https://eventos.utfpr.edu.br//sicite/sicite2020

Aprendizado de máquina aplicado à análise de amostras de açúcar mascavo adulteradas

Machine learning applied to analysis of adulterated brown sugar samples

Enrico Gaetano Corrêa Portolann
enricogaetano@hotmail.com
Universidade Tecnológica Federal do Paraná, Pato Branco, Paraná, Brasil

Vanderlei Aparecido de Lima
valima@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Pato Branco, Paraná, Brasil

RESUMO

Uma das principais aplicações do aprendizado de máquina está relacionada com técnicas de regressão de dados, ou seja, com a criação de modelos matemáticos capazes de prever valores referentes a conjuntos de dados numéricos. É possível usar tais métodos para analisar amostras adulteradas em alimentos. O objetivo deste trabalho foi comparar três modelos de regressão, todos testados em um mesmo conjunto de dados (amostras de açúcar mascavo adulterado com açúcar branco) com o intuito de estimar, a partir dos valores dos níveis de cinza das imagens das amostras, as porcentagens aproximadas de açúcar branco. Os modelos utilizados foram: regressão por mínimos quadrados parciais (*partial least squares* ou PLS), rede neural artificial (*multilayer perceptron* ou RNA) e floresta randômica (*random forest* ou FR). A regressão linear por mínimos quadrados parciais obteve a melhor precisão para o conjunto de teste, com coeficiente de correlação de 92,4%.

PALAVRAS-CHAVE: Aprendizado de máquina; Análise de regressão; Rede neural artificial.

ABSTRACT

One of the main applications of machine learning is related to regression techniques, that is, to the generation of mathematical models capable of predicting values regarding numerical data sets. It is possible to apply these models in order to analyse adulterated food samples. The goal of this project was to compare three regression models, which were tested in the same dataset (brown sugar samples adulterated with white sugar), in order to estimate the percentage of white sugar in each sample, based on the grayscale values of the images of the samples. Tested models include partial least squares regression, artificial neural network and random decision forest. Partial least squares method got the best precision for the test set, with a correlation coeficient of 92,4%.

KEYWORDS: Machine learning; Regression analysis; Artificial neural network.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

Do ponto de vista do aprendizado de máquina, analisar um conjunto de dados significa obter, a partir de algoritmos, modelos capazes de adquirir informações do conjunto e fazer previsões acertadas. Um modelo é construído a partir de um conjunto de treinamento, ou seja, um conjunto de dados que servem como exemplo inicial para que possam ser obtidos valores iniciais dos seus parâmetros. Comparando o resultado previsto com o valor real, os parâmetros são otimizados. O tipo de parâmetro usado depende de cada modelo, bem como o método para ajustar os valores dos parâmetros (WITTEN, 2016, p.9).

Em seguida, o modelo é validado a partir de um segundo conjunto de dados, chamado conjunto de validação. O processo em si é conhecido como validação cruzada (*cross-validation*). Um terceiro conjunto de dados, chamado de conjunto de teste, com valores diferentes dos utilizados anteriormente, é usado para avaliar a precisão final do modelo (WITTEN, 2016, p.149).

Uma das inúmeras possibilidades para esse tipo de regressão é analisar amostras adulteradas em alimentos. Neste trabalho, foram testados três algoritmos de aprendizado de máquina para prever as porcentagens aproximadas de açúcar mascavo adulteradas com açúcar cristal branco. Os métodos utilizados foram: mínimos quadrados parciais (PLS), rede neural artificial (RNA) e floresta aleatória (FR), os quais são explicados brevemente em seguida.

A regressão por mínimos quadrados parciais (PLS), ao contrário da regressão por mínimos quadrados, pode ser feita em um conjunto de dados com muitas variáveis dependentes. A técnica baseia-se em reduzir as variáveis de interesse a um conjunto menor de componentes não relacionados entre si, e realizar uma regressão por mínimos quadrados nesses componentes, em vez do conjunto original (WITTEN, 2016, p.341).

Rede neural artificial (*multilayer perceptron*) é uma rede composta de uma camada de entrada, em que os dados são inseridos, e de uma camada de saída, onde é feita a previsão sobre os dados. Entre a primeira e a última camada, existem camadas ocultas que se comunicam por pesos e funções de ativação. Treinar a rede neural artificial significa ajustar os pesos. O treinamento de redes neurais artificiais é feito por retropropagação, onde o cálculo é realizado minimizando o erro total a partir da atualização, em cada época (ou seja, em cada instância do conjunto de dados), dos pesos da rede (WITTEN, 2016, p.227).

Na regressão por floresta randômica (FR), são construídas várias árvores de decisão, e o resultado final (previsão) é a média das previsões todas as árvores individuais. As árvores de decisão, por sua vez, partem de observações sobre os valores de entrada e obtém conclusões sobre os valores finais previstos (WITTEN, 2016, p.243).

MATERIAIS E MÉTODOS

Foram obtidas 85 amostras de misturas de açúcar mascavo e açúcar cristal branco com diferentes porcentagens. Cada porcentagem corresponde a um grupo de cinco amostras, estando presentes amostras com concentrações de 0%, 2,5%, 5%, 7,5%, 10%, 12,5%, 15%, 17,5%, 20%, 22,5%, 25%, 27,5%, 30%, 32,5%, 35%,

37,5% e 40% de açúcar cristal branco em açúcar mascavo. Foi tirada uma foto de cada uma das amostras, com cuidado para que cada foto fosse tirada nas mesmas condições de luminosidade e resolução/ângulo da câmera. Para cada imagem, foram obtidos três recortes 500x500 pixels de partes diferentes de cada amostra, totalizando 255 recortes.

Os valores dos níveis de cinza de cada banda RGB das imagens foram obtidos pelo programa Chemostat (256 bandas R, 256 bandas G e 256 bandas B). Esses valores serviram de base para o posterior processamento dos dados amostrais.

Com os valores dos níveis de cinza já obtidos, o Excel foi utilizado para pré-processar os dados e separar as amostras em dois conjuntos. Um dos conjuntos, correspondente a 204 dos 255 recortes, foi utilizado como base para a geração dos modelos e o outro, correspondente aos 51 recortes restantes, foi utilizado para testar e validar os modelos.

Para realizar a modelagem dos dados propriamente dita, foi utilizado o software Weka 3.8. Foram gerados três modelos distintos de regressão: um por PLS, outro por rede neural artificial com 5 camadas ocultas, e outro por floresta randômica. Para cada modelo de regressão foram avaliados os valores dos coeficientes de correlação e do desvio médio.

RESULTADOS E DISCUSSÃO

Constam aqui três tabelas (Tabelas 01, 02 e 03) com os valores dos coeficientes de correlação e dos erros médios para cada modelo testado, bem como uma tabela (Tabela 04) com os valores previstos por cada modelo para o conjunto de teste.

Tabela 01 - Coeficientes de correlação e erros médios para os conjuntos de treinamento, validação interna (*cross validation*) e teste da regressão por mínimos quadrados parciais (PLS).

	Regressão por mínimos quadrados parciais (PLS)	
	Coeficiente de correlação	Desvio médio (%)
Conjunto de treinamento	0,952	2,9
Conjunto de validação	0,912	3,9
Conjunto de teste	0,924	3,6

Fonte: Autores (2020).

Tabela 02 - Coeficientes de correlação e erros médios para os conjuntos de treinamento, validação interna e teste da regressão por rede neural artificial (RNA).

	Regressão por rede neural artificial (RNA)	
	Coeficiente de correlação	Desvio médio (%)

Conjunto de treinamento	0,908	6,1
Conjunto de validação	0,886	4,7
Conjunto de teste	0,858	6,2

Fonte: Autores (2020).

Tabela 03 - Coeficientes de correlação e erros médios para os conjuntos de treinamento, validação interna e teste da regressão por floresta randômica (FR).

	Regressão por floresta randômica (FR)	
	Coeficiente de correlação	Desvio médio (%)
Conjunto de treinamento	0,990	1,4
Conjunto de validação	0,914	3,8
Conjunto de teste	0,873	4,3

Fonte: Autores (2020).

Tabela 04 - Valores previstos e desvios de cada método para as porcentagens das amostras do conjunto de teste.

Amostra do conjunto de teste	**Valor verdadeiro da porcentagem da amostra (%)**	**Valor previsto da porcentagem da amostra**		
		Mínimos quadrados parciais (%)	Rede neural (%)	Floresta aleatória (%)
1	**0,0**	1,9	1,2	1,9
2	**0,0**	1,2	1,2	2,2
3	**0,0**	2,2	1,2	3,6
4	**2,5**	16,3	18,7	22,1
5	**2,5**	9,9	12,4	13,0
6	**2,5**	8,0	12,4	12,5
7	**5,0**	4,2	4,4	5,4
8	**5,0**	6,5	1,2	5,5
9	**5,0**	3,7	1,2	5,4
10	**7,5**	8,0	12,4	6,2
11	**7,5**	6,8	3,1	7,9
12	**7,5**	9,9	12,4	13,6
13	**10,0**	11,6	12,4	13,4
14	**10,0**	12,8	12,4	12,6
15	**10,0**	16,4	12,8	18,1
16	**12,5**	10,0	12,4	13,9
17	**12,5**	10,0	12,4	15,1
18	**12,5**	13,1	12,4	12,0
19	**15,0**	11,0	12,4	14,5
20	**15,0**	11,8	12,4	14,5
21	**15,0**	8,0	12,4	12,0
22	**17,5**	22,7	18,8	25,7
23	**17,5**	20,6	12,5	19,8

24	**17,5**	20,4	12,5	18,4
25	**20,0**	14,7	12,4	14,3
26	**20,0**	12,6	12,4	13,6
27	**20,0**	14,8	18,8	20,4
28	**22,5**	12,1	12,4	12,8
29	**22,5**	18,4	18,8	20,4
30	**22,5**	10,6	12,4	11,4
31	**25,0**	30,2	22,3	33,2
32	**25,0**	27,6	27,9	36,4
33	**25,0**	29,6	23,9	33,9
34	**27,5**	23,8	16,9	23,6
35	**27,5**	20,9	12,4	18,0
36	**27,5**	30,5	20,7	22,6
37	**30,0**	35,5	27,6	35,0
38	**30,0**	32,4	27,6	32,0
39	**30,0**	32,5	26,1	34,9
40	**32,5**	33,9	22,3	30,7
41	**32,5**	32,7	23,9	32,5
42	**32,5**	33,3	22,7	32,3
43	**35,0**	32,5	22,4	31,7
44	**35,0**	34,6	23,6	31,3
45	**35,0**	25,4	25,0	20,9
46	**37,5**	36,4	27,9	34,5
47	**37,5**	35,3	22,3	34,5
48	**37,5**	37,5	22,4	38,3
49	**40,0**	39,5	27,9	36,2
50	**40,0**	38,9	27,9	38,5
51	**40,0**	37,8	27,9	38,4

Fonte: Autores (2020).

Observa-se, pelos valores dos coeficientes de correlação, que todos os modelos testados apresentaram-se bastante razoáveis. No entanto, com a tabela que mostra os valores previstos por cada modelo para o conjunto de treinamento, é possível perceber que os modelos têm desempenho mediano em porcentagens mais baixas (0% a 10%), justamente as mais prováveis de ocorrer adulteração.

CONCLUSÃO

Todos os três modelos utilizados (PLS, RNA E FR) foram úteis para prever adulterações de açúcar mascavo com açúcar cristal branco. Uma opção para melhorar ainda mais o desempenho dos algoritmos seria aumentar o número de amostras utilizadas na criação dos modelos e/ou distribuir as porcentagens das amostras em um intervalo mais amplo, de modo a minimizar os efeitos de sobre-ajuste. Observa-se por fim que a regressão por meio dos valores dos níveis de cinza das imagens, apesar de ser bastante simples em relação a outros métodos,

constitui um possível recurso a ser utilizado na área de alimentos para se detectar fraudes.

AGRADECIMENTOS

Ao CNPQ, à UTFPR, ao professor Vanderlei e ao Departamento de Química em geral.

REFERÊNCIAS

WITTEN, I. H; FRANK, E; HALL, M. **Data Mining**: Practical Machine Learning Tools and Techniques. San Francisco: Elsevier, 2016.

BOUCKAERT, R. R et al. **Weka Manual for Version 3.8.1.** Hamilton: University of Waikato, 2016.