classe do tipo quatro (4) não possui nenhum exemplar em todo o conjunto de dados. Para cada amostra existem nove (9) variáveis preditores, que são compostos por: *Refractive Index* (RI), Sódio (Na), Magnésio (Mg), Alumínio (Al), Silício (Si), Potássio (K), Cálcio (Ca), Bárium (Ba) e Ferro (Fe). Todas as análises gráficas não adicionadas nesse artigo estão disponíveis para acesso [2].

A. Análise mono-variável Não Condicional

No primeiro momento foi realizado uma análise mono-variável não condicional para cada um dos preditores e para uma melhor análise foi desenvolvido um histograma para a possível observação de ocorrências dos preditores Figura1 assim como foram calculadas suas médias, desvio padrão e *skewness* respectivamente Figura 2, 3 e 4.

À medida que a *skewness* se torna mais inclinada à direita, a probabilidade de *skewness* se torna maior. Da mesma forma, à medida que a distribuição fica mais inclinada à esquerda, o valor se torna negativo.

A partir da Figura 1 é possível perceber que nos elementos Potássio e Bário há uma maior concentração de amostragem de dados em valores relativamente pequenos e pequeno número de valores grandes, caracterizando-se dessa forma como uma *right skewness.* Por outro lado o elemento Magnésio apresenta uma caracteristica diferente dos elementos mencionados anteriormente dessa forma se caracterizando como uma *left skewness.* Os elementos Sódio e Alumínio foram os elementos que apresentaram o valor de *skewness* mais proximo de zero, dessa formar tendo maior probabilidade de uma distribuição semelhante.

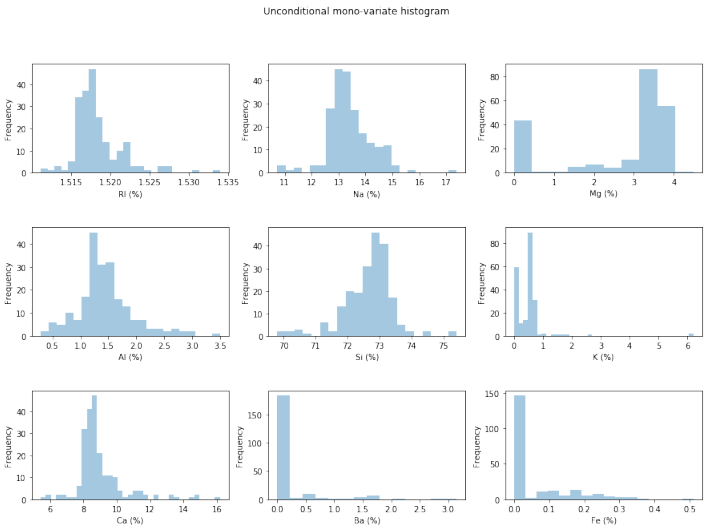

Figura 1. Histogramas monovariável não condicional.

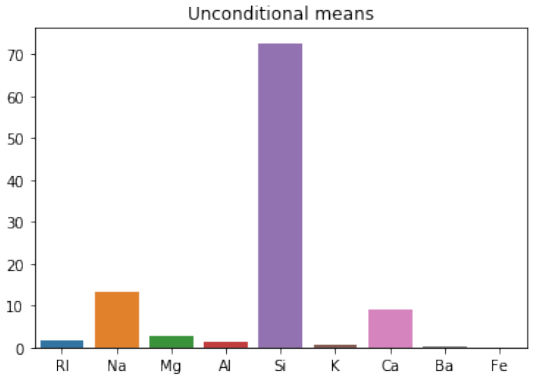

Figura 2. Média monovariável não condicional

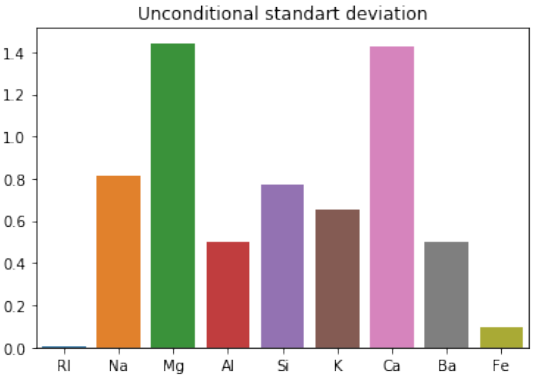

Figura 3. Desvio Padrão monovariável não condicional

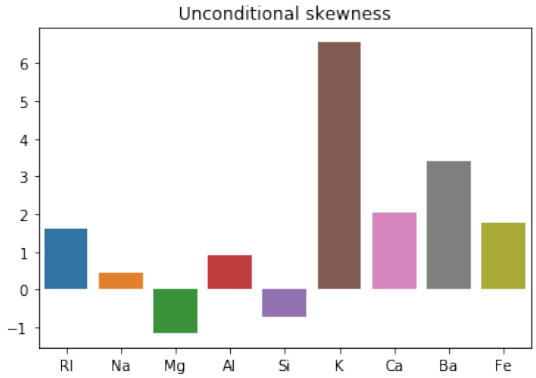

Figura 4. Skewness monovariável não condicional

B. *Análise mono-variável Não Condicional*

Nessa etapa foram realizado os mesmos procedimentos da etapa anterior agrupando os dados para cada uma da 7 classificações, assim gerando para cada classe seus Histogramas bem como calculadas suas características como médias, desvio padrão e *skewness*. Durante essa etapa pode-se perceber que a classe 4 não possui amostras dentro do conjuto de dados.

C. *Análise Bivariável Não Condicional*

Para entender do ponto de vista estatístico como um par de variáveis está relacionado, temos algumas possibilidades como analisar graficamente por meio de um gráfico de dispersão entre os pares de preditores bem como uma análise numérica que pode ser realizada através da Matriz de coeficiente de correlação.

Esses tipos de correlação podem ser, positiva que ocorre quando um atributo tende a aumentar ou diminuir, o outro também tende a aumentar ou diminuir e negativa que ocorre quando um atributo tende a aumentar ou diminuir, o outro possui uma tendência inversa ou nula onde não existe um padrão definido para este tipo de tendência.

A correlação tem como valores limites 1 e -1. Quanto mais próximo de 1 maior é a correlação positiva, quanto mais próximo de -1 maior é a correlação negativa.

Analisando a Figura 6 podemos perceber os elementos mais correlacionados positivamente Ca X RI com um valor de correlação de 0.81 e os elementos com mais relacionados negativamente RI x Si.

Na Figura 5 podemos perceber graficamente que a Relação Ca X RI é positiva e a relação RI x Si é negativa.

D. *Análise Multivariável*

As técnicas de redução de dados são transformações que tem como objetivo reduzir o número de preditores, gerando um conjunto menor de preditores que procuram capturar a maioria das informações nas variáveis originais, dessa forma uma quantidade menor de variáveis podem ser usadas, fornecendo assim uma razoável fidelidade dos dados originais. Para a maioria das técnicas de redução de dados, os novos preditores são funções dos preditores originais e todos os

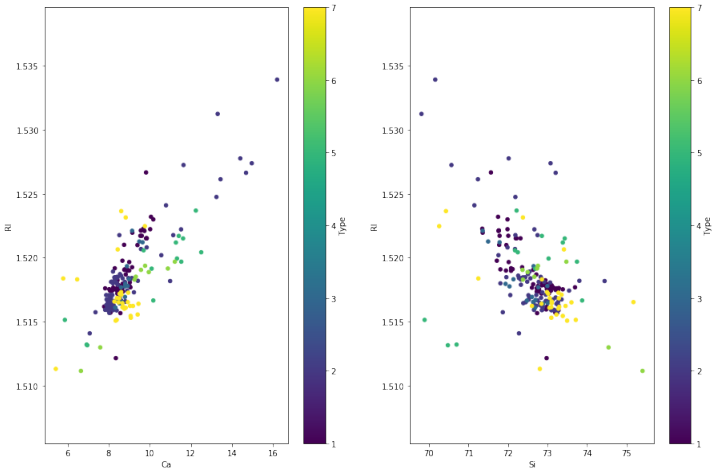

Figura 5. Análise Bivariável Não Condicional

Figura 6. Matriz de correlação

preditores originais ainda são necessários para criar as variáveis substitutas. A técnica aplicada nessa análise foi *Principal Component Analysis* (PCA). O primeiro principal componente (PC) é definido como a combinação linear dos preditores que captura a maior variabilidade de todas as combinações lineares possíveis. Em seguida, os PCs subseqüentes são derivados de modo que essas combinações lineares capturem a maior variabilidade restante, além de não serem correlacionados com todos os PCs anteriores.

A principal vantagem do PCA é que o mesmo cria componentes não correlacionados

A ultima análise desse trabalho consiste em aplicar o PCA para encontra as duas principais componentes PC1 e PC2.

Resultados

E. Análise do conjunto de dados

A partir das médias mono-variáveis foi identificado que todas as amostras possuem como elemento principal o Silício, sendo ele, o principal componente da formação do vidro, compondo aproximadamente 70% da composição das amostras. Os outros elementos encontrados implicam características diferentes de acordo com a sua concentração em cada amostra. Por exemplo, de acordo com Pires et al. (2009) [3] amostras de vidro com um maior teor de Alumínio em sua composição implicam em um aumento da resistência a choques mecânicos, de acordo também com a matriz de correlação na Figura 6, está inversamente correlacionado ao índice de refração (RI) o que implica que quanto maior a concentração de alumínio, menor o índice de refração da luz na amostra de vidro.

Podemos classificar as amostras por tipos, sendo que os elementos encontrados em cada tipo implicam em diversas características. Nas amostras do tipo 1, 2 e 3, foi encontrado uma maior concentração de magnésio o que segundo Kuryaeva (2009) [4] implica em uma maior resistência a mudanças bruscas de temperatura. Todas os tipos de amostras possuem um teor de cálcio dentro da mesma média, pois assim como o Silício, é um dos principais componentes do vidro, sendo responsável por impedir que os átomos de óxido de silício se reorganizem corretamente em cristais de areia, também possuindo uma forte correlação positiva com o índice de refração.

F. Resultados do PCA

Após a utilização do método PCA foram extraídas nove componentes principais, onde cada uma descreve uma certa quantidade das informações que existia nos dados antes da aplicação do método. De acordo com a Tabela I, PC1 e PC2 descrevem juntos, 50,68% dos dados originais, ou seja, usando apenas as duas primeiras componentes principais, 50% da variância dos dados foi perdida ao tentar descrever o dataset original. Esses valores estão diretamente relacionados ao quão correlacionados os dados estão, então quanto maior a correlação entre eles, mais os primeiros componentes irão reter informação sobre os dados.

Tabela I
PERCENTUAL DE VARIÂNCIA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
%	27.90	22.77	15.60	12.86	10.15	5.86	4.09	0.709	0.01

Mesmo com apenas 50% de representação da variância a partir de PC1 e PC2, aumentar indefinidamente o número de componentes principais não significa melhores resultados sempre. Com o aumento do número de componentes, podemos voltar ao mesmo problema inicial, onde trabalhávamos com uma grande quantidade de preditores e de acordo com a Figura 7, a partir do PC5, adicionar outras componentes não significa um aumento considerável na representação da variância dos dados.

Segundo alguns modelos preferem dados com uma baixa correlação para encontrar as soluções e melhorar a estabilidade numérica. A Figura 8 mostra que após a utilização do método, é removida a correlação entre as componentes.

Para verificar a qualidade dos dados adquiridos com o PCA, analisamos PC1 e PC2, onde foi verificado que existe uma

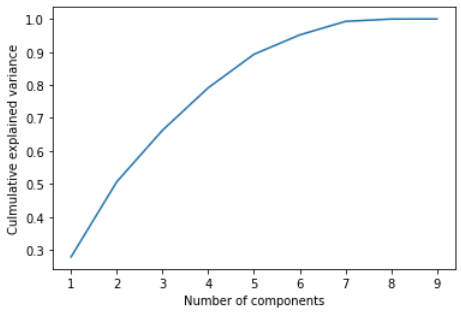

Figura 7. Variância acumulada

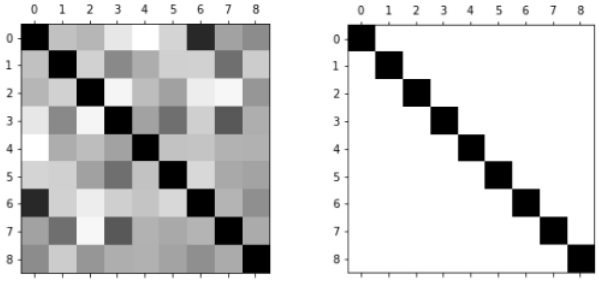

Figura 8. Matrix de correlação antes e depois do PCA

grande concentração das amostras do tipo 1,2 e 3 no centro, mostrando que estes não contribuem com muita informação para as primeiras componentes principais. Já no caso do tipo 7, conseguimos ver uma clusterização dos dados, o que significa que o método consegue representar bem esses dados.

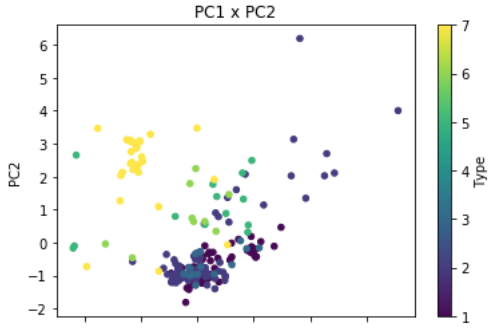

Figura 9. PC1 x PC2

REFERÊNCIAS

[1] A. Hill, P. LaPan, Y. Li. e S. Haney, Impact of Image Segmentation on High Content Screening Data Quality for SK-BR-3 Cells, BMC Bioinformatics, 8: 340, 2007.

[2] Métodos - Desenvolvimento em Python e R[https://github.com/victorlage7/ICA-HW01]

[3] Pires, R. A.; Abrahams, I.; Nunes, T. G.; Hawkes, G. E. The role of alumina in aluminoborosilicate glasses for use in glass–ionomer cements. Journal of Materials Chemistry, 2009.

[4] Kuryaeva, R. G. The state of magnesium in silicate glasses and melts. Glass Physics and Chemistry, 2009.

[5] D. Dua e E. Karra Taniskidou, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], University of California, School of Information and Computer Science, Irvine, CA, 2017.