ConBRepro
X CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO
EVENTO ON-LINE
02 a 04 de dezembro 2020

Clusterização de Clientes: um Modelo Utilizando Variáveis Categóricas e Numéricas

Fernanda Robes de Oliveira
Engenharia da Produção – Universidade Federal do Paraná
Mariana Kleina
Engenharia da Produção – Universidade Federal do Paraná
Marcos Augusto Mendes Marques
Engenharia da Produção – Universidade Federal do Paraná
Jessika Alvares COppi Arruda Gayer
Engenharia da Produção – Universidade Federal do Paraná
Thiago Shoji Obi Tamachiro
Engenharia da Produção – Universidade Federal do Paraná

Resumo: Conhecer o perfil dos clientes é importante em um mercado competitivo, assim a segmentação de clientes por características é imprescindível para um *marketing* de relacionamento eficaz, pois faz com que os clientes certos recebam a comunicação acertada. Porém agrupar os clientes pode não ser uma tarefa fácil, há várias metodologias e algoritmos disponíveis e saber escolher entre elas é crucial para obter resultados favoráveis. Em geral, para esta tarefa as empresas dispõem de um conjunto de dados que contêm tanto variáveis categóricas quanto numéricas, que devem ser processadas de maneira diferente, de acordo com o seu tipo. Este trabalho visa demonstrar uma metodologia de clusterização de clientes que utiliza o algoritmo de clusterização K-medoids e a distância de Gower em uma base de dados heterogênea, ou seja, que contém variáveis categóricas e numéricas, demostrando como processar os dados e aplicar os algoritmos, bem como validá-los por meio dos índices de validação de cluster Davies Bouldin e do Coeficiente da Silhueta, de forma que os resultados possam ser utilizados na comunicação da empresa com o mercado.

Palavras-chave: Clusterização, K-Medoids, variáveis categóricas e numéricas base heterogênea, agrupamento.

Clustering Customers a Model Using Categorical and Continuous Variables

Abstract: Knowing the profile of customers is important in a competitive market, so segmenting customers by characteristics is essential for effective relationship marketing, as it makes sure the right customers receive the right communication. However, grouping customers may not be an easy task, there are several methodologies and algorithms available and knowing how to choose between them is crucial to obtain favorable results. In general, for this task companies have a set of data that contains both categorical and numerical variables, which must be processed differently, according to their type. This work aims to demonstrate a client clustering methodology that uses the K-medoids clustering algorithm and the Gower distance in a heterogeneous database, that is, that contains

categorical and numeric variables, demonstrating how to process the data and apply the algorithms, as well as validating them through the Davies Bouldin cluster validation indexes and the Silhouette Coefficient, so that the results can be used in the company's communication with the market.

Keywords: Clustering, K-Medoid, categorical and numeric variables, heterogeneous basis, grouping.

1. Introdução

As necessidades dos clientes tornam-se cada vez mais complexas e dinâmicas em virtude de novos produtos e serviços que são lançados todos os dias no mercado (WANG; GAO, 2019); a diversificação dos produtos focada na segmentação dos clientes é uma ferramenta essencial para as empresas se manterem em um ambiente de negócios competitivo (LIN et al., 2019) pois contribui para manter e atrair clientes (KEVREKIDIS et al., 2018).

A clusterização é um processo que pode ser definido como a ação de separar os clientes com base nas características em grupos distintos e significativos (ABBASIMEHR; SHABANI, 2019) de modo que os clientes do mesmo agrupamento tenham necessidades e preferências semelhantes. Isto é fundamental quando a empresa está buscando entender o mercado, pois, por meio da sua implementação, há possibilidade de estruturar uma estratégia de *marketing* mais eficiente para cada segmento (BARMAN; CHOWDHURY, 2019). Esta ação que auxilia as empresas a criar relacionamento com o público-alvo (LIU et al., 2017) oferecendo serviços personalizados, bem como ofertas, promoções e recomendações às preferências de cada segmento (JAGABATHULA; SUBRAMANIAN; VENKATARAMAN, 2018), (GRIVA et al., 2018), o que faz os clientes responderem as estratégias de forma mais eficiente (WANG; CHIN, 2017).

Porém, a segmentação pode não ser tão fácil em empresas que tem milhares de clientes, desta forma é fundamental utilizar técnicas que auxiliem neste contexto (CAMERO et al., 2018). Ferramentas e técnicas de análise de negócios e a tomada de decisões orientada a dados estão cada vez mais presentes para suprir as necessidades empresariais (GRIVA et al., 2018).

Uma técnica de mineração de dados utilizada para este fim é o agrupamento ou clusterização, que é um método não supervisionado para reunir os dados em grupos semelhantes (BARMAN; CHOWDHURY, 2019), que tem como objetivo de alocar as observações em agrupamentos homogêneos internamente e heterogêneos entre si com a função de representar o comportamento das variáveis.

Nos agrupamentos, semelhança das observações é baseada em distância, tal que observações mais semelhantes têm distância menor (BUDIAJI; LEISCH, 2019). Desta forma um dos principais requisitos dos algoritmos de agrupamento são as medidas de distância (dissimilaridade) para variáveis numéricas ou semelhança (similaridade) para variáveis binárias (FÁVERO; BELFIORE, 2017). Deste modo a escolha da medida a ser utilizada leva em conta o tipo das variáveis.

Assim, o primeiro passo é analisar as variáveis que serão utilizadas no processo de clusterização. Geralmente os dados são heterogêneos, ou seja, possuem variáveis categóricas e numéricas (HARIKUMAR; SURYA, 2015). Neste caso há algumas abordagens que podem ser utilizadas, como dicotomizar as variáveis numéricas e deste modo utilizar uma medida de semelhança para dados binários gerados. Outra a alternativa é utilizar uma medida híbrida (HUNT; JORGENSEN, 2011).

Porém, Harikumar e Surya (2015) apontam que utilizar a técnica de dicotomização para transformar um conjunto de dados heterogêneos em homogêneos, pode levar a perda de informação.

Após definida a medida de distância, o algoritmo de agrupamento deve ser escolhido. Há vários algoritmos disponíveis, tais como o tradicional *K-means*, que é limitado a dados numéricos (BUDIAJI; LEISCH, 2019), pois a distância euclidiana é falha em capturar a similaridade de atributos categóricos (AHMAD; DEY, 2007).

Assim, neste trabalho, será aplicado o algoritmo *K-medoids* em uma base de dados heterogêneos, pois além de ser menos sensível a *outliers*, é mais flexível a utilização de diferentes distâncias (DE ASSIS; DE SOUZA, 2011).

2. Referencial Teórico

2.1 K-Medoids

O algoritmo é baseado no objeto (*medoids*) localizado mais centralmente em um cluster. É menos sensível a *outliers*, se comparado ao *K-means*, além de ser mais flexível, permitindo diferentes tipos de distância (DE ASSIS; DE SOUZA, 2011).

Usa pontos de referência ao invés do valor médio dos elementos em cada *cluster*, e precede da entrada do número de clusters (UMAMAHESWARI; DEVI, 2018). No entanto para auxiliar na escolha do número de cluster, existem métodos de validação que podem ser utilizados.

Há vários algoritmos para agrupamento de *K-medoids*, porém segundo Park e Jun (2009), o Particionamento em torno de medoids (PAM), proposto por Kauf-man e Rousseeuw (1990), é conhecido por ser mais poderoso, desta forma foi utilizado neste trabalho.

Reynolds, Richards e Rayward-smith (2004) resumiram o algoritmo da seguinte forma:

a) Selecione k objetos aleatoriamente para se tornarem medoides dos clusters iniciais;
b) Calcule a matriz de dissimilaridade se ela não foi fornecida;
c) Atribua cada objeto ao seu medoide mais próximo;
d) Recalcule as posições dos k medoides;
e) Repita as etapas b e c até que os medoides se tornem fixos.

2.2 Distância de Gower

Proposta por Gower em 1971, a distância é capaz de lidar com conjunto de dados heterogêneos e segundo Bektas e Schumann (2019), aborda esta questão de uma maneira eficaz. A distância é calculada pela equação (1):

$$S_{ij} = \frac{\sum_{k=1}^{p} W_k S_k}{\sum_{k=1}^{p} W_k} \quad (1)$$

Em que S_{ij} é a distância entre os elementos x_i e x_j, com $i \neq j$. Se a k-ésima variável é qualitativa tem-se S_k pela equação (2) :

$$S_k = \begin{cases} 0\,, se\ x_{ki} = x_{kj} \\ 1\,, se\ x_{ki} \neq x_{kj} \end{cases} \quad (2)$$

Caso a k-ésima variável for quantitativa, S_k será dado pela equação (3):

$$S_k = \frac{|x_{ki} - x_{kj}|}{max(x_k) - min(x_k)} \quad (3)$$

Onde:

- k : 1, 2, ..., p;
- p : número total de variáveis;
- x_{ki} : é o valor da k-ésima variável para o elemento i;
- i : 1, 2, ..., n;
- n : número de observações;

— W_k : é igual a 1 (um) quando se tem os valores da k-ésima variável para ambos elementos e 0 (zero), quando não se tem os valores da k-ésima variável para quaisquer dos dois elementos.

2.3 Índices de Validação do Cluster

A escolha do número de grupos é de importância central, mas ainda é um dos problemas mais difíceis na análise de *cluster*, porque geralmente é questionável e não é única solução existente (BRENTARI; DANCELLI; MANISERA, 2016). Há vários índices que auxiliam na escolha, tais como o Davies Bouldin e o coeficiente da Silhueta.

2.3.1 Davies-Bouldin

É um índice que avalia a validade do *cluster* (SHEIKH; GHANBARPOUR; GHOLAMIANGONABADI, 2019) em função da proporção da dispersão dentro do *cluster*, e a separação entre *clusters.*

Segundo Ganmawu e Wells (2007), para definir o índice é necessário primeiro definir a medida de dispersão e similaridade do *cluster*.

A medida de dispersão do *cluster* é apresentada na equação (3) :

$$S_i = \left(\frac{1}{|C_i|} \sum_{x \in C_i} d^p(x, c_i) \right)^{\frac{1}{p}}, \qquad p > 0 \tag{4}$$

Onde:
— S_i : é a medida de dispersão do *cluster* C_i;
— $|C_i|$: é o número de pontos no *cluster* C_i;
— c_i : é o centro do *cluster* C_i;
— d : é a distância entre x e c_i.

Normalmente o valor de p é 2, o que torna esta uma distância euclidiana.

Já medida de similaridade entre os *clusters* pode ser definida como R_{ij} , que mensura a similaridade entre os *clusters* C_i e C_j .

Assim R_{ij} é demonstrada na equação (4):

$$R_{ij} = \frac{S_i + S_j}{D_{ij}} \tag{5}$$

Onde:
— D_{ij}: é a distância entre os centroides dos *clusters*.

Como S_i e S_j são as distâncias intra-*cluster* dos *clusters* i e j, ou seja, a distância média de cada instância ao centroide do *cluster* e D_{ij} é a distância entre os centroides dos *clusters* i e j (inter-*cluster*), assim os *clusters* i e j devem ser diferentes, caso contrário a distância entre os *clusters* será zero. Percebe-se que quanto menor R_{ij} , mais distantes e menor a dispersão entre os *cluster*s.

Desta maneira o índice Davies-Bouldin (V_{DB}) é representado na equação (5):

$$V_{DB} = \frac{1}{k} \sum_{k=1}^{k} R_i \tag{6}$$

Onde:

— k: Número de *clusters*;
— R_i: É o valor máximo R_{ij}.

A equação (6) demostra que para cada *cluster* i será selecionado o *cluster* j menos semelhante, e estre valor será dividido pela quantidade de *clusters*. Desta forma quanto melhor o agrupamento, mais próximo a zero será o índice.

2.3.2 Coeficiente da Silhueta

A Silhueta é uma medida da distância entre os grupos utilizada para determinar a qualidade do *cluster*. Para cada observação é definido um índice $\epsilon[-1,1]$, que compara a distância da observação ao *cluster*, com a heterogeneidade do *cluster*.

A análise se dá pela média do índice no *cluster*, assim quanto mais perto de 1, melhor a qualidade do *cluster*, ou seja, sua heterogeneidade é menor. Segundo Chang e Ho (2017), se o coeficiente for maior que 0,5, o cluster é mais eficaz em distinguir entre heterogêneo ou homogêneo. No **Quadro 1**, Kaufman e Rousseeuw (1990), demonstram como pode ser interpretado o coeficiente.

Quadro 1 - Interpretação do Coeficiente De Silhueta

Coeficiente de Silhueta	Interpretação
0,71 a 1,00	Estrutura forte
0,51 a 0,70	Estrutura razoável
0,26 a 0,50	Estrutura fraca
Menor que 0,25	Nenhuma estrutura

Fonte: Kaufman e Rousseeuw (1990)

O Coeficiente da Silhueta é calculado usando a distância intra-*cluster* média e a distância média do cluster mais próximo. O Coeficiente da Silhueta é demonstrado na equação (7):

$$Silhueta = \frac{b_i - a_i}{max\{a_i, b_i\}} \quad (7)$$

Onde:

— a_i: é a distância média entre a observação i às demais observações do *cluster*;
— b_i: é a distância média entre a observação i e todas as observações do *cluster* mais próximo.

3. Metodologia

Este trabalho propõe aplicação da metodologia em um conjunto de dados de 10 mil clientes, pessoas jurídicas, que participaram de treinamentos de gestão empresarial. O objetivo é agrupá-los de forma que a empresa, detentora dos treinamentos, possa efetuar um *marketing* de relacionamento diferenciado em cada grupo.

A metodologia foi implantada e os dados analisados por meio do software R. Para isto utilizou um conjunto de variáveis que contém características dos clientes, detalhadas no Quadro 3.

Quadro 2 – Detalhamento das Variáveis

Nome do Atributo	Tipo de variável	Descrição/Categorias
Porte da empresa	Categórica	MEI - Microempreendedor individual
		ME - Microempresa
		EPP - Empresa de Pequeno Porte

		MGE - Empresa de Médio e Grande Porte
Setor	Categórica	Agronegócios
		Comércio
		Serviços
		Indústria
Quantidade de funcionários	Numérica	Representa a quantidade de funcionários da empresa.
Idade da empresa	Numérica	Representa a idade da empresa em anos

Fonte: Os autores (2020).

Na tabela pode-se observar que há dados categóricos e numéricos, ou seja, a base é heterogênea. Desta forma, antes da aplicação faz-se necessário preparar os dados, de modo que seja possível aplicar a distância de Gower.

Primeiramente os dados categóricos foram dicotomizados, ou seja, transformados em variáveis binárias (*dummy*). Variáveis binárias são as que assumem valor de 0 ou 1, conforme a presença (valor 1) ou ausência (valor 0) da categoria. A quantidade de variáveis *dummies* é determinada pelo número de níveis, ou categorias, da variável original menos um.

Desta forma, por exemplo, a variável porte, que tem quatro categorias (MEI, ME, EPP e MGE), é transformada em três *dummies* (ME, EPP e MGE), conforme demostra a Tabela 1 em que é possível notar como cada variável categórica é representada de maneira dicotomizada. Neste exemplo, nota-se que a categoria MEI não é demostrada como uma variável dummy, isto decorre dela ser vinculada a ausência das demais categorias. Assim, sempre que ME, EPP e MGE forem iguais a zero, significa que o cliente em questão é do porte MEI.

Tabela 1 – Exemplo de dicotoização da variável Porte

	Dummy		
Categórica	**ME**	**EPP**	**MGE**
MEI	0	0	0
ME	1	0	0
EPP	0	1	0
MGE	0	0	1

Fonte: Os autores (2020).

A dicotomização das variáveis é necessária para que se possa analisar a similaridade entre as observações, item necessário na *clusterização* de variáveis qualitativas.

Com relação a quantidade de variáveis formadas, de duas variáveis qualitativas (Porte e Setor), cada uma com quatro categorias, formou-se seis *dummies*.

Já em relação as variáveis numéricas (Idade e Quantidade de funcionários), deve ser aplicada uma técnica de normalização dos dados. Isto segunto Umamaheswari e Devi (2018) pode aumentar a precisão e desempenho de algoritmos de mineração envolvendo distâncias. Desta maneira os atributos numéricos devem ser normalizados para serem considerados na mesma escala (AHMAD; DEY, 2007). Para isso foi considerada a normalização Min-Max, conforme demostrado na equação (8):

$$Z = \frac{y - \min(y)}{\max(y) - \min(y)} \tag{8}$$

Isto posto, as variáveis tiveram a escala alterada para uma variação entre 0 e 1. Na Tabela 2 é possível verificar, para alguns clientes, a normalização das variáveis numéricas.

Tabela 2 – Exemplo de Dados Normalizados

Código do Cliente	Número de Colaboradores	Número Normalizado de Colaboradores	Idade	Idade Normalizada
1	2	0,00033	12	0,21818
2	0	0,00000	4	0,07273
3	1	0,00017	4	0,07273
4	0	0,00000	4	0,07273
5	0	0,00000	2	0,03636
6	0	0,00000	4	0,07273

Fonte: Os autores (2020).

No final da preparação dos dados tem-se uma base com oito variáveis, todas na escala de 0 a 1, conforme a Tabela 3.

Tabela 3 – Exemplo da Base Após Preparação dos Dados

Código do Cliente	Número Normalizado de Colaboradores	Idade Normalizada	Porte ME	Porte EPP	Porte MGE	Setor Comércio	Setor Indústria	Setor Serviços
1	0,00033	0,21818	1	0	0	0	1	0
2	0,00000	0,07273	0	0	0	0	0	1
3	0,00017	0,07273	0	0	0	0	0	1
4	0,00000	0,07273	0	0	0	0	0	1
5	0,00000	0,03636	0	0	0	0	1	0
6	0,00000	0,07273	0	0	0	0	1	0

Fonte: Os autores (2020).

Após a preparação dos dados, a base está pronta para o cálculo das distâncias. Assim é calculada uma matriz com a distância de Gower, que demonstra a distância entre as observações dos clientes; na Tabela 4 pode-se ver o resultado da aplicação da distância de Gower nas observações dos seis primeiros clientes da base.

Tabela 4 – Distância de Gower aplicada nos dados de seis clientes

	1	2	3	4	5	6
1	0,000000	0,393223	0,393203	0,393223	0,147769	0,143223
2	0,393223	0,000000	0,000021	0,000000	0,254545	0,250000
3	0,393203	0,000021	0,000000	0,000021	0,254566	0,250021
4	0,393223	0,000000	0,000002	0,000000	0,254545	0,250000
5	0,147769	0,254546	0,254566	0,254545	0,000000	0,004545
6	0,143223	0,250000	0,250020	0,250000	0,004545	0,000000

Fonte: Os autores (2020).

A matriz de distância é utilizada pelo algoritmo *K-medoids* (PAM) para a determinação dos *clusters*. Porém antes de aplicar o algoritmo de *clusterização* o número de *clusters* precisa ser decidido. Uma estratégia eficaz para definir o número ideal é determinar um intervalo razoável para o número de *clusters* e então aplicar os índices de validação em cada conjunto de dados dentro desta faixa.

Desta forma foi determinado inicialmente o intervalo de 2 a 20 *clusters*, pois como os *clusters* serão utilizados para criar campanhas de *marketing*, ter mais que vinte agrupamentos, torna difícil o gerenciamento e a construção de campanhas pela unidade de marketing.

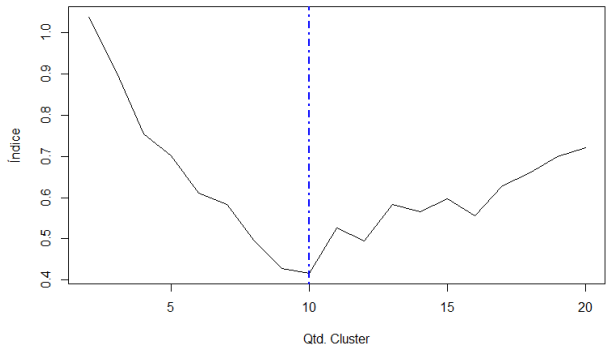
Após a definição do intervalo, aplicou-se o *K-medoids*, uma vez para cada número de *cluster* definido e sobre os resultados foi calculado os índices de validação de *cluster*.

4. Resultados e Avaliações

Para determinar qual o melhor agrupamento, utilizou-se os índices de validação de *cluster* Davies Bouldin e o coeficiente da Silhueta.

Para o índice Davies Bouldin, os resultados são ilustrados no Gráfico 1, que apresenta o índice por quantidade de *cluster*. Como a análise do índice indica que quanto mais próximo de zero melhor o agrupamento, para este conjunto de dados o agrupamento mais adequado se dará com dez *clusters.*

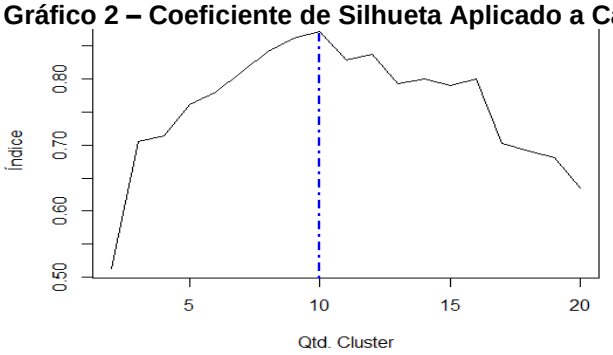
Gráfico 1 – Índice Davies Boldin Aplicado a Cada Cluster



Fonte: Os autores (2020).

Para efetiva comprovação do melhor número de *clusters*, também foi aplicado o coeficiente da Silhueta sobre o resultado de cada agrupamento. A análise do coeficiente é dada pela proximidade do coeficiente do número um, assim quanto mais próximo de um mais forte é a estrutura do *cluster*. Como pode-se analisar no Gráfico 2, em que foram plotados os coeficientes de Silhueta gerados em cada agrupamento, tem-se que o *cluster* cuja estrutura é mais forte é o com dez agrupamentos.

Gráfico 2 – Coeficiente de Silhueta Aplicado a Cada Cluster



Fonte: Os autores (2020).

Como ambos os índices de validade de *cluster* indicaram a utilização de dez *clusters* para o agrupamento dos clientes, este número foi empregado e uma análise descritiva foi desenvolvida para avaliar as características do conjunto de clientes alocados em cada *cluster*.

Determinado o número de *clusters*, pode-se analisar o tamanho de cada agrupamento, a Tabela 5 – Quantidade de Clientes por Cluster demostra a quantidade de clientes em cada um dos *clusters*. Percebe-se que a quantidade de clientes em cada grupo não é homogênea, e isto pode ser atribuído a características distintas de cada grupo.

Tabela 5 – Quantidade de Clientes por Cluster

Cluster	Qtd. Clientes	%
1	460	4,60%
2	2.688	26,88%
3	1.275	12,75%
4	1.751	17,51%
5	1.674	16,74%
6	1.332	13,32%
7	199	1,99%
8	264	2,64%
9	83	0,83%
10	274	2,74%
Total	10.000	100,00%

Fonte: Os autores (2020).

Na Tabela 6 é possível analisar o comportamento da variável Porte em cada Grupo. Nota-se que os *clusters* 2, 3 e 6 tem presença quase exclusiva de empresas do porte MEI, enquanto o porte ME ficou alocado nos grupos 1, 4 e 5. Já as empresas de pequeno porte (EPP) foram designadas aos *clusters* 7, 8 e 9. O porte MGE, embora também presente nos grupos 3 e 6, teve grande maioria no grupo 10.

Tabela 6 – Resultado da Alocação dos Portes por Cluster

Cluster	Porte			
	MEI	ME	EPP	MGE
1	0	460	0	0
2	2.688	0	0	0
3	1.217	0	0	58
4	0	1.751	0	0
5	0	1.674	0	0
6	1.264	0	0	68
7	0	0	199	0
8	0	0	264	0
9	0	0	83	0
10	0	0	0	274

Fonte: Os autores (2020).

Com relação a variável setor, a Tabela 7 apresenta os resultados por grupo. Nota-se que como a variável porte, o setor também teve alocações bem segmentadas. O comércio esteve presente predominantemente nos grupos 5, 6 e 7. A indústria nos grupos 1, 3 e 9. As empresas do setor de serviços foram alocadas nos grupos 2, 4, 8 e 10. Já as empresas de agronegócios não possuíram predominância em nenhum dos grupos.

Tabela 7 - Resultado da Alocação do Setores por Cluster

Cluster	Setor			
	AGRONEGÓCIOS	COMÉRCIO	INDÚSTRIA	SERVIÇOS
1	-	-	460	-
2	25	-	-	2.663
3	30	-	1.245	-
4	7	-	-	1.744
5	9	1.665	-	-

6	-	1.332	-	-
7	-	199	-	-
8	-	-	-	264
9	-	-	83	-
10	4	-	-	270

Fonte: Os autores (2020).

Na Tabela 8 é possível observar a média e mediana das variáveis quantitativas. Nos *clusters* em que o porte MEI está presente, tem-se as empresas mais jovens. Também os grupos 3 e 9, em que estão presentes empresas da indústria, porém não do porte MEI, são os *clusters* que tem as maiores médias de funcionários, exceto pelo cluster 10, este sim com a maior média de funcionários, porém é exclusivo das MGE.

Tabela 8 – Quantidade de Funcionários e Idade da Empresa por Cluster

Cluster	Qtd. de Funcionários		Idade da Empresa	
	Média	Mediana	Média	Mediana
1	3,03	1	14,50	12
2	0,04	0	3,91	3
3	11,80	0	5,04	4
4	2,03	0	12,01	10
5	1,96	1	14,30	12
6	0,98	0	5,43	4
7	6,48	4	18,40	17
8	7,72	2	15,00	12,5
9	10,40	7	18,60	16
10	71,70	1	18,90	16

Fonte: Os autores (2020).

Resumidamente pode-se descrever os grupos da seguinte maneira:

— Grupo 1: Empresa do porte ME, do setor industrial com poucos funcionários e mais velhas;
— Grupo 2: Empresa do porte MEI, alta proporção de Serviço, poucos funcionários e jovens;
— Grupo 3: Alta proporção de empresa do porte MEI, grande concentração de indústria, mediana de funcionários baixa e empresas jovens;
— Grupo 4: Empresa do porte ME, grande concentração do setor de serviços, poucos funcionários e mais velhas;
— Grupo 5: Empresa do porte ME, grande concentração do setor do comércio, poucos funcionários e mais velhas;
— Grupo 6: Empresa do porte MEI, do setor de comércio, poucos funcionários e jovens;
— Grupo 7: Empresa do porte EPP, do setor de comércio, muitos funcionários e mais velhas;
— Grupo 8: Empresa do porte EPP, do setor serviço, muitos funcionários e mais velhas;
— Grupo 9: Empresa do porte EPP, do setor da indústria, muitos funcionários, e mais velhas;

— Grupo 10: Empresa do porte MGE, grande concentração do setor serviços, muitos funcionários e mais velhas.

5. Conclusão

Este artigo apresenta uma metodologia para a *clusterização* de clientes quando os dados apresentam variáveis categóricas e numéricas, demostrando como o algoritmo K-medoids e a distância de Gower podem ser aplicados em uma base de dados híbrida, de maneira que as tipologias de cada variável sejam consideradas. É apresentado também como a base de dados deve ser trabalhada antes do processo de clusterização, o que é fundamental para a correta aplicação dos algoritmos. O resultado do agrupamento determinou dez grupos de clientes, com características distintas. Com este agrupamento de clientes o setor de *marketing* da empresa pode determinar campanhas de relacionamento pautadas nas características dos clientes e assim conseguir maior engajamento nas ações.

Referências

ABBASIMEHR, H.; SHABANI, M. A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. **Kybernetes**, 2019.

AHMAD, A.; DEY, L. A k-mean clustering algorithm for mixed numeric and categorical data. **Data and Knowledge Engineering**, v. 63, n. 2, p. 503–527, 2007.

BARMAN, D.; CHOWDHURY, N. A novel approach for the customer segmentation using clustering through self-organizing map. **International Journal of Business Analytics**, v. 6, n. 2, p. 23–45, 2019.

BEKTAS, A.; SCHUMANN, R. How to Optimize Gower Distance Weights for the k-Medoids Clustering Algorithm to Obtain Mobility Profiles of the Swiss Population. **Proceedings - 6th Swiss Conference on Data Science, SDS 2019**, p. 51–56, 2019.

BRENTARI, E.; DANCELLI, L.; MANISERA, M. Clustering ranking data in market segmentation: a case study on the Italian McDonald's customers' preferences. **JOURNAL OF APPLIED STATISTICS**, v. 43, n. 11, p. 1959–1976, 2016.

BUDIAJI, W.; LEISCH, F. Simple k-medoids partitioning algorithm for mixed variable data. **Algorithms**, v. 12, n. 9, p. 1–15, 2019.

CAMERO, A. et al. Customer segmentation based on the electricity demand signature: The andalusian case. **Energies**, v. 11, n. 7, 2018.

CHANG, C.-I.; HO, J.-C. A Two-Layer Clustering Model for Mobile Customer Analysis. **IT Professional**, v. 19, n. 3, p. 38–44, 2017.

CHEN, T. The RFM-FCM approach for customer clustering. **International Journal of Technology Intelligence and Planning**, v. 8, n. 4, p. 358–373, 2012.

DE ASSIS, E. C.; DE SOUZA, R. M. C. R. A K-medoids clustering algorithm for mixed feature-type symbolic data. **Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics**, p. 527–531, 2011.

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados**. [s.l: s.n.].

GANMAWU, S. A.; WELLS, M. T. **Data Clustering**. [s.l: s.n.].

GRIVA, A. et al. Retail business analytics: Customer visit segmentation using market basket data. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 100, p. 1–16, jun. 2018.

HARIKUMAR, S.; SURYA, P. V. K-Medoid Clustering for Heterogeneous DataSets. **Procedia Computer Science**, v. 70, p. 226–237, 2015.

HUNT, L.; JORGENSEN, M. Clustering mixed data. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, n. 4, p. 352–361, 2011.

JAGABATHULA, S.; SUBRAMANIAN, L.; VENKATARAMAN, A. A Model-Based Embedding Technique for Segmenting Customers. **OPERATIONS RESEARCH**, v. 66, n. 5, p. 1247–1267, 2018.

KEVREKIDIS, D. P. et al. Community pharmacy customer segmentation based on factors influencing their selection of pharmacy and over-the-counter medicines. **SAUDI PHARMACEUTICAL JOURNAL**, v. 26, n. 1, p. 33–43, jan. 2018.

KHALILI-DAMGHANI, K.; ABDI, F.; ABOLMAKAREM, S. Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. **APPLIED SOFT COMPUTING**, v. 73, p. 816–828, 2018.

LIN, Q. et al. A Novel Parallel Biclustering Approach and Its Application to Identify and Segment Highly Profitable Telecom Customers. **IEEE ACCESS**, v. 7, p. 28696–28711, 2019.

LIU, G. et al. Modeling Buying Motives for Personalized Product Bundle Recommendation. **ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA**, v. 11, n. 3, mar. 2017.

PARK, H. S.; JUN, C. H. A simple and fast algorithm for K-medoids clustering. **Expert Systems with Applications**, v. 36, n. 2 PART 2, p. 3336–3341, 2009.

RENUKA DEVI, V.; BHARATHI, G.; PRASAD, G. V. S. N. R. V. Prediction of customer churn in telecom sector using clustering technique. **International Journal of Engineering and Advanced Technology**, v. 8, n. 6 Special Issue 2, p. 826–832, 2019.

REYNOLDS, A. P.; RICHARDS, G.; RAYWARD-SMITH, V. J. The application of K-medoids and PAM to the clustering of rules. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 3177, p. 173–178, 2004.

ROUSSEEUW, PETER J.; KAUFMAN, L. **Finding groups in data**. [s.l.] Hoboken: Wiley Online Library, 1990.

SHEIKH, A.; GHANBARPOUR, T.; GHOLAMIANGONABADI, D. A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. **Journal of Business-to-Business Marketing**, v. 26, n. 2, p. 197–207, 2019.

TSAI, C.-F.; HU, Y.-H.; LU, Y.-H. Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. **EXPERT SYSTEMS**, v. 32, n. 1, p. 65–76, 2015.

UMAMAHESWARI, M.; DEVI, P. I. Prediction of myocardial infarction using K-medoid clustering algorithm. **Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017**, v. 2018- Febru, p. 1–6, 2018.

WANG, A.; GAO, X. Multifunctional Product Marketing Using Social Media Based on theVariable-Scale Clustering. **TEHNICKI VJESNIK-TECHNICAL GAZETTE**, v. 26, n. 1, p. 193–200, fev. 2019.

WANG, C.-H.; CHIN, H.-T. Integrating affective features with engineering features to seek the optimal product varieties with respect to the niche segments. **ADVANCED ENGINEERING INFORMATICS**, v. 33, p. 350–359, 2017.