VI Encontro de Iniciação Científica e Tecnológica
VI EnICT
ISSN: 2526-6772
IFSP – Câmpus Araraquara
21 e 22 de outubro de 2021

Uso de Técnicas de Agrupamento de Dados para Auxílio a Tomada de Decisão de Empresas a partir dos Relatórios de *Business Intelligence* (BI) Gerados

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

[1] Discente do Técnico em Informática Integrado ao Ensino Médio, Bolsista PIBIFSP, IFSP Campus Araraquara, yurikanegae@gmail.com
[2 3] Docente do Instituto Federal de São Paulo – Campus Araraquara, andre.tarallo@ifsp.edu.br, fabiojjs@ifsp.edu.br
[4] Gerente de Inovação - JN Moura Informática, tales@jnmoura.com.br

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

RESUMO: No dia a dia de uma empresa, diversas decisões que impactam o futuro do negócio têm de ser tomadas. Com o passar do tempo, foram criados diversos métodos afim de se dar suporte aos gestores, fornecendo informações e visualizações que não são claras em primeiro momento. Com a evolução tecnológica, esses métodos, denominados de *Business Intelligence*, começaram a se valer da grande quantidade de dados gerados pelos sistemas em suas análises. Muitas vezes, o que confere qualidade e riqueza a essas informações é o pluralismo e a quantidade dos elementos presentes, porém, isso pode causar enganos nos resultados. Ao comparar comércios que tem objetivos, clientelas, métodos e tamanhos diferentes, as análises podem gerar informações errôneas, podendo influenciar aos gestores a tomarem más decisões. Nesse contexto, foram aplicadas três métodos diferentes de pré-processamento e uma técnica de agrupamento de dados para se descrever grupos com comportamento semelhantes entre empresas. Os resultados obtidos contribuem para a identificação desses grupos, fornecendo informações para a melhora dos resultados obtidos com as análises BI.

PALAVRAS-CHAVE: *Business Intelligence*; Mineração de Dados; Agrupamento;

INTRODUÇÃO

De acordo com Howard Dresner, as ferramentas de *Business Intelligence* (BI) são um "conjunto de técnicas e métodos para melhorar as decisões comercias com base em um sistema de suporte de decisão" (POWER, 2007). O conceito de BI precede a existência de computadores e bancos de dados, mas com a evolução tecnológica e aplicação em diversos setores houve o desenvolvimento de métodos que fazem uso da grande quantidade de dados para geração de informações.

Parte fundamental dessas análises é o tratamento que os dados sofrem antes de serem processados. Dentro de uma empresa e suas filias esses dados podem ter diferentes fontes e formas, fazendo necessário a extração, transformação e carregamento (ou ETL – *Extract Transform, Load*), normalizando e carregando os dados em um mesmo local, normalmente um *Data Warehouse* (DW), facilitando o processamento e geração de informações.

Outra parte importante do pré-processamento é a definição de quais dados vão ser comparados uns com os outros. Por mais que o que confere riqueza às informações geradas pelo BI seja a diversidade de objetos, existem situações onde comparar dados totalmente diferentes resulta em informações errôneas. Isso acontece quando comparamos empresas de nichos (*ex*: *pet*, roupa, supermercado, ...) e tamanhos (*ex*: pequeno, médio, grande, ...) distintos, já que essas empresas vão ter comportamentos, objetivos, poder de compra e públicos diferentes.

Baseado nisso, em parceria com a *software house* JN Moura Informática foi desenvolvido um projeto que tem por objetivo definir os atributos e agrupar empresas por meio do algoritmo *K-means* (MACQUEEN,

1967). Os resultados são promissores e passam a contribuir para a melhora das informações geradas pelas as análises de BI, fornecendo uma visão mais realista para o gestor e dando suporte as suas decisões.

FUNDAMENTAÇÃO TEÓRICA

O algoritmo *K-means* é uma técnica de agrupamento particional que procura descrever um número pré-determinado de grupos entre os dados. O seu funcionamento pode ser dividido em 4 etapas principais: a primeira é selecionar *k* elementos aleatórios entre os dados para ser usados como centroides de grupos; a segunda é associar cada objeto entre os dados ao centroide mais próximo ou similar; a terceira é recalcular os centroides com base nos elementos associados a cada grupo; e a quarta é a repetição dos passos 2 e 3 até que não haja mudança nos grupos ou que o número de iterações máximas seja alcançado.

Para definição dos atributos utilizados no agrupamento foram aplicadas duas técnicas diferentes, o *grid seach* e o PCA (*Principal Component Analysis*).

O *grid search* é uma técnica de força bruta utilizada na abordagem de problemas de hiperparâmetros. Em um processo de aprendizado de máquina existem diversos parâmetros, como o número de grupos gerados, que são utilizados para controlar as análises, os hiperparâmetros, e o *grid search* busca definir o melhor valor desses atributos testando todas as combinações possíveis. Ainda existe uma variação para definição de atributos, que testa todas as combinações possíveis de atributos e define aquela que gera melhores resultados.

Já o PCA é uma técnica de análise multivariada que analisa as inter-relações entre um grande número de variáveis. Em suma, seu objetivo é diminuir a quantidade de variáveis as transformando em componentes buscando uma mínima perda de informação. Seu funcionamento pode ser dividido em 6 etapas principais: primeiro, se obtém os *m* dados com *n* atributos; segundo, se calcula a média dos *n* atributos; terceiro, se subtrai a média dos *n* para os *m* dados; quarto, se calcula a matriz de covariância, gerando uma matriz *n* x *n*; quinto, se calcula os autovalores e auto vetores da matriz de covariância; e sexto, se arranja a matriz da Transformada de *Hotelling*.

METODOLOGIA

Esse projeto faz parte de um conjunto maior de projetos, representados na Figura 1, que abrangem desde a transformação e o carregamento dos dados em um DW até a geração dos relatórios de BI. A *software house* JN Moura Informática fornece sistemas gerenciais de *Enterprise Resource Planning* (ERP) para diversas empresas como farmácias (FAR), *pet shops* (PET), agronegócios (AGRO), lojas de roupas (MAG) e varejos (SIS). Os bancos desses clientes sofrem ETL, são carregados em uma STAGE e, posteriormente, um DW. Com esses dados carregados, eles são processados por meio de BI, MD, e *OnLine Analytical Processing (OLAP)*, gerando relatórios para o gerente.

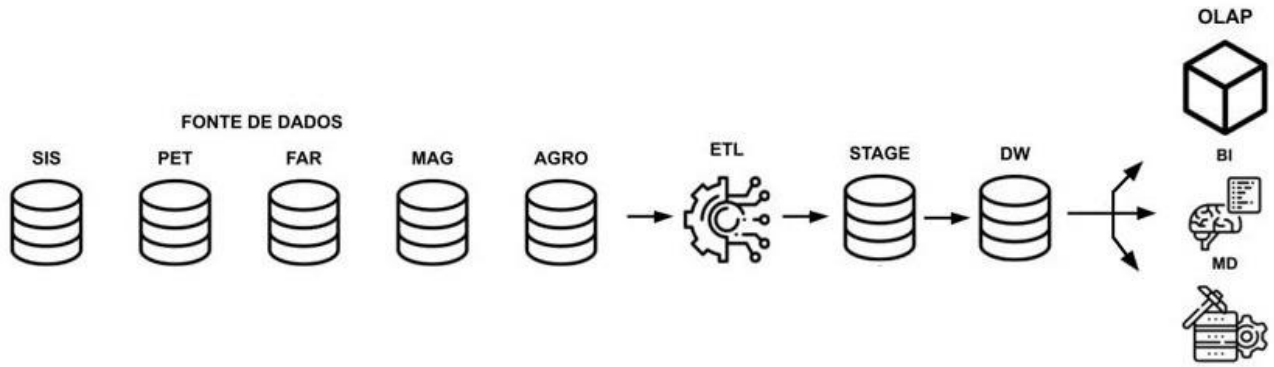

Figura 1 - Diagrama do projeto geral.
Fonte: próprio autor.

Em específico, os processos da mineração de dados (MD), representados na Figura 2, podem ser divididos entre classificação e agrupamento. Os dados presentes no DW sofrem um pré-processamento e passam por uma classificação, que define automaticamente os nichos em que essas empresas atuam, e pelo agrupamento, que define três grupos de empresas com comportamento, ou tamanhos, semelhantes. Essas etapas são necessárias pois nos bancos não existem informações que indiquem o nicho ou o tamanho do cliente.

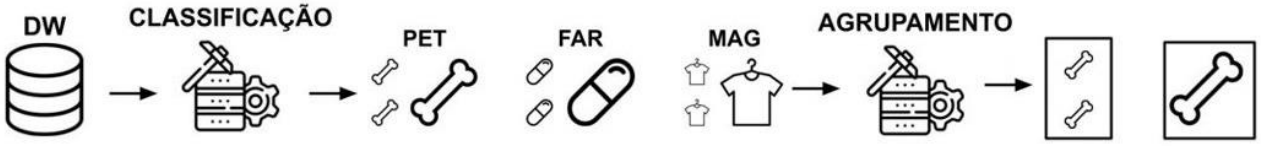

Figura 2 - Diagrama da MD.
Fonte: próprio autor.

Nas análises do agrupamento foram utilizados os dados de 40 empresas diferentes. Foram selecionados os atributos de total de compras em reais (TQ), média de compras em reais (MC), quantidade de compras (QC), total de vendas em reais (TV), média de vendas em reais (MV), quantidade de vendas (QV), lucro em reais (L), quantidade de produtos (QP), quantidade de fornecedores (QF), e quantidade de pontos de venda (QPDV) de cada loja. Os dados foram retirados de um DW e foram armazenados em um arquivo CSV, possibilitando o processamento por meio da linguagem *Python* (VAN ROSSUM e DRAKE, 2009).

Foram aplicadas três metodologias diferentes para efetuar o agrupamento, diferindo somente no pré-processamento que os dados sofrem. Na primeira, os dados sofrem uma normalização e o agrupamento é efetuado com todos os atributos. Na segunda, os dados sofrem uma normalização, é definido o conjunto de atributos que melhor divide os dados e o agrupamento é aplicado com base nesses atributos. No terceiro, os dados sofrem uma normalização, são transformados em duas componentes por meio do PCA e o agrupamento é efetuado nessas duas componentes.

A pré-processamento consistiu em remover valores nulos, transformando-os em zero e remoção de caráteres não numéricos, ambos causados por erros de cadastro e digitação. Além disso, os atributos selecionados apresentam escalas de valores diferentes, portanto, passam por uma normalização baseada na Equação 1.

$$X_{normalizado} = \frac{(X - X_{min})}{X_{max} - X_{\min}} \tag{1}$$

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

A definição dos atributos que melhor dividem os dados é feita com base na técnica *grid search*. Levando em consideração os atributos selecionados, é gerado toda combinação possível entre esses atributos e, para cada uma dessas combinações, é efetuado o agrupamento. Para definir uma pontuação a essas combinações é utilizado a métrica *silhouette* (ROUSSEEUW, 1987), representado pela Equação 2, que mede a similaridade dos objetos com o seu grupo comparado com os demais grupos.

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, se\ a(i) < b(i) \\ 0, se\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, se\ a(i) > b(i) \end{cases} \tag{2}$$

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

O agrupamento de dados é efetuado com o algoritmo *K-means*. A métrica de similaridade utilizada foi a distância euclidiana, já que todos os dados são numéricos e normalizados, e a quantidade de grupos definida foram três, a fim de se definir três tamanhos diferentes entre os objetos.

RESULTADOS E DISCUSSÃO

O resultado da mineração da primeira metodologia está representado na Figura 3 e os centroides dos grupos na Tabela 1.

Tabela 1 – Médias dos grupos gerados com a metodologia 1.
Fonte: próprio autor.

G	TC	QC	MC	TV	MV	QV	L	QPDV	QP	QF
1	8,84e+06	4855,25	1945,15	1,35E+07	96,91	139125,25	5,92e+06	8,5	13722,5	344,25
2	1,08e+06	980,13	1300,76	1,77e+06	55,70	28014,90	5,47e+05	1,86	3795,18	57,77
3	5,42e+05	605,85	560,39	1,92e+06	118,10	15820,42	1,09e+06	1,92	2182,42	55,85

Por meio dos centroides é possível descrever três comportamentos diferentes entre os grupos. O primeiro grupo apresenta centroides elevados para todos os atributos, representando o que seria uma empresa de tamanho grande com produtos diversificados. O segundo grupo apresenta os menores valores para TV, MV, L e QPDV, e os segundos maiores para o restante dos atributos, sendo que uma empresa foca em produtos mais

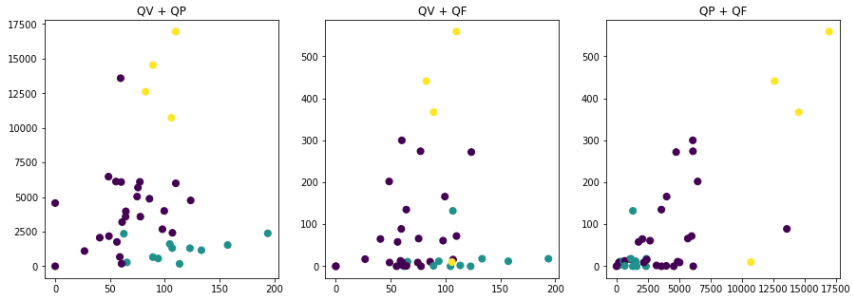

Figura 3 - Visualização de QV, QP e QF da metodologia 1.
Fonte: próprio autor.

baratos e, por isso, compra mais, vende mais, porém, tem menos lucro, representando o que seria a empresa pequena. Já o terceiro grupo apresenta os menores valores para TC, QC, MC, QP e QF, e os segundos maiores para o restante dos atributos, representando uma empresa que foca em produtos um pouco mais caros, porém, tem mais lucro, representando uma empresa média.

Porém, ao analisar como os grupos foram divididos podemos perceber que a primeira metodologia não conseguiu dividir os grupos corretamente, obtendo uma pontuação de 0,36 no *silhouette* (ROUSSEEUW, 1987). A Figura 3 apresenta a visualização dos três atributos com maior desvio padrão (QV, QP e QF), demonstrando que os objetos foram agrupados erroneamente, com grupos "misturados" entre si.

Na aplicação da segunda metodologia, a técnica de *grid search* apontou que um grupo de 5 atributos dividiam melhor o conjunto de dados. Esses atributos são: quantidade de fornecedores (QF), quantidade de produtos (QP), quantidade de compras (QC), quantidade de vendas (QV) e total de vendas (TV). Os resultados estão apresentados na Tabela 2 e Figura 4.

Tabela 2 - Médias dos grupos gerados com a metodologia 2.
Fonte: próprio autor.

G	QF	QP	QC	QV	TV
1	387,25	12082,00	2113,25	61483,50	6,01e+06
2	155,82	5696,41	551,17	12518,88	9,05e+05
3	40,05	1248,57	79,10	2747,63	2,19e+05

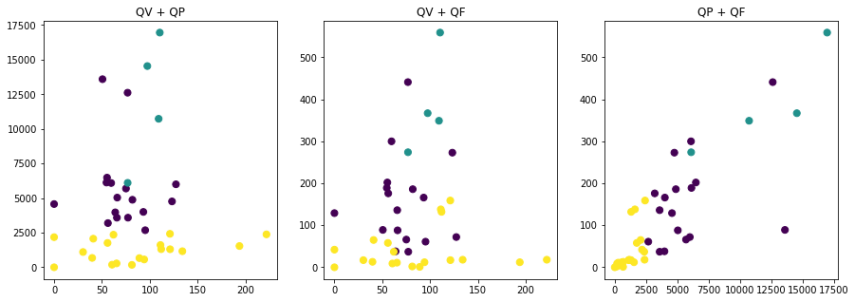

Figura 4 - Visualização de QV, QP e QF da metodologia 2.
Fonte: próprio autor.

Na Figura 4, o primeiro grupo apresenta os maiores valores para todos os atributos, representando o que seria uma empresa de tamanho grande. O segundo grupo apresenta os segundos maiores valores para todos os atributos, representando o que seria uma empresa de tamanho médio. O terceiro grupo apresenta os menores valores para todos os atributos, representando o que seria uma empresa de tamanho pequeno.

Os agrupamentos gerados, representados na Figura 4, obtiveram uma pontuação de 0,41 no *silhouette*, apresentando uma melhora na divisão dos objetos.

Já a terceira metodologia gerou os componentes representados na Tabela 3. Esses componentes não representam diretamente qualquer valor entre os atributos originais, não possibilitando as análises por meio deles. A Tabela 4 representa os centroides dos grupos gerados por meio das componentes levando em consideração os atributos originais.

Tabela 3 - Média dos grupos gerados com a metodologia 3 (PCA).
Fonte: próprio autor

Grupo	Componente 1	Componente 2
1	-0,14	-0,35
2	-0,19	0,21
3	1,24	0,01

Tabela 4 - Média dos grupos gerados com a metodologia 3 (Geral).
Fonte: próprio autor.

G	TC	QC	MC	TV	MV	QV	L	QPDV	QP	QF
1	3,33e+05	251,53	1277,92	4,34e+05	54,52	6852,23	1,71e+05	1,53	2504,46	77,53
2	3,13e+05	323,40	682,82	5,98e+05	94,14	7556,36	2,74e+05	2,13	3426,77	89,13
3	3,78e+06	1788,20	2159,91	4,96e+06	94,17	51141,40	1,91e+06	7,00	12189,40	398,00

Os grupos gerados têm comportamento semelhante aos gerados na primeira metodologia, porém pequenas mudanças. O grupo 1 tem os menores valores para QC, TV, MV, QV, L, QPDV, QP e QF, representando uma empresa de tamanho pequeno. O grupo 2 tem os segundos maiores valores para QC, TV, MV, QV, L, QPDV, QP e QF, representando uma empresa de tamanho médio. E o grupo 3 tem os maiores valores para todos os atributos, representando uma empresa de tamanho grande.

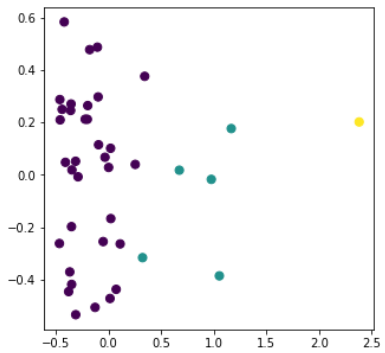

Figura 5 - Visualização PCA da metodologia 3.
Fonte: próprio autor.

A Figura 5 representa os grupos gerados pelas componentes, obtendo uma pontuação de 0,65 no *silhouette* e demonstrando grupos bem divididos. Porém, quando a análise é feita em relação aos dados originais, representados na Figura 6, a pontuação atingida é de 0,31, gerando os centroides representados na Tabela 4.

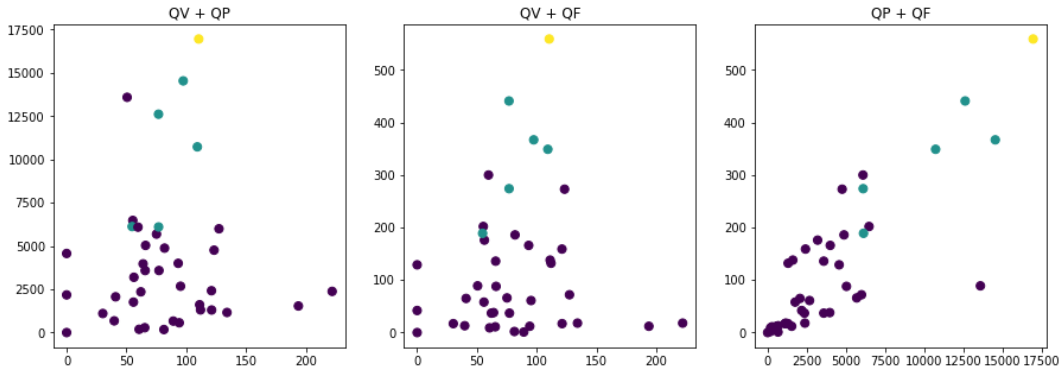

Figura 6 – Visualização de QV, QP e QF da metodologia 3.
Fonte: próprio autor.

CONCLUSÕES

Essa pesquisa procurou definir tamanhos para as empresas dentro de um DW. Os dados utilizados tinham como fonte as transações e compras feitas por empresas que utilizam os sistemas da *software house*. Sobre esses dados, foram aplicadas três metodologias diferentes de pré-processamento e o agrupamento por meio do algoritmo *K-means.*

Os resultados obtidos mostraram a possiblidade da utilização dessas técnicas para a melhora dos relatórios BI.

Para trabalhos futuros, é indicado a realização de testes com os parâmetros do agrupamento, como número de grupos e métricas de distâncias, e tipos de algoritmos diferentes, como algoritmos baseados em hierarquias e densidade. Também, é de suma importância a análise dos atributos definidos e o pré-processamento aplicado, já que os resultados são altamente influenciados pelos atributos fornecidos ao algoritmo.

AGRADECIMENTOS

Agradecemos a *software house* JN Moura Informática e o IFSP pelo auxílio financeiro, intelectual e pelos dados fornecidos a este trabalho de pesquisa.

REFERÊNCIAS

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, v. 1, n. 14, p. 281-297, 1967.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin philosophical magazine and journal of science**, v. 2, n. 11, p. 559-572, 1901.

POWER, D. J. A Brief History of Decision Support Systems. **DSSResources**, 2007. Disponivel em: <http://dssresources.com/history/dsshistory.html>. Acesso em: 11 ago. 2021.

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, 1987.

VAN ROSSUM, G.; DRAKE, F. **Python 3 Reference Manual**. 1. ed. Scotts Valley, CA: CreateSpace, v. 1, 2009.