Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite

Enio Júnior Seidel,Fernando de Jesus Moreira Júnior,
Angela Pelegrin Ansuj, Maria Rosane Coradini Noal

Departamento de Estatística/CCNE
Universidade Federal de Santa Maria/Santa Maria, RS
e-mail: ejrseidel@hotmail.com

Resumo

Este trabalho tem por objetivo comparar os resultados obtidos pelos métodos Ward e K-médias, no agrupamento por similaridade de produtores de leite da região de Santa Maria, RS. Foram utilizadas 231 amostras de leite *in natura* de 63 produtores, coletadas no período de 6 a 30 de setembro de 2004. As variáveis analisadas foram: percentual de água (%), percentual de gordura; acidez em graus Dornic e densidade em g/cm^3. Em ambos os métodos, os 63 produtores foram agrupados em três *clusters*, dos quais 52 produtores ficaram nos mesmos *clusters* nos métodos analisados.
Palavras chave: Produtores de leite, método Ward, método K-médias.

Abstract

In this paper method Ward and method K-Means have been used to compare the result of the clusters to the producers of milk in the region of Santa Maria, RS. Were used 231 sample of raw milk from the 63 producers colleted from 6 to 30 of September of 2004. Were analyzed the flowing variables: percentage of water; percentage of fat, acidity in degree Dornic; density in g/cm^3. Both two methods show three clusters for the producers which 52 producers stay in the same clusters in both two methods.
Key-words: Producers of milk, method Ward, method K-means.

1. Introdução

A composição do leite varia com a espécie, raça, individualidade, alimentação, tempo de gestação e muitos outros fatores inerentes ao local de produção do leite (Valsechi, 2001). Assim, para que se tenha um melhor gerenciamento do processo de produção dos laticínios de modo que a estrutura do processo de transformação esteja de acordo com as caracterísi-

cas dos lotes recebidos, é necessário analisar e conhecer o tipo de leite que cada fornecedor dispõe.

Porém, um dos grandes problemas encontrados é a escolha de um método estatístico que possa definir da melhor forma possível os grupos de fornecedores conforme as características encontradas nas amostras de leite. Por isso, esse trabalho tem por objetivo comparar os resultados obtidos pelo método de Ward (aglomeração hierárquica) e pelo método K-médias (aglomeração não-hierárquica), no agrupamento de produtores de uma indústria de laticínios localizada na cidade de Santa Maria/RS, de aço rdo com as características do leite fornecido. Esses métodos foram escolhidos por serem os mais usados e apresentarem bons resultados. A importância do trabalho está no fato de que, através da comparação dos métodos de agrupamentos citados, pode-se obter uma melhor maneira para formar os conglomerados.

2. Metodologia

Utilizaram-se 231 amostras de leite *in natura* de 63 produtores de uma indústria de laticínios de Santa Maria/RS, coletadas no período de 6 a 30 de setembro de 2004, onde foram analisadas as seguintes variáveis: Porcentagem de água, Porcentagem de gordura, Acidez em graus Dornic e Densidade em g/cm^3. A medida de tendência central utilizada para tratar os dados de um mesmo produtor foi a mediana dos dados de cada variável de suas amostras, a fim de obter um valor que representasse cada produtor. Para o tratamento estatístico foram utilizados os Métodos Ward e o Método K-médias com o auxílio do software Statistica 7.0. Para comparar os dois métodos de agrupamento, foi utilizado o coeficiente de concordância de Kappa, que verifica a concordância entre os resultados obtidos pelos dois métodos.

3. Análise de agrupamentos (AA)

Segundo Malhotra (2006), a análise de agrupamento, ou análise de *clusters*, é uma técnica usada para classificar objetos ou casos em grupos relativamente homogêneos chamados de agrupamentos ou conglomerados. Assim, os objetos em cada agrupamento tendem a ser semelhante entre si, mas diferentes de objetos em outros agrupamentos.

Por isso, três questões fundamentais devem ser consideradas na aplicação da análise de agrupamento: primeira, como será medida a similaridade dos dados; segunda, como formar os agrupamentos; e por fim como decidir quantos grupos formar. Um procedimento para se efetuar a análise de agrupamentos é mostrado na Figura 1.

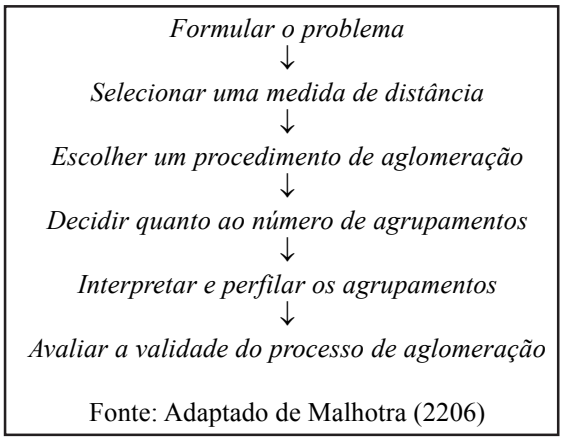

Figura 1. Etapas para efetuar a análise de agrupamento.

Conforme Hair *et al* (2005), as características de cada objeto são combinadas em uma medida de semelhança, que pode ser de similaridade ou dissimilaridade, calculada para todos os pares de objetos, possibilitando a comparação de qualquer objeto com outro pela medida de similaridade e a associação dos objetos semelhantes por meio da análise de agrupamento. As medidas de distância representam a similaridade, que é representada pela proximidade entre as observações ao longo das variáveis.

A distância euclidiana é a medida de distância mais freqüentemente empregada quando todas as variáveis são quantitativas. A distância euclidiana é utilizada para calcular medidas específicas, assim como a distância euclidiana simples e a distância euclidiana quadrática ou absoluta, que consiste na soma dos quadrados das diferenças, sem calcular a raiz quadrada.

A distância euclidiana quadrática é definida por:

$$DE = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \qquad (3.1)$$

Onde:

x_{ij} é a j-ésima característica do i-ésimo indivíduo;

$x_{i'j}$ é a j-ésima característica do i'-ésimo indivíduo.

Quanto mais próximo de zero for a distância euclidiana, mais similares são os objetos comparados.

3.1. Procedimentos de aglomeração

A aglomeração hierárquica se caracteriza pelo estabelecimento de uma hierarquia ou estrutura em forma de árvore. A aglomeração hierárquica interliga os objetos por suas associações, produzindo uma representação gráfica chamada de dendrograma, onde os objetos semelhantes, segundo as variáveis estudadas, são agrupados entre si. Já na aglomeração não-hierárquica, assume-se um centro de agrupamento e, em seguida, agrupam-se todos os objetos que estão a menos de um valor pré-estabelecido do centro.

Para os procedimentos de aglomeração deve-se optar por um método específico. Neste trabalho serão utilizados o método Ward para a aglomeração hierárquica e o método K-médias para a aglomeração não-hierárquica.

3.2. Método Ward

Segundo Hair *et al* (2005), o método de Ward consiste em um procedimento de agrupamento hierárquico no qual a medida de similaridade usada para juntar agrupamentos é calculada como a soma de quadrados entre os dois agrupamentos feita sobre todas as variáveis. Esse método tende a resultar em agrupamentos de tamanhos aproximadamente iguais devido a sua minimização de variação interna. Em cada estágio, combinam-se os dois agrupamentos que apresentarem menor aumento na soma global de quadrados dentro dos agrupamentos.

3.3. Método K-médias

É um método de partição que fornece indicações mais precisas sobre o número de conglomerados a ser formado. Este método talvez seja um dos mais utilizados quando se têm muitos objetos para agrupar, com pequenas variações. O critério mais utilizado de homogeneidade dentro do grupo e heterogeneidade entre os grupos é o da soma dos quadrados residual baseado na Análise de Variância. Assim, quanto menor for este valor, mais homogêneos são os elementos dentro de cada grupo e melhor será a partição (Bussab *et al*, 1990).

4. Resultados e discussão

Para a análise dos dados seguiu-se a sugestão de Malhotra (2006), onde utiliza-se inicialmente o procedimento hierárquico e após a aglomeração não-hierárquica. Assim, utilizou-se o método Ward, que tem se reve-

lado um dos melhores, e mais usados, métodos hierárquicos de aglomeração (Malhotra, 2006) (Kubrusly, 2001).

Os grupos formados estão apresentados na Figura 2, que evidencia a formação de três agrupamentos, baseados no corte feito na maior distância entre grupos. Dentro de cada agrupamento temos produtores com características similares e entre os agrupamentos verificamos características distintas para os produtores de leite.

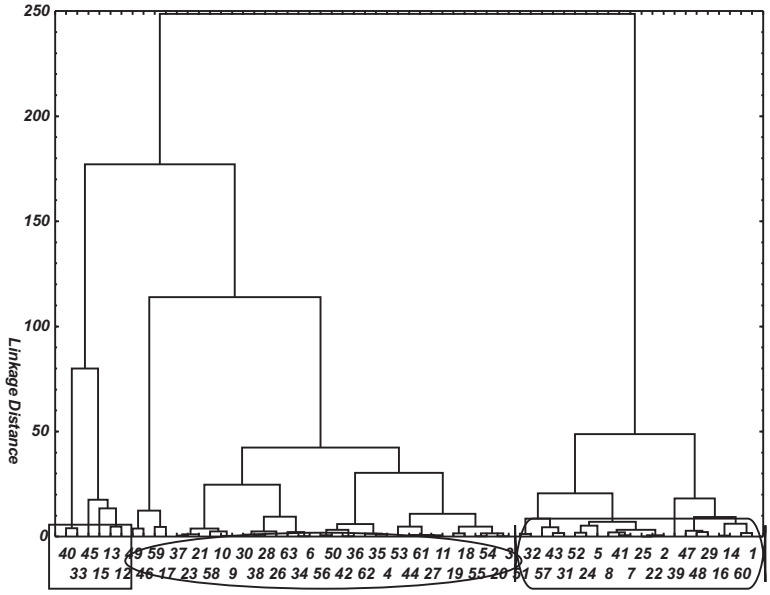

Figura 2. Dendrograma para os 63 produtores de leite *in natura*.

No primeiro agrupamento formado, podemos identificar os fornecedores 12, 13, 15, 33, 40 e 45. Através de uma análise descritiva podemos observar que todos esses produtores fornecem leite com baixas taxas de densidade (em média 1025,73g/cm^3) e acidez (em média 15,03°D).

O segundo grupo é composto pelos fornecedores 3, 4, 6, 9, 10, 11, 17, 18, 19, 20, 21, 23, 26, 27, 28, 30, 34, 35, 36, 37, 38, 42, 44, 46, 49, 50, 53, 54, 55, 56, 58, 59, 61, 62 e 63. Este conglomerado se caracteriza por altas taxas de água excedente (em média 4,66%) e baixo teor de gordura (em média 3,88%). Os fornecedores 17, 46, 49, 59 e 63 tiveram as maiores taxas de água excedente, em média, 7,25%.

E o terceiro agrupamento foi formado pelos fornecedores 1, 2, 5, 7, 8, 14, 16, 22, 24, 25, 29, 31, 32, 39, 41, 43, 47, 48, 51, 52, 57 e 60. Neste grupo podemos observar amostras de leite com alto teor de acidez (em

média 16,53°D) e densidade elevada (em média 1030,48 g/cm^3). Os fornecedores 1, 5, 7, 31, 32, 41, 43, 51 e 52 apresentaram as taxas mais elevadas de acidez e densidade, onde a acidez variou de 16 a 18°D, e a densidade oscilou entre 1030 e 1032g/cm^3. A análise descritiva destes agrupamentos é apresentada na Tabela 1.

Tabela 1. Análise descritiva dos agrupamentos encontrados pelo método Ward.

	Cluster 1		*Cluster* 2		*Cluster* 3	
Variáveis	Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
Água %	4,18	1,18	4,66	1,56	2,66	1,01
Acidez °D	15,02	1,16	15,32	0,82	16,52	0,96
Gordura %	5,39	2,72	3,88	0,66	4,10	0,72
Densidade	1025,73	0,95	1029,26	0,88	1030,48	0,97

Com o objetivo de verificar os agrupamentos formados, recorreu-se ao método K-médias para constituir grupos de fornecedores a partir das mesmas variáveis usadas no método anterior. Para proceder aos agrupamentos, optou-se por escolher três agrupamentos, com base no número de agrupamentos encontrados no dendrograma da Figura 2. Os três agrupamentos formados pelo método K-médias estão na Tabela 2.

Podemos observar que no primeiro agrupamento, se encontram os fornecedores 6, 9, 10, 12, 13, 15, 21, 23, 26, 30, 33, 34, 37, 40, 45 e 58. Este agrupamento se destaca pela baixa densidade (em média 1027,23g/cm^3) e acidez (em média 15,12 °D) no leite.

Já no segundo agrupamento temos os fornecedores 3, 4, 11, 17, 18, 20, 27, 28, 35, 36, 38, 42, 44, 46, 49, 50, 53, 54, 55, 56, 59, 61, 62 e 63. Neste agrupamento temos amostras com baixas percentagens de gordura (em média 3,70%) e altas percentagens de água excedente (em média 5,25%).

E no último agrupamento podemos ver os fornecedores 1, 2, 5, 7, 8, 14, 16, 19, 22, 24, 25, 29, 31, 32, 39, 41, 43, 47, 48, 51, 52, 57 e 60. Este agrupamento é caracterizado por apresentar altas taxas de acidez (em média 16,59°D) e densidade (em média 1030,43g/cm^3) no leite. A análise descritiva destes agrupamentos é apresentada na Tabela 3.

Tabela 2. Resultado da aglomeração pelo método K-médias.

Cluster 1	*Cluster* 2		*Cluster* 3	
6	3	53	1	43
9	4	54	2	47
10	11	55	5	48
12	17	56	7	51
13	18	59	8	52
15	20	61	14	57
21	27	62	16	60
23	28	63	19	
26	35		22	
30	36		24	
33	38		25	
34	42		29	
37	44		31	
40	46		32	
45	49		39	
58	50		41	

Tabela 3. Análise descritiva dos agrupamentos encontrados pelo método K-médias.

Variáveis	*Cluster* 1		*Cluster* 2		*Cluster* 3	
	Média	Desvi padrão	Média	Desvio padrão	Média	Desvio padrão
Água %	3,82	1,02	5,25	1,51	2,66	1,00
Acidez °D	15,12	0,83	15,46	0,91	16,59	0,98
Gordura %	4,64	1,75	3,70	0,59	4,12	0,74
Densidade	1027,23	1,64	1029,68	0,72	1030,43	1,00

Para uma comparação dos resultados obtidos pelos dois métodos utilizados nessa análise, passou-se a uma avaliação das classificações encontradas. Na Tabela 4, estão as freqüências de produtores classificados pelos dois métodos.

Tabela 4. Classificação de produtores em agrupamentos pelos métodos Ward e K-médias.

Método Ward	Método K-médias			Total
	Cluster 1	*Cluster* 2	*Cluster* 3	
Cluster 1	**6**	0	0	6
Cluster 2	10	**24**	1	35
Cluster 3	0	0	**22**	22
Total	16	24	23	63

Podemos perceber que dos 63 produtores classificados, 52 (82,54%) foram classificados nos mesmos agrupamentos pelos dois métodos. O Coeficiente de Concordância de Kappa, estatisticamente significativo (p=0,000) foi de 0,726, indicando uma concordância considerada entre moderada e forte entre os agrupamentos formados pelos métodos Ward e K-médias. Porém, comparando as análises descritivas das Tabelas 1 e 3, verifica-se que o método K-médias representou melhor o agrupamento 2 (baixas percentagens de gordura e altas percentagens de água excedente) e o agrupamento 3 (altas taxas de acidez e de densidade no leite).

Percebe-se que a utilização dos dois métodos conjuntamente é o melhor, pois utilizando o número de agrupamentos encontrado pelo método Ward, podemos definir quantos agrupamentos devem ser formados pelo método K-médias, de modo a encontrar grupos bem homogêneos internamente.

5. Conclusão

Na aplicação do método Ward foram encontrados três agrupamentos. No primeiro tivemos baixas taxas de densidade e acidez; no segundo, altas taxas de água excedente e baixo teor de gordura; e no terceiro agrupamento tivemos alto teor de acidez e densidade elevada.

Para o método K-médias, adotou-se três agrupamentos. O pri-

meiro definido por baixa densidade no leite; o segundo com baixas percentagens de gordura e altas percentagens de água excedente; e o terceiro agrupamento definido por apresentar altas taxas de acidez e densidade no leite.

Os resultados encontrados, pelos métodos Ward e K-médias, tiveram alta concordância, pois 82,54% dos produtores foram agrupados nos mesmos agrupamentos em ambos os métodos, evidenciando a eficiência da robustez dos agrupamentos formados pelos dois métodos.

Na utilização dos dois métodos conjuntamente para a formação de grupos de produtores de leite, o método Ward definiu com mais eficiência a quantidade de agrupamentos que devem ser utilizados, enquanto que o método K-médias classificou de forma mais adequada os produtores dentro dos agrupamentos.

Para futuros trabalhos, sugere-se relacionar a composição do leite com as características locais dos produtores como, por exemplo, raça do animal, alimentação, clima e outras. Também se recomenda o aumento do número de lotes a serem amostrados por produtor, de modo a aumentar a eficiência da análise.

Referências

BUSSAB, W.O.; MIAZAK, E.S.; ANDRADE, D.F. *Introdução à Análise de Agrupamentos*. 9º Simpósio Brasileiro de Probabilidade e Estatística. São Paulo: IME – USP, 1990.

HAIR, J. F., et al. *Análise multivariada de dados*. Trad. Adonai S. Sant'Anna e Anselmo C. Neto. 5 ed. Porto Alegre: Bookman, 2005.

KUBRUSLY, L. S. Um procedimento para calcular índices a partir de uma base de dados multivariados. *Pesquisa Operacional*, Rio de Janeiro, v. 21, n. 1, 2001.

MALHOTRA, N. *Pesquisa de marketing: uma orientação aplicada*. Trad. Laura Bocco. 4 ed. Porto Alegre: Bookman, 2006.

VALSECHI, O. A. O leite e seus derivados. *Tecnologia de Produtos Agrícolas de Origem Animal*. Araras, 2001. Disponível em: http://www.cca.ufscar.br/docentes/vico/O%20LEITE%20E%20SEUS%20DERIVADOS.pdf. Acesso em: 18 Dez 2006.