V Encontro de Iniciação Científica e Tecnológica
V EnICT
ISSN: 2526-6772
IFSP – Câmpus Araraquara
22 e 23 de outubro de 2020

Aplicação de Agrupamento de Dados para a Identificação de Perfis Gerais Entre Clientes

YURI HENRY TAKASHI QUEIROZ KANEGAE[1], FÁBIO JOSÉ JUSTO DOS SANTOS[2], MARCELO CRISCUOLO [3], ANDRÉ DE SOUZA TARALLO[4;] TALES BOALIM[5]

[1]Discente do Técnico em Informática Integrado ao Ensino Médio, Bolsista Fundação, IFSP Campus Araraquara, yurikanegae@gmail.com

[2,3,4]Docente do Instituto Federal de São Paulo – Campus Araraquara, [2]fabiojjs@ifsp.edu.br, [3]criscuolo@ifsp.edu.br, [4]andre.tarallo@ifsp.edu.br

[5]Gerente de Tecnologias e Inovação, Moura Informática

Área de conhecimento (Tabela CNPq): Banco de Dados – 1.03.03.03-0

RESUMO: Com a evolução tecnológica, novos métodos de análise de dados surgiram. Nos setores de varejo, por exemplo, os dados de transações como o preço total do *ticket*, itens e quem comprou passaram a ser armazenados em vastos bancos de dados. Por conta do grande volume de informações, a análise manual desses dados se tornou inviável, fazendo necessária a criação de uma nova área de estudo que busca criar técnicas automatizadas para tal. Essa é a mineração de dados. Desta forma, com o crescimento do volume de dados se tornou mais complexo o processo de gerência de um negócio. O empresário tem que tomar uma série de decisões com impacto direto em seus lucros. Nesse contexto, foi aplicada a técnica de agrupamento de dados a fim de descrever grupos entre clientes de *pet shops*, considerando os seus dados transacionais no estabelecimento. Os resultados obtidos contribuíram para a identificação de perfis gerais de clientes, fornecendo assim informações que possibilitam novas visões e análises do gestor de um negócio.

PALAVRAS-CHAVE: *Data Warehouse*; Mineração de Dados; Agrupamento de Dados.

INTRODUÇÃO

Diariamente, gerentes de varejo tomam decisões que afetam vários campos de seu negócio. Esse processo complexo de tomada de decisão tem um papel fundamental nas empresas, decidindo os possíveis rumos que elas podem tomar, que variam desde a falência até o seu crescimento. Com a evolução tecnológica, surgiram novos métodos de suporte para gestores. Entre esses métodos, é possível citar os algoritmos que auxiliam a tomada de decisão considerando diversos aspectos, que descrevem padrões entre as compras, que tentam prever comportamentos futuros dos clientes ou que fazem recomendações com base no comportamento dos clientes, por exemplo.

Todos esses métodos fazem parte de um campo de estudo chamado Mineração de Dados (MD). MD é definida como a exploração e análise de uma grande quantidade de dados a fim de descobrir padrões e regras significativas (LINOFF e BERRY, 2011). É uma área que abrange diferentes campos de estudo com objetivos diferentes, mas com os seus métodos é possível descrever padrões úteis para dar suporte à gestão.

Entre as áreas de estudo da MD existe o agrupamento de dados. Essa tarefa de mineração busca a criação de grupos contendo os objetos de um determinado banco. Para isso, é necessário o uso de métricas para determinar a similaridade do objeto com o grupo.

Baseado nisso, nosso projeto tem por objetivo agrupar dados transacionais de clientes de *pet shop* por meio do algoritmo K-Means (MACQUEEN, 1967). Os resultados devem contribuir para a identificação de perfis dos clientes, servindo como apoio ao gestor em suas decisões nos âmbitos com influência de sua clientela.

FUNDAMENTAÇÃO TEÓRICA

Os algoritmos de agrupamento tem como objetivo particionar um conjunto de objetos em grupos ou *clusters*, colocando em um mesmo *cluster* objetos semelhantes e em *clusters* distintos objetos não semelhantes (MANNING e SCHUTZE, 1999). Em outras palavras, eles usam métricas de similaridade para associar um objeto ao seu grupo mais parecido, ou menos distante.

Entre os principais algoritmos de agrupamento tem-se o K-Means (MACQUEEN, 1967). O K-Means é um dos mais populares métodos para essa tarefa, o seu funcionamento é simples e confiável, procurando agrupar os dados em um número pré-definido de grupos. Na Figura 1 é apresentado o algoritmo de agrupamento K-Means (MACQUEEN, 1967).

Algoritmo K-Means	
	Entrada: Conjunto de dados X Número de grupos k
1	Escolher k objetos aleatórios em X para centroides dos *clusters*
2	**Repita**
3	**para cada** objeto $x_i \in$ X e *cluster* C_j, j = 1, ..., k **faça**
4	Calcular a distância entre x_i e o centroide do *cluster* $x^{(j)}$: $d\ (x_i, x^{(j)})$, utilizando uma métrica de distância
5	**Fim**
6	**para cada** objeto x_i **faça**
7	Associar x_i ao *cluster* com centroide mais próximo
8	**Fim**
9	**para cada** cluster C_i, j = , ..., k **faça**
10	Recalcular o centroide
11	**Fim**
12	**até** não haver mais alteração na associação dos objetos aos *clusters* ou atingir um número máximo de iterações;

Figura* 1*. Algoritmo K-Means. Fonte: (DOMINGUES, 2019)

Com base na Figura ***1*** é possível definir 4 etapas principais no funcionamento do algoritmo. O primeiro passo é selecionar k objetos aleatórios entre os dados para servir de centroides iniciais dos k *clusters*. O segundo passo trata de associar cada objeto ao grupo com o centroide mais próximo. O terceiro é recalcular o centroide do *cluster* baseado nos itens que foram associados a ele. Por fim, é realizada a repetição das etapas 2 e 3 até que não haja mais mudanças nos grupos ou até que um número máximo de iterações seja alcançado.

O método de associação de cada dado com um grupo é dado pela distância entre cada centroide. Na literatura, existem diferentes métricas sendo utilizadas para medir a similaridade entre os objetos. Dentre elas, para atributos numéricos é possível citar a distância Euclidiana (BALL, 1960) e a distância de Manhattan (KRAUSE, 1986) apresentadas, respectivamente, nas Equações 1 e 2.

$$\boldsymbol{euclidiana}(\boldsymbol{A},\boldsymbol{B}) = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{A_i} - \boldsymbol{B_i})^2} \quad (\mathbf{1})$$

$$\boldsymbol{manhattan}(\boldsymbol{A},\boldsymbol{B}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{A_i} - \boldsymbol{B_i}| \quad (\mathbf{2})$$

onde,
n – número de atributos do objeto
A_i - valor do atributo i do objeto A
A_i - valor do atributo i do objeto B

A definição do número de *clusters* que o algoritmo deve gerar tem profundo impacto nos resultados obtidos e em sua análise. Para definir o número de clusters foi utilizado o método *silhouette* (ROUSSEEUW, 1987), representado pela equação (3), que nos indica o número otimizado de grupos para o conjunto de dados indicado.

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, se\ a(i) < b(i) \\ 0, se\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, se\ a(i) > b(i) \end{cases} \quad (3)$$

onde,
i – um objeto dentro dos dados
a(i) – média da distância de i em relação a todos os elementos do seu *cluster*
b(i) – menor distância de i para qualquer elemento que não esteja no seu *cluster*

Para efetuar as análises foi utilizada a linguagem de programação R (R-PROJECT), de origem na Nova Zelândia, que permite efetuar diversos tipos de análise de dados, partindo desde o processamento deles até a visualização dos resultados.

METODOLOGIA

Nas análises foram utilizados 4 conjuntos de dados de diferentes lojas de pet shop que, que juntas totalizam 16.603 clientes e 442.107 registros de vendas. Os dados extraídos foram referentes as compras dos clientes no estabelecimento. Foram selecionados os atributos de média e total gasto e a quantidade de compras de cada cliente. Esses dados foram extraídos de cada banco e armazenados separadamente em arquivos CSV, possibilitando assim a análise por meio do algoritmo e da linguagem de programação R (R-PROJECT).

Em (FACELLI, LORENA, *et al.*, 2011) são definidas cinco etapas para o efetuar o agrupamento de dados, elas são: pré-processamento de dados, definição da métrica de proximidade, agrupamento, validação e interpretação dos resultados.

O pré-processamento é responsável pela limpeza e transformação de dados a fim de remover incongruências entre eles. Para atingir esse objetivo, foi necessária a padronização dos atributos, tirando letras minúsculas/acentuadas e unidades numéricas diferentes. Foi observado que existiam valores em escalas diferentes, fazendo necessária a normalização dos atributos numéricos para que não haja impacto no agrupamento. O processo de normalização pode ser descrito por (4).

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (4)$$

onde,
X_{min} - Valor mínimo que o atributo atinge
X_{max} - Valor máximo que o atributo atinge

A métrica de proximidade utilizada foi a distância euclidiana. Essa métrica é constantemente utilizada nos agrupamentos e descreve a menor distância entre dois pontos em um plano cartesiano, uma reta. A sua forma multidimensional, definida em (1), descreve a distância entre dois pontos em um plano com *n* dimensões. No caso do agrupamento, cada dimensão é representada por um atributo.

Para a mineração foi utilizado o algoritmo K-Means (MACQUEEN, 1967). Os experimentos foram feitos inicialmente com os dados de cada banco isolados. O número de *clusters* utilizado foi 2, assim como indica o resultado do *silhouette* representado na Figura 2. Porém, após isso foi realizado experimentos com k como 4, a fim de descobrir a influência que a mudança de grupos pode causar nos resultados.

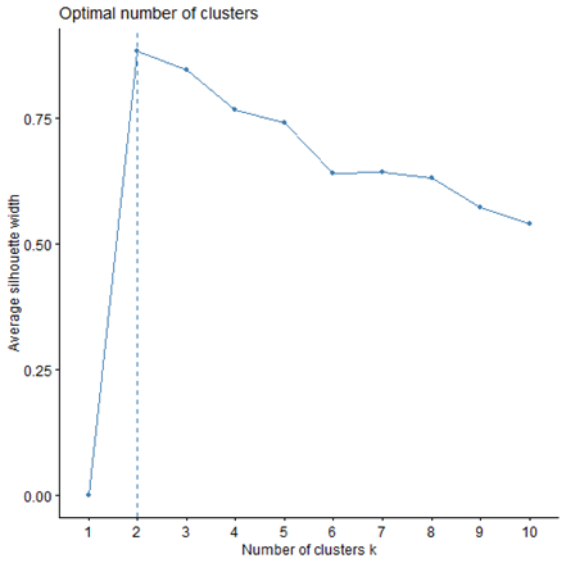

Figura 2. Resultados do silhouette
Fonte: próprio autor

As etapas de validação e interpretação de dados são discutidas na próxima seção.

RESULTADOS OBTIDOS E DISCUSSÃO

Os resultados da mineração com as métricas citadas na seção anterior estão representados na Tabela 1, Tabela 2, Tabela 3 e Tabela 4.

Tabela 1. Resultados com dois grupos na empresa 1

Fonte: próprio autor

Grupo	População	Média	Total	Quantidade
1	208	131,27 R$	4899,30 R$	67,62
2	5426	73,44 R$	308,01 R$	5

Tabela 2. Resultados com dois grupos na empresa 2

Fonte: próprio autor

Grupo	População	Média	Total	Quantidade
1	5946	82,69 R$	301,25 R$	3,45
2	88	963,56 R$	6361,94 R$	30,10

Tabela 3. Resultados com dois grupos na empresa 3

Fonte: próprio autor

Grupo	População	Média	Total	Quantidade
1	4796	107,27 R$	595,21 R$	6,84
2	1	39,47 R$	462843,47 R$	11726

Tabela 4. Resultados com dois grupos na empresa 4

Fonte: próprio autor

Grupo	População	Média	Total	Quantidade
1	134	155,34 R$	232,58 R$	1,67
2	4	377,90 R$	3533,71 R$	12

De modo geral é possível identificar dois tipos de grupos diferentes. O primeiro tipo de grupo é aquele em que as pessoas têm os atributos mais baixos, como os grupos 2, 1, e 2, respectivamente, da Tabela 1, da Tabela 2 e da Tabela 4. Já o segundo tipo de grupo é aquele que tem os atributos mais elevados, como os grupos 1, 2, 1, respectivamente da Tabela 1, da Tabela 2 e da Tabela 4.

Podemos observar que na Tabela 3 há uma discrepância enorme na população entre os grupos, com o grupo 1 tendo 4796 clientes e o grupo 2 1 cliente. Isso se deve a um erro de registro do varejo, que registra todas as vendas de pessoas não cadastradas como "Cliente", causando esse total e quantidade elevados. Para resolver isso o objeto "Cliente" foi deletado do banco, gerando os resultados na Tabela 5.

Tabela 5. Resultados com dois grupos na empresa 4 (corrigido)

Fonte: próprio autor

Grupo	População	Média	Total	Quantidade
1	4785	106,91 R$	535,82 R$	6,56
2	11	264,10 R$	26429,47 R$	125,18

A diferenciação desses dois grupos permite a identificação dos clientes que são consumidores fixos da loja e os consumidores que pouco compram no estabelecimento. Porém, a visão com dois *clusters* pode ser simplista, e um aumento do número de *clusters* pode gerar resultados mais interessantes como o representado na Tabela 6.

Tabela 6. Resultados com quatro grupos da empresa 1

Fonte: próprio autor

Grupo	Média	Total	Quantidade
1	114,03 R$	1722,83 R$	26,92
2	118,99 R$	5123,46 R$	75,49
3	70,28 R$	196,01 R$	3,23
4	122,56 R$	21334,41 R$	176

Com 4 *clusters* é possível fazer uma separação não somente dos clientes fixos e passageiros do estabelecimento como também conseguimos destacar duas novas categorias de clientes que tendem a gastar mais. O grupo 1 contém os clientes que são fixos na loja, já o grupo 3 contém os clientes passageiros do estabelecimento.

No grupo 2 temos pessoas que tem os seus atributos mais elevados, porém não tanto quanto as pessoas do grupo 4. Estas pessoas podem representar os clientes fixos e compram há mais tempo ou que tem um intervalo de compra menor.

Já o grupo 4 é aquele que contém os maiores atributos, com todos eles superando os demais grupos. Estes podem ser clientes que são um pouco mais antigos na loja, e que tendem a gastar em maior volume e quantidade.

CONCLUSÕES

Essa pesquisa buscou a identificação de perfis gerais para os clientes de estabelecimentos de *pet shop*. Os dados utilizados foram oriundos das transações registradas em caixa, sendo associados com o cliente em

questão. Com base nesses dados, foi realizado um pré-processamento nos dados e aplicou-se o algoritmo K-Means (MACQUEEN, 1967), buscando um agrupamento a partir dos dados numéricos.

Os resultados obtidos mostram a viabilidade da aplicação dessa técnica com esse objetivo, conseguindo gerar uma separação dos clientes que são fixos e os clientes temporários do estabelecimento, além de apresentar os clientes que gastam mais, por exemplo. Esses dados podem servir de suporte para decisões que o gestor tem de tomar.

Para trabalhos futuros, é indicado a realização de mais testes com diferentes números de grupos, para assim verificar a viabilidade e as vantagens que essa mudança pode gerar. Também seria de suma importância a criação de uma ferramenta que sirva como *interface* para o usuário, facilitando assim a interação com o algoritmo de mineração, a sua regulação de definições de métricas e a análise dos dados gerados. Também devem ser realizados estudos para identificar a possibilidade de uso de novos atributos e de atribuição de diferentes pesos para cada um dos atributos analisados pelo algoritmo de agrupamento K-Means (MACQUEEN, 1967).

AGRADECIMENTOS

Agradecemos a JN Moura Informática (JN Moura Informática) pelo auxílio financeiro e intelectual a este trabalho de pesquisa.

REFERÊNCIAS

BALL, W. W. R. **A short account of the history of mathematics**. 1°. ed. [S.l.]: New York, Dover Publications, v. 1, 1960.

DOMINGUES, J. M. **Agrupamento de Dados para Análise de Satisfação de Estudantes da Educação Professional e Tecnológica de Nível Superior**. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo-IFSP. Araraquara, p. 21. 2019.

FACELLI, K. et al. **Inteligência Artificial:** Uma Abordagem de Aprendizado de Maquina. 1. ed. Rio de Janeiro: LTC, v. 1, 2011.

JN Moura Informática. **JN Moura Informática**. Disponível em: <https://www.jnmoura.com.br/pt-br/>. Acesso em: 17 set. 2020.

KRAUSE, E. F. **Taxicab Geometry: An Adventure in Non-Euclidean Geometry**. Courrier Corporation. [S.l.], p. 88. 1986. (978-0486252025).

LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics - Doklady**, v. 163, n. 4, p. 845-848, February 1966.

LINOFF, G. S.; BERRY, M. J. A. **Data Mining Techniques:** For Marketing, Sales, and Customer Relationship Management. 3°. ed. Indianapolis: Wiley Publishing, Inc., 2011.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **Proceedings of the fifth berkeley Symposium on Mathematical Statistics and Probability.**, 1967. 281-297.

MANNING, C. D.; SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. 2°. ed. Massachusetts: MIT Press, v. I, 1999.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53-65, November 1987.

R-PROJECT. **R:** The R Project for Statistical Computing. Disponível em: <https://www.r-project.org/>. Acesso em: 17 set. 2020.