Estudos de caso em pré-processamento para modelagem preditiva

Lia Sucupira Furtado e Gabriel Alves das Nevesthanks

Universidade Federal do Ceará, Dept. de teleinformática
Campus do Pici, Av. Mister Hul Brasil

Universidade Federal do Ceará, Dept. de teleinformática
Campus do Pici, Av. Mister Hul Brasil

Resumo.
A modelagem preditiva é uma coleção de técnicas matemáticas que visa encontrar uma relação lógica entre uma resposta (variável dependente) e vários preditores (variáveis independentes) com a intenção de prever valores futuros da resposta. Este documento aborda alguns conceitos de modelagem preditiva para efetuar uma análise exploratória dos dados de amostras de vidro para, em seguida, pré-processá-los. O conjunto de dados é composto por nove preditores. Foram calculados a assimetria, média e desvio padrão, a fim de aplicar métodos estatísticos para reduzir a dimensionalidade da informação e usar este conjunto de dados para construir um modelo que prevê resultados precisos.

1 Introdução

Devida a grande quantidade de dados em um dataset,é importante explorar e analisar os dados para compreender melhor o que eles representam. Muitas vezes, podem ocorrer incoerências, como o aparecimento de outliers ou dados redundantes que atrapalham o processamento e podem originar em um modelo preditivo sem qualidade.Além disso, segundo M. Kuhn e K. Johnson [1] algumas técnicas de modelagem podem ter requisitos restritos, como precisar que os preditores tenham uma escala comum. Por isso aqui discutimos as transformações de centralização, escalonamento e dimensionamento, que são técnicas de pré-processamento desses dados, para que eles fiquem aptos para serem usados nos mais diversos modelos preditivos.

2 Métodos

Análise exploratória dos dados

Este trabalho analisa um conjunto de dados de N = 214 observações de amostras de vidro, no qual nove preditores são considerados (o índice de refração, as percentagens de oito elementos: Na, Mg, Al, Si, K, Ca, Ba e Fe. Também é considerado o rótulo da classe correspondente, que representa sete tipos de vidro.

No primeiro contato com o conjunto de dados fizemos uma análise exploratória dele, analisando a média, o desvio padrão e a obliquidade (Tabela 1).

Dados	Média	Desvio Padrão	Obliquidade
RI	1.518365	0.003036864	1.614015
Na	13.40785	0.8166036	0.4509917
Mg	2.684533	1.442408	-1.144465
Al	1.444907	0.4992696	0.9009179
Si	72.65093	0.7745458	-0.7253173
K	0.4970561	0.6521918	6.505636
Ca	8.956963	1.423153	2.032677
Ba	0.1750467	0.4972193	3.392431
Fe	0.05700935	0.0974387	1.742007

Tabela 1:

Além disso, plotamos o histograma de cada um dos preditores (como, por exemplo, na Figura 1). Primeiramente nos ateremos a obliquidade que é uma informação muito importante sobre um conjunto de dados pois quando os dados estão distribuídos de forma desigual pode prejudicar o resultado do modelo. A maioria das obliquidades dos nossos preditores são valores próximos de zero que de acordo com M. Kuhn e K. Johnson [1] isso indica que a distribuição do preditor é aproximadamente simétrica. No entanto, podemos notar que os valores de obliquidade de K e Ba são muito alto comparado com os demais então seria favorável o uso de algum método estatístico como Box and Cox para reduzir esses valores.

O desvio padrão dos valores são pequenos então podemos concluir que os dados são concentrados em uma mesma faixa de valores, especialmente no índice de refração que tem um desvio quase zero. Fato que pode ser evidenciado no seu histograma em que a faixa de valores vai de 1.510 até 1.530.

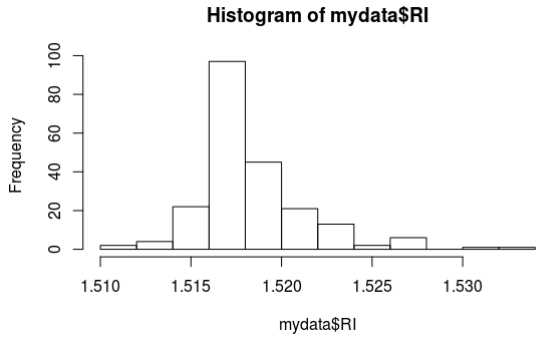

Figura 1: Plot do histograma do índice de refração

Em seguida, fizemos uma análise condicional dos preditores com relação aos tipos de vidro.

Há alguns preditores que demonstram um poder discriminativo, por exemplo: Bário, Potássio e Ferro não influenciam muito no índice de refração das amostras. Outros elementos podem auxiliar na identificação de alguns tipos de vidro, como Sódio (tipos 1, 2 e 5). O Ca possui uma distribuição de frequência parecida com a do índice de refração (figura 2).

Todas as médias, desvios-padrão e valores de assimetria das condições de classe estão registrados na tabela abaixo (figura 3):

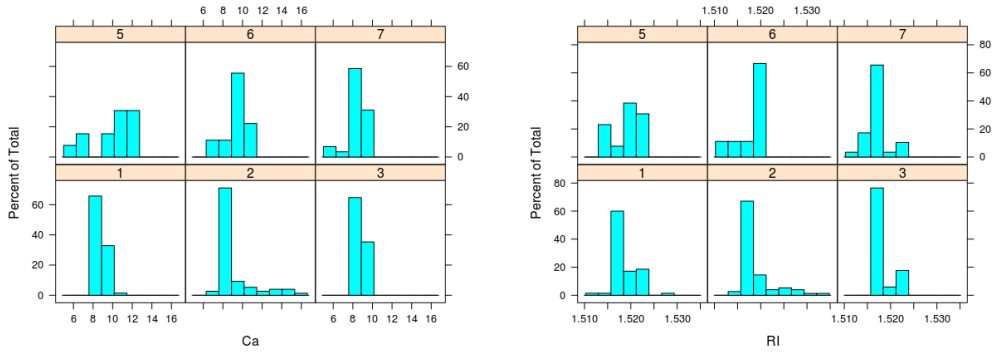

Figura 2: plot do gráfico classe condicional do Ca (esquerda) e do RI (direita).

Predictor	Type of glass	Mean	Standar deviation	Skewness	Predictor	Type of glass	Mean	Standard deviation	Skewness
Na	1	13,24229	0,4993015	0,7702698	Ba	1	0,012714286	0,08383769	7,7269786
	2	13,11171	0,6641594	-1,07061		2	0,050263158	0,36234044	8,4038562
	3	13,43706	0,5068871	-0,5059208		3	0,008823529	0,03638034	3,75
	4	0	0	0		4	0	0	0
	5	12,82769	0,7770366	-1,0541427		5	0,187692308	0,60825096	3,1127799
	6	14,64667	1,0840203	2,0030736		6	0	0	NaN
	7	14,44207	0,6863588	-1,5255252		7	1,04	0,66534094	0,4749361
Mg	1	3,5524286	0,247043	-0,691491	Fe	1	0,057	0,08907496	1,3325706
	2	3,0021053	1,2156615	-1,8091801		2	0,07973684	0,10643275	0,9680155
	3	3,5435294	0,1627859	0,6594818		3	0,05705882	0,10786361	1,8530118
	4	0	0	0		4	0	0	0
	5	0,7738462	0,9991458	0,6608308		5	0,06076923	0,15558821	2,2735474
	6	1,3055556	1,0971339	-0,2618982		6	0	0	NaN
	7	0,5382759	1,1176828	1,7202834		7	0,01344828	0,02979404	1,8768323
Si	1	72,61914	0,5694842	-0,566362	Ca	1	8,797286	0,5748066	0,7013528
	2	72,59803	0,7245726	-1,4035111		2	9,073684	1,9216353	2,1234209
	3	72,40471	0,5122758	-0,7572195		3	8,782941	0,3801112	0,8601291
	4	0	0	0		4	0	0	0
	5	72,36615	1,2823191	-0,7260664		5	10,123846	2,1837908	-0,8905913
	6	73,20667	1,0794675	1,1902438		6	9,356667	1,4499483	-0,654601
	7	72,96586	0,9402337	-1,277012		7	8,491379	0,9735052	-2,0401486
K	1	72,61914	0,214879	-0,9194037	Al	1	1,163857	0,2731581	-1,1036012
	2	72,59803	0,2137262	-0,9894552		2	1,408158	0,3183403	-0,3789803
	3	72,40471	0,2298897	-0,6984326		3	1,201176	0,3474889	-0,3642043
	4	0	0	0		4	0	0	0
	5	72,36615	2,1386951	1,7950462		5	2,033846	0,6939205	1,1256152
	6	73,20667	0	NaN		6	1,366667	0,571861	-0,799801
	7	72,96586	0,6684931	2,2551433		7	2,122759	0,4427261	-0,3086643
R.I.	1	1,518718	0,002268097	0,7599551					
	2	1,518619	0,003802126	2,0989229					
	3	1,517964	0,00191636	1,0632146					
	4	0	0	0					
	5	1,518928	0,003345355	-0,6418643					
	6	1,517456	0,003115783	-1,3495777					
	7	1,517116	0,002545069	1,0293014					

Figura 3: Tabela de médias, desvio padrão e obliquidade da calsse condtional

Relação entre pares de preditores

O gráfico par a par dos preditores (Figura 4) e sua matriz de correlação (Figura 5) são formas de analisar a correlação entre os preditores.Isto é um passo importante do pré-processamento, pois preditores que são altamente correlacionados se forem usados em conjunto no modelo, por terem informações muito parecidas,produzem uma informação redundante na predição. Dessa forma, pode ser até mais vantajoso retirar alguns desses preditores para não prejudicar os resultados do modelo.

Na análise bivaridada, podemos verificar que a relação do índice de refração com o cálcio tem uma alta covâriancia pois os preditores formam uma linha diagonal na Figura 4, que indica que eles variam proporcionalmente . Isso também pode ser verificado na matriz de correlação (Figura 5) pois existe uma a cor azul que indica uma correlação positiva.

Além disso, para relações que no gráfico de preditores em pares tiveram um carácter horizontal ou vertical, percebemos uma baixa correlação, pois eles não variam proporcionalmente. Como exemplo da relação do potássio com o ferro que seu gráfico é horizontal e na matriz de correlação (Figura 5) a cor é branca, o que indica que não existe uma relação entre eles.

Ademais, podemos verificar na matriz de correlação uma correlação negativa entre o RixSi, o MgxAl e o MgxBa.

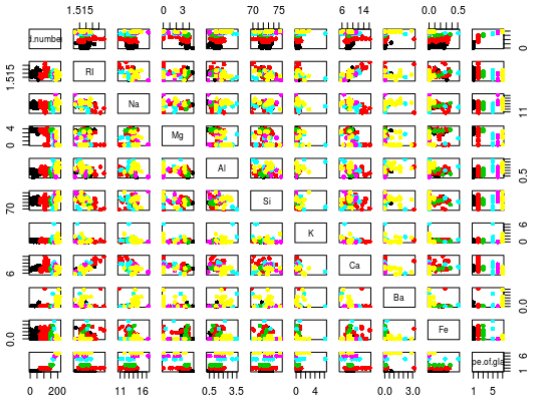

Figura 4: Plot dos preditores em pares

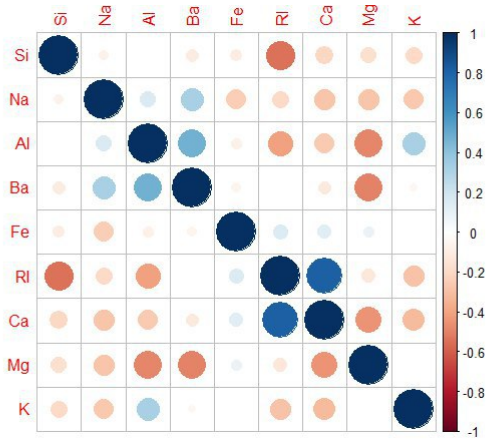

Figura 5: Plot da matriz de correlação

Concluímos que existem preditores correlacionados no nosso conjunto de dados e precisamos minimizar isso pra que ao aplicar o modelo tenhamos um bom resultado e além de que se utilizarmos menos preditores consegue-se reduzir o custo computacional.

Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é uma técnica usada para reduzir a dimensão dos nossos dados. Para escolher os componentes principais, que vão ser os novos preditores, precisamos transformar nossos dados na mesma escala, para nenhum ter um peso maior que o outro na análise de dados, então centralizamos e escalamos os dados diminuindo da média e dividindo pelo desvio padrão.Utilizamos esses valores escalados para fazer a PCA.

A PCA calculará os componentes principais, escolhemos os grupos com os dois maiores valores de variância, pois representavam bem o nosso conjunto de dados (figura 6).

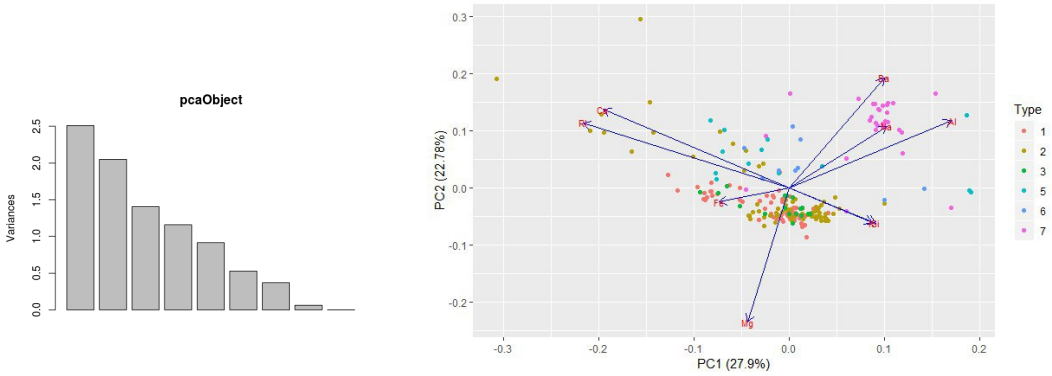

Figura 6: variância dos grupos da PCA (esquerda) e gráfico de dispersão dos dois componentes principais selecionados (direita)

No gráfico de dispersão dos componentes selecionados pela PCA (figura 6 à direita), percebemos

classes distantes e outras que estão demasiadamente próximas (concentradas nas proximidades da média), dificultando a separação delas e caracterizando um limite não linear entre as classes. Percebe-se que as classes com alto grau de sobreposição são: 1, 2 e 3.

3 Resultados

Ao analisar os dados antes do pré-processamento notamos que alguns preditores eram correlacionados e isto resultaria em uma redundância nos resultados. Além disso, os dados possuíam as mais variadas escalas o que podia fazer com que um preditor influenciasse mais que outro ao aplicar o modelo preditivo.

Com o pré-processamento diminuimos a dimensionalidade dos preditores e escolhemos componentes não correlacionadas para representar os dados.Também melhoramos a estabilidade numérica ao transformar os dados em centralizados e na mesma escala. Ademais, exploramos os nossos dados e fizemos alguma análise intuitivas que vão auxiliar na escolha do modelo preditivo. Dessa forma, temos um dataset preparado para ser treinado.

Referências

[1] M. Kuhn and K. Johnson. *Applied Predictive Modeling* , 2014.

[2] G.James, D. Witten et al *An Introduction to Statistical Learning with Applications in R*, 2013.