UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ALLANE RÉGIS PINHEIRO GUIMARÃES

AGRUPAMENTO DE DADOS COMO INSTRUMENTO DE APOIO A ESTRATÉGIAS DE NEGÓCIO

São Luís - Maranhão
2019

Allane Régis Pinheiro Guimarães

AGRUPAMENTO DE DADOS COMO INSTRUMENTO DE APOIO A ESTRATÉGIAS DE NEGÓCIO

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Ivo José da Cunha Serra

São Luís - Maranhão

2019

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Guimarães, Allane Régis Pinheiro.
Agrupamento de dados como instrumento de apoio a estratégias de negócio / Allane Régis Pinheiro Guimarães. – 2019.
65 f.

Orientador(a): Ivo José da Cunha Serra.
Monografia (Graduação) – Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2019.

1. Agrupamento de Dados. 2. Algoritmo Genético. 3. Algoritmo K-means. 4. Método de Ward. 5. Mineração de Dados. I. Serra, Ivo José da Cunha. II. Título.

Allane Régis Pinheiro Guimarães

AGRUPAMENTO DE DADOS COMO INSTRUMENTO DE APOIO A ESTRATÉGIAS DE NEGÓCIO

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado. São Luís, 8 de Julho de 2019:

BANCA EXAMINADORA

_____

Prof. Dr. Ivo José da Cunha Serra
(Orientador)

_____

Prof. Me. Carlos Eduardo Portela Serra de Castro
(Membro da Banca Examinadora)

_____

Prof. Dr. Tiago Bonini Borchatt
(Membro da Banca Examinadora)

São Luís – Maranhão

2019

AGRADECIMENTOS

Agradeço a Deus por ter me dado saúde e força em meio as dificuldades e não por não me deixar esmorecer ou desistir, pelo seu infinito amor e bênçãos.

Aos meus pais, Valdemir e Aliosandra, pelo incentivo, apoio, amor, dedicação e compreensão.

As minhas irmãs, Agda e Amanda, pelo apoio e companheirismo nos momentos felizes e tristes.

Aos meus amigos e colegas de curso, Joysiane Lima, Jodiel Fabricio e Danilo Batista, por todos os momentos compartilhados, felizes e tristes, pelo incentivo e motivação.

Ao meu orientador, Professor Ivo Serra, pelo suporte no pouco tempo que lhe coube, paciência e incentivo.

E finalmente, a esta Universidade, seu corpo docente, direção e administração, pela oportunidade que me foi concedida.

RESUMO

Com o grande volume de informações que são armazenadas em base de dados corporativas sobre produtos, clientes e fornecedores, o mercado cada vez mais competitivo e consumidores mais exigentes, a aplicação de técnicas de descoberta de conhecimento em base de dados corporativas se tornou uma arma poderosa para a estratégia de negócios. Este trabalho apresenta a técnica de agrupamento de dados, que descobre grupos de objetos semelhantes em um conjunto de dados a partir, somente, de informações contidas nos próprios dados. E apresenta também três técnicas de agrupamento, o método de agrupamento hierárquico aglomerativo, o Algoritmo K-means e Algoritmo Genético. Em seguida, são apresentados três estudos de caso no contexto de negócios com aplicação das técnicas estudadas com o objetivo de traçar perfis de clientes e classificar fornecedores para descoberta de conhecimentos úteis que sirvam de apoio a estratégias de negócios, e é feita uma discussão sobre as características de cada problema que motivaram a escolha de cada técnica adotada e resultados alcançados.

Palavras-Chave: Mineração de Dados, Agrupamento de Dados, Método de Ward, Algoritmo K-means, Algoritmo Genético;

ABSTRACT

With the high volume of information that is stored in corporate databases on products, customers and suppliers, alied to the increasingly competitive market and the most demanding consumers, the application of knowledge discovery techniques in corporate databases has become a powerful weapon for business strategy. This work presents the technique of data clustering, which discovers groups of similar objects in a data set from, only, information contained in the data itself. Three clustering techniques are discussed: the agglomerative hierarchical clustering method, the K-means Algorithm and Genetic Algorithm. Next, three case studies are presented in the context of business with application of the techniques studied with the objective of tracing customer profiles and classifying suppliers to find useful knowledge to support business strategies. Finally, and a discussion is made on the characteristics of each problem that motivated the choice of each technique adopted and results achieved.

Keywords: Data Mining, Data Clustering, Ward's method, K-means Algorithm, Genetic Algorithm.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	*Knewledge Discovery in Databases*
SGBD	Sistema de Gerenciamento de Banco de Dados
SSE	Soma do Erro Quadrado

SUMÁRIO

1 INTRODUÇÃO

Avanços rápidos na tecnologia de coleta e armazenamento de dados permitiram que as organizações acumulassem uma grande quantidade de dados (TAN, 2006). Essas bases de dados podem conter registros de negociações, transações de vendas, movimentações de estoques, dados de clientes, dados de matérias-primas usadas nos processos de produção por indústrias, dados de produtos e serviços, dentre outros. De acordo com Goldschmidt, Passos e Bezerra (2015), com a disponibilidade de tantos dados, surgem os questionamentos, o que fazer com todos esses dados? Como analisar e utilizar todo o volume de dados disponível?

Segundo Goldschmidt, Passos e Bezerra (2015) o valor desses dados armazenados está ligado a capacidade de se extrair informação útil que sirva de apoio à tomada de decisão, e/ou que possibilite a exploração e entendimento do que gerou esses dados. A partir dessas bases de dados podemos identificar padrões, distribuições e relacionamentos úteis entre os dados, que se descobertos, podem servir de diferencial mercadológico no contexto de negócios, por exemplo, como descobrir padrões de comportamento dos clientes, ou ainda, em bases de dados científicas, podem ajudar na compreensão de resultados de experimentos, além de outras utilidades.

Contudo, a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Diante disso, é necessária a aplicação de técnicas e ferramentas capazes de extrair desses dados armazenados, informações e conhecimentos implícitos que podem servir para apoiar decisões e embasar o desenvolvimento de estratégias de ação relacionadas ao domínio em que esses dados estão inseridos.

De forma simplificada, o processo automático ou semiautomático de explorar analiticamente grandes bases de dados, com a finalidade de descobrir padrões relevantes que ocorrem nos dados e que sejam importantes para embasar a assimilação de informação, suportando a geração de conhecimento, é denominado de mineração de dados (SILVA, 2016).

A mineração de dados é parte integrante da Descoberta de Conhecimento em Bases de Dados (*Knewledge Discovery in Databases* - KDD), que é o processo geral de conversão de dados brutos em informações úteis. Esse processo consiste em uma série de etapas de transformação, desde o pré-processamento de dados até o pós-processamento dos resultados de mineração de dados (TAN, 2006). A mineração de dados consiste na aplicação de técnicas, implementadas por meio de algoritmos computacionais, capazes de receber, como entrada, um conjunto de fatos ocorridos no mundo real e devolver, como saída, um padrão de

comportamento, o qual pode ser expresso, por exemplo, como uma regra de associação, uma função de mapeamento ou modelagem de um perfil (SILVA, 2016).

Dentre várias tarefas desempenhadas em mineração de dados, o Agrupamento de Dados é um dos problemas centrais, o qual consiste em determinar um conjunto de categorias para descrever uma coleção de objetos de acordo com as suas similaridades ou inter-relacionamentos (KAUFMAN; ROUSSEEUW, 2009 apud MENDES, 2017).

Técnicas de agrupamento fornecem um meio de explorar e verificar estruturas presentes nos dados, organizando-os em grupos de objetos similares (JAIN; DUBES, 1988). É conhecido como aprendizado não supervisionado porque as informações de rótulo de classe não estão presentes. Por essa razão, o agrupamento é uma forma de aprendizado por observação, em vez de aprender por exemplos (HAN; KAMBER; PEI, 2012).

Conforme Han, Kamber e Pei (2012), ao agrupar dados podemos descrever de forma eficiente as características dos diversos grupos encontrados possibilitando assim, maior compreensão da relação entre os dados, além da criação de modelos ou esquemas que venham ser úteis para prever o comportamento de novos dados. Por exemplo, uma base de dados de um sistema de *help desk*, contendo reclamações de clientes, poderia ser usada para geração de perfis de consumidores, onde as reclamações sobre produtos ou serviços registrados no sistema gerariam o perfil de cliente para cada tipo de reclamação, permitindo que a organização atenda cada parcela de consumidores de forma customizada, evitando aquele tipo de reclamação para cada perfil identificado.

Devido a importância da tarefa de agrupamento para a mineração de dados, neste trabalho serão apresentados os principais conceitos e métodos de agrupamento de dados, seus algoritmos e suas características, sendo ilustrados através de três aplicações no contexto de negócios, como um meio de extrair conhecimento de bases de dados corporativas contendo informações de clientes e fornecedores. A primeira aplicação, é feita em uma base de dados de uma agência de viagens online, a segunda aplicação em uma base de clientes de uma empresa fornecedora de produtos de limpeza, ambas com o objetivo de identificar perfis de clientes, e, a terceira aplicação em uma base de dados de uma indústria de laticínios, buscando categorizar fornecedores de leite de acordo com as características dos leites produzidos por esses.

1.1 Motivação

De acordo com Welge et al., (2001), as organizações em geral, possuem uma enorme disponibilidade de dados descritivos de clientes e fornecedores em suas bases de dados e o

uso desses dados, muitas vezes está limitado ao suporte as atividades tradicionais dentro da empresa. Isso levou muitas organizações a perceberam que há nesses dados um grande potencial para gerar conhecimentos que sejam úteis e sirvam de base para apoiar as diversas decisões organizacionais.

Dessa forma, consoante o que explica Jain e Dubes (1998), o agrupamento é uma ferramenta de exploração de dados, e quando aplicado em bases de dados corporativas é capaz de descobrir grupos semelhantes de clientes, fornecedores ou produtos que compartilham algumas características ou propriedades que façam sentido e sejam relevantes para o domínio dos dados. O agrupamento facilita a compreensão dos dados, ao permitir descrever mais eficientemente as características dos grupos formados, focando na análise de cada grupo de clientes, fornecedores ou produtos semelhantes individualmente, de forma a gerar conhecimentos que sirvam de base para estabelecer as melhores políticas da empresa, embasando o planejamento de ações da empresa, gerenciamento de processos de produção, ou publicidades personalizadas a cada grupo. Por exemplo, conforme Pacheco, Capella e Goldschmidt (2010), a partir do agrupamento aplicado em bases de dados contendo informações cadastrais de clientes e suas transações de compras, pode-se identificar padrões no comportamento de consumo dos clientes de cada grupo criado, através dos quais, podem-se traçar o perfil dos clientes, que servirão como base para o desenvolvimento de estratégias personalizadas e adoção de políticas de vendas com foco em cada grupo de clientes, e consequentemente, levando a otimização das vendas junto aos mesmos.

Diante do que foi mencionado, esse trabalho foi motivado pelo interesse em discutir o uso de técnicas de agrupamento de dados como um meio de extrair conhecimentos de bases de dados corporativas que sirvam de apoio a estratégias de negócios visando gerar conhecimentos práticos para solucionar problemas específicos.

1.2 Objetivos do Trabalho

O objetivo principal desse trabalho é realizar um estudo sobre a tarefa de agrupamento de dados, suas teorias e aplicações para descoberta de conhecimento em bases de dados corporativas que sirvam de apoio a estratégias de negócios.

Os objetivos específicos são:

- Apresentar agrupamento de dados e tipos de grupos;
- Introduzir e descrever três técnicas de agrupamento de dados, a técnica de

agrupamento hierárquico aglomerativo, e duas técnicas particionais, o Algoritmo K-means e Algoritmo Genético;

- Apresentar três aplicações do agrupamento de dados em negócios, evidenciando o que motivou a adoção da técnica aplicada em cada uma delas.

1.3 Organização do Trabalho

Este trabalho é composto, além deste capítulo, de três outros capítulos que estão organizados da seguinte forma:

O capítulo 2 faz uma abordagem sobre agrupamento, tipos de grupos e apresenta três técnicas de agrupamento, o método hierárquico aglomerativo, K-means e Algoritmo Genético.

No capítulo 3 são apresentados três estudos de caso com aplicação das técnicas de agrupamento estudadas no capitulo 2, em bases de dados corporativas. Em cada aplicação, é realizada uma discussão dos motivos que levaram a escolha da técnica utilizada.

No capítulo 4 tratam-se de conclusões a respeito dos estudos realizados nesse trabalho, contribuições desse estudo, limitações e possibilidades de novos estudos.

2 AGRUPAMENTO DE DADOS

Este capítulo apresenta a fundamentação teórica necessária para compreensão da discussão apresentada no decorrer do desenvolvimento do trabalho.

2.1 Definições

A análise de grupos ou simplesmente agrupamento é o processo de particionamento de um conjunto de objetos de dados (ou observações) em subconjuntos. Cada subconjunto é um grupo, de modo que os objetos em um grupo são semelhantes uns aos outros, mas diferentes de objetos em outros grupos. O conjunto de grupos resultantes de uma análise de grupos pode ser referido como um agrupamento (HAN; KAMBER; PEI, 2012).

Segundo Dunham (2002), o agrupamento de dados é similar a classificação na finalidade de agrupar dados, mas afirma que é diferente, por não conter grupos pré-definidos. E afirma ainda, que nesse caso, o agrupamento pode ser visto como o processo de criar classes usando apenas informações contidas nos próprios dados.

Para Castro (2013), o agrupamento consiste em criar classes ao formar grupos de dados que possuam valores próximos em certos atributos, e deve ser usado quando o intuito da aplicação é descobrir grupos de dados semelhantes que compartilham propriedades comuns sem qualquer conhecimento prévio do que possa ser a similaridade.

O objetivo do agrupamento é que os objetos dentro de um grupo sejam semelhantes (ou relacionados) entre si e diferentes de (ou não relacionados aos) outros objetos de outros grupos. Quanto maior a semelhança (ou homogeneidade) dentro de um grupo e maior a diferença entre grupos, melhor ou mais distinto será o agrupamento (TAN, 2006).

Consoante o que explica Han, Kamber e Pei (2012), a tarefa de agrupamento é utilizada nas mais variadas áreas, entre elas, a inteligência de negócios. Em inteligência de negócios, o agrupamento pode ser uma ferramenta poderosa para desenvolvimento de estratégias de negócio para melhorar o relacionamento com os clientes e ainda aumentar o faturamento. O agrupamento possibilita organizar clientes em grupos de acordo com características semelhantes compartilhadas por eles. Por exemplo, no próximo capítulo desse trabalho, veremos uma aplicação da tarefa de agrupamento de dados em uma base de dados de uma agência de viagens online. Essa aplicação buscou agrupar os clientes da agência de acordo com seu histórico de compras no site. Através da partição criada, podemos traçar o perfil dos clientes de cada grupo e usar como base para envio de ofertas de viagens personalizados para cada perfil.

Dessa forma, de acordo com Pacheco, Capella e Goldschmidt (2010), o agrupamento aplicado em bases de dados de clientes permite identificar os perfis distintos de clientes a partir dos grupos de clientes semelhantes encontrados e, classificá-los com base em seu padrão de compras facilitando assim, a adoção de estratégias de marketing direcionadas a cada grupo. Podemos ainda, através do agrupamento em bases de fornecedores, classificar fornecedores de acordo com características dos produtos fornecidos por estes, possibilitando controle de qualidade de matérias-primas, por exemplo.

Algoritmos de agrupamento agrupam objetos, ou item de dados, com base em índices de proximidade entre pares de objetos (JAIN; DUBES, 1988). De acordo com Han, Kamber e Pei (2012), em agrupamento de dados, a semelhança entre os objetos é avaliada com base na distância entre os valores dos atributos que descrevem os objetos, e é determinada usando-se, geralmente, medidas de distâncias. De forma que, quanto maior a distância entre dois objetos em questão, mais diferentes eles são, e quanto menor a distância, mais similares eles são.

A análise de grupos pode ser usada como uma ferramenta autônoma para obter informações sobre a distribuição de dados, observar as características de cada grupo e focar em um determinado conjunto de grupos para análise posterior (HAN; KAMBER; PEI, 2012). Porém, Han, Kamber e Pei (2012) destacam que, em alguns casos, a análise de grupos pode ser usada apenas como uma ferramenta de apoio a outras técnicas, como caracterização, seleção de subconjunto de atributos e classificação, de maneira que, após o agrupamento, o usuário pode, posteriormente, focar em cada grupo encontrado individualmente, para analisar suas características e relacionamentos entre os dados em cada grupo.

Além disso, uma característica importante da tarefa de agrupamento, é que ela possui um componente subjetivo. Pois o agrupamento aplicado ao mesmo conjunto de dados pode obter resultados diferentes para diferentes propósitos. Por exemplo, se considerarmos uma baleia, um elefante, e um atum, as baleias e os elefantes formam um grupo de mamíferos. Porém, se o usuário estiver interessado no agrupamento com base no habitat do animal, então baleias e atuns farão parte de um grupo e o elefante de outro grupo (CASTRO, 2013). Dessa maneira, consoante o que explica Castro (2013), a subjetividade está fortemente presente no processo de agrupamento devido a necessidade de decisões serem tomadas pelo usuário, como a escolha do modelo de representação dos dados, escolha da medida de similaridade e escolha do algoritmo a ser utilizado, e as vezes, a escolha do número de grupos desejado.

Segundo Han, Kamber e Pei (2012), os métodos de agrupamento devem atender a alguns requisitos, como: possuir escalabilidade, para suportar bases de dados de grande porte; habilidade de descobrir grupos de forma arbitrária, visto que, os grupos podem ser das mais

diversas formas; ter a capacidade de lidar com diferentes tipos de atributos; ter capacidade de lidar com conjunto de dados que apresentem ruídos, já que a presença de ruídos é algo comum em bases de dados, logo a presença deles não deve comprometer a qualidade do agrupamento; e ter capacidade de lidar com dados de diferentes dimensionalidades.

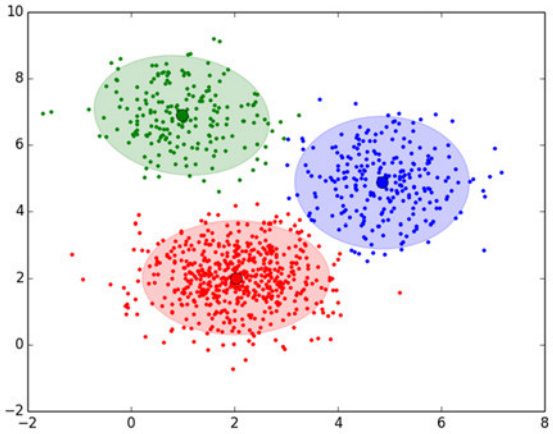
No entanto, conforme Agrawal et al. (1998 apud CARLANTONIO, 2001), atualmente, nenhuma técnica de agrupamento satisfaz todas essas condições. Em vez disso, há métodos apropriados para cada tipo de aplicação. Como exemplo temos a técnica de agrupamento K-means, que será abordada na Seção 2.5.1, que divide o conjunto de dados em um número k especificado pelo usuário. O K-means é indicado quando a base de dados contém grupos de formatos convexos ou globulares e de tamanhos parecidos. Além disso, essa técnica é sensível a ruídos, uma vez que esses dados influenciam no cálculo do centroide de um grupo.

2.2 Processo de Agrupamento

A análise de grupos pode ser definida como o processo de determinação de *k* grupos em um conjunto de dados (VALE, 2005). Desse modo, dado uma amostra de *n* objetos, cada um medido segundo *p* variáveis (ou atributos), o agrupamento procura um esquema de classificação que agrupe os objetos em *k* grupos mutuamente exclusivos baseado nas similaridades entre os objetos (NOLETO, 2007).

A Figura 1 ilustra o agrupamento do conjunto de dados que resultou em três grupos distintos. Os pontos pintados da mesma cor pertencem ao mesmo grupo.

Figura 1 - Exemplo de agrupamento



Fonte: Moya, 2016.

Conforme Mendes (2017), o processo de agrupamento é feito em diversas etapas, que compreendem desde a preparação dos dados para a aplicação da técnica de agrupamento até a interpretação dos resultados. Dessa maneira, para realizar o agrupamento de um conjunto de objetos de dados, devem ser seguidos alguns passos, que segundo Vale (2005), podem ser realizados em três etapas que são: seleção e tratamento de dados, agrupamento de dados e análise dos resultados.

2.2.1 Seleção e tratamento de dados

O resultado de uma análise de agrupamentos deve ser um conjunto de grupos que podem ser consistentemente descritos por meio de suas características (CARVALHO, 2001 apud NOLETO, 2007). Assim, o sucesso do processo de agrupamento depende diretamente das características dos dados que foram selecionadas para representa-los no processo de agrupamento. Segundo Macedo et al. (2013), o objetivo do passo de seleção de características consiste na seleção da parcela de características originais mais representativas para o objetivo da aplicação.

Após a seleção das características deve ser feito o tratamento dos dados, que consiste em preparar os dados para ser aplicado o algoritmo de agrupamento, de maneira a garantir a qualidade dos resultados obtidos no processo de agrupamento. Essa etapa envolve a eliminação de dados duplicados ou corrompidos, tratamento de *outliers* (dados com comportamento fora do esperado) remoção de dados com valores faltantes e transformação dos dados, que compreende etapas de tratamento de atributos e normalização (DONI, 2004).

O tratamento de atributos tem como objetivo adequar os diferentes tipos de atributos para o processo de agrupamento (VALE, 2005). Embora os atributos dos dados possam ser dos mais diversos tipos, Castro (2013) afirma que para um problema de agrupamento de dados podem ser dos tipos quantitativos e qualitativos. Conforme Han, Kamber e Pei (2012), os

atributos quantitativos, são variáveis que possuem uma quantidade mensurável, representada em valores reais ou inteiros. Podem ser do tipo contínuo (variáveis de valores reais ou intervalares) e do tipo discreto (variáveis de valores inteiros). Exemplos de atributos do tipo quantitativo são altura, salário, idade e CPF. Já os atributos qualitativos ou categóricos, são símbolos ou nomes de coisas. Podem ser do tipo binário (quando possuem apenas duas categorias ou estados) ou nominal (quando possui mais que dois tipos de categorias). São exemplos de atributos categóricos: sexo (feminino ou masculino), estado civil (casado, solteiro, divorciado), cor do cabelo (castanho, loiro, ruivo), tipo de pagamento (cartão de crédito, boleto, cartão de débito). Segundo Vale (2005), os atributos do tipo categórico devem ser representados de forma numérica para serem tratados pelo algoritmo de agrupamento. Por exemplo, o atributo cor do cabelo do exemplo anterior pode ser representado pelos valores 1, 2 e 3 sendo, castanho, loiro, ruivo respectivamente.

Conforme Castro (2013), a dimensão dos atributos é um fator determinante para o processo de agrupamento, visto que, atributos de dimensões diferentes podem ter influência proporcional ao tamanho dos valores que pode assumir.

A unidade de medida usada para representar os atributos pode afetar a análise de dados. Em geral, expressar um atributo em unidades menores levará a um intervalo maior para esse atributo e, portanto, tenderá a atribuir um maior efeito ou "peso" a esse atributo. Para ajudar a evitar a dependência da escolha de unidades de medida, os dados devem ser normalizados ou padronizados (HAN; KAMBER; PEI, 2012). Diante disso, Castro (2013) explica que, antes da realização do agrupamento deve ser feita a normalização dos valores para uma escala comum, de forma que cada variável tenha o mesmo peso na execução do algoritmo, ou seja, cada variável terá a mesma influência no processo de agrupamento.

2.2.2 Agrupamento de dados

Essa etapa é onde é feita a aplicação de uma técnica de agrupamento adequada, que deve ser escolhida levando em consideração o tipo de dado e objetivo especifico que se deseja alcançar com o agrupamento de dados. Segundo Dunham (2002), os algoritmos de agrupamento de dados podem ser classificados como hierárquicos e particionais.

No **método hierárquico**, o agrupamento é um conjunto de grupos aninhados organizados como uma árvore. Cada nodo (grupo) na árvore é a união dos seus filhos (subgrupos) e a raiz da árvore é o grupo contendo todos os objetos (TAN, 2006).

No **método particional**, o conjunto de objetos de dados é dividido em subconjuntos (grupos) não interseccionados de modo que cada objeto de dado esteja exatamente em um

subconjunto (TAN, 2006).

As técnicas de agrupamento são categorizadas de acordo com o método de agrupamento usado para agrupar os dados, embora os particionais sejam os mais comuns de acordo com Castro (2013). Conforme Bussab, Miazaki e Andrade (1990), para usar as técnicas de agrupamento, algumas decisões devem ser tomadas pelo usuário, como quantidade de grupos que deseja encontrar, medida de distância e função objetiva, dessa forma, o usuário deve ter conhecimentos das propriedades dessas técnicas e das necessidades da aplicação.

Nos estudos de caso do Capítulo 3, foi realizada uma discussão sobre as características de cada aplicação que motivaram a escolha da técnica utilizada.

2.2.2.1 Medidas de proximidade

Ao longo desse trabalho, pode-se perceber a importância da medida de semelhança entre dois objetos para a definição de agrupamento. As medidas de semelhança podem ser divididas em duas categorias: medidas de similaridade e de dissimilaridade. Na primeira, quanto maior o valor da medida de similaridade, mais semelhantes são os objetos. Já para a segunda, quanto maior o valor da medida, menos semelhantes (mais dissimilares) serão os objetos (BUSSAB; MIAZAKI; ANDRADE, 1990). As diferenças e semelhanças são avaliadas com base nos valores de atributos que descrevem os objetos e geralmente envolvem medidas de distância (HAN; KAMBER; PEI, 2012).

Devido à grande quantidade de tipos de características e escalas, a medida de distância deve ser escolhida cuidadosamente. O mais comum é calcular a dissimilaridade entre dois padrões utilizando uma medida de distância definida no espaço de características (CASTRO, 2003).

Deste modo, para escolher a medida de distância mais apropriada para a aplicação, deve-se levar em consideração principalmente os tipos de características dos dados disponíveis. Embora exista uma quantidade considerável de medidas de proximidade disponíveis na literatura para cada tipo de característica de dado e escala de atributo, Han, Kamber e Pei (2012) destacam que a Distância Euclidiana é usada frequentemente para pontos no espaço Euclidiano. Segundo Hair et al. (2005), corresponde a medida do comprimento de um segmento de reta desenhado entre dois objetos.

A função de Distância Euclidiana entre duas amostras x_i e x_j, ambas de dimensão d (quantidade de características de uma amostra), é dada por (AGUIAR et al., 2018):

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{d} (x_{i,p} - x_{j,p})^2} \quad (1)$$

Considerando x_i e x_j, com apenas dois atributos, temos:

$$d(x_i, x_j) = \sqrt{(x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2} \quad (2)$$

Neste trabalho será utilizada a Distância Euclidiana como medida de proximidade nos estudos de caso do Capítulo 3, sobre agrupamento de dados aplicado em bases de dados corporativas para a identificação dos perfis de clientes e classificação de produtores de leite para apoio a estratégias negócios.

2.2.3 Análise dos resultados

Nesta etapa, deve ser feita primeiramente a avaliação da qualidade do agrupamento, chamada de **validação**. Esta etapa refere-se aos procedimentos que avaliam os resultados da análise de grupos de maneira quantitativa e objetiva. Uma estrutura de agrupamento é válida se não ocorreu por acaso ou se é "rara" em algum sentido, visto que, um algoritmo de agrupamento sempre encontrará grupos, independentemente de existir ou não similaridade nos dados (JAIN; DUBES, 1988).

A **análise dos resultados** é uma etapa tão importante quanto a do próprio agrupamento, visto que é nela que se faz a observação da qualidade dos grupos criados e a análise dos significados de cada um. O agrupamento corresponde apenas a um esboço da possível relação entre os dados, dessa forma é necessário se fazer a interpretação dos resultados de maneira a extrair regras ou o resumo de características que tornem possível explicar os grupos formados. Logo, o especialista no domínio precisará interpretar os grupos de dados formados, atrelando às suas observações outros conhecimentos sobre os dados para dar sentido aos resultados. Em um agrupamento de clientes por exemplo, os grupos criados podem significar o perfil dos clientes que consomem um determinado tipo de produto, ou ainda, o perfil dos clientes que fazem um tipo específico de reclamação sobre um produto.

2.3 Diferentes tipos de Grupos

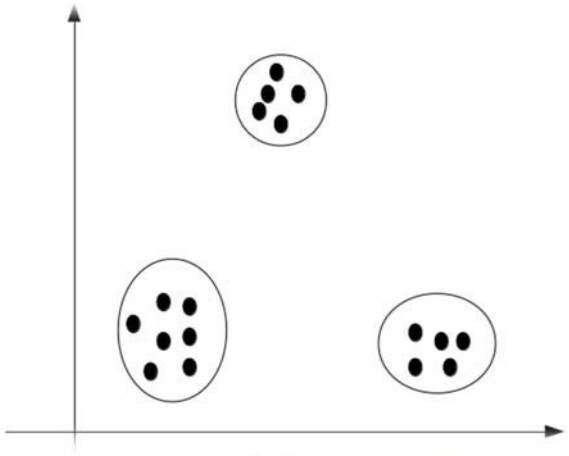
De acordo com Tan (2006), o agrupamento é útil à medida em que pode levar a descoberta de grupos de objetos previamente desconhecidos e potencialmente proveitosos

para os objetivos da análise de dados. Nesta seção, serão apresentadas algumas definições importantes de grupos que podem ser úteis na prática.

2.3.1 Bem separados

Um grupo é um conjunto de objetos no qual cada objeto está mais próximo (ou é mais semelhante) a cada um dos outros objetos no grupo do que de qualquer outro objeto que não esteja nesse grupo. Às vezes um limite é usado para especificar que todos os objetos em grupos devem estar suficientemente próximos (ou serem semelhantes) entre si (TAN, 2006). A Figura 2 ilustra três grupos bem separados. Pode-se perceber que a distância entre os pontos em grupos diferentes é bem maior do que a distância entre os pontos no mesmo grupo.

Figura 2 - Três grupos bem separados

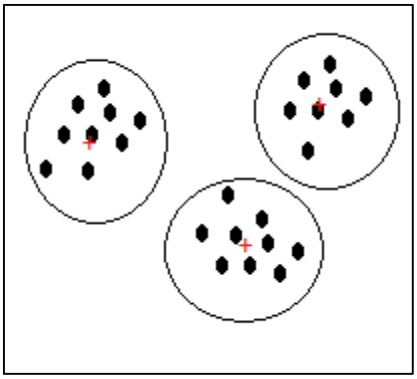


Fonte: Compilação do autor.

2.3.2 Baseados em Protótipos

Consoante o que explica Tan (2006), os grupos são representados por protótipos, onde os elementos de um grupo são mais próximos do protótipo do seu grupo, e mais distantes do protótipo de outros grupos, indicando maior semelhança entre os pontos de um mesmo grupo. O protótipo de um grupo, pode ser um centroide ou um medoide. O centroide corresponde à média das características de todos os objetos de um grupo, deve ser usado quando os dados possuem valores contínuos. Já o medoide, corresponde ao objeto mais próximo do centro do grupo, é usado quando os dados possuem valores categóricos. Tan (2006) explica que, em situações em que o protótipo do grupo corresponde ao ponto central, nos referimos a esses grupos como grupos baseados em centro. A Figura 3 ilustra três grupos baseados em protótipos. A cruz vermelha no centro representa o protótipo do grupo.

Figura 3 - Exemplo de grupos baseados em protótipo



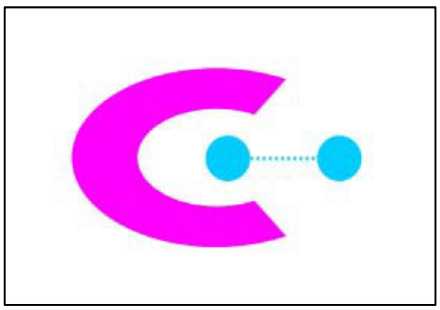
Fonte: Compilação do autor.

2.3.3 Baseados em Grafo

Em agrupamentos baseados em grafo, um grupo é um conjunto de objetos de dados em que objetos do mesmo grupo são conectados entre si, e não tenham conexão com objetos fora do grupo (TAN, 2006). Conforme Tan (2006), se pensarmos nos dados como um grafo, os nodos sendo os dados e as arestas a relação entre eles, o grupo é então um componente conectado. Dessa forma, um exemplo de grupos baseados em grafo segundo Tan (2006), são grupos baseados em contiguidade, onde dois objetos estão conectados apenas se estiverem dentro de uma distância especificada. Ou seja, um ponto em um grupo está mais próximo (mais similar) a um ou mais pontos no mesmo grupo, do que a qualquer ponto que não está no grupo. A Figura 4 ilustra dois grupos baseados em contiguidade. Pode-se perceber que uma pequena ponte de grupos que estão relacionados une dois grupos distintos.

A definição de grupos baseados em grafo é útil quando os grupos tiverem formas irregulares ou estiverem entrelaçados, mas pode ter problema quando ruído estiver presente (TAN, 2006).

Figura 4 - Exemplo de 2 grupos contíguos



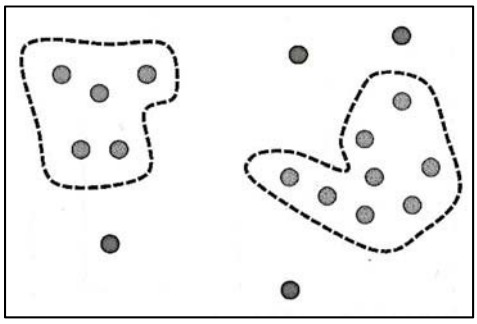
Fonte: Nievola, 2019.

2.3.4 Baseados em Densidade

Em agrupamentos baseados em densidade, um grupo é uma região densa de objetos que seja rodeada por uma região de baixa densidade (TAN, 2006). A Figura 5 ilustra dois grupos baseados em densidade, e quatro pontos que não foram agrupados (ruídos).

Uma definição baseada em densidade de um grupo muitas vezes é empregada quando os grupos são irregulares ou entrelaçados e, quando há ruídos ou elementos externos (TAN, 2006).

Figura 5 - Exemplo de grupos baseados em densidade



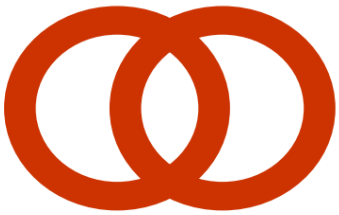
Fonte: Amaral, 2016.

2.3.5 Grupos Conceituais

Também chamado Propriedade Compartilhada. De modo mais geral, podemos definir um grupo como um conjunto de objetos que compartilham alguma propriedade (TAN, 2006). No entanto, Tan (2006) alerta para o fato de que essa definição pode englobar todas as definições prévias de grupo, visto que, objetos em um mesmo grupo sempre compartilham alguma propriedade. Dessa forma, é importante esclarecer que essa abordagem, isto é, da propriedade compartilhada, inclui novos tipos de grupos.

A Figura 6 ilustra dois círculos sobrepostos (grupos), onde os pontos na intersecção dos círculos pertencem a ambos os grupos. Nesse caso, um algoritmo precisaria de um conceito muito específico de grupo para detectar os mesmos. O processo de encontrar esses grupos é chamado de agrupamento conceitual.

Figura 6 - Exemplo de agrupamento conceitual com grupos circulares sobrepostos



Fonte: Nievola, 2019.

2.4 Métodos de Agrupamento

Existem muitos algoritmos de agrupamento na literatura. É difícil fornecer uma categorização nítida dos métodos de agrupamento, pois um método pode ter recursos de várias categorias (HAN; KAMBER; PEI, 2012). Dunham (2002) afirma que os algoritmos de agrupamento podem ser vistos em si como hierárquicos ou particionais.

2.4.1 Métodos Hierárquicos

Os métodos hierárquicos produzem um conjunto aninhado de grupos. Cada nível na hierarquia tem um conjunto separado de grupos. No nível mais baixo, cada item está em seu próprio grupo exclusivo. No nível mais alto, todos os itens pertencem ao mesmo cluster. Com o agrupamento hierárquico, o número desejado de grupos não é inserido (DUNHAM, 2002).

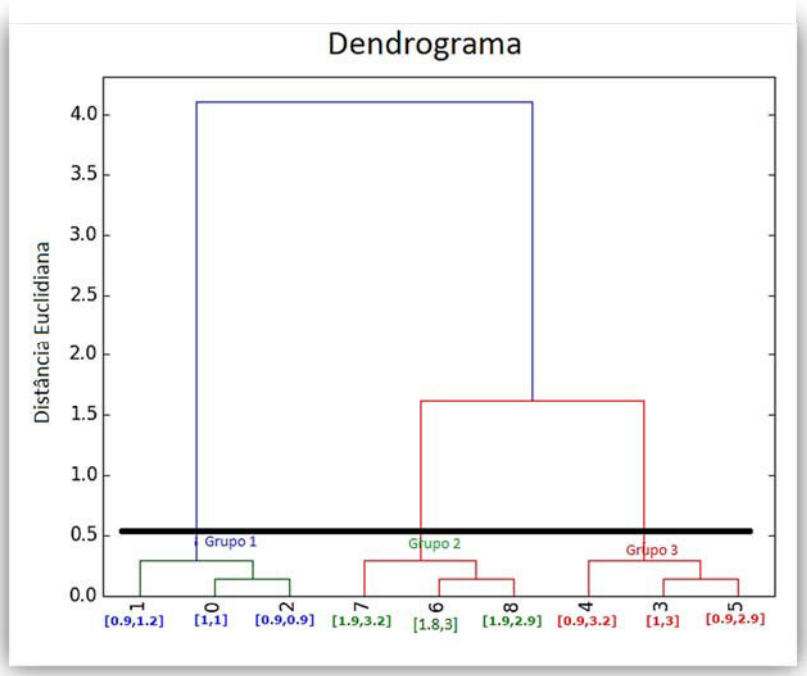
De acordo com Tan (2006), o agrupamento hierárquico é exibido usando uma estrutura de dados chamada dendrograma. É um diagrama que exibe a hierarquia entre os grupos formandos criando uma árvore de grupos, através da qual, podemos observar os relacionamentos entre grupo e subgrupo quanto a ordem no qual os grupos foram formados em cada estágio do agrupamento e os níveis de similaridade ou distância entre os grupos.

As técnicas hierárquicas não necessitam da informação prévia do nenhum número grupos para o particionamento dos dados, em vez disso, o número de grupos pode ser obtido a partir da análise do dendrograma, já que essas técnicas produzem uma solução para cada quantidade de grupos possível de particionar o conjunto de dados. Dessa forma, para obter a quantidade de grupos desejada deve-se fazer um corte no dendrograma no nível desejado. Na Figura 7, a distância entre os grupos é medida ao longo do eixo vertical, e os diferentes objetos do conjunto de dados, ao longo do eixo horizontal.

Para escolher a solução final, examinamos a mudança na medida de heterogeneidade para identificar grandes aumentos indicativos da fusão de agrupamentos distintos (HAIR et al., 2005). Dessa maneira, conforme Hair et al. (2005), o passo anterior a uma mudança

abrupta na similaridade, ou seja, quando a distância entre os grupos aumenta significativamente ao unir dois grupos, indicando que dois grupos distintos foram unidos, pode fornecer um bom ponto de corte para a partição final, onde os grupos serão provavelmente mais homogêneos. No entanto, essa é uma decisão subjetiva, pois devem ser usados também, conhecimentos práticos dos dados para determinar a partição final que faz mais sentido para o objetivo da aplicação.

Figura 7 - Dendrograma



Fonte: Moya, 2016.

O dendrograma da Figura 7 exibe uma partição final de 3 grupos com distância aproximada de 0,5 (linha de corte horizontal preta). O primeiro grupo é formado pelos objetos 1, 0 e 2. O segundo grupo, pelos objetos 7, 6 e 8. E por fim, o terceiro grupo é formado pelos objetos 4, 3 e 5. É possível perceber que quanto mais próximo o corte do dendrograma for da raiz da árvore, menor é o número de grupos, e mais dissimilares poderão ser os objetos. Além disso, pode-se perceber ainda como o nível de similaridade diminui conforme os grupos vão sendo combinados (aumentando a distância), até alcançar a solução de somente um grupo, indicando que os grupos estão se tonando menos homogêneos.

Os métodos hierárquicos utilizam uma matriz de similaridade, conhecida como matriz de similaridades entre agrupamentos, contendo as métricas de distância entre os agrupamentos em cada estágio do algoritmo (VALE, 2005).

Imaginando um estágio do algoritmo onde o número de agrupamentos corrente é três

(G1, G2, G3), pode-se supor a seguinte matriz de similaridades entre os agrupamentos (VALE, 2005). Observe o exemplo da Tabela 1:

Tabela 1 - Exemplo de matriz de similaridades entre grupos

	G1	**G2**	**G3**
G1	0	0,1	0,3
G2	0,1	0	0,4
G3	0,3	0,4	0

Fonte: Vale, 2005.

A partir da Tabela 1, podemos observar que os agrupamentos G1 e G2 são os mais similares, enquanto G1 e G3 são menos similares (VALE, 2005).

Conforme Han, Kamber e Pei (2012), os métodos hierárquicos são classificados de acordo a abordagem de decomposição hierárquica que utiliza para agrupar os dados. Podem ser de dois tipos: aglomerativo ou divisivo.

O **método hierárquico aglomerativo,** também chamado *bottom-up*, começa com cada objeto formando um grupo separado. Mescla sucessivamente os objetos ou grupos mais próximos um ao outro, até que todos os grupos sejam fundidos em um nível mais alto da hierarquia (HAM; KAMBER; PEI, 2012).

O **método hierárquico divisivo**, também chamado top-down, inicia com todos os objetos formando um único grupo, e em cada iteração um grupo é dividido em grupos menores, até que só exista um objeto em cada grupo (HAM; KAMBER; PEI, 2012).

2.4.1.1 Método Hierárquico Aglomerativo

Segundo Tan (2006), a abordagem aglomerativa é a mais comum para a aplicação de agrupamento hierárquico, e será abordada mais detalhadamente nesse trabalho.

Neste método, no estágio inicial, conforme Tan (2006), cada um dos objetos da base de dados representa um grupo separado. Durante o processo de agrupamento hierárquico aglomerativo, em cada estágio os grupos mais próximos vão sendo fundidos até que reste somente um grupo englobando todos os objetos da base. Esse procedimento está formalmente descrito no Algoritmo 1.

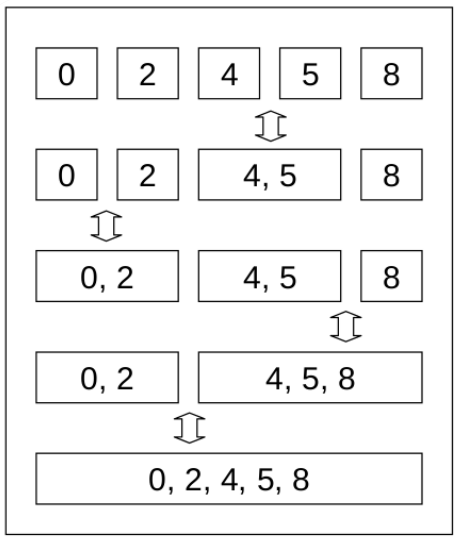
Algoritmo 1 – Algoritmo Hierárquico Aglomerativo

Algoritmo de agrupamento hierárquico aglomerativo básico
1. **Inicio:** cada grupo contém um único objeto.
2. Calcule a matriz de proximidades
3. **repita**
4. Funda os dois grupos mais próximos.
5. Atualize a matriz de proximidades.
6. **até que** que reste apenas um grupo.

Fonte: Tan, 2006.

O processo de execução do agrupamento hierárquico aglomerativo acontece da seguinte forma, de acordo com Noleto (2007): sendo *n* o número de objetos da base de dados, em cada estágio, os grupos vão sendo fundidos, e vão se obtendo *n*-1, *n*- 2, …, e assim sucessivamente, até que reste apenas um grupo com todos os objetos. A Figura 8 ilustra o procedimento do método hierárquico aglomerativo aplicado ao conjunto {0, 2, 4, 5, 8}, baseado em Noleto (2007).

Figura 8 - Fluxo de execução do método hierárquico aglomerativo



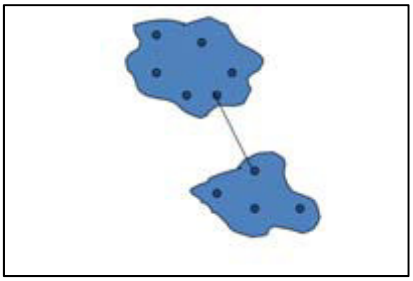
Fonte: Noleto, 2007.

Segundo Tan (2006), a maioria dos algoritmos de agrupamento hierárquico aglomerativo são variações do Algoritmo 1, ou seja, iniciam com pontos iniciais como grupos, em seguida funde os dois grupos mais próximos até que reste apenas um grupo. O que distingue uma

técnica de outra é a forma como se define a proximidade entre grupos. As técnicas mais comuns são: conexão única, conexão completa, média de grupo e método de Ward.

A técnica de **conexão única** ou ligação por vizinho mais próximo define proximidade de entre dois grupos como o mínimo da distância (máximo de semelhança) entre dois pontos quaisquer nos dois grupos diferentes (TAN, 2006). Conforme Doni (2004), essa técnica é boa para lidar com grupos de formas não elípticas, mas é sensível a ruídos, visto que, os ruídos podem ser incorporados a grupos já existentes, e pode ainda, gerar encadeamento de elementos, pois um grupo de um ou mais elementos pode incorporar em cada estágio do agrupamento um grupo de apenas um elemento, ou seja, unindo elementos bem diferentes. O encadeamento forma uma longa cadeia de agrupamentos que dificultam a definição de um nível de corte para a partição final. A Figura 9 ilustra a distância entre dois grupos por conexão única.

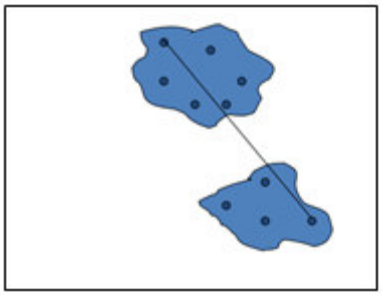
Figura 9 - Conexão única



Fonte: Laureto, 2019.

A técnica de **conexão completa** ou ligação por vizinho mais distante define proximidade entre dois grupos como o máximo da distância (mínimo da semelhança) entre quaisquer dois pontos nos dois grupos diferentes (TAN, 2006). Segundo Tan (2006), essa técnica, é menos susceptível a ruídos e elementos externos, já que essa técnica leva mais tempo para fundir grupos com esses elementos aos outros grupos, e tende a criar grupos compactos, favorecendo formatos globulares. A Figura 10 ilustra a distância entre dois grupos por conexão completa.

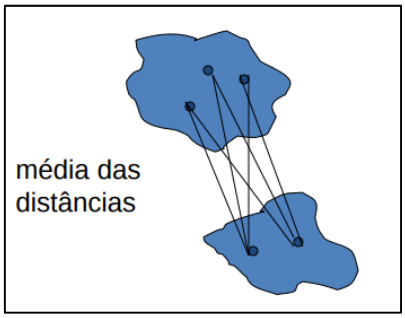
Figura 10 - Conexão completa



Fonte: Laureto, 2019.

A técnica da **média do grupo** define proximidade entre dois grupos como a proximidade média de pares entre todos os pares de pontos nos diferentes grupos (TAN, 2006). Segundo Doni (2004), essa técnica possui menor sensibilidade a ruídos do que os métodos de conexão única e conexão completa e tende a formar grupos de tamanhos similares. A Figura 11 ilustra a distância entre dois grupos por média do grupo.

Figura 11 - Média do grupo



Fonte: Laureto, 2019.

O **Método de Ward** supõe que um grupo seja representado pelo seu centroide, e mede a proximidade entre dois grupos em termos do aumento no erro quadrado que resulta da fusão dos dois grupos (TAN, 2006). Dessa maneira, conforme Hair et al., (2005), os dois grupos a serem fundidos, em cada estágio do algoritmo, serão aqueles que apresentarem o menor aumento na soma total do erro quadrado de todos os grupos.

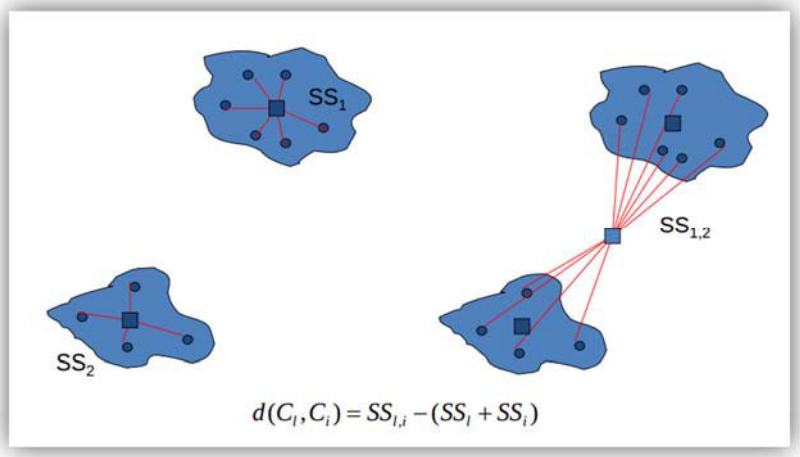
A soma do erro quadrado é definida formalmente como (TAN, 2006):

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2 \qquad (3)$$

Onde *SSE* é a soma do erro quadrado para todos os objetos no conjunto de dados C_i, k é o número de grupos, x é o ponto no espaço representando um dado objeto, c_i é o centroide do grupo C_i e $dist(c_i, x)$ é a distância Euclidiana entre o objeto x e o representante de um grupo c_i. Em outras palavras, para cada objeto em cada grupo, a distância do objeto ao centro do grupo é quadrada e as distâncias são somadas (HAN; KAMBER; PEI, 2012).

O método de Ward é sensível a presença de *outliers* e tem a tendência a combinar grupos com poucos elementos (DONI, 2004). A Figura 12 ilustra o processo de cálculo da distância entre dois grupos pelo método de Ward, onde *SS* é a mesma *SSE*.

Figura 12 - Método de Ward



Fonte: Laureto, 2019.

Como bem explica Tan (2006), as técnicas de agrupamento hierárquico aglometativo costumam tomar boas decisões locais de junção de dois grupos, visto que para decidir se combinam dois grupos ou não, as técnicas usam informações da distância de todos os pares de todos os grupos, ou seja, mensuram a semelhança entre eles em cada estágio do algoritmo para identificar os dois grupos que devem ser fundidos, no entanto, uma vez que a decisão de combinação dos dois grupos é tomada, ela não pode ser desfeita, assim, o algoritmo evita que um critério de otimização local, se torne critério de otimização global.

As técnicas hierárquicas são vantajosas por permitir determinar experimentalmente o número de grupos desejado a partir da análise dos resultados através do dendrograma e, como afirma Tan (2006), a escolha dos pontos iniciais não é um problema para essas técnicas. Contudo, possuem uma grande desvantagem segundo Doni (2004), o custo computacional elevado, pois requerem grande quantidade de espaço de armazenamento e tempo de processamento por causa das matrizes de similaridade, que é calculada em cada estágio do algoritmo para todas as combinações dos dados, tornando a sua aplicação impraticável em grandes bases de dados.

Uma das principais críticas ao método de agrupamento hierárquico é seu custo computacional elevado, que é de pelo menos $O(N^2)$, o que limita o seu uso para aplicações com grandes conjuntos de dados (XU; WUNCH, 2009).

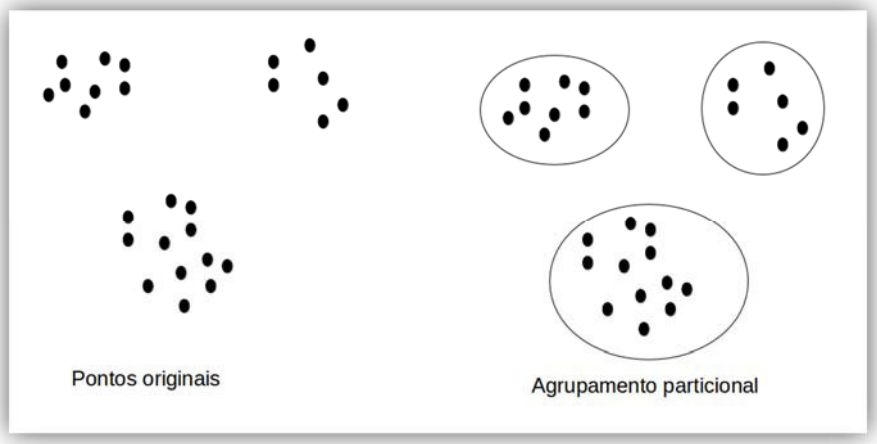
2.4.2 Métodos Particionais

A versão mais simples e fundamental da análise de grupos é o particionamento, que divide o conjunto de objetos de dados em subconjuntos (grupos) não interseccionados, de modo que cada objeto de dado esteja em exatamente um subconjunto (TAN, 2006).

Conforme Vale (2005), os métodos particionais buscam minimizar uma função objetiva, organizando os objetos da base em um número k de grupos escolhido previamente. Cada objeto de dado é atribuído ao grupo em que essa função objetiva é minimizada. Dessa forma, esses métodos particionais produzem somente uma solução de agrupamento para o conjunto de dados a partir da quantidade k de grupos, diferente dos métodos hierárquicos que produzem uma solução para cada quantidade de grupos possível de se particionar o conjunto de dados.

A Figura 13 ilustra um agrupamento particional, que resultou em três grupos não interseccionados, ou seja, cada ponto pertence a somente um grupo.

Figura 13 – Exemplo de agrupamento particional



Fonte: Compilação do autor.

As técnicas particionais possuem um desempenho superior as técnicas hierárquicas quando aplicadas em bases de dados de grande porte, devido ao fato de obterem como resultado do agrupamento somente uma partição do conjunto de dados, ou seja, somente uma solução em vez de criar soluções para cada quantidade de grupos possível.

Segundo Ankerst et al. (1999) essas técnicas são eficientes se o número de grupos puder ser estimado e o grupos tiverem formatos convexos, forem de tamanhos parecidos e tiverem densidades semelhantes.

Além disso, como já foi possível perceber, esses métodos necessitam que o número de

grupos que se deseja particionar o conjunto de dados seja informado previamente, o que segundo Fung (2001), poderá levar a interpretações equivocadas sobre a estrutura dos dados caso o número de grupos não seja o mais adequado para representar o conjunto de dados e no fato da qualidade dos resultados do algoritmo depender do estado inicial do algoritmo, podendo gerar resultados diferentes a cada rodada. Isso acontece porque o algoritmo é obrigado a condicionar os dados a uma estrutura específica, ao invés de encontrar a estrutura mais adequada para o particionamento dos dados.

Segundo Bussab (1990), conforme citado por Doni (2004), na prática esses métodos são aplicados múltiplas vezes para valores diferentes de k, com o intuito de escolher posteriormente a quantidade de grupos que melhor represente o conjunto de dados.

Para Fung (2001), uma das principais vantagens das técnicas particionais, é a possibilidade de mover os objetos de dados entre os grupos durante a execução do processo de agrupamento, permitindo a correção dos grupos, e além disso, devido a sua complexidade computacional, que é linear, os métodos particionais possuem desempenho superior no agrupamento de grandes volumes de dados, em relação aos métodos hierárquicos, isso acontece devido ao fato de não precisar calcular e armazenar a matriz de similaridades no processo.

Estes algoritmos são relativamente eficientes e têm uma complexidade temporal equivalente ao número de objetos e número de grupos a criar, isto é: O ($t \cdot k \cdot n$) onde *n* é o número de objetos, *k* o número de grupos e *t* o número de iterações (CASTRO, 2003).

Os algoritmos particionais mais comuns são K-means e K-medoid. O K-means será abordado detalhadamente na Seção 2.5.1, pois foi usado nos estudos de caso das Seções 3.1 e 3.3 do Capítulo 3 sobre agrupamento de clientes de um site de reservas de hotéis, e agrupamento de produtores de leite de uma indústria de laticínios, respectivamente. Será abordado também, neste trabalho (Seção 2.5.2) o Algoritmo Genético, que assim como o K-means, obtém um agrupamento particional como resultado do agrupamento e foi usado no estudo de caso da Seção 3.2.

2.5 Algoritmos de Agrupamento

Nesta seção serão abordadas duas técnicas de agrupamento particional, K-means e Algoritmo Genético, que foram utilizadas nos estudos de caso do Capitulo 3, ambas no contexto de negócios, com o objetivo de traçar o perfil de clientes como forma de descoberta de conhecimento para apoio a estratégias de negócios.

2.5.1 Algoritmo K-means

Este algoritmo, com pequenas variações, talvez seja um dos mais usados em análise de agrupamentos (BUSSAB et al., 1990 apud NOLETO, 2007). É uma técnica particional de agrupamento baseada em protótipos, que particiona os dados em k grupos mutuamente exclusivos, sendo k o valor desejado de grupos informado pelo usuário. Cada grupo é representado pelo seu centro, que corresponde ao centroide do grupo, e cada objeto de dado é atribuído ao grupo que possui o centro mais próximo.

Uma função objetiva é usada para avaliar a qualidade do particionamento para que objetos dentro de um grupo sejam semelhantes uns aos outros, mas dissimilares a objetos de outros grupos (HAN; KAMBER; PEI, 2012). Geralmente, a Soma do Erro Quadrado, apresentada na Seção 2.4.1 (Equação 3), é usada como função objetiva pelo Algoritmo K-means, de acordo com Tan (2006).

Han, Kamber e Pei (2012) explicam que a Soma do Erro Quadrado como função objetiva tenta tornar os k grupos tão compactos e separados quanto possível, visto que, o algoritmo divide o conjunto de dados em k partições que minimizem o erro quadrado, de forma que os objetos de um mesmo grupo sejam bem semelhantes entre si, e bem diferentes de objetos de outros grupos.

O algoritmo trabalha da seguinte forma, conforme Tan (2006): inicialmente, o K-means escolhe k objetos de dados do conjunto de dados a ser agrupado para serem os representantes de grupo, k corresponde ao número de grupos desejado, que deve ser informado pelo usuário previamente. A partir daí, em cada estágio do algoritmo, o K-means atribui cada objeto de dado ao grupo em que o centroide está mais próximo do objeto, e então atualiza os representantes de grupo recalculando os centroides. Os passos de atribuição e atualização dos centroides é repetido até que a função objetiva venha a convergir ou um critério de parada seja alcançado. A execução termina, quando os centroides não mudarem mais, ou seja, quando não houver nenhuma mudança de objetos de um grupo a outro capaz de minimizar a função objetiva. O K-means é formalmente descrito no Algoritmo 2.

Algoritmo 2 – Algoritmo K-means

Algoritmo K-means

Entrada:

k: o número de grupos

D: conjunto de dados contendo *n* objetos.

Saída: um conjunto de k grupos

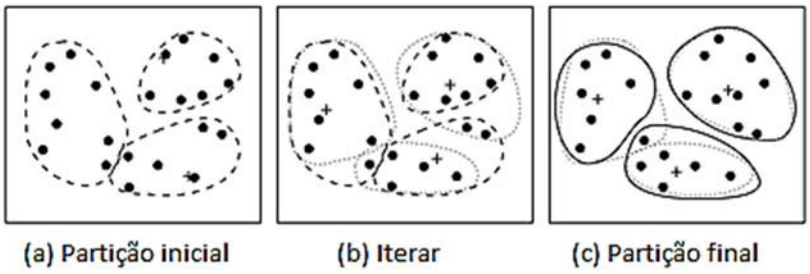
Método:

1. Escolha arbitrariamente *k* objetos de D como os centros iniciais dos grupos;
2. **Repita**
3. (re) atribua cada objeto ao grupo ao qual o objeto é mais semelhante, com base no valor médio dos objetos do grupo;
4. Atualize os centroides dos grupos;
5. **Até que** não haja mudanças.

Fonte: Han, Kamber, Pei, 2012.

A operação de K-means é ilustrada na Figura 14, começando com 3 centroides. Considere um conjunto de objetos, considerados no espaço 2D como ilustrado na Figura 14(a). Seja k=3, ou seja, o número de grupos desejado pelo usuário. De acordo com o Algoritmo 2, escolhemos arbitrariamente 3 objetos para serem os representantes de grupos iniciais, de forma que, cada objeto vai corresponder ao centro, onde são marcados por um "+". Cada objeto é atribuído a um grupo com base no centro do grupo ao qual ele é mais próximo. Essa distribuição forma as silhuetas circundadas por curvas pontilhadas, como mostra a Figura 14(a). Em seguida, os centros dos grupos são atualizados. Usando os novos centros, os objetos são redistribuídos para os grupos com base nos centros a que são mais próximos. Essa redistribuição forma novas silhuetas circundadas por curvas pontilhadas, como mostra a Figura 14(b). Esse processa itera, levando a Figura 14(c), até que eventualmente nenhuma redistribuição de objetos ocorra, finalizando o processo (HAN; KAMBER; PEI, 2012). Segundo Han, Kamber e Pei (2012), esse processo de redistribuição iterativa de objetos a grupos para melhorar o particionamento é chamado de relocação iterativa.

Figura 14 - Agrupamento de um conjunto de objetos usando o Algoritmo k-means



Fonte: Han, Kamber e Pei, 2012.

Para medir a distância entre um objeto e um representante de grupo, deve-se usar uma medida de proximidade para determinar a semelhança entre eles. Podem haver diversos tipos de medidas de proximidade para cada tipo de dado. Por exemplo, Tan (2006) afirma que para dados no espaço Euclidiano podemos usar tanto a distância Euclidiana como também a distância de Manhattan. E afirma ainda que, geralmente, são usadas medidas simples no K-means, visto que, o algoritmo calcula repetidamente a distância entre os pontos com cada centroide.

O algoritmo K-means nem sempre converge para o ótimo global, mas geralmente converge para o ótimo local (HAN; KAMBER; PEI, 2012). Isso acontece porque, como já vimos no algoritmo básico, os centroides iniciais são escolhidos aleatoriamente. Tan (2006) alerta que essa abordagem pode gerar centroides pobres, ou seja, o agrupamento obtido pode não ser o ideal de forma que o erro quadrado ainda não seja o mínimo. Dessa maneira, uma solução comumente empregada para esse problema descrita por Tan (2006) é a execução do algoritmo múltiplas vezes, cada uma com um conjunto diferente de centroides iniciais escolhidos aleatoriamente, e posteriormente o conjunto de grupos com o erro quadrado mínimo pode ser selecionado como partição final. No entanto, Tan (2006) afirma, que embora simples, essa estratégia pode não funcionar muito bem dependendo do conjunto de dados e do número de grupos procurados.

Uma abordagem efetiva para o problema do uso de centroides iniciais selecionados aleatoriamente, é pegar uma amostra de pontos e agrupá-los usando uma técnica de agrupamento hierárquico. Então, k grupos são extraídos do agrupamento hierárquico e os centroides desses grupos são usados como centroides iniciais. Essa abordagem muitas vezes funciona bem, mas é prática somente se a amostra for relativamente pequena e k for relativamente pequeno comparado com o tamanho da amostra (TAN, 2006).

O K-means possui uma complexidade de tempo da ordem de $O(n \cdot k \cdot l)$ e uma

complexidade de espaço da ordem de $O(k+n)$, onde n é o número de elementos, k é o número de grupos e l é o número de iterações do algoritmo (JAIN; DUBES, 1988). Normalmente, $k \leq n$ e $l \leq n$. Portanto, K-means é linear em *n,* e é eficiente e simples desde que k, o número de grupos sejas significativamente menor que n (HAN; KAMBER;PEI, 2012).

Diante do que foi mencionado, o K-means é simples e pode ser usado para uma ampla variedade de tipos de dados (TAN, 2006). É eficiente em encontrar grupos compactos e bem separados, no entanto, múltiplas execuções podem ser necessárias a fim de encontrar a partição que melhor represente o conjunto de dados. Além disso, como bem explica Tan (2006), essa técnica não é indicada para lidar com grupos de formatos não globulares e com grupos de tamanhos muito diferentes, pois tende a dividir grupos grandes na tentativa de minimizar o erro quadrado. E devido a sua complexidade computacional, podemos concluir que possui desempenho eficiente no agrupamento de grandes volumes de dados. Entretanto, esse método só pode ser aplicado a conjuntos de dados em que exista uma noção de centro, devido ao cálculo do centroide.

2.5.2 Algoritmos Genéticos

Algoritmos Genéticos são técnicas estocásticas de busca e otimização global, poderosas e largamente aplicáveis, inspiradas nos mecanismos naturais da evolução e da genética (LACERDA; CARVALHO, 1999 apud CARLANTONIO, 2001). Computacionalmente, o processo acontece a partir um conjunto de soluções potenciais (população de indivíduos) de um problema, onde a computação evolutiva expande essa população criando soluções novas e potencialmente melhores, usando para isso as melhores soluções correntes, de acordo com Dunham (2002).

Em agrupamento, o Algoritmo Genético pode ser visto como uma técnica para encontrar a melhor partição de um conjunto de dados, ou seja, a melhor organização do conjunto de dados em grupos. O algoritmo genético assume um modelo para representar a solução de particionamento do conjunto de dados, e cria iterativamente novas soluções a partir desses modelos, combinando os modelos (soluções) correntes, de forma que as novas soluções (modelos) criadas sejam melhores que as anteriores. Ou seja, combina as soluções correntes iterativamente para criar soluções melhores com o intuito de encontrar a melhor solução para o problema. Conforme Dunham (2002), essas técnicas usam uma função de avaliação para determinar as melhores soluções, de forma que, os melhores modelos sejam escolhidos para dar origem a novos modelos. E explica ainda que os algoritmos genéticos diferem em como a

solução é representada, como indivíduos diferentes no modelo são combinados para a geração de novos modelos e como a função de avaliação é usada.

Os Algoritmos Genéticos utilizam metaforicamente alguns termos da biologia, são eles: população (conjunto de soluções), geração (cada iteração do processo evolutivo), cromossomo (representação de uma solução), indivíduo (um cromossomo e o valor da sua aptidão), gene (variável codificada no cromossomo) (CARLANTONIO, 2001).

Assim, de forma análoga ao princípio da evolução e da genética, um algoritmo genético parte de uma população inicial, gerada aleatoriamente, realiza a avaliação de cada indivíduo calculando a função objetiva, seleciona os melhores (escolhe aqueles cuja função objetiva tenha os maiores valores, se for um problema de maximização, ou menores, no caso de minimização), realiza manipulações genéticas (cruzamento e mutação) a fim de criar uma nova população com soluções melhores, a partir da qual o processo reinicia-se iterativamente. Esse procedimento adaptativo pode ser usado para resolver qualquer problema de otimização. (MONTENEGRO; BRITO, 2006). Dessa maneira, o Algoritmo Genético busca a solução ótima para o problema mediante a avaliação e evolução de várias soluções possíveis para o problema.

Segundo Carlantonio (2001), os Algoritmos Genéticos desenvolvidos para serem aplicados ao problema de agrupamento são conhecidos como Algoritmos de Clustering Genéticos. E, o uso dessas técnicas em agrupamento de dados foi motivado pela capacidade dessas técnicas de testar um conjunto grande de soluções ao explorar grandes regiões do espaço de busca.

Para usar Algoritmo Genético, a primeira coisa, e talvez a tarefa mais difícil, é determinar como modelar um problema como um conjunto de indivíduos (soluções potenciais do problema) (Dunham, 2002).

Representação do problema

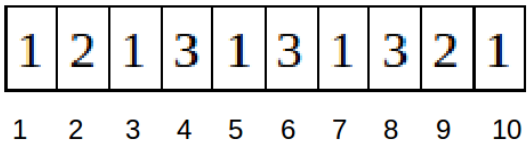
A representação de uma solução no Algoritmo Genético é chamada de cromossomo. Um cromossomo é uma estrutura de dados, geralmente vetor ou cadeia de bits, que representa uma possível solução do problema a ser otimizado. O conjunto de todas as configurações que o cromossomo pode assumir forma o seu espaço de busca (CARLANTONIO, 2001).

Conforme Montenegro e Brito (2006), uma solução para o problema de agrupamento considerando *n* objetos a serem agrupados, e um número *k* de grupos, a solução é representada por um vetor com *n* posições e em cada posição conterá um valor entre 1 e k obtido por meio de sorteio. Por exemplo, para um conjunto de objetos contendo 10 elementos, e k=3 grupos, a

solução será (Figura 15):

Na Figura 15, cada posição de 1 a *n* no cromossomo corresponde a um objeto do conjunto de dados, e o valor de cada gene corresponde ao grupo ao qual ele pertence. Desse modo, o grupo cujo o rótulo é 1 será formado pelos objetos 1, 3, 5, 7 e 10, o grupo cujo rótulo é 2 pelos objetos 2 e 9, e o grupo de rótulo 3 pelos objetos 4, 6 e 8.

Figura 15 - Exemplo de cromossomo



Fonte: Compilação do autor.

Dunham (2002) explica que, embora a representação mais frequentemente usada seja a sequência de bits, outras representações são possíveis, desde que os operadores genéticos usados por essas técnicas sejam definidos. E afirma ainda que um *array* com caracteres não-binários ou até mesmo estruturas mais complicadas podem ser usadas.

A cada iteração de um Algoritmo Genético, uma nova população de indivíduos deve ser criada, sobre a qual serão aplicados os operadores genéticos para criar os novos indivíduos. Esse processo é chamado de reprodução, e envolve passos de seleção dos indivíduos e reprodução por meio de cruzamento e mutação.

Seleção

O processo de seleção do algoritmo genético é baseado no processo de seleção natural de seres vivos. Consiste em selecionar os melhores cromossomos (soluções) da população inicial, ou seja, aqueles de alta aptidão, para gerar cromossomos filhos aplicando os operadores genéticos de cruzamento e mutação (CARLANTONIO, 2001). Diante disso, Dunham (2002) explica que, para determinar os melhores (ou mais aptos) indivíduos a sobreviver é usa-se uma função de aptidão, que geralmente está relacionada a função objetiva do problema. Os melhores indivíduos a partir do valor da função de avaliação calculado sobre cada um deles, são então escolhidos como cromossomos pais, ou seja, cromossomos que vão dar origem a nova população de indivíduos.

Existem vários métodos de seleção disponíveis. Segundo Souza (2008), os mais comuns são: Método da Roleta, Método do Torneio e o Método da Amostragem Universal

Estocástica. Contudo, para selecionar os melhores indivíduos baseado em sua aptidão, um dos métodos mais usados, como afirma Souza (2008), é o método da roleta, no qual, indivíduos de uma geração são escolhidos para fazer parte da próxima geração, através de um sorteio de roleta. De acordo com Dunham (2002), é dado por:

$$P_{I_i} = \frac{f(I_i)}{\sum_{I_i \in P} f(I_i)} \quad (4)$$

Onde P é a população dada, f é a função de aptidão, e P_{I_i} é a probabilidade do I_i indivíduo ser selecionado. Cada indivíduo recebe um setor da roleta proporcional a sua aptidão relativa, que corresponde a sua probabilidade de seleção (DUNHAM, 2002). Dessa forma, os indivíduos com melhor aptidão recebem uma porção maior da roleta do que os indivíduos menos aptos (SOUZA, 2008).

A roleta é rodada tantas vezes quanto for o número de indivíduos da população, escolhendo-se assim aqueles indivíduos que darão origem à próxima geração (SOUZA, 2008).

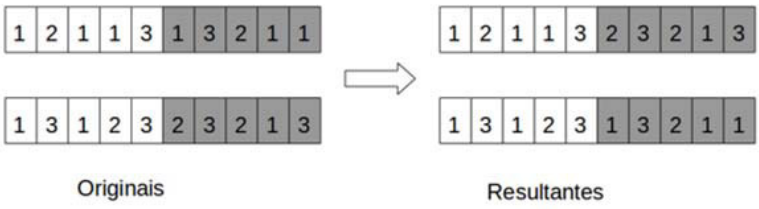
Reprodução (cruzamento e mutação)

Segundo Dunham (2002), a reprodução em algoritmos genéticos é realizada por algoritmos que indicam como combinar estruturas genéticas de pares de indivíduos (cromossomos pais) da população, ou seja, como combinar as soluções potenciais do problema, com o objetivo de gerar soluções (indivíduos) melhores, de forma a garantir a diversificação da população no espaço de soluções ao gerar configurações diferentes. Para a reprodução dos indivíduos, dois operadores genéticos devem ser aplicados, são eles: cruzamento e mutação.

A Figura 16 ilustra o processo de cruzamento por ponto único, onde o ponto de corte é aleatoriamente gerado, e este é menor ou igual ao tamanho do cromossomo. Os caracteres que precedem o ponto de corte são preservados, e os caracteres posteriores são trocados entre o par participante do processo (TEIXEIRA, 2005).

Na Figura 16, os locais que indicam os pontos de cruzamento são destacados com a cor cinza. O cruzamento é obtido trocando os últimos cinco bits das duas cadeias.

Figura 16 - Exemplo de cruzamento por ponto único



Fonte: Compilação do autor.

Além do cruzamento por ponto único, existem outras abordagens, incluindo a determinação de pontos de cruzamento múltiplos (DUNHAM, 2002).

O cruzamento usa uma probabilidade (taxa de cruzamento) para determinar quantos novos descendentes (indivíduos) serão criados por cruzamento (DUNHAM, 2002). Não ocorrendo o cruzamento, os filhos serão iguais aos pais, isto permite que algumas soluções sejam preservadas (CARLANTONIO, 2001).

Após o cruzamento, um operador de mutação deve ser aplicado, que também está sujeito a uma taxa de aplicação, que deve ser definida conforme características do problema, todavia, deve ser aplicado com uma probabilidade muito pequena (OCHI et al., 2004). Montenegro e Brito (2006) explicam que em agrupamento, a mutação consiste em alterar informações em alguns indivíduos com o objetivo de tentar recuperar valores entre 1 e k que podem ter sido eliminados no processo de reprodução. De acordo com Dunham (2002), a mutação altera informações presentes nos indivíduos aleatoriamente, e por isso deve ser aplicada com uma probabilidade pequena.

A seguir o processo do Algoritmo Genético, conforme Almeida (2015):

1. Represente a solução do domínio de problema como um cromossomo de tamanho fixo, e escolha o tamanho N da população de cromossomos;
2. Defina a função de aptidão para medir o desempenho de um cromossomo individual no domínio do problema.
3. Gerar aleatoriamente uma população inicial de cromossomos de tamanho N.
4. Calcule a aptidão de cada cromossomo da população.
5. Selecione um par de cromossomos da população atual para o cruzamento.
6. Criar um par de cromossomos filhos, aplicando os operadores genéticos (cruzamento, mutação).

7. Coloque os cromossomos filhos criados na nova população (população intermediária).
8. Repetição passo 5 até que o tamanho da nova população se torne igual ao tamanho da população inicial.
9. Substitua a população atual (pai) de cromossomos com a nova população (filho).
10. Ir para passo 4, e repita o processo até que o critério de parada seja satisfeito.

Os dois principais critérios de parada são um número máximo de gerações (cada passo do processo evolutivo é uma geração) ou a convergência da população (OCHI et al., 2004). No entanto, Ochi et al. (2004) alertam que ambos são critérios difíceis de definir, visto que, como não sabemos o número máximo ideal de gerações, o número de gerações escolhido pode ser insuficiente para atingir a melhor solução. E alertam ainda, que no caso da convergência como critério de parada, pressupõe-se que algoritmo vai parar quando todos os indivíduos forem iguais, no entanto, não há garantias de que isso pode vir a ocorrer, ou ainda quando atingir um percentual de indivíduos iguais.

Algumas vantagens dos algoritmos genéticos são: tem a capacidade de lidar com diferentes tipos de parâmetros, tanto com parâmetros contínuos, como discretos, ou uma combinação deles; realizam buscas simultâneas em várias regiões do espaço de busca ao que trabalhar com população, não somente com um ponto, ou seja, trabalha com um conjunto de soluções em cada estágio do processo de execução; otimizam um número grande de variáveis; fornecem uma lista de parâmetros ótimos, e não uma simples solução; trabalham com dados gerados experimentalmente, pois usam representação da solução e não o conjunto de dados em si, e são tolerantes a ruídos e dados incompletos devido a abordagem evolutiva que utilizam (CARLANTONIO, 2001).

Apesar disso, Carlantonio (2001) afirma que, eles não são eficientes para muitos problemas devido a sua lentidão. Diante disso, são voltados para aplicações que envolvam um grande conjunto de soluções que são inviáveis de serem analisadas por algoritmos tradicionais.

Neste trabalho, no estudo de caso da Seção 3.2 foi utilizado como função de avaliação a soma do erro quadrado apresentada na Seção 2.4.1 (Equação 3).

3 ESTUDOS DE CASO

Este capítulo ilustra a aplicação dos conceitos sobre agrupamento de dados apresentados no Capítulo 2, apresentando três estudos de caso com aplicações de algoritmos de agrupamento para extração de conhecimento de três bases de dados corporativas para o apoio a estratégias de negócios. Na Seção 3.1, é discutida uma aplicação do algoritmo K-means em uma base de dados de um site de reservas de hotéis, visando identificar perfis de clientes de uma agência de viagens online. Na Seção 3.2, é discutida uma aplicação do Algoritmo Genético na base de dados de uma empresa especializada em fornecimento de materiais de limpeza, visando traçar o perfil de seus clientes. E na Seção 3.3, é discutida uma aplicação de dois métodos de agrupamento em conjunto, o método de agrupamento hierárquico aglomerativo de Ward e o algoritmo K-means, aplicados na base de dados de uma indústria de laticínios com o objetivo de classificar seus fornecedores de leite de acordo com as características do leite produzidos por eles.

Para escolher qual a técnica de agrupamento é adequada para a aplicação, alguns fatores devem ser considerados como: tamanho da base de dados, tipos de dados disponíveis e objetivo da aplicação.

Em aplicações cujo o objetivo é a criação de uma taxonomia lógica dos dados, o mais indicado é usar métodos hierárquicos de agrupamento, uma vez que, produzem uma relação hierárquica dos grupos. Também são indicados se não houver uma noção do número de grupos existente, visto que, pode-se chegar a esse número ideal através da análise do dendrograma. Todavia, esses métodos não são viáveis de serem aplicados em grandes conjuntos de dados devido ao seu custo computacional elevado. Segundo Hair et al., (2005), soluções hierárquicas são preferidas quando muitas ou todas as soluções alternativas devem ser examinadas, o tamanho da amostra é moderado (abaixo de 300 – 400, não excedendo 1000) ou uma amostra de um conjunto maior de dados é aceitável.

Em aplicações cujo o objetivo é criar um agrupamento por resumo, como para traçar perfis de clientes, os métodos particionais são os mais indicados, pois eles obtêm uma única partição do conjunto de dados, e devido a isso, possuem melhor desempenho do que os métodos hierárquicos em aplicações com grandes bases de dados. No entanto, Guha et al. (1998) afirmam que eles não têm um comportamento adequado quando a base de dados possui grupos de formatos e tamanhos muito diferentes pois pode dividir os grupos maiores na tentativa de minimizar a função objetiva. Além disso, esses métodos são eficazes quando o número de grupos existente pode ser estimado e são de formatos globulares.

Hair et al. (2005) afirmam que, há casos em que uma combinação de dois métodos, um hierárquico e um método não-hierárquico podem ser o mais adequado, visto que, um método hierárquico pode ser utilizado para selecionar o número de grupos que melhor represente o conjunto de dados e/ou para identificar os centros de grupos que servirão como sementes iniciais para inicializar o método não-hierárquico, e um método particional então é utilizado para agrupar todos os dados usando os pontos sementes para encontrar a partição do conjunto de dados mais precisa.

Nos estudos de caso apresentados nas Seções 3.1, 3.2 e 3.3, foram analisadas algumas características que motivaram a escolha das técnicas de agrupamento utilizadas.

3.1 Agrupamento de clientes de uma Agência de Viagens

Nesta seção, será discutido o trabalho realizado por Aguiar, Santana e Bastos (2018) no agrupamento de uma base de dados de Reserva de Hotéis, buscando identificar perfis de clientes de uma agência de viagens online.

Uma agência de viagens do Brasil, possui um site de reservas de hotéis por onde oferece hospedagens em hotéis de todo o Brasil. Para divulgar seus serviços, o departamento de marketing da agência envia ofertas por e-mail para seus clientes de forma arbitrária, ou seja, sem nenhum filtro. Essa abordagem acarreta um certo custo com os envios do e-mail marketing, uma vez que o valor do serviço é determinado pela quantidade de e-mails destinatários (AGUIAR; SANTANA; BASTOS, 2018).

Visando melhorar a eficiência do uso do dinheiro investido em e-mail marketing, foi pensado em realizar uma classificação dos usuários do site, com base em seu histórico de compras, com o objetivo de identificar perfis de clientes e, dessa maneira, enviar oferta de maneira mais eficaz, ou seja, apenas aos usuários com mais probabilidade de se interessar pelo anúncio (AGUIAR; SANTANA; BASTOS, 2018).

Segundo Aguiar, Santana e Bastos (2018), para o agrupamento dos clientes, foram selecionadas na base de dados 2.959 vendas realizadas entre os anos de 2016 e 2017 pelo site da agência. Cada cliente possui treze características, são elas: Estado onde o cliente reside, Idade, Total comprado, Valor Médio de compra do cliente, Quantidade de compras, Tipo de pagamento, Quantidade de parcelas, Cidade onde reside o cliente, Valor da venda específica, Hotel, Destino, Mês de estadia, Quantidade de diárias.

O intuito dessa aplicação é agrupar os clientes do site da agência de viagens conforme o comportamento de consumo de cada um. Pode-se ver o problema como o particionamento de um conjunto de clientes em grupos menores e homogêneos, com alto grau de semelhança

entre clientes de um mesmo grupo. Cada grupo corresponde a um perfil de cliente com comportamento de consumo semelhante. Nesse caso, é interessante que todos os clientes façam parte de um grupo, visto que, toda empresa deseja fidelizar clientes que já possui e conquistar novos, logo, as características de todos os clientes devem ser consideradas na hora de resumir as características dos elementos de cada grupo encontrado para que todos os clientes se encaixem em algum perfil.

Além disso, como os resultados do agrupamento serão usados para identificar o perfil dos clientes para enviar ofertas personalizadas por e-mail, uma grande quantidade de grupos (perfis) seria difícil de gerenciar pela equipe de marketing e despenderia mais esforço. O ideal é encontrar um número consideravelmente pequeno de grupos que seja suficiente para representar a base de dados analisada e gerar perfis consistentes, e como não se sabe o número exato de grupos, pode ser obtido através de execuções sucessivas de algum algoritmo particional para quantidades de grupos diferentes.

Como a taxonomia lógica dos grupos não é o propósito da aplicação e o número de grupos deve ser pequeno, podendo ser obtido através de algumas execuções de algum algoritmo particional, um algoritmo hierárquico não é indicado, pois embora seja vantajoso por não necessitar informar previamente o número de grupos e permitir escolher a quantidade que melhor satisfaça o problema através do dendrograma, o mesmo possui, como visto na Seção 2.4.1, um custo computacional elevado devido ao cálculo e armazenamento das matrizes de similaridades, não sendo aconselhável para grandes volumes de dados.

Diante das características do problema mencionadas, tamanho da base de dados, quantidade de grupos estimada, classificação de todos os elementos, interesse em uma única partição dos dados capaz de resumir a base de dados em uma quantidade suficientemente pequena de grupos, um algoritmo particional é mais apropriado. Nesse caso, a primeira escolha é o algoritmo K-means por ser o mais simples, e segundo Braga (2005) um dos mais recomendados para a identificação do perfil de clientes. O K-means particiona os dados em k grupos de tal forma que a similaridade intra-cluster (distância entre elementos de um mesmo grupo) seja alta e a similaridade inter-cluster (distância entre elemento de grupos distintos) seja baixa, e todos os pontos são agrupados. Embora, a quantidade de grupos não possa ser definida inicialmente nesta aplicação, o K-means pode ser aplicado várias vezes para valores de k diferentes, e posteriormente o usuário poderá escolher o k ideal.

O algoritmo aplicado por Aguiar, Santana e Bastos (2018) foi o Algoritmo K-means. Como a quantidade ideal de grupos não é conhecida e a qualidade do agrupamento depende das condições iniciais do K-means, o algoritmo foi aplicado várias vezes com o valor de k

variando de 2 a 10 grupos, realizando 30 execuções do algoritmo para cada valor de k. Posteriormente, foram analisadas as médias das distância inter-cluster e intra-cluster relativas às 30 execuções do algoritmo para cada k (Tabela 2).

Conforme Aguiar, Santana e Bastos (2018), a distância intra-cluster é a distância entre dois elementos pertencentes ao mesmo grupo, dessa forma, como a intenção é que elementos de um mesmo grupo sejam altamente semelhantes, o ideal para essa medida é um valor baixo, indicando nível alto de similaridade entre os clientes. Já a distância inter-cluster é a distância entre dois grupos, dessa forma, como o pretendido é encontrar clientes de grupos diferentes bem dissimilares, o valor ideal para essa medida é um valor alto, indicando baixa similaridade entre clientes de grupos diferentes ou grupos bem separados.

A Tabela 2 apresenta as médias de 30 execuções das distâncias inter-cluster e intra-cluster.

Tabela 2 - Resumo dos resultados da simulação do algoritmo

Quantidade de grupos	Distância inter-cluster	Distância intra-cluster
2	0,58	1595,98
3	2,14	1480,49
4	4,77	1407,52
5	8,49	1354,52
6	13,81	1306,17
7	19,76	1264,06
8	27,84	1233,99
9	36,94	1205,38
10	47,34	1183,41

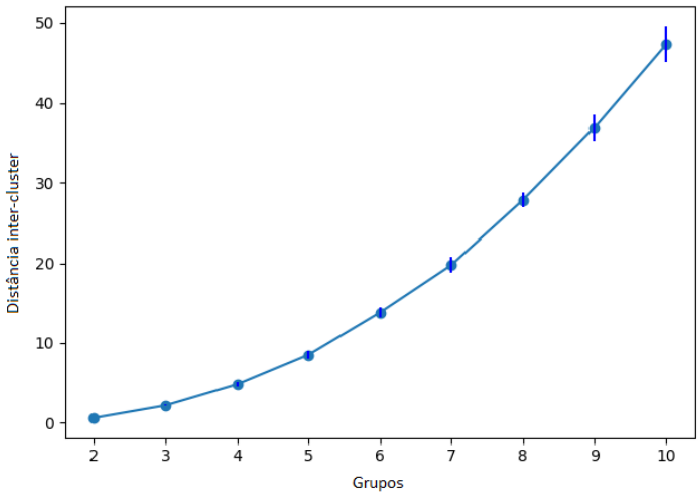
Fonte: Aguiar, Santana e Bastos, 2018.

Os resultados da execução do K-means, estão exibidos na Tabela 2. A partir dela, pode-se perceber que o algoritmo conseguiu cumprir o que se esperava dele, ao maximizar a distância inter-cluster e minimizar a distância intra-cluster, conforme o número de k aumenta (AGUIAR; SANTANA; BASTOS, 2018). Ou seja, no intervalo de k=2 a k=10, conforme o valor de k aumenta, os grupos vão se tornando mais separados uns dos outros e mais homogêneos internamente, de forma que, os clientes em um mesmo grupo são mais semelhantes entre si, e mais diferentes aos clientes de outros grupos.

Nas Figuras 17 e 18 podemos ver a variação das métricas das distâncias inter-cluster e

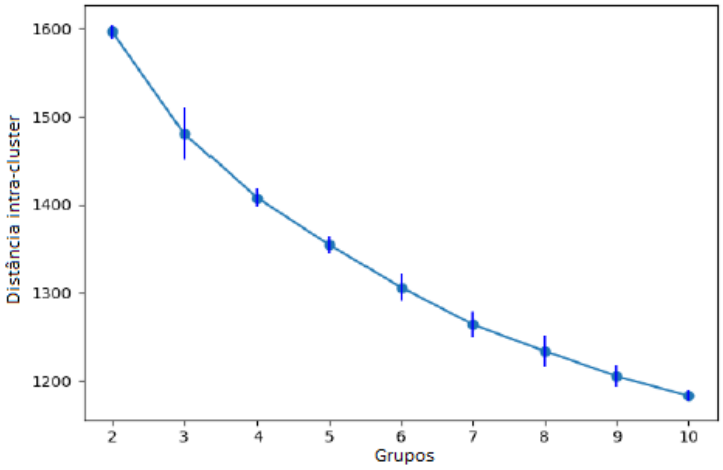
intra-cluster (ordenadas) em relação a quantidade de grupos (abscissas). No gráfico da Figura 17, a distância inter-cluster aumenta (distância entre os grupos) conforme o número de grupos aumenta. E no gráfico da Figura 18, a distância intra-cluster (distância entre cada elemento do grupo e o seu centroide) decai conforme o número de grupos aumenta.

Figura 17 - Variação da distância inter-cluster para valores crescentes de k



Fonte: Aguiar, Santana e Bastos, 2018.

Figura 18 - Variação da distância intra-cluster para valores crescentes de k



Fonte: Aguiar, Santana e Bastos, 2018.

Em relação ao número de grupos ideal, as métricas inter-cluster e intra-cluster apontam 10 como a quantidade ideal (AGUIAR; SANTANA; BASTOS, 2018). Porém, esse valor é muito alto para se trabalhar na prática, visto que, a equipe de marketing da empresa teria que lidar com 10 perfis diferentes de clientes, tendo que criar ofertas personalizadas para cada um

deles.

Dessa forma, pode-se indicar 3 como a quantidade ideal de grupos, uma vez que, as distâncias intra-cluster e inter-cluster têm a tendência de se otimizarem com a quantidade de k aumentando. A quantidade 3 foi apontada, também, pela equipe de negócios do site de viagens, uma vez que, a granularização de 10 grupos não seria necessária para o direcionamento das ações de e-mail marketing (AGUIAR; SANTANA; BASTOS, 2018).

Após escolhida a quantidade ideal de grupos como 3, ao analisarem uma das partições obtidas para identificar as características em comum entre os clientes de cada grupo, constataram que o K-means criou os grupos considerando principalmente a variável “Mês de Estadia”, sendo um grupo de usuários que se hospedam mais nos primeiros meses do ano, outro grupo de usuários que se hospedam mais próximo ao meio do ano e o último mais próximo do final do ano (AGUIAR; SANTANA; BASTOS, 2018). A forma que essa análise foi feita, não foi especificada pelos autores.

Diante do que foi mencionado, podemos concluir que o algoritmo K-means obteve sucesso em particionar os dados, pois através da partição criada foi possível extrair conhecimento útil para a empresa.

3.2 Agrupamento de Clientes Johnson

Nesta seção, ser o trabalho realizado por Pacheco, Capella e Goldshmidt (2010) no agrupamento de clientes da empresa Johnson, especializada no fornecimento de materiais de limpeza, utilizando Algoritmos Genéticos.

A Johnson é uma empresa tradicional no mercado de fornecimento de materiais de limpeza. Ela fornece produtos de limpeza para clientes de diversos segmentos, tais como: hospitais, empresas limpadoras, farmácias, hotéis, restaurantes, dentre outros. Em seu banco de dados, a Johnson mantém um histórico de todas as vendas realizadas no ano fiscal vigente (últimos 11 meses) (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Essa aplicação teve início a partir do desejo da Johnson de criar um conceito de Perfil de Cliente para servir de base para futuros trabalhos de marketing planejados pela empresa, explorando a potencialidade de compra de seus clientes. Através desse Perfil de Cliente, pretende-se adotar estratégias de vendas direcionadas segundo cada grupo (perfil) de clientes. E esse trabalho representa a primeira etapa de um projeto piloto corporativo em desenvolvimento pela Johnson com a finalidade de melhorar seus serviços e alavancar suas vendas (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

A base de dados selecionada para agrupamento, por orientação da Direção do Grupo

Johnson no Brasil, é referente aos clientes do segmento "*Floor Care*" (Tratamento de Piso), contendo 119 clientes ativos, ou seja, que têm operado efetivamente junto à Johnson ao longo do ano fiscal vigente (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

O segmento *Floor Care* é composto por cinco Sistemas de Limpeza, cada um agrupa produtos com aplicações complementares (base seladora e acabamento) que possuem preços e qualidade compatíveis (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Dessa forma, percebe-se que o que essa aplicação buscou fazer, foi resumir a base de dados de clientes, particionando os dados em uma quantidade razoável e consideravelmente pequena de grupos (perfis) que possam representar de forma eficiente todos os clientes da base de acordo com o consumo dos produtos do segmento *Floor Care*. Assim como na aplicação anterior (Seção 3.1), deverá ser feita uma análise da base de dados, separando clientes que possuem comportamento de consumo semelhantes. E, como a linha de produtos *Floor Care* possui cinco Sistemas de Limpeza, somos levados a estimar que a quantidade de grupos ideal não deve ser maior que cinco.

Como a quantidade de grupos pode ser estimada e obter a estrutura hierárquica dos grupos não é a finalidade dessa aplicação, os métodos hierárquicos não são indicados, além de serem menos aconselháveis para grandes conjuntos de dados devido ao seu alto custo computacional.

Além disso, como a base de dados é referente a um conjunto de clientes de uma mesma empresa, classificados de acordo com o comportamento de consumo de produtos do mesmo segmento, contendo apenas cinco sistemas de limpeza com a mesma finalidade variando apenas em qualidade, então provavelmente esses clientes não terão um comportamento muito distinto tendendo a formar grupos globulares, indicando o uso de métodos particionais.

Para essa aplicação o ideal é que todos os clientes sejam classificados, visto que as informações de consumo de cada cliente são de grande importância para elaboração do conceito de Perfil de Cliente que se deseja criar, de forma que todos os clientes se encaixem em algum perfil, para apoio a adoção de políticas de vendas direcionadas a cada grupo.

Diante das características de agrupamento particional mencionadas, quantidade de grupos estimada, grupos de formatos globulares, o algoritmo K-means, por ser o mais simples que satisfaz essas características e mais indicado para esse tipo de tarefa, é o primeiro a ser considerado. No entanto, conforme vimos na Seção 2.5.1, o K-means pode apresentar alguns problemas quando aplicado em bases de dados mais complexas, como, sua tendência em convergir para ótimos locais e o fato da qualidade do agrupamento depender significativamente da escolha dos centroides inicias, sendo necessário ser executado várias

vezes a fim de encontrar a melhor partição dos dados. Uma alternativa ao K-means que também satisfaz as características do problema, é o Algoritmo Genético, que dentre outras vantagens mencionadas na Seção 2.5.2, utiliza uma abordagem evolutiva, selecionando sempre as melhores soluções para gerar novas, através de operadores genéticos, e ainda trabalha com população de indivíduos (conjunto de soluções), ou seja, explora muito mais o espaço de busca, tendo então, mais chances de alcançar a solução ótima, obtendo, geralmente, melhores resultados do que o K-means.

Dessa forma, a técnica de agrupamento escolhida por Pacheco, Capella e Goldshmidt (2010) foi Algoritmo Genético para agrupamento dos clientes em 4 grupos, utilizando como função de avaliação a Soma do Erro Quadrado apresentada na Seção 2.4.1 (Equação 3). O método de seleção utilizado foi o da roleta, e o cruzamento foi feito usando o método de ponto único apresentado na Seção 2.5.2, com uma taxa de 65% e probabilidade de mutação de 8%. O tamanho da população foi definido em 200 indivíduos e como critério de parada foi definido 30.000 indivíduos processados.

O objetivo do modelo proposto consistiu em minimizar a função de avaliação, identificando o cromossomo que representasse uma distribuição de clientes em grupos, onde a similaridade de cada cliente em relação ao centroide de seu grupo fosse a maior encontrada (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Os dados, extraídos da base de dados operacional da Johnson, foram convertidos para uma estrutura que consolida as informações de cada cliente em uma linha (registro). As informações disponíveis de cada cliente são: Cliente (identificação do cliente), Faturamento, Tempo de Operação junto a Johnson, Metragem (área de exposição do produto), Quantidade de Produtos Comprados por Sistema de Limpeza e Frequência de Compra por Sistema de Limpeza. Cada uma dessas informações está expressa de forma percentual proporcional ao cliente que possua o maior valor para aquele atributo, exceto o atributo Cliente (identificação do cliente). Por exemplo, o Faturamento de um determinado cliente é a proporção do faturamento desse cliente em relação ao maior faturamento encontrado no conjunto de clientes da base em análise. E como o segmento *Floor Care* possui cinco Sistemas de Limpeza, consequentemente, os atributos Quantidade de Produtos Comprados por Sistema de Limpeza e a Frequência de Compra por Sistema de Limpeza contém as informações para cada um dos cinco sistemas de limpeza, formando no total 13 características para cada cliente (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Além disso, os Sistemas de Limpeza foram concebidos de forma a agrupar produtos em ordem decrescente de qualidade e preço. Assim, o Sistema 1 foi composto por produtos de

melhor qualidade, e consequentemente, maior preço, enquanto o Sistema 5 englobou produtos de qualidade e preços inferiores (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Na Tabela 3 temos o resumo de como ficou a distribuição dos clientes nos 4 grupos.

Tabela 3 - Distribuição dos clientes no melhor cromossomo

Grupo	Total de Clientes
1	28
2	28
3	58
4	5

Fonte: Pacheco, Capella e Goldshmidt, 2010.

A soma do quadrado da distância entre cada cliente e o centroide do *cluster* onde o cliente tenha sido enquadrado foi de 1.67 no melhor indivíduo encontrado (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

A Tabela 4 mostra um resumo das características em comum entre clientes de cada grupo.

Tabela 4 - Resumo das características dos clientes por grupo

Grupo	Faturamento	Tempo de Operação	Área de exposição	Sist. Mais Comprado
1	Até 30%	De 20 a 60%	De 20 a 60%	5
2	De 30 a 100%	De 60 a 100%	De 60 a 100%	4
3	Até 70%	De 20 a 100%	Até 60%	5
4	Até 30%	De 60 a 100%	Até 20%	5

Fonte: Pacheco, Capella e Goldshmidt, 2010.

Pacheco, Capella e Goldshmidt (2010) explicam que, apesar de todos os clientes terem sido agrupados, há 18 casos de clientes em que suas características não se enquadram no perfil da maioria dos clientes do seu grupo. Esses casos foram enviados para o departamento de marketing da empresa Johnson para serem avaliados e, se possível, reatribuídos ao grupo que melhor se encaixe.

Além disso, fica claro que nenhum dos grupos produzidos concentrou clientes que

comprassem apenas os Sistemas de Limpeza 1 e 2, que são os Sistemas de maior qualidade e mais caros. Isso chama a atenção para necessidade de um maior esforço da equipe de marketing na divulgação e incentivo dos clientes para a compra de tais produtos (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Considerando o perfil geral dos clientes enquadrados no grupo 2, concentram-se os clientes com alto faturamento, alto tempo de operação junto à empresa e uma grande área de exposição. Diante disso, podemos perceber que os clientes nesse perfil possuem um potencial natural para compras dos produtos de melhor qualidade, (PACHECO; CAPELLA; GOLDSHMIDT, 2010). Ou seja, produtos dos sistemas de limpeza 1 e 2 do segmento *Floor Care*.

O perfil geral dos clientes enquadrados no Grupo 3 levou a uma política de vendas cuja principal argumentação se baseia no alto tempo de negociação da Johnson junto aos referidos clientes e na confiabilidade dos serviços e produtos oferecidos ao longo do tempo. Tal histórico deverá ser utilizado como uma forma de incentivar os clientes na experimentação de Sistemas de Limpeza de qualidade 1 ou 2. Os clientes que já utilizam tais Sistemas deverão ser trabalhados no sentido de preservar seu interesse pelos produtos oferecidos (PACHECO; CAPELLA; GOLDSHMIDT, 2010).

Pacheco, Capella e Goldshmidt (2010) afirmam que através dos resultados alcançados por essa aplicação, o departamento de Marketing da Johnson foi capaz de estudar a classificação dos clientes obtida e definir políticas de venda específicas para cada grupo de clientes, além de identificar um grupo com forte potencial de venda (clientes do grupo 2), que corresponde aos clientes com faturamento de 30% a 100% e área de exposição de 60% a 100%, em relação ao maior faturamento e maior área de exposição presentes no rol de clientes analisados.

Podemos concluir então, que o Algoritmo Genético obteve sucesso em agrupar os dados, visto que, conhecimento útil foi extraído da partição criada.

3.3 Agrupamento de produtores de leite

Nesta seção apresentaremos uma comparação entre os resultados obtidos pelos métodos Ward e K-means no trabalho realizado por Seidel et al. (2008) no agrupamento de produtores de leite da Região de Santa Maria, RS, buscando definir os grupos de fornecedores conforme as características encontradas nas amostras de leite produzidas por eles.

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

apud SEIDEL et al., 2008). 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 características dos lotes recebidos, é necessário analisar e conhecer o tipo de leite que cada fornecedor dispõe (SEIDEL et al., 2008). Diante disso, conforme explica Seidel et al. (2008), uma indústria de laticínios localizada na cidade de Santa Maria, RS, deseja classificar os seus fornecedores de leite de acordo com as características do leite fornecido por eles.

O intuito dessa aplicação é agrupar os fornecedores de leite da indústria de laticínios conforme as características dos leites produzidos por cada um. Para isso, foram selecionadas 231 amostras de leite *in natura* (sem processamento industrial) de 63 produtores, coletadas no período de 6 a 30 de setembro de 2004. As variáveis analisadas foram: Porcentagem de água, Porcentagem de gordura, Acidez em graus Dornic e Densidade em g/cm^3 (SEIDEL et al., 2008).

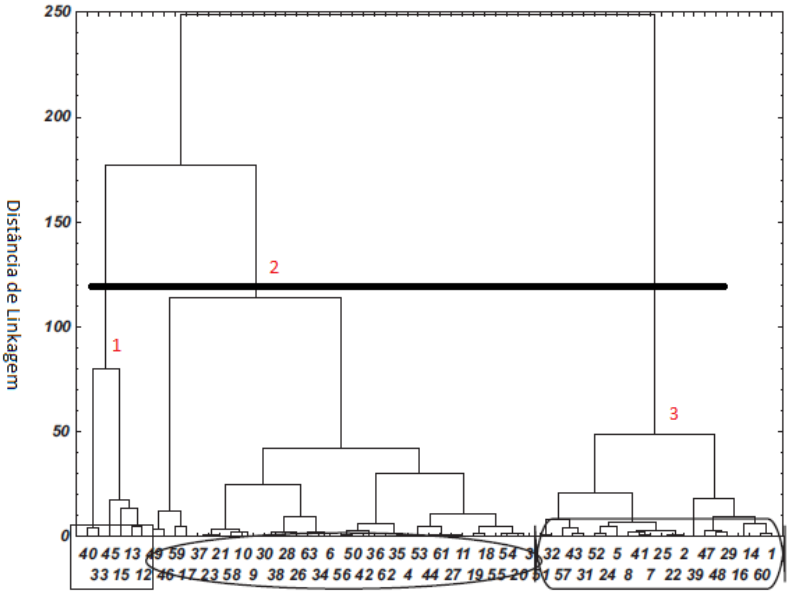
Diante disso, percebemos que o que essa aplicação buscou fazer, foi particionar o conjunto de dados em grupos menores de fornecedores, onde cada grupo deverá conter fornecedores cujo os leites produzidos por eles possuam características semelhantes. Podemos perceber o interesse em somente uma partição do conjunto de dados, indicando o uso de algum método paticional. No entanto, não se tem dados que indiquem a quantidade de grupos que melhor satisfaça o problema. E, como a base de dados pode ser considerada pequena, com apenas 63 elementos (fornecedores de leite) compostos por 4 características cada um, podemos aplicar um método hierárquico, analisar os agrupamentos formados através do dendrograma e definir o número de grupos que melhor represente o conjunto de dados. Ou, podemos ainda, aplicar um método hierárquico para definir a quantidade de grupos e, posteriormente aplicar um método particional com a quantidade de grupos encontrada pelo método hierárquico, para obter a partição mais precisa do conjunto de dados.

Dessa forma, para obter a melhor maneira de formar os grupos, Seidel et al., (2008) optaram por aplicar dois métodos de agrupamento, um hierárquico aglomerativo e um particional, e posteriormente comparar os resultados. Os métodos escolhidos por Seidel et al., (2008) foram o método de Ward, para o agrupamento hierárquico, e o K-means para o agrupamento particional. Segundo Seidel et al. (2008), esses métodos foram escolhidos por serem os mais usados e apresentar bons resultados.

Para tratar os dados de um mesmo produtor, foi utilizada a mediana dos dados de cada variável de suas amostras de leite, a fim de obter um valor que representasse cada produtor (SEIDEL et al., 2008).

A Figura 19 apresenta o dendrograma dos três grupos formados baseado no corte feito na distância entre grupos de aproximadamente 120. A distância é medida ao longo do eixo vertical (ordenadas), e os diferentes produtores ao longo do eixo horizontal (abscissas).

Figura 19 – Dendrograma dos resultados encontrados pelo método de Ward



Fonte: Seidel et al., 2008.

Verificamos a divisão dos fornecedores em 3 grupos. No primeiro grupo formado, podemos identificar os fornecedores 12, 13, 15, 33, 40 e 45. 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. E o terceiro grupo 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 (SEIDEL et al., 2008).

Observando o dendrograma da Figura 19, quando o número de grupos muda de 3 para 2 (distância de aproximadamente 200), percebemos um grande aumento na distância entre os grupos, indicando que ao unir esses elementos formaram-se grupos bem menos homogêneos do que no passo anterior. No entanto, se cortássemos o dendrograma mais abaixo (distância de aproximadamente 80), quando o número de grupos é igual a 4, a similaridade entre elementos de um mesmo grupo seria maior, mas haveria mais grupos finais e alguns grupos seriam muito pequenos. Diante disso, um corte quando o número de grupos é igual a 3 parece o mais adequado para essa aplicação.

A Tabela 5 apresenta a análise descritiva dos grupos encontrados pelo método de Ward, que foi feita calculando a média das características dos elementos em cada grupo e o desvio

padrão.

O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados. Quanto mais próximo de 0 for o desvio padrão, mais homogêneos são os dados. (GOUVEIA, 2019).

Tabela 5 - Análise descritiva dos grupos encontrados pelo método de Ward

Variáveis	**Grupo 1**		**Grupo 2**		**Grupo 3**	
	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

Fonte: Seidel et al., 2008.

Analisando a Tabela 5, percebemos que no grupo 1 concentram-se os produtores cujo o leite possui baixas taxas de densidade (em média 1025,73 g/cm^3) e acidez (em média 15,02°D). O grupo 2, se caracteriza por altas taxas de água excedente (em média 4,66%) e baixo teor de gordura (em média 3,88%). No grupo 2, os fornecedores 17, 46, 49, 59 e 63 tiveram as maiores taxas de água excedente, em média 7,25%. No grupo 3, concentram-se os produtores de leite com alto teor de acidez (em média 16,52 °D) e densidade elevada (em média1030, 48 g/cm^3) (SEIDEL et al., 2008).

No grupo 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 (SEIDEL et al., 2008).

Após a análise dos resultados obtidos pelo método de Ward, Seidel et al. (2008) decidiram aplicar o algoritmo K-means para criar grupos de produtores de leite usando as mesmas variáveis consideradas no processo de agrupamento hierárquico, com o intuito de verificar os grupos formados pelo método de Ward. Diante do número de grupos estabelecidos definidos a partir do dendrograma, Seildel et al. (2008) optaram por escolher 3 como quantidade de grupos para o agrupamento particional.

A Figura 20 ilustra como os produtores de leite foram distribuídos nos 3 grupos pelo Algoritmo K-means.

Figura 20 - Distribuição dos produtores pelo Algoritmo K-means

Grupo 1	Grupo 2		Grupo 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	

Fonte: Seidel at al., 2008.

Na Figura 20, pode-se perceber que os produtores de leite foram distribuídos nos 3 grupos de forma mais uniforme pelo K-means. A diferença na quantidade de elementos em cada grupo é menor em comparação com os grupos formados pelo método de Ward.

A Tabela 6, apresenta a análise descritiva dos produtores de leite em cada um dos 3 grupos.

Tabela 6 - Análise descritiva dos grupos encontrados pelo K-means

Variáveis	Grupo 1		Grupo 2		Grupo 3	
	Média	Desvio 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

Fonte: Seidel et al., 2008.

A partir da análise descritiva apresentada na Tabela 6, perce-se que, no grupo 1 concentram-se os produtores cujo o leite possui baixa densidade (em média 1027,23 g/cm^3) e acidez (em média 15,12°D). O grupo 2 concentra produtores cujo o leite possui alta porcentagem de água excedente (em média 5,25%) e baixa porcentagem de gordura (em média 3,70%). E finalmente, o grupo 3 concentra produtores cujo o leite possui com altas taxas de acidez (em média 16,59°D) e densidade (1030,43 g/cm^3) (SEIDEL et al., 2008).

A Tabela 7 apresenta as frequências de produtores classificados pelos dois métodos, para uma comparação dos resultados obtidos.

Tabela 7 - Classificação de produtores pelos métodos Ward e K-means

Ward	K-means			Total
	Grupo 1	Grupo 2	Grupo 3	
Grupo 1	6	0	0	6
Grupo 2	10	24	1	35
Grupo 3	0	0	22	22
Total	16	24	23	63

Fonte: Seidel et al., 2008.

A partir da Tabela 7, podemos perceber que os resultados encontrados pelos métodos, Ward e K-means, tiveram alta concordância, pois dos 63 produtores, 52 (82,54%) foram classificados nos mesmos grupos pelos dois métodos, evidenciando a eficiência e robustez dos agrupamentos formados por ambos (SEIDEL et al., 2008).

Fazendo uma comparação entre os resultados obtidos pelos dois métodos, dispostos nas Tabelas 5 e 6, verifica-se que o K-means representou melhor o grupo 2 (baixas porcentagens

de gordura e altas porcentagens de água excedente) e o grupo 3 (altas taxas de acidez e de densidade no leite) (SEIDEL et al., 2008). O grupo 2 criado pelo K-means possui 11 produtores a menos do que o grupo 2 criado pelo método de Ward, ainda assim a porcentagem de água apresentou um aumento em relação ao grupo 2 no método de Ward, e a porcentagem de gordura apresentou uma redução. O grupo 3 criado pelo K-means possui um produtor a mais que estava classificado no grupo 2 pelo método de Ward, no entanto, a taxa de acidez que já era a maior em relação aos três grupos no método de Ward, apresentou um pequeno aumento, indicando que esse produtor que foi classificado no grupo 2 pelo método de Ward, se enquadra melhor no grupo 3.

Diante disso, percebe-se que o K-means classificou os produtores de leite de forma mais adequada dentro dos grupos (SEIDEL et al., 2008), em conformidade com as características de cada um. Dessa forma, os grupos criados pelo K-means possuem elementos mais semelhantes internamente. Isso acontece porque o K-means utiliza uma abordagem de relocação iterativa, que muda os elementos de grupo durante a execução do algoritmo na tentativa de minimizar a função objetiva globalmente, ao contrário do método de Ward, no qual as decisões de fusão de grupos são finais, de forma que, uma vez que dois grupos são fundidos, essa operação não pode ser desfeita. Devido a isso, pode acontecer de um elemento em um grupo estar mais próximo do centroide de algum outro grupo do que do centroide do seu grupo, ou seja, é mais semelhante aos elementos de outro grupo, podendo formar grupos menos homogêneos que o K-means, como aconteceu nessa aplicação.

Diante do que foi mencionado, Seidel et al., (2008) concluíram que a utilização dos dois métodos conjuntamente, isto é, método de Ward e K-means, é o mais indicado para essa aplicação, já que, usando o método de Ward foi possível definir com mais eficiência o número de grupos que devem ser formados, e aplicar no K-means, que por sua vez, obteve grupos bem homogêneos internamente.

4 CONCLUSÃO

Neste trabalho foi realizado um estudo sobre a tarefa de agrupamento de dados e da aplicação dessa técnica em bases de dados corporativas visando a descoberta de conhecimento para apoio a estratégias de negócios. Descrevemos o que é agrupamento de dados, o processo de agrupamento, três técnicas de agrupamento, o agrupamento hierárquico aglomerativo, o K-means e Algoritmos Genéticos, e apresentamos três estudos de caso no contexto de negócios.

A tarefa de agrupamento de dados é uma ferramenta poderosa de descoberta de conhecimentos, visto que, é capaz de descobrir no conjunto de dados grupos semelhantes de objetos, usando para isso somente informações contidas nos próprios dados, sem qualquer conhecimento prévio ou com poucos conhecimentos sobre o conjunto de dados. Vimos que, através dessa, podemos identificar de forma eficiente perfis distintos de clientes e classificá-los de acordo com seu padrão de consumo, e ainda categorizar fornecedores de acordo com as características dos produtos fornecidos por eles, dentre outras aplicações.

Para entender o funcionamento das três técnicas utilizadas nos estudos de caso do Capitulo 3, apresentamos alguns conceitos de tipos de grupos. Entre eles, os grupos baseados em protótipos na Seção 2.3.2, em que um grupo é representado por um protótipo, onde os elementos de um grupo são mais próximos do protótipo do seu grupo do que do protótipo de outros grupos.

Para entender como pode ser realizado o agrupamento de dados, descrevemos três técnicas de agrupamento, o agrupamento hierárquico aglomerativo, o Algoritmo K-means e o Algoritmo Genético. Destacamos suas vantagens, suas desvantagens, o uso mais adequado para os mesmos, e conhecimento prévio exigido para utilizar os métodos.

O agrupamento hierárquico aglomerativo, refere-se a um conjunto de técnicas caracterizadas por produzir uma composição hierárquica dos grupos representada por uma estrutura de árvores, através da qual, é possível definir experimentalmente a quantidade de grupos desejada fazendo a análise exploratória dos resultados.

O K-means e Algoritmo Genético, são técnicas semelhantes por particionar o conjunto de dados em um número k de subconjuntos exclusivos especificado pelo usuário, e cada configuração obtida é avaliada por uma função objetiva. E ao contrário das técnicas hierárquicas, os grupos podem ser melhorados gradativamente.

Apesar das características em comum entre o K-means e o Algoritmo Genético, algumas particularidades podem ser observadas. O K-means, se destaca em agrupamento de dados devido a sua eficiência e simplicidade em obter grupos compactos e bem separados. O

Algoritmo Genético, por sua vez, se destaca pela abordagem evolutiva que utiliza, realizando buscas simultâneas em várias regiões do espaço de busca ao trabalhar com população, tendo então mais chances de chegar em regiões mais promissoras do espaço de busca evitando, assim, a convergência do algoritmo para um ótimo local.

Foram apresentados no capítulo 3, três estudos de caso, a fim de ilustrar o contexto de utilização de cada técnica, evidenciando fatores significativos que levaram a escolha do K-means, Algoritmo Genético ou método hierárquico aglomerativo.

A primeira aplicação (Seção 3.1), teve como objetivo agrupar os usuários do site de uma agência de viagens online com base em seu histórico de compras, e entregar ao usuário da aplicação uma partição do conjunto de dados de forma que se possa traçar o perfil dos usuários resumindo as características em comum entre os elementos de cada grupo. Para esse caso, o k-means foi escolhido como técnica mais adequada, conforme as características do problema, como, número de grupos estimado, classificação de todos os elementos, além de ser a técnica mais simples e indicada para esse tipo de aplicação.

A segunda aplicação, apresentada na Seção 3.2, a mineração de dados teve como objetivo agrupar clientes de uma empresa especializada no fornecimento de produtos de limpeza com base no consumo de cada cliente dos produtos do segmento *Floor Care* (Tratamento de Piso). Para esse caso, a técnica escolhida para o agrupamento dos dados foi o Algoritmo Genético. A partir das características particionais do problema, o número de grupos estimado, grupos de formatos globulares, classificação de todos os pontos, percebemos que o K-means poderia ser aplicado a esse problema, contudo, a possibilidade de obter resultados de melhor qualidade utilizando o Algoritmo Genético, estimulou o uso dessa técnica.

Na terceira aplicação, apresentada na Seção 3.3, a mineração de dados teve como objetivo classificar produtores de leite de uma indústria de laticínios localizada em Santa Maria, RS, de acordo com as características das amostras de leite fornecidas por eles. Para esse caso, foi feita uma comparação entre duas técnicas de agrupamento, a técnica de agrupamento aglomerativo de Ward e o algoritmo K-means. Diante das informações disponíveis, tamanho da base de dados e nenhuma estimativa do número de grupos adequado para a classificação dos produtores de leite, percebemos que o mais indicado seria a aplicação de um método hierárquico aglomerativo para identificarmos experimentalmente o número de grupos e, posteriormente, aplicar um método particional para verificar os grupos encontrados pelo método hierárquico. E por fim, concluiu-se que o uso das duas técnicas em conjunto, Método de Ward e K-means, era o mais indicado para essa aplicação.

As três aplicações apresentadas nesse trabalho contribuem para gerar conhecimento para a aplicação prática de técnicas de agrupamento de dados para solução de problemas específicos no contexto de negócios. Podemos perceber a importância de analisar cuidadosamente os objetivos e os dados disponíveis para a escolha da técnica mais apropriada para cada aplicação.

Quanto as limitações desse trabalho, embora, tenhamos realizado o estudo de apenas três técnicas, o agrupamento hierárquico aglomerativo, K-means e Algoritmo Genético, podem ser encontradas várias outras na literatura. Além disso, os estudos de caso apresentados estão limitados ao mesmo contexto de aplicação, e possuem objetivos parecidos, no entanto, a tarefa de agrupamento de dados é utilizada no desenvolvimento de aplicações em diversas áreas como segmentação de imagens em visão computacional e análise de textos. E por fim, a falta de descrições mais detalhadas sobre as aplicações nas fontes consultadas, como ferramentas utilizadas, ou detalhes de como foi feita a análise dos resultados.

Como possibilidade de novos estudos, seria interessante abordar outras técnicas de agrupamento, por exemplo, o algoritmo Fuzzy C-means que utiliza o conceito de que determinada amostra pode pertencer a mais de um grupo de acordo com um grau de pertinência. Além disso, poderíamos estudar variações do Algoritmo Genético de Agrupamento, como a apresentada em Carlantonio (2001), que descreve uma abordagem capaz de encontrar o número de grupos e a partição correta para o conjunto de dados, sem a necessidade de informar previamente qualquer parâmetro de entrada.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; GEHRKE, J.; GUNOPULOS, D.; RAGHAVAN, P. **Automatic Subspace Clustering on High Dimensional Data for Data Mining Applications**. In: Proceedings of the ACM SIGMOD Conference on Management of Data , p. 94-105, Seattle, Washington, USA, June. 1998.

AGUIAR, P.; SANTANA JÚNIOR, C. J.; BASTOS FILHO, C. J. A. B. **Aplicação de Algoritmos de Clusterização em uma Base de Dados de Reservas de Hotéis.** In: Revista de Engenharia e Pesquisa Aplicada. 2018.

ALMEIDA, J. D. S. **Computação Evolucionária.** 2015. 40 slides.

AMARAL, F. **Aprenda Mineração de Dados: Teoria e Prática**. 1 Edição. Rio de Janeiro: Alta Books Editora, 2016.

ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H.-P.; SANDER J. **OPTICS: Ordering Points to Identify the Clustering Structure**. In: Proceedings of the ACM SIGMOD Conference on Management of Data, p. 49-60, Philadelphia, PA, USA, June. 1999.

BUSSAB, W. O.; MIAZAKI, É. S; ANDRADE, D. F. de. **Introdução à Análise de Agrupamentos**. 9o Simpósio Nacional de Probabilidade e Estatística. São Paulo:ABE, 1990.

BRAGA, L. P. V. B. **Introdução à Mineração de Dados.** 2 Edição Revista e Ampliada. Rio de Janeiro: E-Papers Serviços Editoriais, 2005.

CARLANTONIO, L. M. **NOVAS METODOLOGIAS PARA CLUSTERIZAÇÃO DE DADOS**. 2001. Tese (Mestrado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001. Disponível em:

http://wwwp.coc.ufrj.br/teses/mestrado/inter/
2002/teses/%20CARLANTONIO_LM_02_t_M_int.pdf. Acesso em: 4 abr. 2019.

CARVALHO, L. A. V. C. **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. 8. ed. São Paulo: Érica, 2001.

CASTRO, M. A. T. A. **Agrupamento "Clustering"**, Instituto Superior de Engenharia do Porto, 2013.

DONI, M. V. **ANÁLISE DE CLUSTER: MÉTODOS HIERÁRQUICOS E DE PARTICIONAMENTO**. 2004. Monografia (Bacharelado em Sistemas de Informação) - Faculdade de Computação e Informática da Universidade Presbiteriana Mackenzie, São Paulo, 2004. Disponível em: http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF. Acesso em: 27 mar. 2019.

DUNHAM, M. H. **Data Mining: Introdutory and Advanced Topics**. 3. ed. New Jersey: Prentice Hall, 2002.

FUNG, G. **A Comprehensive Overview of Basic Clustering Algorithms**, 2001. Disponível em: <file:///C:/Users/Asus/Desktop/clustering%20(1).pdf>. Acesso em: 16 de jul. 2019.

GOLDSCHMIDT, R.; PASSOS, E**. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações**. Rio de Janeiro: Campus, v. 1, 2015.

GUHA, S.; RASTOGI, R.; SHIM, K. **CURE: An Efficient Clustering Algorithm for Large Databases**. In: Proceedings of the ACM SIGMOD Conference on Management of Data , pp. 73-84, Seattle, Washington, USA, June. 1998.

HAIR, J. F.; BLACK W. C.; BABIN B. J.; ANDERSON R. E.; TATHAN R. L. **Análise multivariada de dados**. Trad. Adonai S. Sant'Anna e Anselmo C. Neto. 5 ed. Porto Alegre: Bookman, 2005.

HAN, J; KAMBER, M.; PEI, J. **Data Mining: concepts and techniques**. 3 ed. Wyman

Street: Elsevier, 2012.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Prentice-Hall, Inc., 1988.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis.** New York: Wiley, 2009.

LACERDA, E., G., M., CARVALHO, A., C., P., L., F. de**. Introdução aos Algoritmos Genéticos**. In: Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação, Vol. II, pp. 51-126, Rio de Janeiro, Brasil, Julho. 1999.

MACEDO, D. C. de. et al. **Reclamações de clientes como fonte de inovações a partir de uma base de Help Desk utilizando Data Mining – Um exemplo de aplicação**. In: Biblioteca Digital de Periódicos. 2013.

MENDES, J. C. **Agrupamento deDadosesuasAplicações**. 2017. Trabalho de conclusão de curso (Bacharelado em Ciência da Computação) - Universidade Federal do Maranhão, São luís, 2017.

MONTENEGRO, F. M. T.; BRITO, J. A. M. **UM ALGORITMO GENÉTICO PARA O PROBLEMA DE AGRUPAMENTO.** In: XXXVIII SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL. 2006.

MOYA, R. **Seleção do Número Ideal de Clusters**. Disponível em: <https://jarroba.com/seleccion-del-numero-optimo-clusters/>. Acesso em: 14 junho de 2019.

NIEVOLA, J. C. **Análise de Agrupamentos**. 99 slides. Disponível em:<http://www.ppgia.pucpr.br/~fabricio/ftp/Aulas/Mestrado/IA/Nievola/MD/MD-06-Agrupamento.pdf>. Acesso em: 3 de abril de 2019.

NOLETO, L. F. **Métodos De Segmentação De Mapas Auto-Organizáveis Para Análise De Agrupamento.** Monografia (Bacharelado em Ciência da Computação) – UFSC. Santa Catarina, 2007.

OCHI, L. S.; DIAS, C. R.; SOARES, S. S. F. **Clusterização em Mineração de Dados.** 2004.

PACHECO, M.; CAPPELA, A.; GOLDSCHMIDT, R. **Clusterização de Clientes Johnson utilizando Algoritmos Genéticos.** In: Inteligência Computacional Aplicada. 2010.

SEIDEL, E. J. et al. **Comparação Entre o Método Ward e o Método k-médias no Agrupamento de Produtores de Leite.** Ciência e Natura, [S.l.], p. 07-15, june 2008. ISSN 2179-460X. Disponível em: <https://periodicos.ufsm.br/cienciaenatura/article/view/9737/5830>. Acesso em: 2 de junho de 2019.

SILVA, L. A.; SARAJANE, M. P.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**: 1. ed. Rio de Janeiro: Elsevier, 2016.

SOUZA, S. A. **Algoritmos Genéticos Aplicados à Proteção e Estimação de Harmônicos em Sistemas Elétricos de Potência**, Tese apresentada a Escola de Engenharia de São Carlos, 2008.

TAN, Pang-Ning. **Introdution to Data Mining**. 2006.

TEIXEIRA, O. **Proposta de um novo Algoritmo Genético baseado na Teoria dos Jogos.** Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal do Pará. 2005.

VALE, M. N. **Agrupamentos de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos**. 2005. Dissertação (Mestrado em Engenharia Elétrica) - PUC-Rio, Rio de Janeiro, 2005. Disponível em: https://www.maxwell.vrac.pucrio.br/Busca_etds.ph pstrSecao=resultado&nrSeq=7975@1. Acesso em: 27 mar. 2019.

XU, R.; WUNSCH, D. **Clustering.** John Wiley & Sons, 2009.

WELGE, M. E.; Shaw, M. J.; Subramaniam, C.; Tan, G. W. **Knowledge management e data mining for marketing.** In: Decision Support Systems, vol. 31 n 1 p127-137, 2001.