Melhora da Acurácia de um Modelo SVM com a Normalização dos Dados

Adilson Chrestani Junior

Pós Graduação em Inteligência Artificial Aplicada – Universidade Federal do Paraná

Resumo. *Cientes que de todo o processo de concepção de um algoritmo de Machine Learning a etapa de ETL consome pelo menos 50% do tempo, é fundamental compreender as tratativas dos dados a fim de evitar o efeito GiGo (Garbage in, Garbage Out) que enviesa negativamente a qualidade de predição do modelo. O intuito deste short paper é analisar uma base de dados pública referente à classificação de vidros e tentar melhorar a acurácia de um modelo SVM aplicado à ela apenas com modelagem de dados, através da Normalização dos atributos. Com o pré-processamento, foi possível aumentar a acurácia em 35 pontos percentuais.*

1. Introdução

Apesar de títulos sensacionalista que apontam que as máquinas já possuem autonomia e até um nível de consciência própria dado o volume de informação e bases de dados que temos hoje, é inviável pensar que os algoritmos que "lideram" estas máquinas são treinados com dados diretamente da fonte por tanto questões de custo quanto principalmente questões de negócio e conhecimento acerca da estrutura de tais dados, que hoje estão na cabeça de nós humanos. É por isso que a limpeza, filtragem, separação e abstração realizada por nós é de suma importância para direcionar os algoritmos à convergência e melhores resultados.

O objetivo deste short paper é mostrar através de uma base disponível no repositório de Machine Learning da UCI como podemos dobrar a capacidade de acurácia de um algoritmo de SVM com poucos ajustes nos dados de entrada.

2. Metodologia

O repositório de dados contendo 214 instâncias de modelos de vidro com suas respectivas composições elementais (Rl, Na, Mg, Al, Si, K, Ca, Ba, Fe) pode ser consultado pelo link https://archive.ics.uci.edu/static/public/42/glass+identification.zip. Com o objetivo de criar um modelo para encontrar a classe de um determinado vidro com base nas proporções de elementos contido nele, foi-se utilizado um algoritmo de SVM (Máquinas de Vetores de Suporte) que em sua essência tem como função principal separar conjuntos de dados espacialmente para identificar grupos distintos dentre os dados disponibilizados. Dado que o objetivo principal do paper é melhorar as

métricas do modelo através do pré-tratamento dos dados, não foram utilizadas manipulações nos hiperparâmetros do SVM.

Como técnicas de pré-tratamento foram utilizadas as Normalização por MinMax, que redimensiona os atributos conforme manipulação junto do valor mínimo e máximo daquela coluna e o Forward Selection que, através de uma combinação dos atributos, retorna o conjunto deles que tenham maximizado a média de qualidade. A utilização do Forward Selection surgiu como solução para tratar os diversos outliers por atributo que poderiam prejudicar a acurácia do modelo.

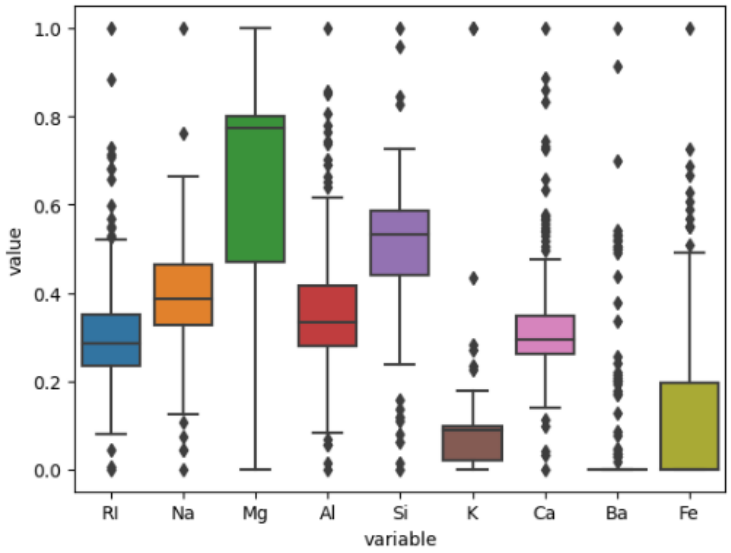

Figura 01. Distribuição de Outliers por Atributos (Normalizados)

Com o intuito de validar se a intervenção na base dos dados foi eficiente, foi-se comparadas as matrizes de confusão e score seguindo as três modelagens de SVM abaixo:

1. SVM com a base de classificação dos vidros “crua” (sem ajuste nenhum)
2. SVM com a base de classificação dos vidros Normalizada
3. SVM com a base de classificação dos vidros Normalizada e Otimizada (Forward Selection)

3. Resultados e Discussões

Preliminarmente aplicando o SVM na base de dados “crua” nós obtivemos uma acurácia de 0.35% (Figura 02). Pela análise da matriz de confusão percebe-se que o algoritmo errou grosseiramente principalmente as previsões da classe 1 (assumindo ser a 2) – comportamento esse que pode ser explicado pelas diferenças de grandezas entre os atributos da Figura 03.

Ao realizarmos a normalização das colunas e reaplicarmos o SVM, a acurácia sobe para 71.6% (Figura 02 – quadrado em amarelo), comprovando a eficácia da normalização das grandezas antes de alimentar o modelo com os dados.

Matriz de confusão - com os dados ORIGINAIS usados para TESTE

	precision	recall	f1-score	support
1	0.00	0.00	0.00	21
2	0.35	1.00	0.52	23
3	0.00	0.00	0.00	5
5	0.00	0.00	0.00	4
6	0.00	0.00	0.00	3
7	0.00	0.00	0.00	9
accuracy			0.35	65
macro avg	0.06	0.17	0.09	65
weighted avg	0.13	0.35	0.18	65

f1_score Teste
0.5227272727272727
[[0 21 0 0 0 0]
[0 23 0 0 0 0]
[0 5 0 0 0 0]
[0 4 0 0 0 0]
[0 3 0 0 0 0]
[0 9 0 0 0 0]]

Matriz de confusão - com os dados NORMALIZADOS usados para TESTE

	precision	recall	f1-score	support
1	0.67	0.86	0.75	21
2	0.71	0.74	0.72	23
3	0.00	0.00	0.00	5
5	1.00	0.50	0.67	4
6	0.00	0.00	0.00	3
7	0.82	1.00	0.90	9
accuracy			0.71	65
macro avg	0.53	0.52	0.51	65
weighted avg	0.64	0.71	0.66	65

f1_score Teste
0.7192494089834516
[[18 3 0 0 0 0]
[5 17 0 0 1 0]
[4 1 0 0 0 0]
[0 0 0 2 0 2]
[0 3 0 0 0 0]
[0 0 0 0 0 9]]

Matriz de confusão - com os dados FORWARD SELECTION usados para TESTES

	precision	recall	f1-score	support
1	0.64	0.86	0.73	21
2	0.71	0.65	0.68	23
3	0.00	0.00	0.00	5
5	1.00	0.50	0.67	4
6	0.00	0.00	0.00	3
7	0.64	1.00	0.78	9
accuracy			0.68	65
macro avg	0.50	0.50	0.48	65
weighted avg	0.61	0.68	0.63	65

f1_score Teste
0.7161497287355411
[[18 3 0 0 0 0]
[6 15 0 0 0 2]
[4 1 0 0 0 0]
[0 0 0 2 0 2]
[0 2 0 0 0 1]
[0 0 0 0 0 9]]

Figura 02. Resultados das Matrizes de confusão para os três testes realizados

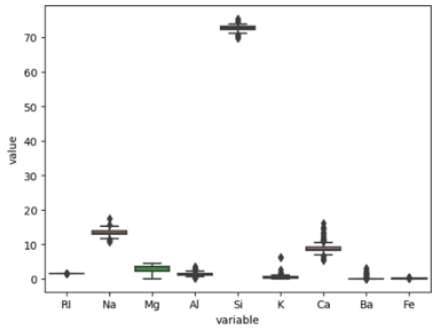

Figura 03. Boxplots por atributo pré normalização

Todavia, ao unirmos o método de Normalização com o método de Forward Selection, perdemos 3 pontos na acurácia ao considerarmos como atributos válidos apenas as composições de 3 dos 9 elementos.

4. Conclusão

A geração dos gráficos Boxplot de atributos foi essencial para identificar o problema de desbalanceamento de dimensões de atributos e a Normalização por MinMax foi a chave para dobrar a acurácia do Modelo.

Apesar de não melhorar a capacidade de classificação, o método de Forward Selection nos deu um insight poderoso em termos de otimização de recurso: apesar de neste dataset ser necessário olhar para os outliers dos atributos de forma não exclusiva porque eles compõem critérios importantes na definição das Classes de vidros, é possível trocarmos 3% de acurácia por uma redução de 66% na quantidade de dados utilizadas no treinamento ao reduzirmos os 9 atributos para 3. Embora para uma base de 214 instâncias não faça diferença, com o aumento exponencial destas leituras em outras bases mais pesadas esse trade-off pode ser a chave para otimização de custo.