Cadernos do IME - Série Matemática
e-ISSN: *2236-2797 (online)*
DOI: 10.12957/cadmat.2022.66664
Recebido em 21 Abr. 2022 | Aceito em 07 Set. 2022

ANÁLISE DISCRIMINANTE COMO CRITÉRIO DE VALIDAÇÃO DE AGRUPAMENTOS EM MINERAÇÃO DE TEXTOS

DISCRIMINANT ANALYSIS AS A VALIDATION CRITERIA FOR CLUSTERING IN TEXT MINING

CARLA CRISTINA P. CRUZ[a] REGINA S. LANZILLOTTI[b]
HAYDÉE S. LANZILLOTTI[c]

Resumo

O presente artigo optou-se por escolher, inicialmente, três textos oriundos de sites que abordavam o tema e a agregação deles materializou o processo de Descoberta de Conhecimento Textual. Sob esta orientação, foi realizado o pré-processamento, conjunto de procedimentos para extrair e recuperar dados textuais considerados relevantes, etapa que permitiu o processamento sob a ótica da *Quadratic Discriminant Analysis* (QDA) com uso da abordagem equidistante proposta pela covariância ajustada, em que foram dispostos 30 parágrafos dicotomizados em 20 e 10. A acurácia do modelo apresentou probabilidade discriminatória de 87 chances em 100, sendo que o primeiro grupo mostrou 17 acertos e o segundo nove, reconhecendo palavras-chave adequadas sob a ótica semântica da Nutrigenômica e Nutrigenética, resultados esperados pelos especialistas que buscam conglomerados cognitivos textuais.

Palavras-chave: Mineração de texto, nutrigenômica, análise discriminante.

Abstract

[a]Instituto de Matemática e Estatística - IME, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil; ORCID: https://orcid.org/0000-0003-3656-4492; **E-mail:** carlapassos2889@gmail.com

[b]Instituto de Matemática e Estatística - IME, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil; ORCID: https://orcid.org/0000-0001-7789-6843; **E-mail:** reginalanzillotti@gmail.com

[c]Instituto de Nutrição - NUT, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil; ORCID: https://orcid.org/0000-0002-7252-4415; **E-mail:** haydeelan@gmail.com

This article was chosen to choose, initially, three texts from websites that addressed the topic and the aggregation of them materialized the process of Discovery of Textual Knowledge. Under this guidance, pre-processing was carried out, a set of procedures to extract and retrieve textual data considered relevant, a step that allowed processing from the perspective of Quadratic Discriminant Analysis (QDA) using the equidistant approach proposed by covariance adjusted, in which 30 were provided for paragraphs dichotomized into 20 and 10. The accuracy of the model showed a discriminatory probability of 87 in 100, with the first group showing 17 hits and the second nine, recognizing appropriate keywords from the semantic perspective of Nutrigenomics and Nutrigenetics, results expected by experts who seek conglomerates textual cognitive.

Keywords: Text mining, nutrigenomic, discriminant analysis.

MSC2010: 3B52, 03E72, 94D05, 62A86

1 Introdução

A Mineração de Texto (MT) é um processo de Descoberta de Conhecimento que utiliza técnicas de análise e extração de dados a partir de diferentes tipos de textos, além de envolver a aplicação de algoritmos computacionais que os processam. Identifica-se informações úteis e implícitas que não poderiam ser recuperadas por métodos tradicionais de consulta, uma vez que a informação obtida se encontra em formato não-estruturado [14].

A Mineração de Texto pode ser utilizada em diversas áreas do conhecimento, como Direito, Medicina, Negócios, dentre outros [2]. Neste artigo optou-se por utilizá-la em textos que tratam da Nutrigenômica e Nutrigenética. A primeira corresponde a uma área da Nutrição que permite estratégias eficazes de intervenção dietética para recuperar a homeostase normal e prevenir doenças relacionadas à alimentação [16]. A segunda, visa compreender como a composição genética de um indivíduo coordena a resposta à dieta, ou seja, gerenciar dietas para prevenir acometimentos à saúde [17]. Nesse contexto, a Mineração de Texto ajudará no reconhecimento de padrões nos termos encontrados gerando palavras-chave para o pesquisador, tornando mais rápida a busca por textos relacionados.

A fim de transformar as informações não-estruturadas em estruturadas, ou seja, em dados quantitativos, os textos passam por um processo de coleta, limpeza e tratamento. Esta etapa, chamada de pré-processamento, é considerada a etapa mais importante do processo de Mineração de Texto. Após esta etapa, os dados estruturados ficam dispostos em uma matriz do modelo Termo-Documento que armazena

a frequência absoluta do termo relativo a cada documento. Seja n (conjunto de documentos) e m (termos), pode ser possível modelar cada documento como um vetor v no espeço m-dimensional onde cada documento é armazenado em cada coluna da matriz e cada termo, na linha [1].

Em prosseguimento algumas técnicas podem ser aplicadas em Mineração de Textos, com o intuito de reconhecer padrões, agrupar e classificar os dados [12]. Diante do exposto, o objetivo deste artigo é a utilização da Análise Discriminante (AD) para a separação de termos referentes à classe Nutrigenômica e Nutrigenética para a verificação da classificação dos termos nos grupos.

2 Metodologia

2.1 Descrição dos dados utilizados

O banco de dados foi composto por três textos coletados do *blog* Sophie Deam [3], *site* NutMed [5] e *site* Profissão BioTec [20], sob o tema Nutrigenômica e Nutrigenética.

Após a coleta dos textos, os mesmos passam por um processo transformação dos dados textuais para dados numéricos, a qual se denomina *KDT - Knowledge Discovery in Text*, a ser explicada na subseção 2.2. Todas as análises foram realizadas pelo *software RStudio* [18].

2.2 *KDT - Knowledge Discovery in Text*

O método *KDT*, também conhecido como Descoberta de Conhecimento em Textos, consiste em um conjunto de procedimentos para extrair e recuperar dados textuais considerados relevantes [4]. Em outras palavras, transforma dados textuais em dados numéricos, sendo composto por quatro etapas, Figura 1:

1 **Coleta das informações**. Também conhecida como Recuperação da Informação ou Identificação do Problema [19], é a etapa de busca, coleta e armazenamento dos dados que serão analisados, além de definir o tipo de abordagem a ser aplicada nos textos: semântica ou estatística [12], para a realização da etapa posterior;

2 **Pré-processamento**. Também conhecida como Extração de Conhecimento ou Extração da Informação, é o conjunto de ações tomadas sobre os documentos textuais a fim de torná-los manipuláveis. Considerada a parte mais importante do processo;

3 **Mineração da informação**. Uma vez que os textos foram transformados para dados estruturados, os dados ficam compatíveis para aplicação das técnicas de Mineração de Textos [19];

4 **Análise dos resultados**. Etapa final, também conhecida como pós processamento [12], tem como objetivo a interpretação e análise dos resultados obtidos na fase anterior. Neste artigo serão descritos com mais detalhes as etapas dois e três do *KDT*.

Figura 1: Etapas do *KDT*

Fonte: Autoral, 2021.

De acordo com as etapas descritas, no processamento é necessário realizar o cálculo da relevância dos termos, pois o fato da coletânea textual possuir muitas palavras, não significa que todas tenham a mesma importância.

O grau de associação da palavra com o texto corresponde ao peso, que indica a importância da palavra [14]. Dessa forma, faz-se necessário calcular a relevância de cada termo e, dentre os vários métodos existentes, indica-se o *Term Frequency* (TF) que usa a frequência absoluta, na qual calcula quantas vezes o termo aparece em um documento [23]. Sua fórmula é descrita por:

$$F_{abs} = \sum_{i=1}^{N} \frac{X_i}{N} \tag{1}$$

O cálculo baseado na frequência relativa corresponde a um índice obtido pela razão entre a Frequência Absoluta [23] e o número total de palavras no documento:

$$F_{rel} = \frac{F_{abs}}{N} \tag{2}$$

Após o cálculo das frequências relativas, selecionam-se os termos a fim de minimizar o desafio da dimensionalidade dos dados. Este procedimento visa obter um subconjunto conciso e representativo de termos da coleção textual. Pode-se usar ou não, métodos para se reduzir a quantidade de termos presentes na coletânea textual. No entanto, é necessário definir um ponto de corte, denominado Limiar (*threshold*), definido pelo usuário ou pelo próprio método adotado.

Uma vez que os textos foram transformados para dados estruturados, parte-se para a etapa de Processamento ou Mineração da Informação, pois já há compatibilidade para aplicação das técnicas de Mineração de Textos. Nesta fase, armazenam-se os termos selecionados em uma matriz termo-documento designada por um conjunto de N documentos e M termos [1], conforme Quadro 1.

Quadro 1: Matriz termo-documento

Termos	**Documento 1**	**Documento 2**	**Documento 3**
1	1	0	0
2	0	2	1
3	3	0	2
⋮	⋮	⋮	⋮
N	3	0	2

Fonte: Autoral, 2021.

Em seguida, se aplicam técnicas de Mineração de textos com o intuito de descobrir os padrões textuais para agrupar e classificar os termos [12], conforme ilustrada na Figura 1. A Análise Discriminante (AD), subseção 2.3, técnica estatística multivariada utilizada para discriminar e classificar objetos [21] foi utilizada para grupar termos da coleção textual.

2.3 Análise Discriminante

A Análise Discriminante (AD) é uma técnica de classificação que utiliza a inferência estatística multivariada para discriminar e classificar objetos [21]. Tem como objetivo usar informações das variáveis explicativas (independentes) para se alcançar a variável resposta categórica (dependente) que naturalmente traduz conjuntos mutu-

amente exclusivos que efetiva a separação ou discriminação de forma clara entre dois ou mais grupos, enquanto a variável independente pode explicar a máxima variação possível da categorização [11].

Segundo Johnson e Wichern [9] a classificação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos. O problema da discriminação entre dois ou mais grupos visando posterior classificação, foi inicialmente abordado por Fisher [7]. A classificação considerada ideal depende de dois fatores:

- Exige que a classificação deva resultar em pouca ocorrência de má classificação, sendo necessário que se considere as probabilidades a *priori* e os custos de uma classificação errônea [9];

- Deve considerar se as variâncias amostrais (ou populacionais) são ou não iguais. Caso sejam, as funções discriminantes são ditas lineares (*Linear Discriminant Analysis* - LDA) e, caso contrário, são denominadas funções discriminantes quadráticas (*Quadratic Discriminant Analysis* - QDA) [21].

Optou-se pela *Quadratic Discriminant Analysis* (QDA) para dois grupos, que se desenvolve com a Distância Quadrática de Mahalanobis para encontrar os pontos equidistantes dos centroides destes grupos, considerando matrizes de covariâncias diferentes, isto é, $Cov_1 \neq Cov_2$. A QDA foi tratada com o pacote *MASS* [22] presente no *software RStudio* [18], que permitiu obter as médias das frequências relativas.

O cálculo das Distâncias Quadráticas de Mahalanobis [11] foi realizada pelas expressões (2.3) e (2.4) para os **Grupos 1** e **2**, respectivamente, onde $\overline{x}_1$ e $\overline{x}_2$ são as médias e Cov_1 e Cov_2 são as estimativas das matrizes de covariâncias populacionais dos mesmos grupos.

$$Grupo1 : d_1^2 = (x - \overline{x}_1)^T Cov_1^{-1}(x - \overline{x}_1) \tag{3}$$

$$Grupo2 : d_2^2 = (x - \overline{x}_2)^T Cov_2^{-1}(x - \overline{x}_2) \tag{4}$$

A covariância combinada permite fazer o teste de igualdade das matrizes de médias com o uso da estatística teste T^2 de Hotelling [15] que é explicitada por:

$$T^2 = \left[\frac{n_1 \times n_2}{n_1 + n_2}(md_1 - md_2)(InvCOV_{comb})(md_1 - md_2)\right] \tag{5}$$

que adota a estatística teste da distribuição *F de Snedecor* pela expressão:

$$F = [(n_1 + n_2 - p - 1)/(n_1 + n_2 - 2)]T^2 \tag{6}$$

tendo $(p, (n_1 + n_2 - 2))$ graus de liberdade.

Assumindo a condição de Normalidade, a função é definida por:

$$Q(x) = ln(p_i) - \frac{1}{2} \times ln|Cov_i| - \frac{1}{2} \times [(x - \overline{x}_i)^T Cov_i^{-1}(x - \overline{x}_i)] \tag{7}$$

onde:

- p_i, probabilidade a *priori* de ser do Grupo 1 ou 2;
- Cov_i, matriz de covariância;
- $\overline{x}_i$, média do grupo.

O teste estatístico de significância é proposto para avaliar se a media da função discriminante varia de grupo para grupo [13], obtida por:

$$\Phi_j^2 = \frac{\left((n-1) - \left(\frac{p+m}{2}\right)\right)}{ln(1 + \lambda_j)} \tag{8}$$

onde:

- n, tamanho da amostra;
- p, termos considerados;
- m, grupos;
- λ_j raiz característica.

Esta estatística teste é confrontada com o valor Qui-quadrado tabelado com $(p + m - 2)$ graus de liberdade. A proposta de Morrison [15] quanto à regra de classificação para dois grupos consiste em estabelecer uma designação específica a uma das unidades de observação da população e, se a equação linear discriminante for positiva para todos os dois grupos, estabelece-se uma tabela de contingência cruzada.

O *Cross Validation* avalia a validade dos resultados obtidos. No entanto, neste estudo será usado o *leave-one-out-Cross-Validation - LOOCV*, caso especial do *Cross Validation*, onde a partir das n observações da amostra, realiza-se n treinamentos e

n testes de validação. Para cada treinamento de tamanho $n-1$ observações, uma delas fica de fora do teste de validação [10].

O LOOCV usa como estimativa o *Mean Square Error* (MSE), explicitada na equação (2.9). A seguir se utilizou toda a amostra para treinamento e teste das funções discriminantes [8]. Assim, o erro de classificação é calculado pelo número de observações classificadas incorretamente em relação ao número de observações da amostra [10].

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2 \tag{9}$$

onde:

- n, tamanho da amostra;
- $y_i - \hat{y}_i$, *Mean Square Error* (MSE);
- h_i, calcula a distância entre os valores da amostra (x_i) e a média ($\overline{x}$) através da equação:

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{10}$$

Por fim, para visualização pictográfica dos resultados foi utilizado o *Wordcloud* [6], gráfico que mostra o grau de frequência dos termos em um texto, foi gerado também pelo *software R* [18].

3 Resultados

Primeiramente foram computadas as frequências dos termos obtidos em cada Documento, obtendo-se um total de 185, 209 e 265, para os Documentos 1, 2 e 3, respectivamente. A Nuvem de Palavras, Figura 2, permitiu a visualização dos termos de maior frequência.

Em razão de uma melhor visualização dos termos, os 100 termos mais frequentes da agregação dos Documentos 1, 2 e 3 foram utilizados. O termo "nutrigenoma" é o mais frequente, seguido de "alimento", "gene", "genoma" e "expressao". No entanto, "nutrigenetica" não surgiu.

Dentre os 10 termos julgados de maior relevância, Tabela 1, os quatro mais expressivos foram "nutrigenoma", "alimento", "gene" e "genoma", com as frequências 21, 15, 14 e 12, respectivamente, que subsidiaram a aplicação da QDA.

Figura 2: Nuvem de Palavras

Fonte: Autoral, 2021.

Tabela 1: Termos mais relevantes nos textos escolhidos para Análise Discriminante em Mineração de textos

Termos	Documento 1	Documento 2	Documento 3	Total
nutrigenoma	4	8	9	**21**
alimento	11	1	3	**15**
gene	3	5	6	**14**
genoma	0	7	5	**12**
expressao	3	4	7	**12**
estudo	1	5	4	**10**
nutricao	3	5	3	**10**
dieta	0	5	4	**9**
ciencia	4	3	1	**8**
individuo	1	5	1	**7**

Fonte: Autoral, 2021.

Os textos dos três documentos foram unificados obtendo-se 30 parágrafos e a classificação dos grupos foi feita da seguinte forma: se o parágrafo continha apenas um dos quatro termos, ele seria classificado como **Grupo 1**; caso contrário, foi classificado no **Grupo 2**. A Tabela 2 mostra os resultados obtidos, onde foram estabelecidos 19 e 11 parágrafos para o **Grupo 1** e **Grupo 2**, respectivamente.

Tabela 2: Termos x Parágrafos e seus respectivos Grupos

Parágrafo	**nutrigenoma**	**alimento**	**gene**	**genoma**	**Grupo**
1	1	0	0	0	**1**
2	2	1	1	0	**2**
3	0	1	1	0	**2**
4	0	1	0	0	**1**
5	0	2	1	0	**2**
6	0	2	0	0	**1**
7	1	1	0	0	**2**
8	0	3	0	0	**1**
9	1	0	0	0	**1**
10	0	0	1	0	**1**
11	0	0	1	0	**1**
12	1	0	0	0	**1**
13	1	0	0	0	**1**
14	2	0	1	0	**2**
15	0	0	1	0	**1**
16	1	0	0	0	**1**
17	0	0	0	3	**1**
18	0	0	0	2	**1**
19	1	1	1	2	**2**
20	1	0	0	0	**1**
21	1	0	0	0	**1**
22	1	1	0	1	**2**
23	1	0	2	2	**2**
24	0	0	1	1	**2**
25	1	1	1	0	**2**
26	0	0	2	0	**1**
27	1	0	0	0	**1**
28	2	1	0	0	**2**
29	2	0	0	0	**1**
30	0	0	0	1	**1**

Fonte: Autoral, 2021.

A QDA permitiu obter as médias das frequências relativas 0.55, 0.25, 0.35 e 0.20 dos termos "nutrigenoma", "alimento", "gene" e "genoma" no **Grupo 1** e similarmente, 1.00, 1.00, 0.70 e 0.80 para estes termos no **Grupo 2**.

A matriz de covariâncias amostrais dos **Grupos 1** e **2** inserida no Quadro 2, indica a associação entre os termos, sendo que todas as covariâncias do **Grupo 1**, agregado de frequências pouco expressiva, mostraram associação inversa para os pares de termos mas, no **Grupo 2**, os pares nutrigenoma/gene e gene/genoma tiveram associação direta. Na diagonal principal do **Grupo 1** foi indicado que

maior variabilidade segundo a frequência para o termo "nutrigenoma", enquanto no **Grupo 2** para "genoma".

Quadro 2: Matrizes de Covariâncias

Termos	nutrigenoma	alimento	gene	genoma
		GRUPO 1		
nutrigenoma	**0.3743**	-0.1754	-0.1462	-0.1754
alimento	-0.1754	**0.6725**	-0.0877	-0.1053
gene	-0.1462	-0.0877	**0.3158**	-0.0877
genoma	-0.1754	-0.1053	-0.0877	**0.6725**
		GRUPO 2		
nutrigenoma	**0.6000**	-0.1000	-0.1000	-0.1000
alimento	-0.1000	**0.3636**	-0.1364	-0.1909
gene	-0.1000	-0.1364	**0.3636**	0.2091
genoma	-0.1000	-0.1909	0.2091	**0.6727**

Fonte: Autoral, 2021.

A covariância combinada permitiu fazer o teste de igualdade dos vetores das médias pela estatística T^2 de *Hotelling* que assumiu valor de 0.1840, conduzindo a um risco de 0.34% (P_{valor}) de acreditar nesta hipótese. Assumindo condição de Normalidade, a função quadrática obteve os valores do log do determinante da matriz covariância por grupo e a Distância Quadrática de Mahalanobis. As equações discriminantes assumiram as expressões:

$$Z_1 = (1.4167 \times t_1) + (0.7857 \times t_2) + (1.5714 \times t_3) + (0.7143 \times t_4) \tag{11}$$

$$Z_2 = (1.2690 \times t_1) + (2.1764 \times t_2) + (1.6626 \times t_3) + (2.5690 \times t_4) \tag{12}$$

onde as variáveis $t_{1,2,3,4}$ representam os termos "nutrigenoma", "alimento", "gene" e "genoma", respectivamente. Observa-se que na primeira equação, os termos "nutrigenoma" e "gene" são os mais citados com coeficientes 1.4167 e 1.5714, respectivamente. Na segunda equação lideraram os termos "genoma", seguido de "alimento" com maiores participações, 2.5690 e 2.1764.

Os coeficientes lineares discriminatórios inferiram as raízes características, sendo que no **Grupo 1** os valores foram 0.3156, 0.1751, 0.3501, 0.1592, enquanto para o **Grupo 2** assumiram os valores 0.1653, 0.2835, 0.2166 e 0.3346. As raízes características acumuladas do **Grupo 1** inferem que três termos representariam 84.08% da variabilidade das frequências tratadas, sendo que no **Grupo 2** este indicador seria de 66.54%.

A Matriz de Classificação, Quadro 3, correspondente à Validação Cruzada, a qual teve o intuito de validar os resultados e verificar se a classificação prévia ser considerada errônea. Este procedimento enfatiza a sua importância, pois verifica se os grupos foram classificados com maior precisão. Observa-se que a Validação Cruzada faz uma possível identificação de classificações de grupos realizadas previamente de forma errônea, o que significa que seu uso faz com o que os grupos sejam classificados de forma mais precisa.

Quadro 3: Matriz de Classificação

	Atual	**Previsto**
	Grupo 1	**Grupo 2**
Grupo 1	17	3
Grupo 2	1	9

Fonte: Autoral, 2021.

Foram calculadas as taxas de erro e acurácia do modelo, cujos valores foram de 0.13 (4/30) e 0.87 (26/30), respectivamente, derivados da soma dos erros e acertos dos **Grupos 1** e **2**. Por fim, foi confrontado o resultado da Matriz de Classificação com os da Validação Cruzada, acompanhadas da probabilidade de pertencer ou não aquele grupo proposto pela validação, Tabela 3.

Observa-se que mesmo se a Validação Cruzada forneça o mesmo grupo que a classificação prévia, algumas probabilidades apontaram que havia pequena possibilidade de pertencimento em outro grupo.

Dentre os casos destacam-se os ocorridos nas linhas **3**, **4** e **22**, sendo que nesta última ocorre quase um empate entre as probabilidades, apontando que, ou a classificação prévia foi realizada de maneira equivocada ou os termos deste parágrafo não pertencem a nenhum dos dois grupos.

Tabela 3: Classificação e probabilidade de pertencimento ao grupo proposto pela Validação

Parágrafo	Inicial	Previsto	Prob. Previsto	Prob. Não Previsto
1	1	1	0.9841	0.0159
2	2	2	0.0000	1.0000
3	**1**	**1**	**0.8027**	**0.1973**
4	**1**	**1**	**0.7873**	**0.2127**
5	2	2	0.9694	0.0306
6	1	1	0.6478	0.3522
7	1	1	0.7632	0.2368
8	2	1	0.0001	0.9999
9	1	1	0.9841	0.0159
10	1	1	0.9618	0.0382
11	1	1	0.9618	0.0382
12	1	1	0.9841	0.0159
13	1	1	0.9841	0.0159
14	2	2	0.0005	0.9995
15	1	1	0.9618	0.0382
16	1	1	0.9841	0.0159
17	2	1	0.0026	0.9974
18	1	1	0.9938	0.0062
19	2	2	0.0000	1.0000
20	1	1	0.9841	0.0159
21	1	1	0.9841	0.0159
22	**2**	**2**	**0.5138**	**0.4862**
23	2	2	0.0000	1.0000
24	1	1	0.9960	0.0040
25	2	2	0.0064	0.9936
26	1	1	0.0000	1.0000
27	1	1	0.9841	0.0159
28	2	2	0.0040	0.9960
29	1	2	0.0023	0.9977
30	1	1	0.9892	0.0108

Fonte: Autoral, 2021.

4 Considerações Finais

O suporte computacional para a análise dialógica do discurso da Nutrigenômica e Nutrigenética na busca de palavras-chaves tornou-se viável pelo suporte computacional dado à adoção da Análise Discriminante (AD).

Neste tema ocorreram resultados relevantes no contexto da Nutrição, uma vez que o modelo proposto caracterizou conglomerados cognitivos textuais inferindo os

atributos verbais julgados mais importantes que permitiram reconhecer palavras julgadas padrões textuais, tendo avaliação técnico-científico de um especialista na área de Nutrição.

A aplicação da Análise Discriminante (AD) impôs um ajuste no banco de dados para que esta metodologia se tornasse viável. A verificação se houve classificação correta dos grupos mostrou que 87% de acertabilidade.

Logo, observa-se que o procedimento pode ser útil para o auxílio na verificação de classificação dos grupos em diferentes estudos, possibilitando verificar se a alocação dos objetos nos grupos se realizou de forma correta ou não. O destaque do presente estudo foi apresentar um procedimento de validação necessário quando considerado as particularidades de outras propostas de coletâneas textuais.

Agradecimentos

Agradecimento à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001, processo 88881.506840/2020-01.

Referências

[1] BARANAUSKAS, J. A.: **Mineração de Textos**. Relatório Técnico. Departamento de Física e Matemática da Universidade de São Paulo, São Paulo, 2017. Disponível em: http://dfm.ffclrp.usp.br/~augusto.

[2] CARRILHO JUNIOR, J. R.: **Desenvolvimento de uma Metodologia para Mineração de Textos**. 96f. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1.

[3] DERAM, S.: **O que é Nutrigenômica?** Blog Sophie Deram, 2016. Disponível em: https://sophiederam.com/br/nutricoaching/nutrigenomica/#:~:text=Para%20voc%C3%AA%20entender%20o%20que,determinado%20para%20a%20vida%20inteira.

[4] DIXON, M.: **An Overview of Document Mining Technology**, 1997. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download.

[5] EQUIPE NUTMED: **O que é Nutrigenômica?** Site NutMed, s.d. Disponível em: https://nutmed.com.br/blog/novidades/noticia-67#:~:

text=Mas;%20voc%C3%AA%20sabe%20o%20que,entre%20indiv%C3%ADduos%20ou%20grupos%20populacionais.

[6] FELLOWS, I. *et al.*: **Package 'wordcloud'**. R package version, v. 2, p. 331, 2018. Disponível em: ftp://cran.wu-wien.ac.at/pub/R/web/packages/wordcloud/wordcloud.pdf.

[7] FISHER, R. A.: **The use of multiple measurements in taxonomic problems**. Annals of Eugenics, 179-188, 1936.

[8] GARETH, J. *et al*: **An introduction to statistical learning: with applications in R**. Spinger, 2013.

[9] JOHNSON, R. A.; WICHERN, D. W.: **Applied multivariate statistical analysis**. Prentice-Hall, 2007.

[10] KHOURY JUNIOR, J. K.; *et al.*: **Análise Discriminante Paramétrica para Reconhecimento de Defeitos em Tábuas de Eucalipto Utilizando Imagens Digitais**. Revista Árvore, p. 299-309, 2015. Disponível em: https://www.scielo.br/j/rarv/a/Zx9fbLrhMgjfYbkq33jYwHb/abstract/?lang=pt.

[11] LATTIN, J.; CARROLL, J. D.; GREEN, P. E.: **Análise de Dados Multivariados**, Cengage Learning, 2011.

[12] MADEIRA, R. de O. C.: **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais**. 68f. Dissertação (Mestrado em Ciências, ênfase em Modelagem Matemática da Informação) - Fundação Getúlio Vargas, Rio de Janeiro, 2015. Disponível em: http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/14593/TEXTO%20DISSERTA%c3%87%c3%83O%20VFINAL1.pdf?sequence=1&isAllowed=y.

[13] MANLY, B. J. F.: **Métodos Estatísticos Multivariados - uma introdução**. New York: Bookman, 2008.

[14] MORAIS, E. A. M.; AMBRÓSIO, A. P. L.: **Mineração de Textos**. Goiânia: Universidade Federal de Goiás, 2007. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf.

[15] MORRISON, D. F.: **Multivariate Statistical Methods**. New York: McGraw-Hill, 1976.

[16] MÜLLER, M.; SANDER, K.: **Nutrigenomics: goals and strategies**. Nature Reviews Genetics, v. 4, n. 4, p. 315-322, 2003.

[17] MUTCH, D. M.; WAHLI W.; WILIAMSON, G.: **Nutrigenomics and nutrigenetics: the emerging faces of nutrition**. The FASEB journal, v. 19, n. 12, p. 1602-1616, 2005.

[18] RSTUDIO TEAM: **RStudio: Integrated Development Environment for R**. RStudio, Inc., Massachusetts: Boston, 2015. Disponível em: https://www.r-project.org/conferences/useR-2011/abstracts/180111-allairejj.pdf.

[19] STAUDT JUNIOR, J. L.: **Text Mining Utilizando o Software R: um estudo de caso de uma biblioteca Americana**. 49f. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: https://lume.ufrgs.br/handle/10183/149102.

[20] TONIAL, G.: **Nutrigenômica: nutrição a nível molecular**. Site Profissão BioTec, 2016. Disponível em: https://profissaobiotec.com.br/nutrigenomica-nutricao-a-nivel-molecular/#:~:text=A%20nutrigen%C3%B4mica%20pode%20ser%20entendida,cont%C3%A9m%20aproximadamente%2025%20mil%20genes.

[21] VARELLA, C. A. A.: **Análise Multivariada Aplicada as Ciências Agrárias - Análise Discriminante**. Notas de aula. Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, 2016. Disponível em: http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/ANALISE%20DISCRIMINANTE.pdf.

[22] VENABLES, W. N.; RIPLEY, B. D.: **Modern Applied Statistics with S**. Springer, 2002.

[23] WIVES, L. K.: **Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva**. 116f. Tese (Doutorado em Computação), Universidade Federal do Rio Grande do Sul, 2002. Disponível em: https://seer.ufrgs.br/cadernosdeinformatica/article/view/v1n1p25-28.