ESTUDO SOBRE TÉCNICAS DE PRÉ-PROCESSAMENTO PARA MELHOR VISUALIZAÇÃO E UTILIZAÇÃO DE DADOS

Rômulo Freire Férrer Filho[1] and Marcus Vinicius L M Andrade [2] and Pablo Roberto Grisi[3] and Guilherme Alves[4] and Rhaniel Magalhães Xavier [5]

Universidade Federal do Ceará - Departamento de Engenharia de Teleinformática
R. Cinco, 100 - Pres. Kennedy, Fortaleza - CE, 60355-636 - Brasil

Resumo. Desenvolver aplicações que causem impacto na sociedade costuma requerer a utilização de dados reais, porém é muito raro tais dados chegarem até os desenvolvedores de forma perfeita e prontos para o uso. Faz-se então necessário a aplicação de técnicas de pré-processamento, em cima do conjunto de dados a ser trabalhado, com a finalidade de se obter informações concisas, sem redundância, fazendo com que a aplicação final tenha um ótimo rendimento. Nesse trabalho estão apresentadas algumas técnicas possíveis de serem aplicadas em um determinado conjunto de dados utilizando a linguagem R. Ao final, tais técnicas propiciaram uma redução considerável na quantidade de dados à serem trabalhados, tornando possível uma aplicação mais efetiva e menos custosa.

1 Introdução

O pré-processamento de dados, objeto de estudo deste trabalho, consiste na preparação dos dados para uso em projetos computacionais na grande área da Ciência dos Dados. Durante o processo de coleta de dados, diversas falhas costumam ocorrer, pois os dados do "mundo real"são, em geral, inconsistentes, incompletos, por exemplo tabelas onde faltam dados em algumas linhas. Também pode-se verificar dados "sujos", contendo *outliers* ou erros.

Por estes motivos, faz-se necessário o trabalho de pré-processamento dos dados antes de aplicá-los em projetos, como aprendizado de máquina e reconhecimento de padrões. Este trabalho pode ser dividido em diferentes passos, são estes: *Data Cleaning, Data Integration, Data Transformation, Data Reduction* e *Data Discretization*.

O processo de limpeza dos dados consiste em remover ou preencher valores em falta, diminuir a "sujeira"dos dados e resolver inconsistências. Em seguida, o processo de integração dos dados é realizado e consiste em integrar dados diferentes e resolver os possíveis conflitos entre estes. Posteriormente tem-se a transformação dos dados, por meio da normalização, agregação e generalização. Prossegue-se então com a redução, para obter-se uma representação fiel dos dados, porém em tamanho menor, para facilitar o processamento e enfim, tem-se a discretização dos dados.

2 Metodologia

2.1 Unconditional mono-variate analysis

Primeiramente, foram plotados histogramas das concentrações de cada elemento químico e do índice de refração em cada amostra. Além disso, também foram calculados: média, desvio padrão e *skewness* de cada preditor. Os cálculos e plots foram feitos utilizando a linguagem **R**; o script e os dados gerados estão anexados ao trabalho. A Tabela 1 mostra o resultado dos cálculos. A Figura 1 mostra um dos histogramas, para o índice de refração(RI), enquanto a Figura 2 mostra

o histograma para o elemento Si. Todos os histogramas mencionados acima estão anexados ao trabalho e podem ser encontrados no diretório:

$$/Figures/unconditional$$

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Mean	1.518365	13.40785	2.684533	1.444907	72.650935	0.497056	8.956963	0.175047	0.057009
SD	0.003037	0.816604	1.442408	0.49927	0.774546	0.652192	1.423153	0.497219	0.097439
Skewness	1.602715	0.447834	-1.136452	0.89461	-0.720239	6.460089	2.018446	3.36868	1.729811

Tabela 1: Tabela que mostra o resultado dos cálculos de Média, Desvio Padrão e Skewness

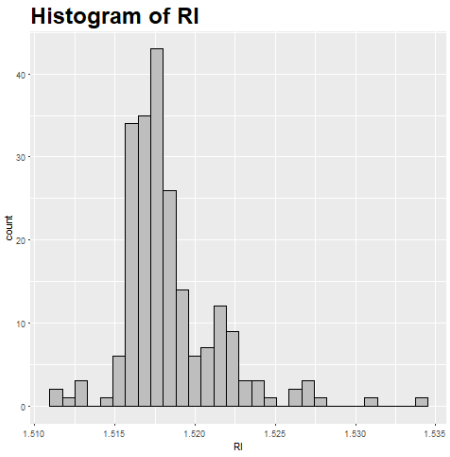

Figura 1: Histograma que mostra a distribuição da concentração de RI nas amostras

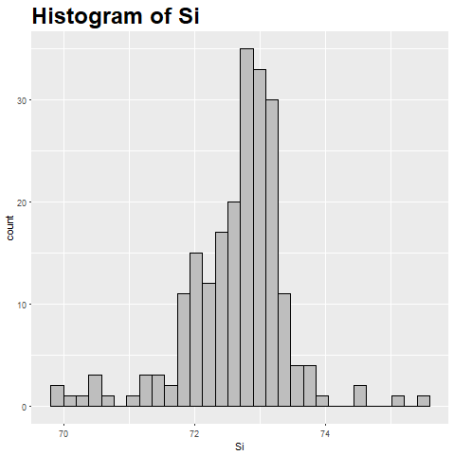

Figura 2: Histograma que mostra a distribuição da concentração de Si nas amostras

Agora, é possível extrair algumas informações dos dados. Por exemplo, a Tabela 1 mostra que a média, em porcentagem, ludo RI é de, aproximadamente, 1.52%. Mostra também um desvio padrão pequeno, de aproximadamente 0.003. Com isso, é possível afirmar que a maioria das ocorrências de RI possuem valores próximos à sua média. Por último, temos informação sobre a *skewness* de RI, por ser positiva, ela nos mostra que há uma maior distribuição de amostras na parte esquerda do histograma. Com a Figura 1, é possível verificar e validar todas essas informações.

Também é possível fazer essa análise para os outros preditores. O Si, por exemplo, possui desvio padrão aproximadamente igual à 0.774, isso indica que os valores das ocorrências estão próximos à sua média (72.65%), porém menos concentradas do que no caso do RI. Ademais, há um *skewness* de -0.720, que, por ser negativo, indica uma distribuição de amostras maior no lado direito do histograma e, por ser próximo de 0, indica que esse acúmulo é perto do centro. Mais uma vez, é possível verificar e validar todas essas informações com a Figura 2.

2.2 Class-conditional mono-variate analysis

Nessa seção serão apresentados os resultados para a análise feita levando em consideração as classes das amostras apresentadas no conjunto de dados. Possuir esse tipo de informação ajuda a

compreender como cada classe de amostra interfere na ocorrência de cada preditor, possibilitando uma melhor interpretação do conjunto de dados. As Tabelas 2, 3 e 4 mostram os cálculos de média, desvio padrão e Skewness dos preditores nas Classes 1, 2 e 3, respectivamente.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Mean	1.518718	13.242286	3.552429	1.163857	72.619143	0.447429	8.797286	0.012714	0.057
SD	0.002268	0.499301	0.247043	0.273158	0.569484	0.214879	0.574807	0.083838	0.089075
Skewness	0.743729	0.753823	-0.676726	-1.080037	-0.554269	-0.899773	0.686378	7.561993	1.304118

Tabela 2: Cálculo de média, desvio padrão e skewness para as amostras de Classe 1

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Mean	1.518619	13.111711	3.002105	1.408158	72.598026	0.521053	9.073684	0.050263	0.079737
SD	0.003802	0.664159	1.215661	0.31834	0.724573	0.213726	1.921635	0.36234	0.106433
Skewness	2.057633	-1.049549	-1.77359	-0.371525	-1.375901	-0.969991	2.081649	8.238537	0.948973

Tabela 3: Cálculo de média, desvio padrão e skewness para as amostras de Classe 2

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Mean	1.517964	13.437059	3.543529	1.201176	72.404706	0.406471	8.782941	0.008824	0.057059
SD	0.001916	0.506887	0.162786	0.347489	0.512276	0.22989	0.380111	0.03638	0.107864
Skewness	0.970795	-0.461944	0.602157	-0.332546	-0.691398	-0.637722	0.785363	3.424032	1.691939

Tabela 4: Cálculo de média, desvio padrão e skewness para as amostras de Classe 3

É possível identificar alguns detalhes interessantes, por exemplo: as médias e os desvios-padrões do elemento **Na** são bem parecidos nas classes 1 e 3. Entretanto, os *skewness* são bem diferentes entre si. Na Classe 1 temos uma distribuição de amostras mais à esquerda, enquanto que na Classe 3 essa distribuição está mais à direita. Essa diferença de *skewness* sugere que no primeiro caso temos algumas amostras com porcentagens mais altas que a média aparecendo com menor frequência, enquanto que no segundo caso temos algumas algumas poucas amostras com porcentagens mais baixas que a média aparecendo com menor frequência. Através dos histogramas é possível obter uma outra forma de visualização das informações mencionadas. Todas as tabelas e histogramas estão disponíveis e separadas por classe no seguinte diretório:

$$/Figures/class - conditional$$

2.3 Unconditional bi-variate analysis

Nessa parte do processo, foi feita uma análise entre pares de preditores a fim de verificar a existência de algum tipo de relacionamento entre os preditores. Esse parte é importante, pois se os dados estão relacionados pode haver redundância de informação, o que não é desejável, pois aumenta apenas a complexidade do modelo sem trazer ganho significativo de informação.

Assim é feito esse pré-processamento para verificar os possíveis relacionamentos, e, quando possível, reduzir a complexidade do modelo. A figura 3 mostra um plot de todos os preditores, dois a dois, com cada elemento da diagonal principal sendo o histograma do elemento correspondante. As amostras estão identificados por cores de acordo o tipo de vidro correspondente, como é mostrado na legenda ao lado da figura.

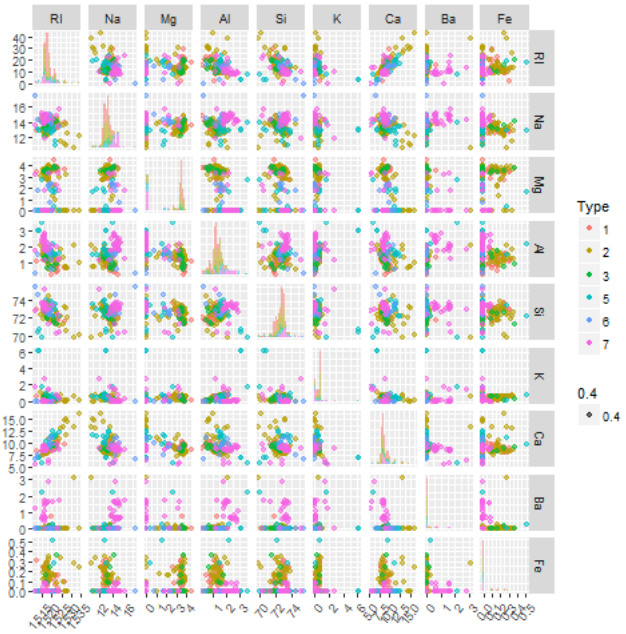

Figura 3: Plot dos preditores em pares

Analisando a figura 3 suspeita-se de um forte relacionamento linear posistivo entre a concentração de Cálcio(Ca) e o Índice Refrativo, do inglês *Refractive Index(RI)* e um relacionamento linear negativo, porém mais fraco, entre a concentração de Silício(Si) e o Índice Refrativo. Para uma melhor análise e confirmação dessas suspeitas, foi calculada a matriz de correlação apresentada na figura 4, que apresenta a correlação entre cada dois dos preditores ($\rho(Pred_1, Pred_2)$).

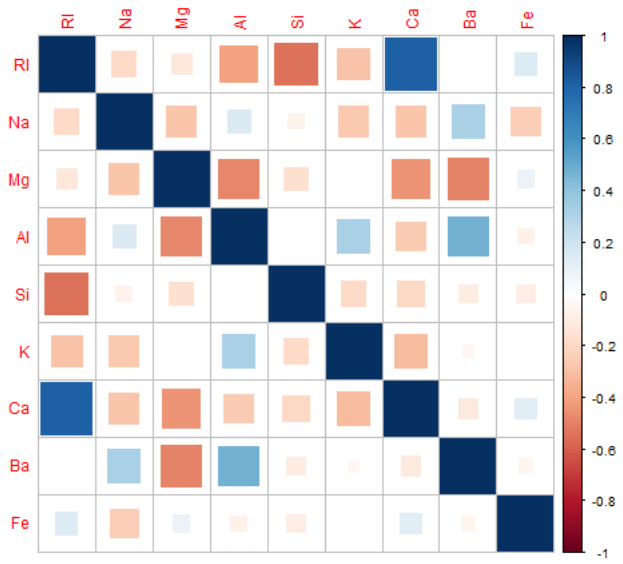

Figura 4: Plot da Matriz de Correlação entre os preditores

Devido às suspeitas da existência de relacionamentos lineares entre os preditores Ca e RI e os preditores Si e RI, foram analisadas as células correspondentes da matriz de correlação confirmando-se as suspeitas, visto que a célula referente à $\rho(RI, Ca)$ está quase completamente preenchida com um azul forte, indicando uma alta correlação positiva, e a célula referente à $\rho(RI, Si)$ está um poucos menos preenchida em um tom de vermelho mais fraco, indicando correlação negativa mais fraca.

2.4 Unconditional multi-variate analysis

O último passo do pré-processamento e análise foi a análise incondicional e multivariada dos preditores. Isto foi feito por meio da aplicação do método de redução de dimensionalidade chamado PCA, do inglês *Principal Component Analysis*, estudado em [1] e em [2]. Este método consiste em realizar uma transformação linear dos preditores, gerando os chamados Componentes Principais. A soma total da variância nos preditores não é alterada, mas os valores individuais da variância de cada preditor é alterada de forma que o k-ésimo componente possua a k-ésima maior variância que um preditor sozinho pode conter após a transformação linear, assim os k primeiros preditores devem possuir a maior parcela de variância que pode ser explicada por um grupo de k preditores como é visto na figura 5, onde cada barra reprensenta um componente e sua variância.

Após essa transformação, pode-se reduzir a quantidade p original de preditores para uma quantidade $k < p$, em que k representa os k primeiros preditores que expliquem um total de variância acima de um certo limite pré-estabelecido. Quanto mais componentes se perserva, maior é a representação dos dados, como o objetivo deste método é a redução do conjunto de dados, não faria sentido utilizar todas as componentes. Por esse motivo deve-se escolher aquelas que melhor representam as informações contidas no conjunto de dados, estas são as que estão relacionadas aos maiores autovalores, ou seja, os k primeiros componentes.

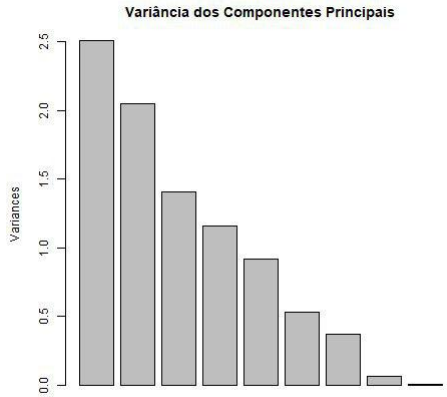

Figura 5: Componentes Principais

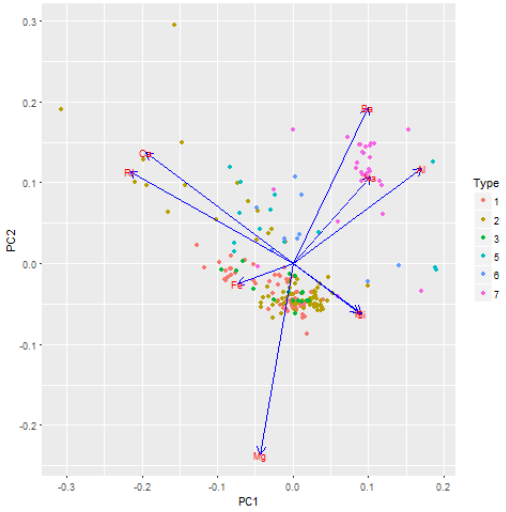

Figura 6: Resultado do PCA

A Figura 6 mostra o resultado da aplicação do PCA. Foram utilizados apenas o PC1 e o PC2, para uma boa visualização dos dados transformados pelos componentes principais, apesar de os dois juntos possuírem apenas, aproximadamente 50%, da variância dos dados. O *scatter plot* pode ser visto no diretório $/Figures/unconditional/multi-variate/PCA1.jpg$. Analisando o *scatter plot* é possível verificar que as classes ficaram mais separadas, possibilitando uma melhor visualização. Apesar disso, as classes 1, 2 e 3 ainda estão bem sobrepostas, tornando uma boa separação delas uma tarefa difícil. Além disso as fronteiras não estão definidas linearmente, o que torna a tarefa de escolher uma região de fronteira entre as classes ainda mais difícil.

2.5 Resultados

Ao longo do desenvolvimento deste trabalho, pode-se obervar a importância do pré-processamento de dados para o processo de mineração de dados. Como dito anteriormente, faz-se necessário o uso desta técnica para evitar redundância e transformar nossos dados em análise da forma mais real possível para obtermos os melhores e mais fiéis resultados. Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados. Foi visto, pelos experimentos descritos acima, que existe redundância de dados, analisados na correlação existentes entre preditores, o que não é desejável em uma análise crítica. Reduzindo essa redundância existente pode proporcionar uma aplicação menos custosa computacionalmente, o que é de suma importância quando se trata de um conjunto muito grande de dados.

Referências

[1] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[2] Michela Mulas. Data pre-processing. Slides de Aula, 2018.