Jakelson Carreiro Mendes

Agrupamento de Dados e suas Aplicações

São Luís

2017

Jakelson Carreiro Mendes

Agrupamento de Dados e suas Aplicações

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, para aprovação no componente curricular Monografia II.

Universidade Federal do Maranhão – UFMA

Curso de Ciência da Computação

Orientador: Prof. Dr. Ivo José da Cunha Serra

São Luís

2017

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

Mendes, Jakelson Carreiro.
Agrupamento de Dados e suas Aplicações / Jakelson Carreiro Mendes. – 2017.
52 p.

Orientador(a): Ivo José da Cunha Serra.
Monografia (Graduação) – Curso de Ciência da Computação, Universidade Federal do Maranhão, Universidade Federal do Maranhão, São Luís, 2017.

1. Agrupamento de Dados. 2. Aplicações de Agrupamento. 3. Mineração de Dados. 4. Técnicas de Agrupamento. I. Serra, Ivo José da Cunha. II. Título.

Jakelson Carreiro Mendes

Agrupamento de Dados e suas Aplicações

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, para aprovação no componente curricular Monografia II.

Trabalho aprovado. São Luís, 18 de julho de 2017:

Prof. Dr. Ivo José da Cunha Serra
(Orientador)
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão – UFMA

Prof. Dr. Tiago Bonini Borchartt
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão – UFMA

Prof. MSc. Carlos Eduardo Portela Serra de Castro
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão – UFMA

São Luís
2017

Este trabalho é dedicado à minha mãe Maria da Natividade e em memória ao meu pai Carlos Mendes. A todos aqueles que de alguma forma estiveram e estão próximos.

Agradecimentos

Em primeiro lugar, ao professor e orientador Ivo Serra que gentilmente aceitou ser meu orientador, demostrando sempre muito interesse e paciência para a construção deste trabalho.

Aos meus pais, Carlos e Natividade, que independente de qualquer obstáculo sempre estiveram comigo, me dando apoio. Agradeço também a Deus (Universo) e a minha família, irmãs Luzia, Aldenira e Marinalva, ao meu cunhado Claudio, sobrinhos Alyne e Alisson, pelo grande apoio, pela paciência e por sempre acreditarem em mim e no meu potencial. Parentes e amigos que me ajudaram dando palavras de incentivo e me reanimando em momentos difíceis também serão para sempre lembrados.

A todos os professores, em especial aos professores Portela e Bonini que gentilmente aceitarem fazer parte da banca e ao professor Hilkias Jordão.

Aos grandes amigos e colegas de curso, como Aronilson Aguiar, Benedito Vieira, Thales Levi, Tiago Ramos, Danilo Carvalho, Glécio Santos, Fernando Beleza, Alessandro Jorge, Márcio Sygeaks e tantos outros que compartilharam comigo momentos de dificuldades e alegrias, para todos o meu grande agradecimento, por todos os grandes momentos.

"Na eternidade onde não existe o tempo, nada pode crescer, nada pode se desenvolver, nada muda. Então, a morte criou o tempo para que as coisas pudessem crescer e para que pudessem morrer.."

(True Detective HBO. No episódio cinco de True Detective da 1ª Temporada, The Secret Fate of All Life (O Destino Secreto de toda a Vida))

Resumo

Nos últimos anos, verificou-se um grande crescimento na quantidade de dados armazenados. A mineração de dados surgiu com o propósito de identificar e extrair informações relevantes baseados nessa base de dados. Avanços nas tecnologias de armazenamento de dados, o aumento na velocidade e capacidade dos sistemas e a melhoria desses sistemas gerenciadores de banco de dados, têm permitido transformar essa enorme quantidade de dados em grandes bases de dados. Este trabalho apresenta os conceitos fundamentais de Agrupamento de Dados (*Clustering*) que é uma técnica de *Data Mining* para fazer agrupamentos automáticos de dados segundo seu grau de semelhança, também serão apresentadas três técnicas/métodos simples, mas muito importantes para introduzir muitos dos conceitos envolvidos no agrupamento de dados. As técnicas de agrupamento são instrumentos valiosos na análise exploratória dos dados e encontram aplicações em várias áreas. Ao final será apresentado alguns dos diversos domínios de aplicações de agrupamento, tais como: biologia, recuperação de informações, medicina, segmentação de imagens e mineração de textos.

Palavras-chave: Mineração de Dados. Agrupamento de Dados. Técnicas de Agrupamento. Aplicações de Agrupamento.

Abstract

In recent years, there has been a large increase in the amount of data stored. Data mining was developed with the purpose of identifying and extracting relevant information based on this database. Advances in data storage technologies, the increase in speed and capacity of systems and the improvement of these database management systems, have allowed to transform this enormous amount of data into large databases. This work presents fundamental concepts of Clustering that is the Data Mining technique to make automatic groupings of data according to their degree of learning, three simple techniques / methods are also presented, but very important to introduce many of the concepts involved In the data grouping. Grouping techniques are valuable tools in the exploratory analysis of data and find applications in several areas. At the end will be presented some of these diverse fields of clustering applications, such as: In biology, information retrieval, medicine, image segmentation and text mining.

Keywords: Data Mining. Data Grouping. Grouping Techniques. Grouping Applications.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DBSCAN	Density Based Spatial Clustering of Application with Noise
Eps	Raio de vizinhança de um ponto
MinPts	Número mínimo de pontos

Sumário

1 Introdução

Nos últimos anos, verificou-se um grande crescimento na quantidade de dados armazenados. Avanços nas tecnologias de armazenamento de dados, o aumento na velocidade e capacidade dos sistemas, o barateamento dos dispositivos de armazenamento e a melhoria dos sistemas gerenciadores de banco de dados, têm permitido transformar essa enorme quantidade de dados em grandes bases de dados (Fayyad, Piatetsky-Shapiro e Smyth (1996)). Estima-se que a cada 20 meses as empresas no mundo dobrem o volume de dados acumulados em seus computadores (Diniz e Neto (2000)).

As técnicas de agrupamento são instrumentos valiosos na análise exploratória dos dados e encontram aplicações em várias áreas, tais como: biologia, medicina, engenharia, marketing, visão computacional e sensoriamento remoto. Uma área de aplicação recente que tem se beneficiado significativamente da análise de agrupamento é a bioinformática. Nessa área, muitos trabalhos têm sido desenvolvidos aplicando-se algoritmos de agrupamento para análise de dados de expressão gênica (Faceli (2006)).

Técnicas de agrupamento fornecem um meio de explorar e verificar estruturas presentes nos dados, organizando-os em grupos de objetos similares (Jain e Dubes (1988)). O agrupamento pode ser visto como pertencente ao paradigma de aprendizado não supervisionado, em que o aprendizado é dirigido aos dados, não requerendo conhecimento prévio sobre as suas classes ou categorias (Mitchell et al. (1997)).

A quantidade de informações disponíveis ultrapassou a capacidade humana de compreensão. Não é viável, sem o auxílio de ferramentas computacionais apropriadas, a análise de grandes quantidades de dados pelo homem. Portanto, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação (GOLDSCHMIDT e PASSOS (2005)).

Dados, produzidos e armazenados em larga escala, são inviáveis de serem analisados por especialistas através de métodos tradicionais, tais como planilhas de cálculos e relatórios informativos operacionais, onde o especialista testa sua hipótese contra a base de dados (AURÉLIO, 1999).

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 e Rousseeuw (2009)). A solução para esse problema consiste frequentemente no objetivo final de mineração de dados, apresentando uma ampla aplicabilidade em diversas áreas (Everitt e Dunn (2001)).

A solução de um problema de agrupamento também pode ajudar a solucionar outros problemas relacionados, tais como classificação de padrões, extração de regras em bases de dados, sumarização de documentos e compressão de dados (Wang & Fu, 2015; Tan et al., 2005). A complexidade do problema de agrupamento de dados advém em boa parte de sua natureza não supervisionada, em que não se dispõe objetivamente de um resultado final desejado. Em outras palavras, em contraste aos problemas supervisionados, como classificação de padrões, não se dispõe, em agrupamento de dados, de uma meta concreta a ser alcançada. O que acontece, é que a dificuldade do problema começa pela própria definição do que se entende por grupo *(cluster)*, conceito com elevado grau de subjetividade. É importante mencionar que, na maioria dos casos, existe uma variedade de categorizações alternativas para um mesmo conjunto de objetos, ou seja, os objetos podem ser agrupados de maneiras diferentes dependendo da perspectiva. Por exemplo, um sistema bancário pode estar interessado em encontrar grupos nos quais os objetos (clientes) do mesmo grupo apresentam informações econômicas similares, levando em conta informações como: renda familiar, quantidade de pessoas na família e quantidade de bens. Só que o mesmo sistema bancário também pode estar interessado em encontrar grupos nos quais os objetos do mesmo grupo apresentem informações como por exemplo: endereço contados.

1.1 Motivação

A análise de agrupamento é de grande utilidade na análise exploratória de dados sobre os quais existe pouco ou nenhum conhecimento prévio disponível (Jain e Dubes (1988)). O objetivo do agrupamento é encontrar uma estrutura de grupos nos dados, em que cada grupo contém objetos que compartilham algumas características ou propriedades consideradas relevantes para o domínio dos dados estudados (Jain e Dubes (1988)). Entretanto, não existe uma definição precisa do que é um grupo e há uma grande variedade de algoritmos de agrupamento descrita na literatura, cada algoritmo com suas próprias características e peculiaridades (Barbara (2000)).

Cada algoritmo é baseado em uma definição de cluster e faz uso de alguma heurística para achar o melhor agrupamento para um determinado conjunto de dados. Assim, cada algoritmo de agrupamento pode apresentar um comportamento superior aos demais para uma conformação específica dos dados no espaço de atributos. Por exemplo, um algoritmo pode ser apropriado para encontrar apenas grupos específicos e outro podem encontrar grupos de formas arbitrárias, mas que possuam a mesma densidade. Nesse ponto surge a primeira dificuldade em análise de agrupamento: mesmo que os dados estejam estruturados idealmente segundo uma das possíveis definições de cluster, como selecionar o algoritmo mais apropriado, uma vez que as características dos dados não são conhecidas previamente?

Além disso, cada algoritmo de agrupamento é capaz de encontrar estruturas com diferentes níveis de refinamento (estruturas com diferentes números de grupo ou com grupos

de densidades diferentes), dependendo dos valores de seus parâmetros (Jain e Dubes (1988)). Por exemplo, o algoritmo k-médias encontra uma estrutura diferente para cada número de clusters, k, dado como parâmetro ao algoritmo. Porém, na análise exploratória de um conjunto de dados, em geral, o número de clusters presentes nos dados não é conhecido previamente. Esse aspecto oferece a segunda dificuldade da análise de agrupamento: dado que se conheça o algoritmo de agrupamento mais apropriado para um determinado conjunto de dados, como estabelecer quais valores de seus parâmetros fornecem as estruturas mais representativas do conjunto de dados? Em outras palavras, em que níveis de refinamento podem ser encontradas as estruturas subjacentes nos dados?

Os sistemas não supervisionados, também conhecidos como agrupamento ou clustering, têm o objetivo de separar um conjunto de observações não classificadas em um número discreto de grupos que são definidos pela estrutura natural dos dados, sem uso de qualquer informação prévia sobre os grupos. Deste modo, quando se tem necessidade de explorar a desconhecida natureza dos dados independente de se ter uma pré-informação de pertinência, a análise de grupos é a ferramenta mais adequada (Xu e Wunsch (2009)).

Este trabalho é motivado pelo interesse em discutir alguns domínios de aplicação de agrupamento de dados.

1.2 Objetivo do Trabalho

O objetivo principal deste trabalho é realizar uma discussão sobre agrupamento, as técnicas mais utilizadas quando se falar em agrupamento de dados, as técnicas de agrupamento hierárquico aglomerativo e os particionas K-means e DBSCAN *(Clusterização Espacial Baseada em Densidade de Aplicações com Ruído)* , e por fim mostrar algumas das diversas áreas onde se pode fazer aplicações utilizando agrupamento de dados particulamente em segmentação de imagens e mineração de textos.

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. Apresenta tópicos sobre o significado de agrupamento; uma exposição sobre os principais métodos utilizados atualmente em agrupamento de dados.

Capítulo 3 apresenta algumas das diversas áreas de aplicações onde pode-se utilizar agrupamento de dados, como na biologia, recuperação de informações, clima, psicologia, medicina, negócios, segmentação de imagens e na mineração de textos.

O capítulo 4 apresenta as conclusões obtidas do que foi realizado no capítulo 2 e 3, além de sugestões para a realização de trabalhos futuros.

2 Agrupamento

2.1 Considerações Iniciais

Neste capítulo, serão descritos os conceitos básicos de agrupamento e as seguintes três técnicas simples: Agrupamento Hierárquico Aglomerativo, K-means e DBSCAN (Tan, Steinbach e Kumar (2009)), porém importantes para introduzir muitos dos conceitos em agrupamento.

2.2 Definições

Agrupamentos são classes, ou grupos conceitualmente significativos de objetos que compartilhem características comuns, desempenham um papel importante em como as pessoas analisam e descrevem o mundo. De fato, seres humanos têm habilidades na divisão de objetos em grupos (agrupamento) e atribuir objetos particulares a esses grupos (classificação). Mesmo crianças relativamente jovens podem rotular rapidamente os objetos em uma fotografia como veículos, prédios, pessoas, animais e dentre outros, ou seja, Os seres humanos estão sempre classificando o que percebem a sua volta, por exemplo: criando classes de relações humanas diferentes e dando a cada classe uma forma diferente de tratamento; formando classes de comportamento em diferentes ambientes; definindo classes sociais; estabelecendo preconceitos e tratando as pessoas segundo estes estereótipos, entre outras formas de classificação (CARVALHO, 2002).

A Mineração de Dados também conhecida pelo termo inglês *Data Mining* é o processo de explorar grandes quantidades de dados. Consiste em uma funcionalidade que agrega e organiza dados, encontrando neles padrões, associações, mudanças e anomalias relevantes.

Agrupamento *(clustering)* é uma técnica de *Data Mining* para fazer agrupamentos automáticos de dados segundo seu grau de semelhança. Por exemplo, procedimento de agrupamento também pode ser aplicado a bases de texto utilizando algoritmos de *Text Mining*, onde o algoritmo procura agrupar textos que falem sobre o mesmo assunto e separar textos de conteúdo diferentes. Agrupamento é uma classificação não supervisionada (sem classes predefinidas). A classificação não-supervisionada baseia-se no princípio de que o algoritmo computacional é capaz de identificar por si só as classes dentro de um conjunto de dados. Esse tipo de classificação é frequentemente realizado através de métodos de agrupamentos.

Dessa forma, o agrupamento procura encontrar conjuntos de dados que se agrupam naturalmente por alguma similaridade gerando diversos grupos menores, sendo muito útil pelo motivo de não conhecer com antecedências as categorias existentes nos conjuntos de dados que serão analisados.

Técnicas ou algoritmos de agrupamento permitem a construção de importantes ferramentas para a análise exploratória de dados para os quais existe pouco ou nenhum conhecimento prévio (Jain e Dubes (1988)).

O objetivo de uma técnica de agrupamento é encontrar uma estrutura de grupos (clusters) nos dados, em que os objetos pertencentes a cada grupo compartilham alguma característica ou propriedade relevante para o domínio do problema em estudo (Jain e Dubes (1988)). Embora a ideia do que constitui um grupo seja intuitiva, não existe uma definição formal única e precisa para esse conceito. Ao contrário, existe uma grande variedade de definições na literatura, ou seja, há á diversas noções de um grupo que se provam úteis na prática. Isso é resultado da grande diversidade de visões/objetivos dos pesquisadores de diferentes áreas que utilizam/desenvolvem técnicas de agrupamento.

Um grupo é um conjunto de entidades semelhantes e entidades pertencentes a grupos diferentes não semelhantes (Jain, Murty e Flynn (1999)).

Grupos podem ser descritos como regiões conectadas de um espaço multidimensional contendo uma alta densidade relativa de pontos, separados de outras regiões por uma região contendo uma baixa densidade relatica de pontos (Everitt, 1993; Mertz, 2006).

Agrupamento também é um processo subjetivo, deste modo, é necessário atenção extra ao se realizar uma análise de grupo nos dados. A subjetividade está presente em diversos aspectos, entre eles nas hipóteses estabelecidas sobre os dados, a definição da medida de proximidade, a determinação do número de grupos, a seleção do algoritmo de agrupamento e a determinação dos índices de validação (Xu e Wunsch (2009)).

Além disso, para o mesmo conjunto de dados, objetivos diferentes geralmente levam a diferentes partições. Um exemplo simples e direto é na partição de animais: uma águia, um canário, um leão, uma pantera e um carneiro. Se os animais são divididos com base no critério de poder ou não voar, temos dois clusters: com a águia e o canário em um grupo e o restante em outro grupo. No entanto, se mudarmos o critério e avaliarmos se eles são ou não carnívoros, temos uma partição completamente diferente com o canário e o carneiro em um cluster e os outros três no segundo grupo (Xu e Wunsch (2009)).

Algumas das definições comuns para grupo são (Barbara (2000)):

Grupos bem separado: um grupo é um conjunto de pontos tal que qualquer ponto em um determinado grupo está mais próximo (ou é mais similar) a cada outro ponto nesse grupo do que a qualquer ponto não pertencente a ele.

Grupos baseado em centro: um grupo é um conjunto de pontos tal que qualquer ponto em um dado grupo está mais próximo (ou é mais similar) ao centro desse grupo do que ao centro de qualquer outro grupo. O centro de um grupo pode ser um centroide, como a média aritmética dos pontos do grupo ou um medóide (isto é, o ponto mais representativo do grupo).

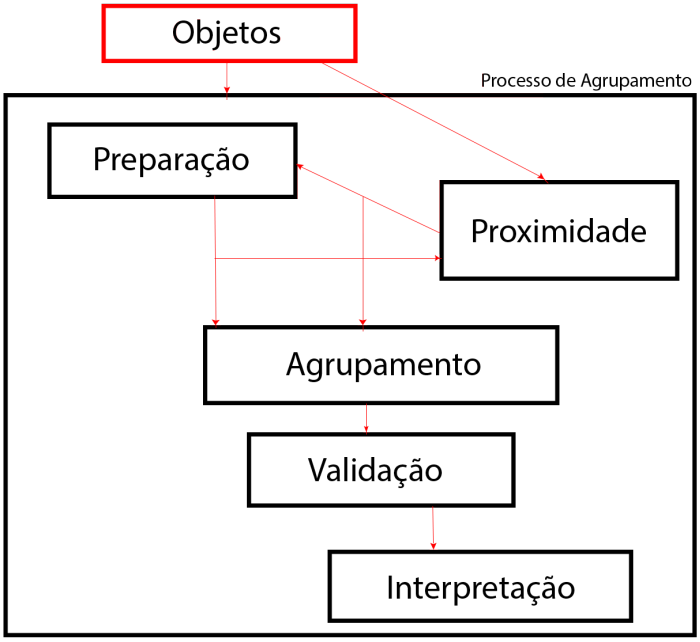
Grupos contínuo (vizinho mais próximo ou agrupamento transitivo): um grupo é um conjunto de pontos tal que qualquer ponto em um dado grupo está mais próximo (ou é mais similar) a um ou mais pontos nesse grupo do que a qualquer ponto que não pertence a ele.

Grupos baseado em densidade: um grupo é uma região densa de pontos, separada de outras regiões de alta densidade por regiões de baixa densidade.

Grupos baseado em similaridade: um grupo é um conjunto de pontos que são similares, enquanto pontos em grupos diferentes não são similares.

O processo de agrupamento compreende diversas etapas que vão desde a preparação dos objetos, até a interpretação dos grupos obtidos. A Figura 1 resume as etapas do processo de agrupamento com as informações utilizadas e geradas em cada etapa. As etapas e a figura apresentada são baseadas nas informações apresentadas por (Jain, Murty e Flynn (1999)) e (Barbara (2000)), cada uma dessas etapas são descritas a seguir.

Figura 1 – Processo de agrupamento.



Fonte: Produzido pelo autor

Preparação: Os objetos a serem agrupados podem representar um objeto físico, como uma cadeira, ou uma noção abstrata, como um estilo de escrita. Tais objetos também são comumente chamados de padrões, exemplos, amostras, instancias ou pontos. A preparação dos dados para o agrupamento envolve vários aspectos relacionados ao seu pré-processamento e à forma de representação apropriada para sua utilização por um algoritmo de agrupamento.

O pré-processamento pode envolver, por exemplo, normalizações, conversão de tipos e redução do número de atributos por meio de seleção ou extração de características (Jain, Murty e Flynn (1999)). Vários trabalhos discutem formas de padronização dos dados, seleção de atributos e outros aspectos relativos à preparação dos dados, como os de (Jain, Murty e Flynn (1999)), (Barbara (2000)).

Proximidade: Esta etapa consiste da definição de uma medida de proximidade apropriada ao domínio da aplicação. Essa medida de proximidade pode ser uma medida de similaridade ou de dissimilaridade entre dois objetos. A escolha da medida de proximidade a ser empregada com um algoritmo de agrupamento deve considerar os tipos e escalas dos atributos que definem os objetos e também as propriedades dos dados que o pesquisador deseja focalizar. Por exemplo, o pesquisador deve ter em mente se a magnitude relativa dos atributos descrevendo dois objetos é suficiente ou seu valor absoluto deve ser considerado (Gordon 1999). As medidas de proximidade, em geral, consideram que todos os atributos são igualmente importantes.

(Jain e Dubes (1988)) descrevem detalhadamente as medidas de proximidade mais apropriadas para cada tipo e escala de atributo possível. Uma das medidas de proximidade mais comumente utilizada é a Distância Euclidiana formalizada pela seguinte equação:

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Em matemática, distância euclidiana (ou distância métrica) é a distância entre dois pontos. A distância entre dois objetos *(x e y)* onde n é o número de dimensões e x_k e y_k são, respectivamente, os atributos(componentes) de índice k de x e y.

Agrupamento: Esta etapa consiste da aplicação de um algoritmo de agrupamento, ou seja, de uma técnica de agrupamento apropriado para agrupar os dados de acordo com um objetivo específico, essas técnicas serão vistas nas Seções 2.3.1 e 2.3.2.

Validação: Esta etapa se refere à avaliação do resultado de um agrupamento e deve, de forma objetiva, determinar se os grupos são significativos, ou seja, se a solução é representativa para o conjunto de dados analisado. Uma estrutura de agrupamento é válida se não ocorreu por acaso ou se é "rara" em algum sentido, já que qualquer algoritmo de agrupamento encontrará grupos, independentemente de existir ou não similaridade nos dados (Jain e Dubes (1988)).

Interpretação: Refere-se ao processo de examinar cada grupo com relação a seus objetos para rotulá-los, descrevendo a natureza do grupo. A interpretação de grupos é mais que apenas uma descrição. Além de ser uma forma de avaliação dos grupos encontrados e da hipótese inicial, de um modo confirmatório, os grupos podem permitir avaliações subjetivas que tenham um significado prático. Ou seja, o especialista pode ter interesse em encontrar diferenças semânticas de acordo com os objetos e valores de seus atributos em cada grupo.

Mais detalhes sobre cada um desses passos do agrupamento é encontrado em (Faceli,

Carvalho e Souto (2005)).

Existe um grande número de algoritmos de agrupamento descritos na literatura, mas porém, não existe um algoritmo de agrupamento universal, capaz de revelar toda a variedade de estruturas que podem estar presentes em um conjunto de dados. Além disso, como lembra (Hartigan (1985)), "diferentes agrupamentos são adequados para diferentes propósitos. Dessa forma, não é possível afirmar que um agrupamento é melhor que outro". Isso tudo leva a dificuldades na escolha do melhor algoritmo a ser aplicado a um problema específico. Apesar de também existir uma grande diversidade de técnicas de validação capazes de auxiliar nessa escolha, em geral, cada uma apresenta uma tendência de favorecer um tipo de algoritmo, por ser baseada no mesmo conceito que o critério de agrupamento dos algoritmos desse tipo (Handl, Knowles e Kell (2005)).

Além da dificuldade da escolha do melhor algoritmo para uma dada aplicação, muitos dos algoritmos apresentam restrições. Alguns dos problemas comuns a vários algoritmos de agrupamento são (Jain e Dubes (1988)) e (Handl, Knowles e Kell (2005)):

- Adequação a domínios e/ou conjuntos de dados restritos.
- Restrição dos formatos da estrutura que pode ser encontrada.
- Necessidade de conhecimento prévio do número de clusters presentes nos dados ou o difícil ajuste de parâmetros.
- Instabilidade dos resultados obtidos. Várias execuções de um algoritmo produzem agrupamentos diferentes, podendo associar um mesmo objeto a clusters diferentes.

2.3 Técnicas de Agrupamento

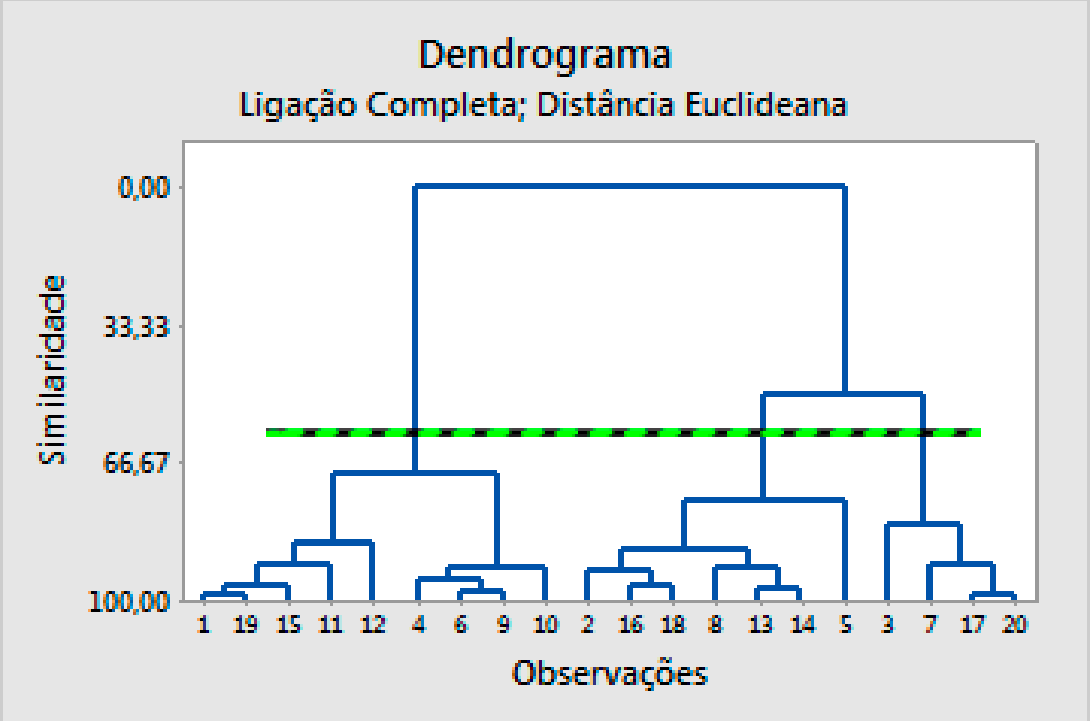
2.3.1 Método Hierárquico

Neste método o processo de identificação de grupos é geralmente realimentado recursivamente, utilizando tanto objetos quanto grupos já identificados previamente como entrada para o processamento. Deste modo, constrói-se uma hierarquia de grupos de objetos, no estilo de uma árvore (Diniz e Neto (2000)).

As técnicas de agrupamento do tipo hierárquico produzem uma hierarquia entre os grupos. Essa hierarquia pode ser representada por uma árvore de grupos, que é conhecida como dendograma (diagrama que mostra a hierarquia e a relação dos agrupamentos em uma estrutura) Figura 2. Nesta representação os dados individuais são as folhas da árvore e os nós do interior são aglomerados de grupos. Por padrão, o nível de similaridade é medido no eixo vertical e as diferentes observações (objetos/elementos) são listadas ao longo do eixo horizontal. O gráfico mostra como os agrupamentos são formados: unindo duas observações individuais ou pareando uma observação individual com um agrupamento existente. É possível ver em que

nível de similaridade os agrupamentos são formados, e a composição dos agrupamentos da partição final. O corte no dendrograma indicado pela linha tracejada representa uma partição do conjunto de objetos {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20} com três grupos: g1 = {1, 19, 15, 11, 12, 4, 6, 9, 10}, g2 = {2, 16, 18, 8, 13, 14, 5} e g3 = {3, 7, 17, 20}. Os métodos hierárquicos permitem, assim, a exploração dos dados em diferentes níveis de granularidade. Ou seja, os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os elementos.

Figura 2 – Dendrograma: diagrama que mostra a hierarquia e a relação dos agrupamentos em uma estrutura.



Fonte: <https:goo.gl/nz3KZ6>

Os métodos hierárquicos não requerem que seja definido um número *a priori* de grupos. Através da análise do dendograma, pode-se inferir no número de agrupamentos adequados.

Os métodos hierárquicos requerem uma matriz Tabela 1 contendo as métricas de distância entre os agrupamentos em cada estágio do algoritmo. Essa matriz é conhecida como matriz de similaridades entre agrupamentos. Dessa forma, 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:

Tabela 1 – Tabela ilustrativa da 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: Produzido pelo autor

Pela tabela ilustrativa acima se pode observar que G1 e G2 são os agrupamentos mais similares, enquanto que G2 e G3 são os menos similares. São utilizados os medidas de distância entre grupos para o cálculo dos valores de proximidade entre os agrupamentos.

As técnicas de agrupamento do tipo hierárquico são subdivididos em métodos Aglomerativos e Divisivos, ou seja, há duas abordagens básicas para gerar um agrupamento hierárquico, os métodos aglomerativos *(bottom-up)* e os métodos de divisão ou divisivos *(top-down)* (Tan, Steinbach e Kumar (2009)):

Aglomerativo: Inicia com os pontos como grupos individuais e, em cada etapa, funde os pares mais próximos de grupos. Isto requer a definição de uma noção de proximidade de grupos.

Divisivo: Inicia com um grupo inclusivo com tudo e, a cada etapa, divide um grupo até que reste apenas grupos únicos de pontos individuais. Neste caso, é preciso decidir qual grupo dividir em cada etapa e como fazer a divisão.

As técnicas de agrupamento hierárquico aglomerativo são as mais comuns entre as técnicas de agrupamento do tipo hierárquico, e nesta seção enfocaremos exclusivamente este método.

Muitas técnicas de agrupamento hierárquico aglomerativo são variações sobre uma abordagem única: iniciando com pontos individuais como grupos, funde sucessivamente os dois grupos mais próximos até que reste apenas um grupo. Esta abordagem é expressada mais formalmente no Algoritmo 1 da Figura 3.

Figura 3 – Algoritmo de Agrupamento Hierárquico Aglomerativo Básico.

Algoritmo 1 Algoritmo de agrupamento Hierárquico Aglomerativo Básico.

```
1:  Calcule a matriz de proximidade, caso necessário.
2:  repita
3:      Funda os dois grupos mais próximos.
4:      Atualize a matriz de proximidade para refletir a proximidade
4:      entre o novo grupo e os grupos originais
5:  até que Reste apenas um grupo
```

Fonte: Produzido pelo autor

2.3.2 Método Particional

Um agrupamento particional é uma divisão do conjunto de objetos de dados em subconjuntos (grupos) não interseccionados de modo que cada objeto de dado esteja exatamente em um subconjunto. No próximo tópico, é feita uma descrição dos algoritmos K-means e DBSCAN (Tan, Steinbach e Kumar (2009)).

2.3.2.1 Algoritmo K-means

O algoritmo K-Means (MacQueen et al. (1967); Duda, Hart e Stork (2001)) também chamado de K-Médias é uma das técnicas de agrupamento particionais mais populares, por possuir o maior número de variações devido à sua simplicidade e facilidade de implementar em linguagens computacionais.

A ideia do algoritmo K-Means é fornecer uma classificação de informações de acordo com os próprios dados. Esta classificação, é baseada em análise e comparações entre os valores numéricos dos dados. Desta maneira, o algoritmo automaticamente vai fornecer uma classificação automática sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhuma pré-classificação existente.

O algoritmo K-Means implementa uma técnica de agrupamento baseada em protótipos (centros). O K-Means define um protótipo em termos de um centroide, que normalmente corresponde à média dos padrões em um grupo. K-Means inicia escolhendo K centroides iniciais, em que K é um parâmetro definido pelo usuário, que representa o número de grupos desejados. Cada padrão é então atribuído ao centroide mais próximo, e cada coleção de padrões atribuída ao centroide forma um grupo. O centroide de cada grupo é então atualizado baseado nos padrões atribuídos ao grupo. O processo de atribuição dos padrões e a atualização dos centroides se repete até que os centroides permaneçam inalterados.

Como é ilustrado no Algoritmo 2 da Figura 4, O algoritmo K-Means pode ser descrito da seguinte maneira. Escolhe-se K distintos valores para os centros dos grupos (isso pode ser feito aleatoriamente), associa cada ponto ao centro mais próximo, (pode-se usar a distância euclidiana), recalcular o centro de cada grupo, e isso ocorre até que as iterações acabem ou não ocorra mais mudança de objetos dos grupos.

Figura 4 – Algoritmo de Agrupamento K-Means

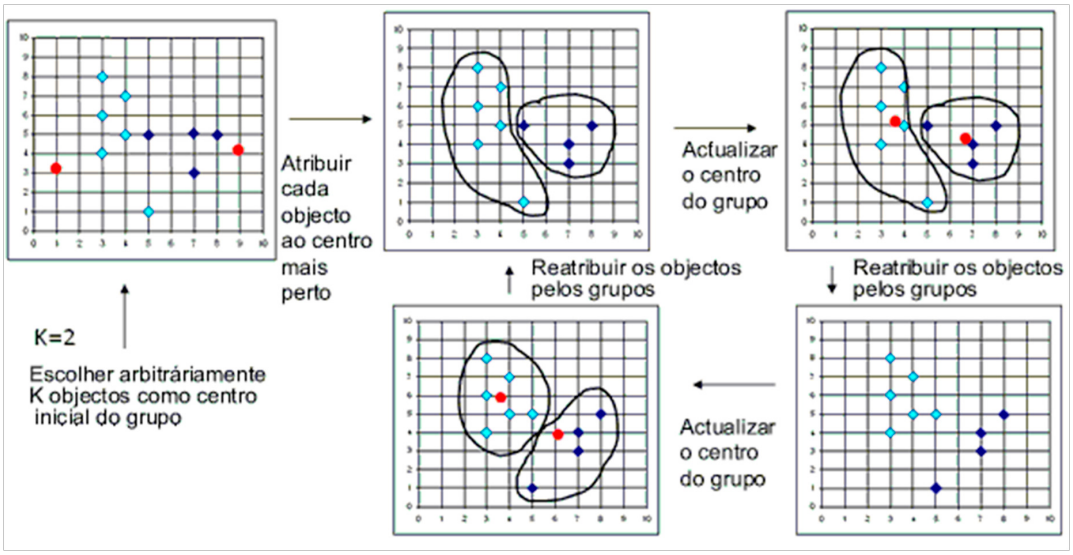
Algoritmo 2 Algoritmo K-Means.

especifique k.
selecione os k objetos que serão os centroides dos agrupamentos.
enquanto
    Para *todos os objetos Faça.*
        Calcule a distância entre o objeto e os centroides.
        Adiciona o objeto ao grupo que possuir a menor distância.
        Recalcule os centroides dos grupos.
    Fim do Para
até que *o número de interações* $= i$ *ou Não ocorra mudança dos centroides.*

Fonte: Produzido pelo autor

A Figura 5 mostra um exemplo prático dos passos da aplicação do algoritmo K-Means: Inicialmente escolhesse arbitrariamente k objetos como centro inicial do grupo, é atribuído cada objeto ao centro mais próximo e atualizado o centro de cada grupo. Novamente, é atribuído os objetos pelos grupos e atualizado o centro do grupo, e isso ocorre até que nenhum objeto mude mais de grupo.

Figura 5 – Passos de aplicação do algoritmo K-médias.



Fonte: <https:goo.gl/G1NNyF>

2.3.2.2 Algoritmo DBSCAN

DBSCAN, abreviação do termo *'Density Based Spatial Clustering of Application with Noise' (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído)* é um método de agrupamento não paramétrico baseado em densidade (número de pontos dentro de um raio específico (Eps), proposto no ano de 1996 por, (Ester et al. (1996)) publicaram o artigo intitulado A Density Based Spatial Clustering of Applications With Noise.

Agrupamentos baseados em densidade localizam regiões de alta densidade que estejam separadas entre si por regiões de baixa densidade. DBSCAN é um algoritmo de agrupamento baseado em densidade simples e eficaz que ilustra uma quantidade de conceitos que são importantes para qualquer abordagem de agrupamento baseado em densidade.

Os dois parâmetros de entrada que o DBSCAN necessita são:

Raio de vizinhança de um ponto: determina o raio de vizinhança (Eps) para cada ponto da base de dados. Dado o parâmetro Eps, o algoritmo DBSCAN verifica a quantidade de pontos contidos no raio (Eps) para cada ponto da base de dados, e se essa quantidade exceder certo número, um cluster é formado;

Número mínimo de pontos (MinPts): parâmetro que especifica o número mínimo de pontos, no dado raio (Eps), que um ponto precisa possuir para ser considerado um ponto central e conseqüentemente, de acordo com as definições de cluster baseado em densidade, inicia a formação de um cluster.

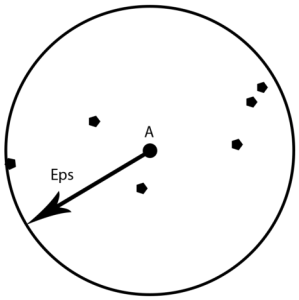
Com os parâmetros Eps e MinPts definidos o algoritmo basicamente realiza a separação do conjunto de observações em três classes:

Pontos Centrais. Estes pontos estão no interior de um grupo baseado em densidade. Um ponto é central se o número de pontos dentro de uma determinada vizinhança em torno do ponto conforme determinado pela função de distância e um parâmetro de distância especificada pelo usuário, Eps, exceder um determinado limite, MinPts, que também é um parâmetro especificado pelo usuário. Na Figura 7, o ponto A é um ponto central, para o raio indicado Eps se MinPts ≤ 7.

Pontos Limites: Um ponto de limite não é um ponto central, mas fica dentro da vizinhança de um ponto central. Na Figura 7, o ponto B é um ponto de limite. Um ponto de limite pode cair dentro das vizinhanças de diversos pontos centrais.

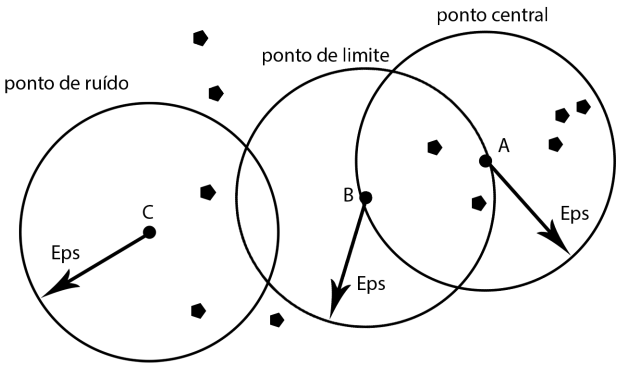
Pontos de Ruídos: Um ponto de ruído é qualquer ponto que não seja nem um ponto central nem um ponto limite. Na Figura 7, o ponto C é um ponto de ruído.

Figura 6 – *Densidade baseada em centro.*



Fonte: Produzido pelo autor

Figura 7 – *Pontos de centro, de limite de ruído.*



Fonte: Produzido pelo autor

Dadas as definições de pontos de centro, de limite e de ruído o algoritmo DBSCAN pode ser descrito informalmente da seguinte maneira. Quaisquer dois pontos do centro que estejam suficientemente próximos, dentro de uma distância Eps entre si, são colocados no mesmo grupo. Da mesma forma, qualquer ponto de limite que esteja suficientemente próximo de um ponto do centro é colocado no mesmo grupo do ponto de centro. Pontos de ruído são descartados. Os detalhes formais apresentados no Algoritmo 3 da Figura 8.

Figura 8 – Algoritmo de Agrupamento DBSCAN

Algoritmo 3 Algoritmo DBSCAN.

```
1:   Rotular todos os pontos como de centro, de limite ou ruído.
2:   Eliminar os pontos de ruído.
3:   Colocar uma aresta entre todos os pontos de centro que...
3:   estejam dentro da Eps uns dos outros.
4:   Tornar cada grupo de pontos de centro conectados um grupo separado.
5:   Atribuir cada ponto de limite a um dos grupos dos seus pontos de centro associados.
```

Fonte: Produzido pelo autor

3 Aplicações de Agrupamento

Neste capítulo, é feita uma descrição de algumas das diversas formas e áreas onde agrupamento de dados são utilizados, com objetivo de se alcançar um determinado fim, ou seja, será comentado onde pode ser aplicado, por exemplo na segmentação de imagens e mineração de texto.

Existe uma ampla variedade de campos onde pode-se aplicar agrupamento de dados como por exemplo: na psicologia e outras ciências sociais, biologia, estatística, reconhecimento de padrões, recuperação de informações, aprendizado de máquina e mineração de dados. A seguir estão alguns exemplos (Tan, Steinbach e Kumar (2009)):

• **Biologia**: Biólogos gastaram muitos anos criando uma taxonomia, ou seja, uma classificação hierárquica de todas as coisas vivas conhecidas pelo homem em: reino, filo, classe, ordem, família, gênero e espécie. Assim, talvez não seja surpreendente que muito do trabalho inicial em análise de grupos procurou criar uma disciplina de taxonomia matemática que pudesse encontrar automaticamente tais estruturas de classificação. Mais recentemente, biólogos aplicaram o agrupamento para analisar as grandes quantidades de informações genéticas que agora estão disponíveis. Por exemplo, o agrupamento tem sido usado para encontrar grupos de genes que tenham funções semelhantes, com isso uma área de aplicação recente tem se beneficiado significativamente da análise de agrupamento, a bioinformática (Baldi and Brunak 1998; Wang et al. 2003; Narayanan 2005). Nessa área, muitos trabalhos têm sido desenvolvidos aplicando algoritmos de agrupamento para análise de dados de expressão gênica (Wang et al. 2003; Lorkowski and Cullen 2003; Zhao and Karypis 2005; Azuaje and Dopazo 2005; Narayanan 2005). Na bioinformática, as investigações em genômica funcional e a análise de dados de expressão gênica têm utilizado técnicas de agrupamento para encontrar grupos de genes, amostras ou ambos. O agrupamento de genes é baseado na similaridade dos perfis moleculares de células em diferentes condições e pode ser utilizado, por exemplo, para estudar os mecanismos regulatórios dos genes ou dividir o genoma em conjuntos de genes que estão envolvidos nos mesmos processos ou em processos relacionados (Eisen et al. 1998; Spellman et al. 1998; Tamayo et al. 1999; Nikkila et al. 2002; Hautaniemi et al. 2003).

• **Recuperação de Informações**. A *Word Wide Web* consiste de bilhões de páginas *Web* e os resultados de consulta a uma "ferramenta de pesquisa" pode retornar milhões de páginas. O agrupamento pode ser usado para agrupar estes resultados de pesquisas em um número pequeno de grupos, cada um dos quais captura um determinado aspecto da consulta. Por exemplo, uma consulta de "filme" poderia retornar páginas *Web* agrupadas em categorias como resenha, trailers, elencos e cinemas. Cada categoria, ou seja, grupo pode ser dividida em subcategorias (subgrupos), produzindo uma estrutura hierárquica que auxilie mais a exploração

de um usuário nos resultados das consultas.

• **Clima**. Compreender o clima da terra requer encontra padrões na atmosfera e no oceano. Para este fim, a análise de grupos tem sido aplicada para encontrar padrões na pressão atmosférica de regiões polares e áreas do oceano que tenham um impacto significativo sobre o clima da terra.

• **Psicologia e Medicina**. Uma doença ou condição possui frequentemente uma quantidade de variantes, e a análise de agrupamentos pode ser usada para identificar essas diferentes subcategorias. Por exemplo, o agrupamento tem sido usado para identificar diferente tipos de depressão. A análise de agrupamentos também pode ser usada para padrões na distribuição espacial ou temporal de uma doença.

• **Negócios**. Negócios juntam imensas quantidades de informações sobre clientes atuais e potencias. O agrupamento pode ser usado para segmentar clientes em um número menor de grupos para análise adicional e atividades de marketing.

A seguir na Seção 3.1 e 3.2, descreve de forma mais detalhada, alguns exemplos de domínio onde se pode aplicar agrupamento de dados, como em, segmentação de imagens (Kolossoski (2007)) e mineração de textos (Rezende, Marcacini e Moura (2011)).

3.1 Segmentação de Imagens

Em visão computacional, segmentação se refere ao processo de dividir uma imagem digital em múltiplas regiões ou objetos, com o objetivo de simplificar e/ou mudar a representação de uma imagem para facilitar a sua análise.

O processamento digital de imagens é uma ferramenta muito importante, pois traz muitas vantagens e melhorias para áreas como a robótica, medicina, fotografia, meteorologia, enfim, tudo o que envolve imagens. Tem como objetivos a manipulação e análise de imagens por computador visando a extração de informação destas, para que os resultados venham a trazer benefícios para as áreas citadas anteriormente, além de muitas outras aplicações.

Dentro do processamento de imagens, levando em consideração a maneira como as imagens são processadas digitalmente, encontra-se o processo de segmentação, que consiste em dividir a imagem em várias partes de acordo com as características dos pontos (conjunto de pixels). Esse é um processo simples para o ser humano, que consegue segmentar uma cena, obtida através da visão, imediatamente, separar todos os objetos e definir seus contornos. Porém, é um processo complicado ao nível de computação.

A visão é um dos mais poderosos e complicados sentidos que o ser humano possui. Através da visão é possível obter as posições e propriedades dos objetos, assim como suas relações entre os mesmos e o ambiente que os cerca. A visão permite realizar três tarefas básicas:

- Percepção do mundo;
- Concepção de uma estratégia para tomada de decisão;
- Execução de uma ação.

Enquanto esta visão é muito natural para o ser humano, foi provado por muitos especialistas que é muito complexo ensinar um computador a funcionar com um sistema de visão (mesmo rudimentar). Uma causa disto é que até hoje é difícil explicar precisamente o processo da percepção. Não se podem recuperar todas as informações de uma cena apenas pela sua intensidade de bordas dos objetos. Esta intensidade é, na verdade, resultado de uma combinação de fatores como a superfície de um objeto, fonte e direção de iluminação, luz ambiente, condição atmosférica, entre outros. O olho humano é sempre comparado a uma câmera e, em muitos casos, as similaridades são bem evidentes. Há, entretanto, uma grande diferença: o processo de visualização. Na câmera a filme, as imagens são produzidas no filme através das mudanças fotoquímicas, enquanto que no olho humano as mudanças fotoquímicas provocam impulsos nervosos que são transmitidos ao cérebro. Assim, o cérebro interpreta estes impulsos como provenientes de objetos situados fora do corpo; sendo impossível perceber as imagens como se estivessem na retina. Até mesmo as imagens geradas pelo efeito pós-imagem são projetadas para fora do corpo (ainda que de olhos fechados).

Apesar da visão humana ser bem descrita num nível neuroanatômico, o processamento da informação realizado pela retina e pelo córtex visual do cérebro permanece, ainda hoje obscuro. O conhecimento acerca de visão biológica ainda é muito limitado, desconexo e especulativo. Esta hipótese tem motivado pesquisadores em visão computacional a propor teorias computacionais sobre o que seria o processo de visão. Tais teorias têm evoluído ao longo dos anos, baseadas na crescente compreensão deste processo. Entre os adventos produzidos nestas últimas décadas está o processamento digital de imagens.

O processamento digital de imagens conhecido também como processamento de imagens, surgiu da teoria de processamento de sinais e se tornou uma das múltiplas facetas da teoria da informação. Imagens são como a representação ou descrição de um objeto, pessoa ou cena, trazendo informações através de distribuições de intensidade de luz.

O interesse em métodos de processamento de imagens digitais decorre de duas áreas principais de aplicação:

- Melhoria da informação visual para a interpretação humana.
- Processamento de dados de cenas para percepção automática através de máquinas.

O processamento digital de imagens é uma área de concentração do conhecimento humano, que tem como objetivos a manipulação e análise de imagens por computador visando a extração de informação destas. Os recursos disponibilizados pelo processamento de imagens são utilizados em várias atividades, entre as quais estão a medicina, a robótica e a meteorologia.

3.1.1 Processamento de Imagens

No Processamento de Imagens, podemos dividir o processamento em algumas etapas. A primeira delas é a aquisição da imagem digital. Para isso, é necessário um sensor para imageamento (Imagear é a capacidade que um sensor tem para discriminar, numa área), que pode ser uma câmera de TV monocromática ou colorida, ou também uma câmera de varredura por linha, que produz uma única linha de imagem por vez, por exemplo. Tão importante quanto o sensor é a capacidade de digitalizar o sinal produzido pelo mesmo.

Depois de obtida a imagem digital, o próximo passo é o pré-processamento da imagem. A função chave no pré-processamento é a melhoria da imagem de forma que as chances de sucesso dos processos seguintes sejam maiores. Por exemplo, o pré-processamento pode envolver técnicas para o realce de contrastes, remoção de ruído e isolamento de regiões cuja textura indique a probabilidade de informação alfanumérica.

A próxima etapa é a segmentação, que divide uma imagem de entrada em várias partes ou objetos constituintes. Em geral, a segmentação automática é uma das tarefas mais difíceis no processamento de imagens digitais.

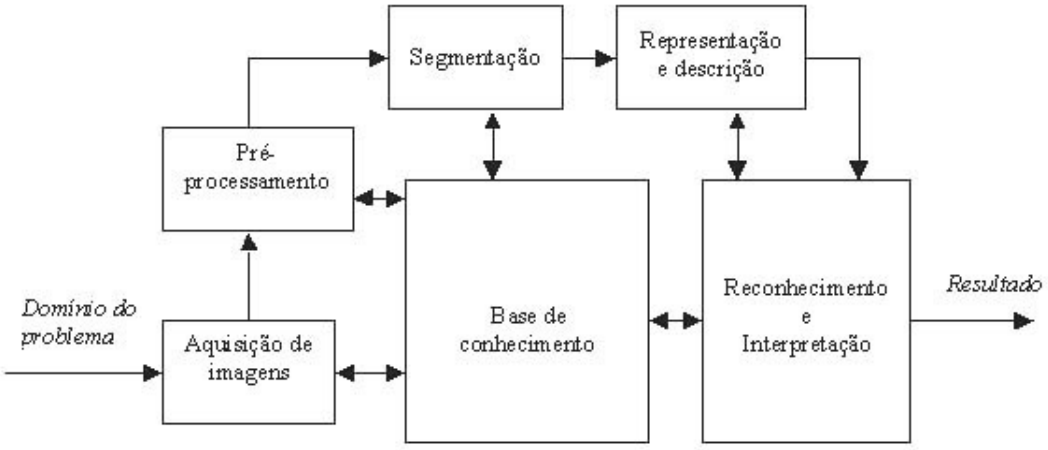
O estágio anterior produz uma saída constituída tipicamente por dados em forma de pixels, que correspondem tanto à fronteira de uma região como a todos os pontos dentro da mesma. É necessário converter esses dados para uma forma adequada ao processamento computacional. A primeira decisão que precisa ser feita é se os dados devem ser representados como fronteiras ou como regiões completas, e também deve ser especificado um método para descrever os dados, de forma que as características de interesse sejam enfatizadas.

O próximo processo, de descrição, também chamado seleção de características, procura extrair características básicas para discriminação entre classes de objetos, ou que resultem em alguma informação quantitativa de interesse. Quando se trata de reconhecimento de caracteres, descritores tais como buracos e concavidades são características poderosas que auxiliam na diferenciação entre uma parte do alfabeto e outra.

O último estágio envolve reconhecimento e interpretação. Reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo seu descritor. A interpretação envolve a atribuição de significado a um conjunto de objetos reconhecidos.

A Figura 9 mostra esquematicamente os passos fundamentais no processamento de imagens:

Figura 9 – Esquema dos passos fundamentais no processamento de imagens.



Fonte: <https://goo.gl/GsJLvq>

3.1.2 O Problema da Segmentação de Imagens

Na área de detecção e reconhecimento de imagens não basta simplesmente representar uma imagem com diferentes cores ou graduações de cinza. Também é necessário identificar regiões e estabelecer subdivisões na imagem em sua unidade básica (*pixel*) para que possa ser interpretada de acordo com uma finalidade específica. A identificação de regiões ou segmentos (não sobrepostos) na imagem é chamada de segmentação.

A segmentação de imagens traz como resultado um conjunto de regiões/objetos ou contornos extraídos da imagem. Assim, os pixels em uma determinada região são similares em alguma característica ou propriedade computacional, tais como cor, intensidade, textura ou continuidade. Em relação às mesmas características, regiões adjacentes devem possuir diferenças significativas entre umas e outras.

Há uma certa dificuldade no processo de segmentação de imagens nos computadores. A identificação dos segmentos deve obedecer a algumas características. Os pixels devem possuir alguma propriedade em comum dentro da imagem. Essa propriedade pode ser, por exemplo, uma superfície que representa um osso dentro de uma radiografia, uma peça sendo submetida a um controle de qualidade ou um mapa ilustrando alguma característica de uma foto. Dentre as propriedades desejáveis de uma imagem são destacadas algumas a seguir:

- Homogeneidade da região representada pelos pixels.
- Os segmentos são regiões fechadas e devem ser delimitadas por bordas ou outros segmentos.
- Cada pixel deve pertencer somente a uma região, não havendo regiões adjacentes com pixels em comum.
- Os segmentos, com relação a níveis de cinza e textura, devem ser uniformes e

homogêneos.

- As regiões devem ser simples e não devem apresentar buracos pequenos.
- Regiões adjacentes devem possuir diferenças significativas.
- Os segmentos devem ter bordas precisas.

As características mencionadas acima, na prática, são utópicas por que superfícies homogêneas são geralmente cheias de furos e as bordas em geral são irregulares. Além disso, podem ocorrer a fusão e perda de bordas nas regiões adjacentes. Como regra geral a identificação de segmentos é especifica e típica para cada aplicação. Geralmente destaca-se a separação da região de interesse buscando-se descontinuidade e similaridade nos diferentes tons da imagem. As descontinuidades são representadas pelas mudanças bruscas nos tons das cores como linhas e bordas. As similaridades baseiam-se nos limiares dos tons, subdivisão da imagem em regiões homogêneas e crescimento de regiões.

3.1.3 Dificuldades Inerentes na Segmentação de Imagens

Existem algumas dificuldades inerentes à segmentação de imagens. Antes que um processo de segmentação seja executado, alguns fatores devem ser considerados. Dentre eles podemos destacar os seguintes:

- Segmentação de forma autônoma em larga escala: Quando a segmentação de imagens envolve processos automáticos, há uma grande necessidade de controlar o ambiente de onde a imagem é retirada. Ambientes bem controlados (grandes contrastes) tendem a facilitar a interpretação das imagens. Ambientes externos, dependentes do clima, iluminação e outros fatores apresentam várias dificuldades.
- Controle da luminosidade: Conforme a aplicação envolvida, a existência de sombras tende a dar uma falsa impressão acerca do tamanho real da região a ser segmentada.
- As bordas das regiões são muitas vezes irregulares e imprecisas.
- A precisão do resultado depende da qualidade da distinção entre os diferentes elementos da imagem.
- Escolha da melhor estratégia e adequação à aplicação que se deseja.

Vários algoritmos e técnicas de segmentação de imagens foram desenvolvidos, não havendo, porém, uma solução geral para este problema. Muitas vezes para a resolução de um problema de segmentação é necessária a combinação das técnicas, para que ocorra a adaptação ao domínio do problema.

3.1.4 Agrupamento na Segmentação de Imagens

Quando se fala em análise de imagens, o primeiro passo geralmente é a segmentação da imagem em subdivisões de suas partes constituintes ou objetos. O quanto a imagem deve ser subdividida depende unicamente do domínio do problema a ser resolvido, por exemplo, uma imagem obtida de um radar de trânsito provavelmente terá como objetivo a identificação de veículos, sendo assim deve ser realizada uma segmentação de objetos que tenham formato e tamanho de um carro.

Um dos métodos de agrupamento desenvolvidos para resolver o problema da segmentação de imagens é o K-Means, o algoritmo k-means é utilizado para segmentar imagens, baseado em seus atributos, em k pedaços. Ele assume que os atributos dos pontos da imagem formam um espaço vetorial. O objetivo do algoritmo é minimizar a variância dos atributos dos pontos que estão dentro de um determinado segmento.

Esse método de agrupamento utilizado na Segmentação de Imagens pode derivar varias outras aplicações com por exemplo:

• Imagens Médicas (Localização de tumores e outras patologias; Medida de volume de tecidos; Cirurgia guiada por computador; Diagnostico de doenças; Planos de tratamento; Estudo da estrutura anatômica.

• Sistemas de reconhecimento de faces;

• Sistemas de controle automático de trafego;

• Sistemas de visão computacional.

• Localização de objetos em imagens de satélite (estradas, florestas, entre outros.)

3.2 Mineração de Textos

O avanço das tecnologias para aquisição e armazenamento de dados tem permitido que o volume de informação gerado em formato digital aumente de forma significativa. Cerca de 80% desses dados estão em formato não estruturado, no qual uma parte significativa são textos (Kuechler (2007)). A organização inteligente dessas coleções textuais é de grande interesse para a maioria das instituições, pois agiliza processos de busca e recuperação da informação. Nesse contexto, a Mineração de Textos permite a transformação desse grande volume de dados textuais não estruturados em conhecimento útil.

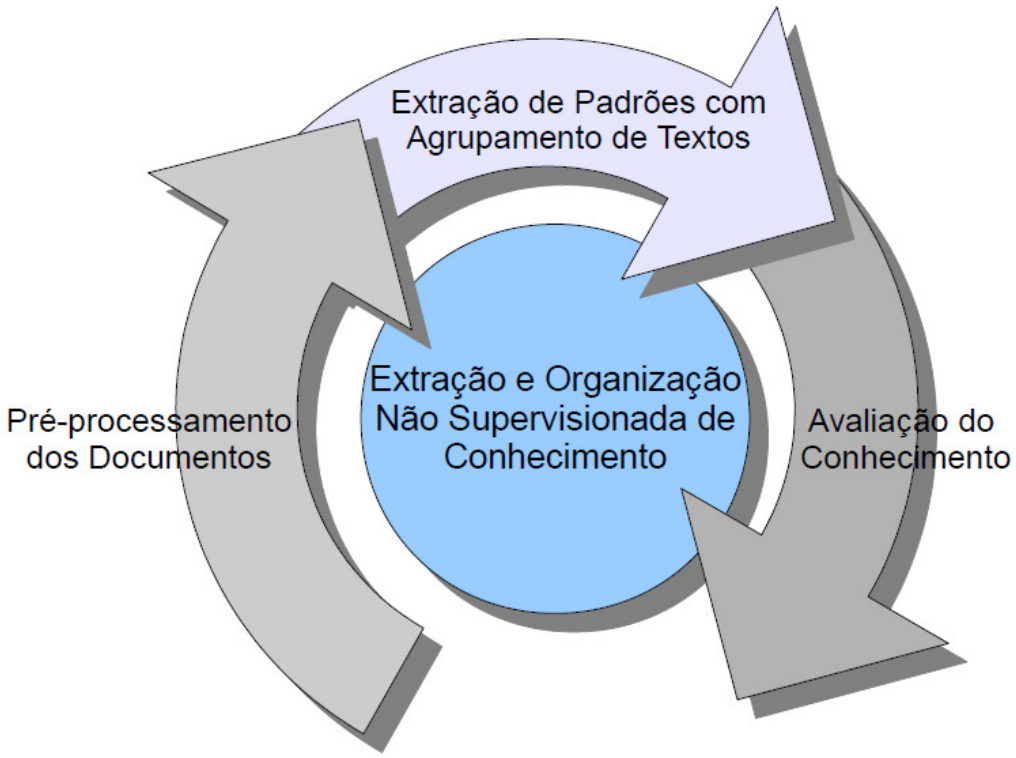
Entre as diversas maneiras de se instanciar um processo de Mineração de Textos, o uso de métodos não supervisionados para extração e organização de conhecimento recebe grande atenção na literatura, uma vez que não exigem conhecimento prévio a respeito das coleções textuais a serem exploradas. Um processo de Mineração de Textos para extração e organização não supervisionada de conhecimento pode ser dividido em três fases principais, com é ilustrado

na Figura 10: Pré-Processamento dos Documentos, Extração de Padrões com Agrupamento de Textos e Avaliação do Conhecimento.

No Pré-processamento dos Documentos os dados textuais são padronizados e representados de forma estruturada e concisa, em um formato adequado para extração do conhecimento. Assim, na extração de padrões, método de agrupamento (K-Means) de textos, descritos na seção 2.3.2.1 podem ser utilizados para a organização de coleções textuais de maneira não supervisionada (Feldman e Sanger (2006)). Em tarefas de agrupamento, o objetivo é organizar um conjunto de documentos em grupos, em que documentos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos documentos de outros grupos. Os métodos de agrupamento também são conhecidos como algoritmos de aprendizado por observação ou análise exploratória dos dados, pois a organização obtida é realizada por observação de regularidades nos dados, sem uso de conhecimento externo. Por fim, na Avaliação do Conhecimento, os resultados obtidos são avaliados de acordo com o contexto do problema, bem como a novidade e utilidade do conhecimento extraído.

Ao final desse processo, as coleções textuais são organizadas em grupos de documentos. Em especial, busca-se uma organização hierárquica da coleção, na qual os documentos são organizados em grupos e subgrupos, e cada grupo contém documentos relacionados a um mesmo tema. Os grupos próximos à raiz representam conhecimento mais genérico, enquanto seus detalhamentos, ou conhecimento mais específico, são representados pelos grupos de níveis mais baixos. Dessa forma, o usuário pode visualizar a informação de interesse em diversos níveis de granularidade e explorar interativamente grandes coleções de documentos. Os resultados obtidos por meio desse processo auxiliam diversas tarefas de organização da informação textual, partindo-se da hipótese que se um usuário está interessado em um documento específico pertencente a um grupo, deve também estar interessado em outros documentos desse grupo e de seus subgrupos.

Figura 10 – Fases para extração e organização não supervisionada de conhecimento.



Fonte: <https://goo.gl/XxPKM9>

Para a extração e organização não supervisionada de informação a partir de dados textuais, o diferencial está na etapa de extração de padrões, na qual são utilizados métodos de agrupamento de textos para organizar coleções de documentos em grupos. Em seguida, são aplicadas algumas técnicas de seleção de descritores para os agrupamentos formados, ou seja, palavras e expressões que auxiliam a interpretação dos grupos. Após validação dos resultados, o agrupamento hierárquico e seus descritores podem ser utilizados como uma hierarquia de tópicos para tarefas de análise exploratória dos textos, além de apoiar sistemas de recuperação de informação.

3.2.1 Agrupamento na Mineração de Textos

Com o objetivo de realizar a extração de padrões, após a representação dos textos em um formato estruturado, utiliza-se métodos de agrupamento de textos para obter a organização dos documentos.

O algoritmo k-means é o representante mais conhecido para agrupamento particional e muito utilizado em coleções textuais (Steinbach, Karypis e Kumar (2000)). Já no agrupamento hierárquico a maioria dos trabalhos relacionados com agrupamento hierárquico na literatura referenciam as estratégias aglomerativas, mostrando pouco interesse nas estratégias divisivas. A possível causa é a complexidade das estratégias divisivas, que cresce exponencialmente em

relação ao tamanho do conjunto de dados, proibindo sua aplicação em conjuntos de dados grandes.

Em um contexto no qual grande parte das informações estão armazenadas na forma textual, faz-se necessário o desenvolvimento de técnicas computacionais para a organização destas bases e a exploração do conhecimento nelas contido. Para tal fim, tarefas eficazes e eficientes de organização do conhecimento textual podem ser aplicadas. Dentre elas, destacam-se iniciativas para extração e organização do conhecimento de maneira não supervisionada, obtendo-se uma organização da coleção em grupos de documentos em temas e assuntos similares. Esta é a forma mais intuitiva de se estruturar o conhecimento para os usuários, uma vez que o agrupamento obtido fornece uma descrição sucinta e representativa do conhecimento implícito nos textos.

4 Conclusão

Este trabalho introduziu uma discussão sobre agrupamento de dados e três técnicas simples, porém importantes para introduzir muitos dos conceitos envolvidos no agrupamento de dados. São eles:

• K-means: Esta é uma técnica particional de agrupamento baseada em protótipos que tenta encontrar *(K)* números especificado pelo usuário e representa o número de grupos, que são representados pelos seus centroides.

• Agrupamento Hierárquico Aglomerativo: Esta abordagem de agrupamento se refere a um conjunto de técnicas de agrupamento intimamente relacionadas que produzem um agrupamento hierárquico iniciando com cada ponto como um grupo único e depois fundindo repetidamente os dois grupos mais próximos até que reste um único grupo englobando tudo. Algumas destas técnicas têm uma interpretação natural em termos de um agrupamento baseado em grafos, enquanto que outras têm uma interpretação em termos de uma abordagem baseada em protótipo.

• DBSCAN: Este é um algoritmo de agrupamento baseado em densidade que produz um agrupamento particional, no qual o número de grupos é determinado automaticamente pelo algoritmo. Pontos e regiões de densidade baixa são classificadas como ruído e omitidas; assim, o DBSCAN não produz um agrupamento totalmente completo.

A preferência pelas técnicas apresentadas, agrupamento hierárquico aglomerativo, K-means e DBSCAN se justifica pelo fato de serem bastante aceitos no meio acadêmico e científico e frequentemente serem usados como padrão em agrupamento de dados em relação a outros algoritmos. Estas técnicas são bastante usadas para o desenvolvimento de aplicações em diversas áreas. A discussão gira em torno da definição de agrupamento de dados, técnicas de agrupamento e alguma das diversas áreas de aplicações como na segmentação de imagens que em visão computacional, se refere ao processo de dividir uma imagem digital em múltiplas regiões (conjunto de *pixels*) ou objetos, com o objetivo de simplificar e/ou mudar a representação de uma imagem para facilitar a sua análise. Segmentação de imagens é tipicamente usada para localizar objetos e formas (linhas, curvas, entre outros) em imagens. O resultado da segmentação de imagens é um conjunto de regiões/objetos ou um conjunto de contornos extraídos da imagem. Como resultado, cada um dos pixels em uma mesma região é similar com referência a alguma característica ou propriedade computacional, tais como cor, intensidade, textura. Regiões adjacentes devem possuir diferenças significativas com respeito a uma mesma característica. Uma outra área de aplicação de agrupamento de dados, é na mineração de texto, conhecida também como mineração de dados textuais e semelhante à análise textual, refere-se ao processo de obtenção de informações importantes

de um texto. Informações importantes são obtidas normalmente pela elaboração de padrões e tendências. Geralmente a mineração de texto envolve o processo de estruturação do texto de entrada, de derivação de padrões dentro da estrutura de dados e, por fim, de avaliação e interpretação do resultado. Geralmente, "importante" em mineração de texto refere-se a algumas combinações de relevância, originalidade e interesse. Tarefas típicas de mineração de texto incluem categorização e agrupamento de texto, extração de conceito/entidade, entre outros.

A análise de texto envolve informações de recuperação, análise lexical a fim de estudar a frequência de distribuição de palavras, reconhecimento de padrões, identificação/anotação, extração de informações, técnicas de mineração de dados que incluem link e associação de análises, visualização e analítica preditiva. O objetivo maior é transformar o texto em dados para análise, por meio da aplicação do processamento de linguagem natural.

Embora técnicas de classificação (ou categorização) e análise de *cluster* tenham um resultado final similar, com a divisão de diferentes elementos em classes, ou agrupamentos, os métodos de agrupamento de dados são mais poderosos e complexos, uma vez que as categorias, ou agrupamentos, não são previamente determinados.

Quanto as limitações desse trabalho, não foram apresentadas outras técnicas utilizadas no agrupamento de dados, dificuldades em encontrar referências bibliográficas e uma falta de descrições mais detalhadas sobre exemplos de aplicações.

Como complemento do trabalho realizado e em continuidade ao estudo de agrupamento de dados, seria relevante abordar:

• Outros métodos de agrupamento de dados, em especial métodos que utilizem técnicas *fuzzy*, redes neurais ou algoritmos genéticos.

Referências

BARBARA, D. An introduction to cluster analysis for data mining. *Retrieved November*, v. 12, p. 2003, 2000. Citado 4 vezes nas páginas 24, 28, 29 e 30.

DINIZ, C. A. R.; NETO, F. L. *Data mining: uma introdução*. [S.l.]: ABE, 2000. Citado 2 vezes nas páginas 23 e 31.

DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification. 2nd. *Edition. New York*, p. 55, 2001. Citado na página 34.

ESTER, M. et al. Density-based spatial clustering of applications with noise. In: *Int. Conf. Knowledge Discovery and Data Mining*. [S.l.: s.n.], 1996. v. 240. Citado na página 36.

EVERITT, B. S.; DUNN, G. *Applied multivariate data analysis*. [S.l.]: Wiley Online Library, 2001. v. 2. Citado na página 23.

FACELI, K. *Um framework para análise de agrupamento baseado na combinação multi-objetivo de algoritmos de agrupamento*. Tese (Doutorado) — Universidade de São Paulo, 2006. Citado na página 23.

FACELI, K.; CARVALHO, A.; SOUTO, M. de. *Análise de dados de expressão gênica*. [S.l.], 2005. Citado na página 31.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado na página 23.

FELDMAN, R.; SANGER, J. Information extraction. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, p. 94–130, 2006. Citado na página 46.

GOLDSCHMIDT, R.; PASSOS, E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. *Rio de Janeiro: Campus*, v. 1, 2005. Citado na página 23.

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, Oxford University Press, v. 21, n. 15, p. 3201–3212, 2005. Citado na página 31.

HARTIGAN, P. Algorithm as 217: Computation of the dip statistic to test for unimodality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, JSTOR, v. 34, n. 3, p. 320–325, 1985. Citado na página 31.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice-Hall, Inc., 1988. Citado 6 vezes nas páginas 23, 24, 25, 28, 30 e 31.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, Acm, v. 31, n. 3, p. 264–323, 1999. Citado 3 vezes nas páginas 28, 29 e 30.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 2009. v. 344. Citado na página 23.

KOLOSSOSKI, G. Segmentação de imagens e algoritmo k-means. In: . [s.n.], 2007. p. 2–8. Disponível em: <https://goo.gl/GsJLvq>. Citado na página 40.

KUECHLER, W. L. Business applications of unstructured text. *Communications of the ACM*, ACM, v. 50, n. 10, p. 86–93, 2007. Citado na página 45.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citado na página 34.

MITCHELL, T. M. et al. *Machine learning. WCB*. [S.l.]: McGraw-Hill Boston, MA:, 1997. Citado na página 23.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informação da FSMA*, n, v. 7, p. 7–21, 2011. Citado na página 40.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering algorithms. In: *KDD-2000 Text Mining Workshop*. [S.l.: s.n.], 2000. Citado na página 47.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciência Moderna, 2009. Citado 4 vezes nas páginas 27, 33, 34 e 39.

XU, R.; WUNSCH, D. C. *Clustering. Hoboken*. [S.l.]: NJ: Wiley, 2009. Citado 2 vezes nas páginas 25 e 28.