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

RICARDO MELO DE MENDONÇA

Uma discussão sobre a análise de grupos em mineração de dados e suas aplicações

São Luís
2018

Ricardo Melo de Mendonça

Uma discussão sobre a análise de grupos em mineração de dados e suas aplicações

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
2018

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

Mendonça, Ricardo Melo de.
Uma discussão sobre a análise de grupos em mineração de dados e suas aplicações / Ricardo Melo de Mendonça. - 2018.
60 f.

Orientador(a): Ivo José da Cunha Serra.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, Auditório do DEINF CCET, 2018.

1. Algoritmos de agrupamento. 2. Análise de Grupos. 3. Mineração de Dados. I. Serra, Ivo José da Cunha. II. Título.

RICARDO MELO DE MENDONÇA

Uma discussão sobre a análise de grupos em mineração de dados e suas aplicações

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. Ivo José da Cunha Serra, Dr

Trabalho aprovado, São Luís – MA, 6 de abril de 2018:

BANCA EXAMINADORA

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

Prof.ª Simara Vieira da Rocha, Dra.
(Membro da Banca Examinadora)

Prof. Carlos Eduardo Portela Serra de Castro, M.Sc
(Membro da Banca Examinadora)

São Luís
2018

AGRADECIMENTOS

Inicialmente, agradeço a Deus, por todas as bênçãos conquistadas e por ter me dado forças para concluir este trabalho.

Aos meus amados pais, Antonio e Ana, meus guerreiros, minha base familiar, pelo amor, carinho, dedicação e apoio. Por terem me proporcionado sem medir esforços essa conquista, e principalmente por terem acreditado em meu potencial.

A minha irmã Nikole, por estar ao meu lado sempre, nos momentos felizes e tristes, estando além da convivência mútua, nos eternos laços de amor.

A minha companheira Vanessa, pela força, carinho e sempre ao meu lado oferendo o apoio, tanto em momentos bons quanto ruins.

Ao meu orientador, professor Ivo que me acompanhou em parte do Curso e me orientou na elaboração deste trabalho, sempre com disponibilidade e atenção.

A todos os familiares e aos amigos próximos, que me ajudaram de forma direta e indireta na concretização deste trabalho.

”Perdoai as nossas ofensas assim como nós perdoamos a quem nos tem ofendido”

RESUMO

Este trabalho discute sobre a tarefa de análise de grupos em mineração de dados e as suas aplicações. Grandes quantidades de dados são gerados todos os dias e bases de dados com terabytes ou petabytes já são comuns. A mineração de dados surgiu com o propósito de extrair e auxiliar o analista de dados a identificar informações relevantes nessas bases de dados. Este trabalho apresenta conceitos de descoberta de conhecimento em base de dados e da técnica de agrupamento de dados, que faz agrupamentos automáticos sem conhecimento prévio da base de dados, também serão apresentados três técnicas para a realização destes agrupamentos e uma discussão comparativa entre elacomparaçãos. Em seguida os conceitos e técnicas apresentados serão utilizados para realizar uma análise de grupos em duas aplicações diferentes, identificando e discutindo os pontos fortes e fracos de cada técnica para cada domínio.

Palavras-Chave: Mineração de Dados, Análise de Grupos, Algoritmo K-Means, Algoritmo DBSCAN, Aplicações de Agrupamento.

ABSTRACT

This work discuss the data mining Clustering task and its applications. Great quantities of data are generated every day and data bases with terabytes or petabytes are common nowadays. Data Mining has come with the purpose of extract and assist the data analist to identify relevant information on database. This work presents concepts of knowlegde discovery on database and the Clusering technique, which makes automatic groupings without previous knowledge of the database, three techniques will also be presented for the accomplishment of these clusterings and also a comparative discussion between them. Then the concepts and techniques presented will be used to perform a group analysis in two different applications, identifying and discussing the strengths and weaknesses of each technique for each domain.

Keywords: Data Mining, Clustering, K-Means, DBSCAN, Clustering Application.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DBSCAN	*Density Based Spatial Clustering of Applications with Noise*
Eps	Raio de vizinhança
GPS	*Global Positioning System*
KDD	*Knowledge Discovery in Databases*
MinPts	Número mínimo de pontos
OMT	Organização Mundial do Turismo
ToPI	*Tourist Place Identification*

SUMÁRIO

1 INTRODUÇÃO

Avanços rápidos na tecnologia de coleta e armazenamento de dados permitiram que as organizações acumulassem uma grande quantidade de dados. No entanto, a extração de informação útil é cada vez mais difícil (TAN, 2006). Sendo assim, existe uma necessidade urgente para uma nova geração de teorias computacionais e ferramentas para auxiliar humanos a extrair informação útil dos crescentes volumes de dados digitais (FAYYAD, 1996).

A mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados (TAN, 2006). Larose (2005) define que a mineração de dados geralmente é classificada de acordo com a sua capacidade em realizar determinadas tarefas. E também, que as tarefas mais comuns são: agrupamento, classificação, descrição, regressão, predição e associação.

Classificar é uma das mais primitivas atividades do ser humano (ANDERBERG, 1973; EVERITT et al., 2001) e representa um papel importante e indispensável para a longa história do desenvolvimento humano. A fim de aprender um novo objeto ou entender um novo fenômeno, as pessoas sempre tentam identificar características descritivas e em seguida comparar tais características com aquelas de objetos e fenômenos já conhecidos, baseado nas suas similaridades ou dessimilaridades, generalizando como proximidade, levando em consideração certos padrões ou regras.

Exemplificando, todos os objetos naturais são basicamente classificados entre três grupos: animal, vegetal, e mineral. De acordo com a taxonomia biológica, todos os animais são então classificados em categorias de reino, filo, classe, ordem, família, genus, e espécie, de geral a específico. Com esta classificação em mãos, podemos inferir as propriedades de um objeto específico de acordo com a sua categoria. Por exemplo, quando vemos um leão marinho deitado na terra, nós sabemos imediatamente que ele é um bom nadador sem realmente olhar ele nadando (XU E WUNCH, 2009).

A prática de classificar objetos de acordo com as suas similaridades perceptíveis é a base de boa parte da ciência, e organizar dados em grupos de acordo com estas classificações é uma das maneiras mais fundamentais que o homem utiliza

para entender e aprender. A análise de grupos é o estudo formal dos algoritmos e métodos para agrupar, ou classificar, objetos. Um objeto é descrito tanto por seu conjunto de medidas ou pelo relacionamento entre o objeto e outros objetos (JAIN E DUBES, 1988).

A análise de grupos pode ser vista como pertencente ao paradigma de aprendizado não supervisionado, em que o aprendizado é dirigido aos dados, não necessitando de conhecimento prévio sobre as suas classes e categorias (MITCHELL et al., 1997).

Na classificação não supervisionada, também denominada de agrupamento ou análise de dados exporatória, nenhum dado rotulado está disponível (EVERITT et al., 2001; JAIN e DUBES, 1988).

O intuito final do agrupamento é distinguir um conjunto finito e não rotulado de dados, separando de um conjunto finito e discreto de estruturas "naturais" ocultadas, em vez de fornecer uma caracterização precisa de amostras não analisadas de alguma distribuição probabilística (XU e WUNCH, 2009).

Existem diferentes métodos para alcançar o objetivo da análise de grupos. Neste trabalho será apresentada uma discussão sobre os principais conceitos e algoritmos de agrupamento de dados, sendo exemplificados através de duas aplicações, uma no domínio da mineração de dados educacionais, outra para identificar pontos de interesse baseado em fotografias geo-referenciadas.

1.1 Motivação

Duda et al (2001) enumera pelo menos cinco razões básicas a respeito da motivação em utilizar procedimentos não supervisionados.

A primeira razão, diz que colher e rotular um amplo conjunto de dados de amostra pode ser surpreendentemente custoso. Já a segunda razão, pode ser preferível proceder pelo caminho inverso: treinar com grandes conjuntos de dados não classificados, e só então utilizar procedimentos supervisionados para classificar os objetos aos grupos encontrados. A terceira razão, que em várias aplicações os padrões podem mudar lentamente com o tempo, logo, se estas alterações puderem ser acompanhadas por um classificador em modo não supervisionado, um

desempenho melhor pode ser alcançado. Como quarta razão, podemos utilizar métodos não supervisionados para achar características, que então serão utilizáveis para a categorização. Finalmente, a quinta razão, aponta que nas primeiras etapas de uma investigação a respeito de um conjunto de dados, pode ser de grande importância e utilidade obter algum esclarecimento a respeito da natureza ou estrutura dos dados.

Este trabalho é motivado pelo interesse em discutir algumas aplicações da análise de grupos, que é um procedimento não supervisionado, dentro do contexto da mineração de dados, e a análise da teoria utilizada em cada uma destas aplicações. E também, tem a motivação de disponibilizar esta discussão para que sirva de referencial para futuros trabalhos na área acadêmica ou de mercado que tenham a necessidade de fazer uso de técnicas de agrupamento de dados.

1.2 Objetivo do Trabalho

O objetivo principal deste trabalho é realizar uma discussão, levando em consideração aspectos teóricos e práticos, sobre a análise de grupos em mineração de dados.

Os objetivos específicos são:

- Apresentar os principais métodos de agrupamento e tipos de grupo
- Introduzir algumas das técnicas de agrupamento e dois algoritmos básicos de agrupamento particional, K-Means e DBSCAN *(Density Based Spatial Clustering of Applications with Noise)*
- Apresentar duas aplicações do agrupamento de dados, evidenciando os motivos da adoção de diferentes tipos de algoritmos em cada uma delas.

1.3 Trabalhos Relacionados

Autores como Jain e Dubes (1988), Duda et al (2001), e Tan (2006), Theodoridis e Koutroumbas (2008) e Xu e Wunch (2009) apresentam em seus livros, conceitos, técnicas e domínios de uso da análise de grupos. Entre as obras citadas,

algumas são específicas sobre agrupamento, e outras são focadas em outras áreas, como na de reconhecimento de padrões e na área de mineração de dados. Ainda que aguns destes livros não sejam primariamente a respeito da análise de grupos, seus conceitos são apresentados, exemplificados e suas características listadas, e assim como este trabalho, promovem uma discussão sobre o agrupamento de dados e as suas utilizações.

O trabalho de Berkhin (2006) teve como principal objetivo fornecer uma revisão para compreender as diversas técnicas de agrupamento em mineração de dados, além de evidenciar os pontos fortes e fracos de cada uma delas, e indicar diferentes propriedades para a avaliação das técnicas de agrupamento, fazendo referências aos tipos de dados e áreas em que cada técnica apresenta maior afinidade. Assim como este trabalho, visa apresentar uma discussão a respeito do agrupamento de dados e o relacionamento das suas técnicas e métodos com diferentes aplicações.

Faz-se menção ainda, o autor Jain et al. (1999) que além de fazer uma revisão dos métodos e técnicas de agrupamento, também apontou aplicações nas áreas de segmentação de imagens, reconhecimento de objetos e de escrita a mão livre, recuperação de informação e mineração de dados. Cada uma das aplicações foram apresentadas com o intuito de evidenciar a utilização de técnicas de agrupamento em diferentes áreas, reforçando que a análise de grupos é uma tarefa com grande abrangência.

1.4 Organização do Trabalho

Este trabalho está organizado em mais três capítulos, sendo o capítulo dois, nomeado por "Fundamentação Teórica sobre Análise de Grupos", no qual serão apresentados os conceitos a respeito da análise de grupos, seus tipos de agrupamento e tipos de grupo, dois algoritmos de agrupamento, o K-Means e o DBSCAN, e uma comparação entre eles. Estes conceitos, técnicas e comparações são essenciais para a compreensão da discussão proposta no capítulo posterior.

Em seguida, no capítulo três, nomeado por "Aplicações de Análise de Grupos", para ilustrar, serão apresentadas duas aplicações, uma no domínio da mineração de dados educacionais, e outra para a identificação de locais de interesse a partir de

fotografias geo-referenciadas, em que ambas utilizaram métodos de agrupamento dedados para atingir seus objetivos. Além disso, de forma concomitante com a apresentação das aplicações no capítulo três, é proposta uma discussão à respeito das técnicas utilizadas para cada aplicação.

Finalmente, no capítulo quatro, serão feitas conclusões a respeito deste estudo, algumas das suas limitações e a apresentação de sugestões para trabalhos futuros.

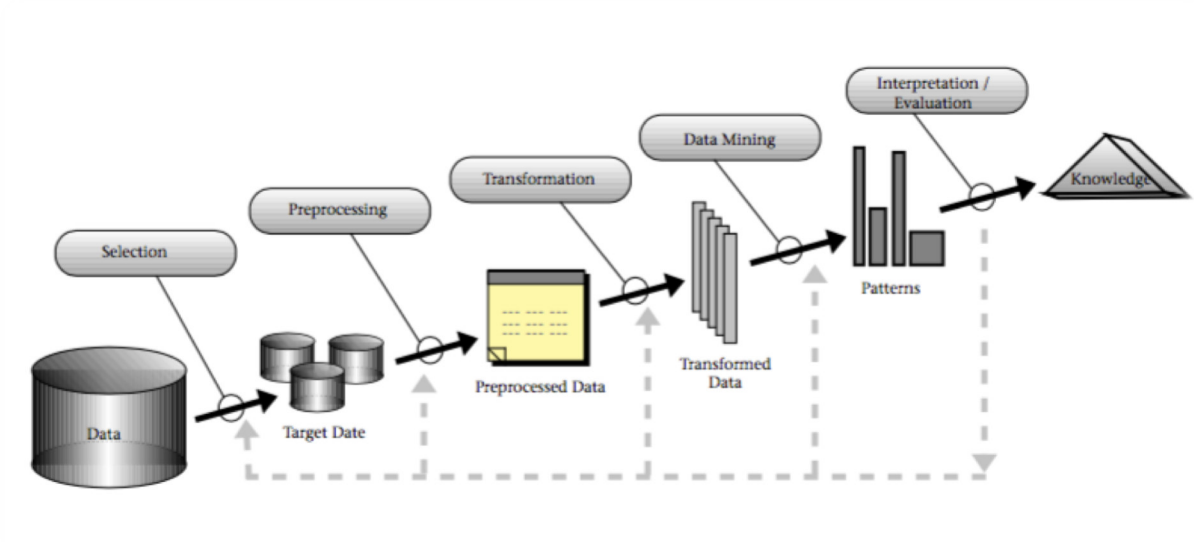
2 FUNDAMENTAÇÃO TEÓRICA SOBRE ANÁLISE DE GRUPOS

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

2.1 Descoberta de Conheciemento em Bases de Dados

A descoberta de conhecimento em bases de dados (KDD – *Knowledge Discovery in Databases*) é o processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados, e pode ser representado conforme a figura 1. Aqui, o padrão é uma expressão em alguma linguagem que descreve um subconjunto ou um modelo que se aplica a algum subconjunto dos dados. O termo processo implica que KDD é composto de várias etapas. Por ser não trivial, é entendível que algum grau de busca ou inferência é envolvido, que não é um processo computacional direto, como computar uma média de um conjunto de números (FAYYAD, 1996).

Figura 1 - Etapas do processo KDD



Fonte: (FAYYAD, 1996)

Segundo Tan (2006), a mineração de dados é uma das etapas do KDD, e possui diferentes tarefas, entre elas: análise de grupos, classificação, descrição, regressão, predição e associação.

Antes de entrarmos em detalhes a respeito da análise de grupos, é pertinente entender que esta é uma tarefa da mineração de dados, que por sua vez é parte do KDD (TAN ,2006).

2.2 Análise de Grupos

Segundo Jain e Dubes (1988), a análise de grupos é o estudo formal dos métodos e algoritmos de agrupamento de objetos, que são descritos por um conjunto de medidas ou pelo seu relacionamento entre outros objetos. Diferente da tarefa de classificação, a análise de grupos não utiliza classes e definições previamente estabelecidas e não tem a pretensão de criar regras para uma separação de dados futuros, mas apenas de organizar de forma válida e conveniente os dados já existentes. O objetivo do agrupamento é separar um conjunto finito de dados não rotulados em estruturas de dados "naturais" escondidas (XU e WUNCH, 2009).

Um grupo é composto de um conjunto de objetos que apresentam uma proximidade entre si. O autor Everitt (1974) apresenta ainda algumas definições de grupo, podendo ser um conjunto de entidades parecidas entre si, e diferentes de entidades de outros grupos. Ou um grupo é uma agregação de pontos no espaço, tal que a distância entre dois pontos dentro de um grupo é sempre menor que a distância entre um ponto de um grupo com um ponto de outro grupo. Ou grupos podem ser definidos como regiões conectadas em um espaço multidimensional contendo uma alta densidade relativa de pontos, separado de outros grupos por regiões com baixa densidade.

Ademais, as duas últimas definições de grupo implicam que os dados podem ser representados como pontos no espaço, e embora seja possível fazer essa separação de grupos de forma intuitiva, ainda não é claro como o fazemos, o que torna muito difícil uma formalização única e definitiva. Além disso, dados podem ser agrupados diferentemente de acordo com o propósito, os grupos podem ter diversos formatos e podem mudar com o passar do tempo.

Xu e Wunch (2009) salienta que o agrupamento é um processo subjetivo, o que pede mais cuidado quando for realizada uma análise de grupo em cima dos dados, a obra orienta também, que para o mesmo conjunto de dados, diferentes objetivos geralmente levam à diferentes partições.

Exemplificando, a saber que na partição de animais, quais sejam, uma águia, um leão, um cardeal, uma pantera e uma rã. Se eles são divididos pelo critério de voar, teríamos dois grupos, um com a águia e o cardeal e outro grupo com o restante dos animais. No entanto se o critério for o da alimentação, se carnívoro ou não, teríamos uma divisão bem diferente da anterior. Técnicas de agrupamento oferecem várias vantagens em comparação ao processo de agrupamento manual.

A primeira vantagem, um programa de agrupamento pode aplicar um critério com objetivo específico de forma consistente para particionar os grupos. Seres humanos são excelentes em procurar grupos em duas ou três dimensões, mas pessoas diferentes nem sempre identificam os mesmos grupos nos dados. A medida de proximidade que define a similaridade entre os objetos depende da educação e da bagagem cultural do indivíduo. E ainda, é bastante comum para diferentes pessoas criarem agrupamentos diferentes se os grupos não forem bem separados.

A segunda vantagem, um algoritmo de agrupamento consegue formar grupos em uma fração de tempo que demoraria no agrupamento manual, possibilitando assim que o profissional aproveite melhor o seu tempo analisando ou interpretando os resultados obtidos. A velocidade, confiabilidade e consistência do algoritmo de agrupamento são motivos mais que suficientes para a sua utilização (JAIN e DUBES, 1988).

Segundo Xu e Wunch (2009), o procedimento de análise de grupos é descrito em quatro etapas, podendo ser visualizadas na figura 2.

Na etapa da Seleção ou Extração de atributos, como apontado por Jain et al. (1999) e Bishop (1995), a seleção de atributos consiste na escolha de características distintas de um conjunto de dados, enquanto a extração utiliza alguma transformação para gerar atributos úteis a partir dos originais.

Claramente, a extração de atributos é potencialmente capaz de produzir atributos mais eficientes para descobrir a estrutura dos dados. No entanto, a extração pode gerar atributos que não são fisicamente interpretáveis, enquanto na seleção é assegurada a retenção do significado original. Na literatura, estes dois termos são

muitas vezes utilizados antes de ser feita esta distinção. Ambas são muito importantes para a efetividade da aplicação de agrupamento. Uma seleção adequada ou uma geração de atributos significativos para representar o conjunto pode diminuir consideravelmente os custos de armazenamento e medição de proximidade entre os objetos, simplificando as seguintes etapas do procedimento e facilitando o entendimento dos dados. Geralmente, os atributos ideais devem ser úteis para distinguir padrões pertencentes à grupos diferentes, imunes à ruídos, e fáceis de se obter e interpretar.

A etapa da Seleção ou Projeto do algoritmo de agrupamento, se baseia em apontar uma medida de proximidade adequada e construir a sua função de critério. Por percepção, os objetos são agrupados em diferentes grupos de acordo com o quanto eles se assemelham ou não, já que quase todos os algoritmos de agrupamento são explicitamente ou implicitamente ligados a alguma definição em particular de medida de proximidade, e alguns algoritmos até trabalham diretamente em uma matriz de proximidade.

Visto que, a partir do momento em que uma medida de proximidade é adotada, o agrupamento pode ser visto como um problema de descobrir a melhor solução de todas as soluções viáveis. Além disso, os grupos obtidos são dependentes da seleção da função de critério, sendo assim, a subjetividade é característica inerente da análise de grupos.

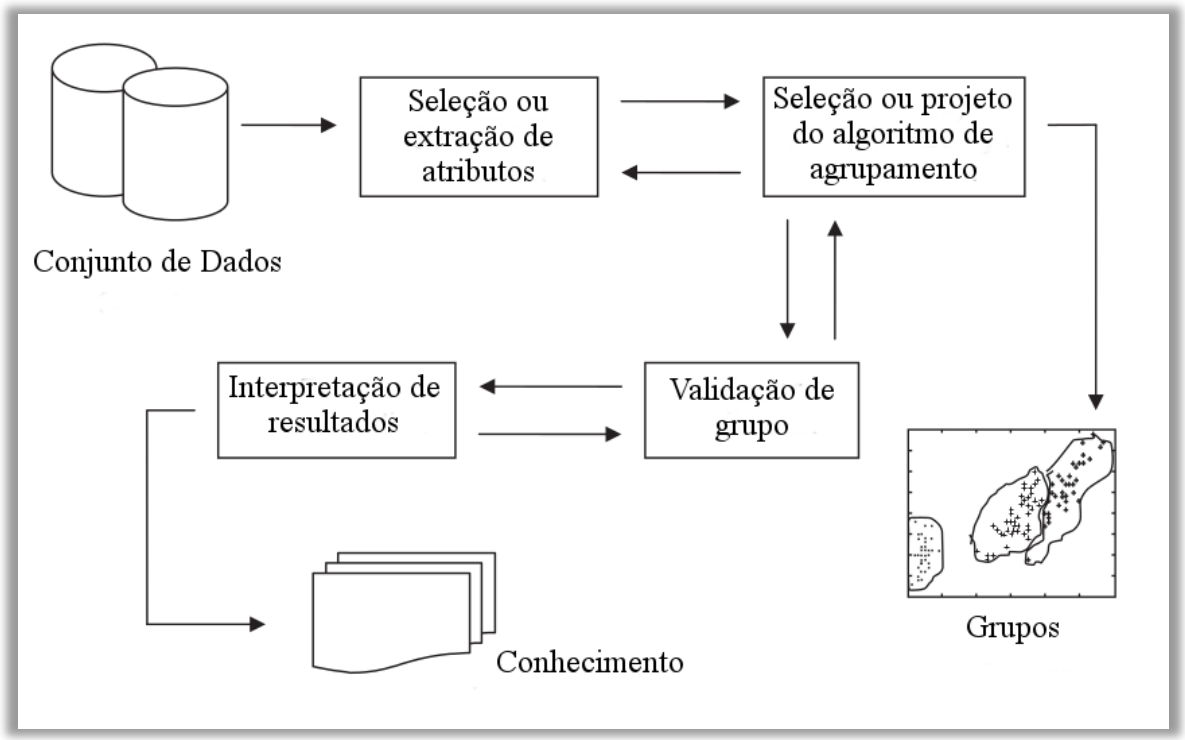
O agrupamento é utilizado em diferentes domínios de aplicação, e uma gama de algoritmos têm sido desenvolvidos para resolver diversos problemas em várias áreas. No entanto, não existe um algoritmo de agrupamento universal para resolver todos os problemas, por isso, é muito importante investigar cuidadosamente as características do problema para selecionar a estratégia apropriada de agrupamento. Nesse sentido, algoritmos de agrupamento que são desenvolvidos para resolver um problema particular em uma área especializada, geralmente partem de pressupostos a favor da aplicação de interesse. Por exemplo, o algoritmo K-means é baseado na medida Euclidiana e consequentemente tende a gerar grupos esféricos, no entanto, se os grupos reais estão em outras formas geométricas, o K-means pode não ser efetivo, e precisaríamos recorrer a outros modelos. Esta discussão será mais aprofundada com o auxílio de exemplos no capítulo três.

Em seguida, quanto a etapa da Validação de Grupo, recebido um conjunto de dados, cada algoritmo de agrupamento sempre pode produzir uma partição, independentemente de existir ou não esta estrutura em particular nos dados. Ainda mais, diferentes abordagens de agrupamento normalmente levam a diferentes grupos de dados, e até mesmo para o mesmo algoritmo, o ajuste dos parâmetros pode afetar o resultado final. Por isso, o problema de avaliar se a estrutura de dados encontrada é significativa, também chamado de validação de grupo, é particularmente importante (GORDON, 1998; HALKIDI ET AL, 2002; JAIN e DUBES, 1988). Por exemplo, se no conjunto de dados não estão presentes nenhuma estrutura de agrupamento, a saída de um algoritmo de agrupamento perde seu significado, sendo assim, é necessário realizar algum tipo de teste para assegurar a existência da estrutura de agrupamento antes de se fazer mais análises. Deste modo, Gordon (1998) e Jain e Dubes (1988) consideram alguns testes para situações em que não existam estruturas no grupo de dados, mas que não são muito utilizadas por que os usuários normalmente são confiantes da presença de grupos nos dados de interesse.

Logo após, tem-se a etapa da Interpretação de Resultados, que tem como objetivo final, a entrega ao usuário de uma compreensão significativa dos dados originais para que ele possa desenvolver um claro entendimento destes dados, e então efetivamente resolver os problemas encontrados. Em vista disso, Anderberg (1973) viu a análise de grupos como "um dispositivo para sugerir hipóteses", e também indicou que "um conjunto de grupos não é por si só um resultado finalizado, mas apenas um possível esboço". Dessa maneira, especialistas nos respectivos domínios são encorajados a interpretar a partição dos dados, integrando outras evidências experimentais e informações do domínio, sem restringir as suas observações e análises a algum resultado de agrupamento específico. Consequentemente, posterior análise e experiências podem ser necessárias.

É interessante observar na figura 2, que o fluxo também inclui um caminho de volta, sendo assim, a análise de grupos não é um processo de uma só execução. Em muitas circunstâncias, o agrupamento requer uma série de tentativas e repetições. Além disso, não existe um critério universal e efetivo para guiar a seleção de atributos e esquema de agrupamento. Um critério de validação dá algum esclarecimento a respeito da qualidade da solução, mas até escolher o critério apropriado é um problema que demanda esforço.

Figura 2 - Procedimento de agrupamento



Fonte: Adaptado de (XU e WUNCH, 2009)

2.2.1 Diferentes tipos de agrupamento

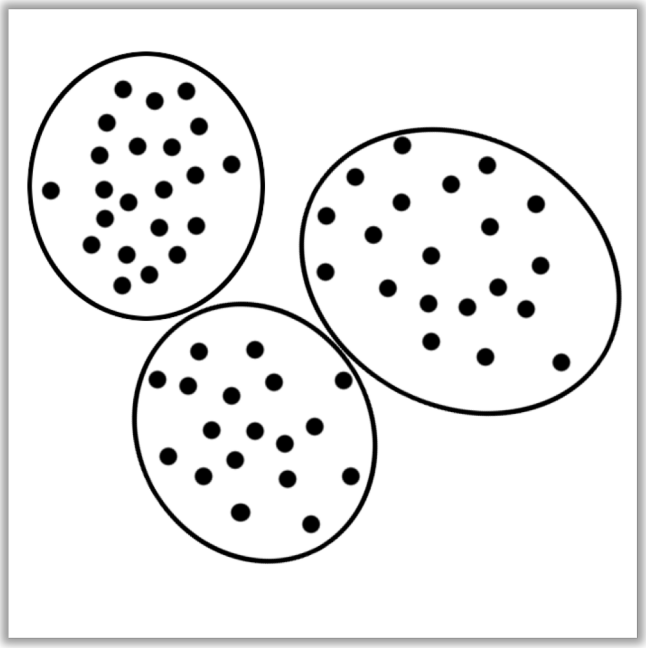
Um conjunto de grupos é chamado comumente de agrupamento e, nesta seção, fazemos a distinção de diversos tipos de agrupamentos: particional, hierárquico, exclusivo, difuso, completo ou parcial.

2.2.1.1 Agrupamento Particional

Um agrupamento particional é uma divisão do conjunto de objetos de dados em subconjuntos (grupos) que não se interseccionam, de modo que cada objeto de dado esteja exatamente em um subconjunto (TAN, 2006). A figura 3, ilustra um

agrupamento particional que resultou em três grupos distintos, e nenhum ponto está em mais de um grupo.

Figura 3 – Agrupamento particional



Fonte: Produzido pelo autor

Segundo Jain e Dubes (1988), o problema do agrupamento particional pode ser formalmente estabelecido como segue: dado *N* padrões em um espaço métrico *d*-dimensional, determinar uma partição destes padrões em *K* grupos, de forma que os padrões dos objetos em um grupo sejam mais similares entre si do que com os padrões em grupos diferentes. O Valor de *K* pode ser, ou não, especificado a priori.

Desta maneira, a solução teórica para o problema particional é direta: selecionar o critério, avaliar este critério para todas as suas possíveis partições contendo *K* grupos, e escolher a partição que otimiza o critério. Assim, a primeira dificuldade encontrada é selecionar um critério que represente as noções intuitivas sobre "grupo" em uma fórmula matemática. Desse modo, o critério deve ser simples por motivos computacionais, mas complexo o bastante para refletir as estruturas de dados existentes. Além disso, outro problema que acompanha o uso de algoritmos particionais é a escolha da quantidade de grupos de saída.

Métodos particionais têm vantagem em aplicações envolvendo grandes quantidades de dados, em que a construção de um dendograma, como no método de

agrupamento hierárquico que será apresentado na seção 2.2.1.2, é computacionalmente proíbitivo (JAIN et al, 1999). Ainda que se conseguisse produzir este diagrama, ele seria impraticável para análise.

Diversos trabalhos enfatizam que a enumeração simples de todas as possíveis partições não é computacionalmente praticável, mesmo para pequenos conjuntos de dados (JAIN e DUBES, 1988; XU e WUNCH, 2009; BERKHIN, 2006). De fato, se analisarmos que existem 34.106 partições distintas entre 10 objetos em quatro grupos, e que esse número sobe para 11.259.666.000 se 19 objetos forem divididos em quatro grupos, esta declaração se torna óbvia.

Certamente, uma abordagem para se particionar os dados é utilizar a definição de função objetiva, em que distâncias ou similaridades podem ser utilizadas para computar medidas de relação dos objetos em um grupo e entre grupos (BERKHIN, 2006). Métodos de otimização iterativa são largamente utilizados para encontrar soluções, e um dos mais utilizados é o algoritmo K-means, baseado na soma do erro quadrado (SSE), que será discutido na seção 2.3.1.

Como vantagens do agrupamento particional pode-se citar:

- São apropriados para aplicações com grandes quantidades de dados (JAIN et al., 1999).
- Por ter uma função objetiva para calcular a proximidade entre os objetos, agrupamentos particionais podem ser vistos como um problema de otimização (problema de encontrar a melhor solução de todas as soluções viáveis) (BERKHIN, 2006).

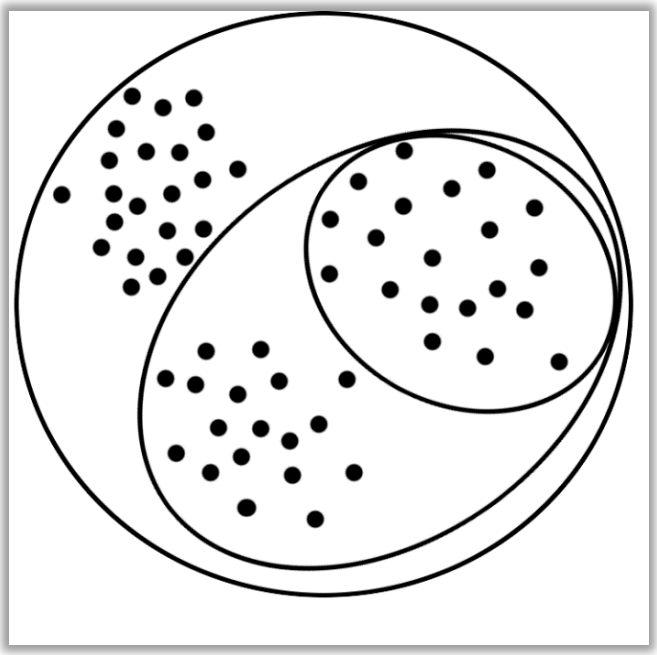
E como desvantagens pode-se citar:

- Métodos de agrupamento particional normalmente assumem que o número de grupos a serem encontrados é conhecido (XU e WUNCH, 2009).
- Selecionar um critério que traduza as definições de "grupo" e formular a função objetiva representa uma dificuldade (XU e WUNCH, 2009).

2.2.1.2 Agrupamento Hierárquico

O agrupamento hierárquico possui uma filosofia diferente do agrupamento particional, no sentido que, enquanto algoritmos particionais analisam os dados dividindo em grupos inteiramente separados, os algoritmos hierárquicos dividem os dados em grupos aninhados (THEODORIDIS e KOUTROUMBAS, 2008). Desta forma, constrói uma hierarquia de grupos, uma árvore, também conhecida como dendograma. A figura 4 mostra um agrupamento hierárquico que resultou em três grupos, porém existe uma relação de hierarquia entre os grupos, sendo que o maior grupo engloba os dois grupos menores, e dentro do grupo intermediário está contido o grupo menor.

Figura 4 – Agrupamento hierárquico.



Fonte: Produzido pelo autor

Jain e Dubes (1988) faz as seguintes definições e considerações a respeito do agrupamento hierárquico. Normalmente é conveniente caracterizar o método escrevendo um algoritmo, mas o algoritmo deve ser separado do método em si. Primeiro vem a noção de sequência de partições aninhadas. Os *n* objetos a serem agrupados são denotados pelo conjunto X.

$$X = \{x_1, x_2, \ldots, x_n\}$$

Onde x_i é o i-ésimo objeto. Uma partição C, de X, quebra X em subconjuntos $\{c_1, c_2, ..., c_m\}$. C satisfaz as seguintes condições:

$$c_i \cap c_j = \emptyset \text{ para } i \text{ e } j \text{ de 1 a } m, \text{ i} \neq \text{j}$$

$$c_1 \cup c_2 \cup \ldots \cup c_m = X$$

Nesta notação, "∩" é a interseção, "∪" é a união e "Ø" representa um conjunto vazio. O agrupamento é uma partição; os componentes da partição são chamados de grupos. Uma partição B está aninhada na partição C se todos os componentes de B forem um subgrupo de um componente de C. Isto é, C é formado pela fusão dos componentes de B. Por exemplo, se o agrupamento C com três grupos e o agrupamento B com cinco grupos são definidos como a seguir, então B está aninhado em C. Ambos C e B são agrupamentos de um conjunto de objetos $\{x_1, x_2, \ldots, x_{10}\}$.

$$C = \{(x_1, x_3, x_5, x_7), (x_2, x_4, x_6, x_8), (x_9, x_{10})\}$$

$$B = \{(x_1, x_3), (x_5, x_7), (x_2), (x_4, x_6, x_8), (x_9, x_{10})\}$$

Nem C e nem B estão aninhados na seguinte partição, e esta partição não está aninhada nem em C ou em B.

$$\{(x_1, x_2, x_3, x_4), (x_5, x_6, x_7, x_8), (x_9, x_{10})\}$$

O agrupamento hierárquico é uma sequência de partições em que cada uma das partições está aninhada na próxima partição da sequência, e pode ser classificado de acordo com a sua metodologia de construção destas partições, como sendo aglomerativo ou divisivo (JAIN e DUBES, 1988; KAUFMAN e ROUSSEEUW, 1990). Desta forma, um algoritmo aglomerativo para agrupamentos hierárquicos começa com os grupos disjuntos, que coloca cada um dos *n* objetos em um grupo individual. O algoritmo de agrupamento a ser empregado dita como a matriz de proximidade, calculada a partir da medida de proximidade entre os objetos, deve ser interpretada para fundir dois ou mais desses grupos triviais, então aninhando o agrupamento trivial em uma segunda partição.

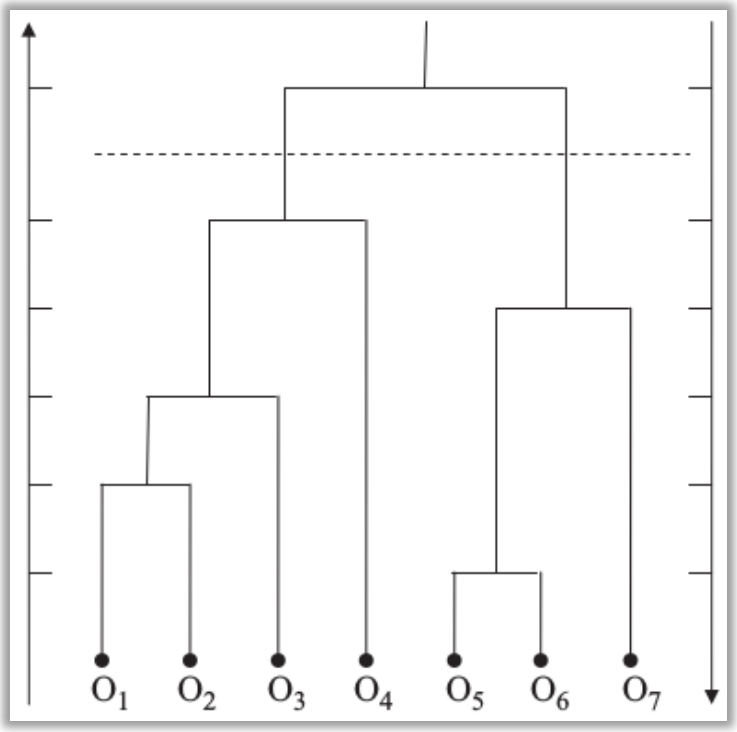
O processo é repetido para formar uma sequência de agrupamentos aninhados em que o número de grupos diminui conforme a sequência progride, até que reste apenas um único grupo contendo todos os *n* objetos. Um algoritmo divisivo realiza

esta tarefa na ordem inversa, começando com apenas um conjunto contendo todos os *n* objetos e terminando com cada um dos objetos em um grupo individual.

Tanto o método aglomerativo quanto o divisivo organizam os dados em uma estrutura hierárquica representável em um dendograma, como ilustrado na figura 5, onde cada "O" representa um único objeto no conjunto de dados. O nó raiz do dendograma representa todo o conjunto de dados, e cada nó folha representa um único objeto dos dados. Os nós intermediários descrevem o quanto os objetos estão próximos uns dos outros; e a altura do dendograma expressa a distância entre cada par de pontos de dados ou grupo, ou um ponto de dado e um grupo. O resultado final do agrupamento pode ser obtido cortando o dendograma em diferentes níveis, a exemplo da linha tracejada na figura 5 (XU e WUNCH, 2009).

Esta representação provê descrições muito informativas e uma visualização do potencial das estruturas de agrupamento de dados, especialmente quando relações de hierarquia realmente existem nos dados, como em dados de pesquisas evolucionárias em diferentes espécies de organismos, ou outras aplicações em medicina, biologia e arqueologia (EVERITT et al., 2001; THEODORIDIS e KOUTROUMBAS, 2008).

Figura 5 - Exemplo de um dendograma de um agrupamento hierárquico.



Fonte: (XU e WUNCH, 2009)

Em comparação com os métodos aglomerativos, os métodos divisivos precisam considerar $2^{N-1} - 1$ divisões possíveis em dois grupos para um conjunto de dados com *N* pontos, o que é computacionalmente muito custoso. Por isso, métodos aglomerativos são mais comumente utilizados. Ainda assim, 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 e WUNCH, 2009).

Como vantagens do agrupamento hierárquico pode-se citar:

- Flexibilidade em lidar com diferentes tipos de dados, sejam dados discretos, contínuos, intervalos ou nominais, contanto que seja possível criar uma matriz de proximidade.
- O resultado de um algoritmo hierárquico normalmente gera um dendograma, que mostra com clareza a distância entre os objetos e grupos (BARBARA, 2000).
- A quantidade de grupos não é predefinida e o número de grupos desejado pode ser obtido ao cortar o dendograma na altura apropriada (BARBARA, 2000).
- Segundo Jain e Dubes (1988), agrupamentos hierárquicos produzem grupos de melhor qualidade.

E como desvantagens pode-se citar:

- Critério vago de finalização por não possuir uma função objetiva (BERKHIN, 2006; BARBARA, 2000).
- O fato de que a maioria dos algoritmos hierárquicos não revisitam os grupos construídos com o propósito de melhoramento (BERKHIN, 2006).
- Alta complexidade computacional, de pelo menos $O(N^2)$, que pode ser proíbitiva para algumas aplicações.
- Sensível a ruídos e elementos externos (dados legítimos, mas que possuem comportamento anormal) (BARBARA, 2000).
- Tende a dividir grupos grandes (BARBARA, 2000).

2.2.1.3 Agrupamento Exclusivo, Não-exclusivo e Difuso

No que diz respeito a relação entre o objeto e a sua associação com o grupos, os agrupamentos podem ser classificados de três formas: exclusivo, não-exclusivo (interseccionado) ou difuso.

Para o agrupamento exclusivo, cada objeto é associado exclusivamente a um único grupo (TAN, 2006; XU e WUNCH, 2009). Para o agrupamento não-exclusivo, cada objeto pode pertencer a mais de um grupo. Existem diversas situações em que a estrutura de agrupamento vai representar mais fielmente a realidade quando permitir que um objeto pertença a mais de um grupo. Por exemplo, uma pessoa em uma universidade poder ser tanto um aluno matriculado quanto um funcionário da universidade (TAN, 2006). Nestes casos um agrupamento não-exclusivo reflete melhor o conjunto de dados.

No agrupamento difuso, cada objeto pertence a todos os grupos ao mesmo tempo, com um peso atribuído a cada um deles, e este peso varia de 0 (não pertence totalmente) a 1 (pertence totalmente). Muitas vezes é imposta a restrição de que a soma dos pesos de um objeto seja igual a 1, também por este fato, um agrupamento difuso não aborda verdadeiras situações multiclasses, onde um objeto pertence a várias classes.

Dessa maneira, a abordagem de agrupamento difuso é mais apropriada quando se deseja evitar a arbitrariedade de atribuir um objeto a apenas um grupo quando pode estar próximo de vários. Na prática, um agrupamento difuso muitas vezes é transformado em um agrupamento exclusivo atribuindo-se cada objeto ao grupo no qual seu peso de ser membro for mais alto (TAN, 2006).

2.2.1.4 Agrupamento Completo

Um agrupamento completo atribui cada objeto do conjunto de dados a um grupo, isto é, nenhum dos objetos deixa de ser colocado em um grupo (TAN, 2006). Sendo assim, em diversos casos é desejável que todos os objetos do conjunto de dados sejam colocados em um grupo. Por exemplo, em uma aplicação para agrupar

pessoas de acordo com o risco de ataque cardíaco, não seria interessante deixar de classificar alguém.

2.2.1.5 Agrupamento Parcial

Um agrupamento parcial não atribui cada objeto a um grupo (TAN, 2006). A motivação para isso é que alguns objetos no conjunto de dados podem não pertencer a grupos bem definidos. Muitas vezes o conjunto de dados pode apresentar ruídos, elementos externos ou simplesmente elementos desinteressantes. Um exemplo poderia ser um agrupamento com interesse de agrupar notícias com temas em comum, assim, ao se realizar uma pesquisa, apenas notícias relevantes à pesquisa seriam retornadas.

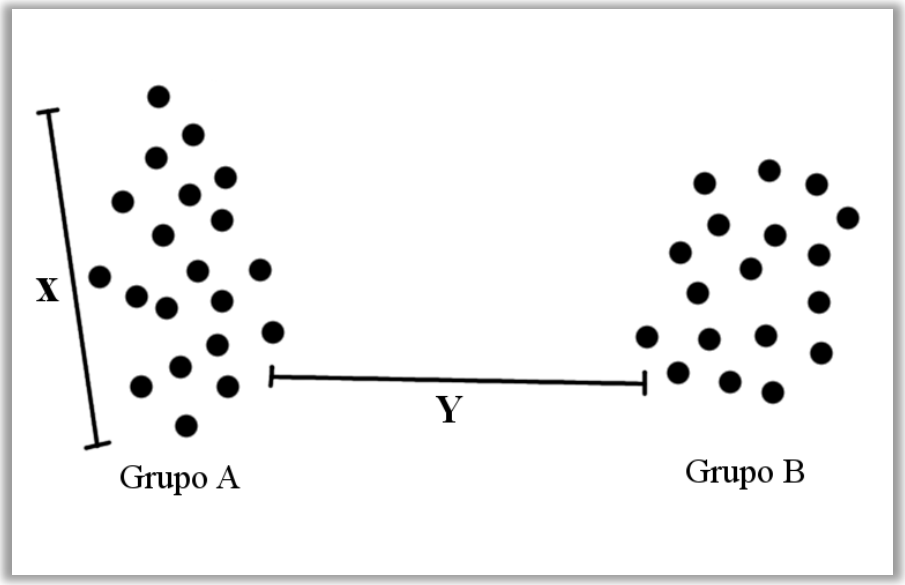
2.2.2 Diferentes tipos de grupo

A análise de grupos tem por objetivo encontrar grupos úteis de objetos no conjunto de dados. De forma não surpreendente, existem noções diferentes de grupo que se provam úteis quando postos em prática. Nesta seção será apresentada algumas definições tipos de grupo: bem separados, baseado em protótipos, baseado em grafos, baseado em densidade, e conceituais.

2.2.2.1 Bem Separados

Em um conjunto de dados em que os grupos são bem separados, cada grupo é um conjunto de objetos no qual cada objeto está mais próximo de cada um dos outros objetos no grupo do que de qualquer outro objeto que não esteja nesse grupo. Esta é uma definição idealista de um grupo e que raramente é satisfeita em aplicações reais. Grupos bem separados não são necessariamente globulares, mas podem ter qualquer formato e dimensão (TAN, 2006). A figura 6 mostra um exemplo de dois grupos bem separados, a distância X entre os pontos mais distantes entre si no grupo A é menor que a distância Y dos pontos mais próximos entre o grupo A e B, $X < Y$.

Figura 6 – Exemplo de grupos bem separados.

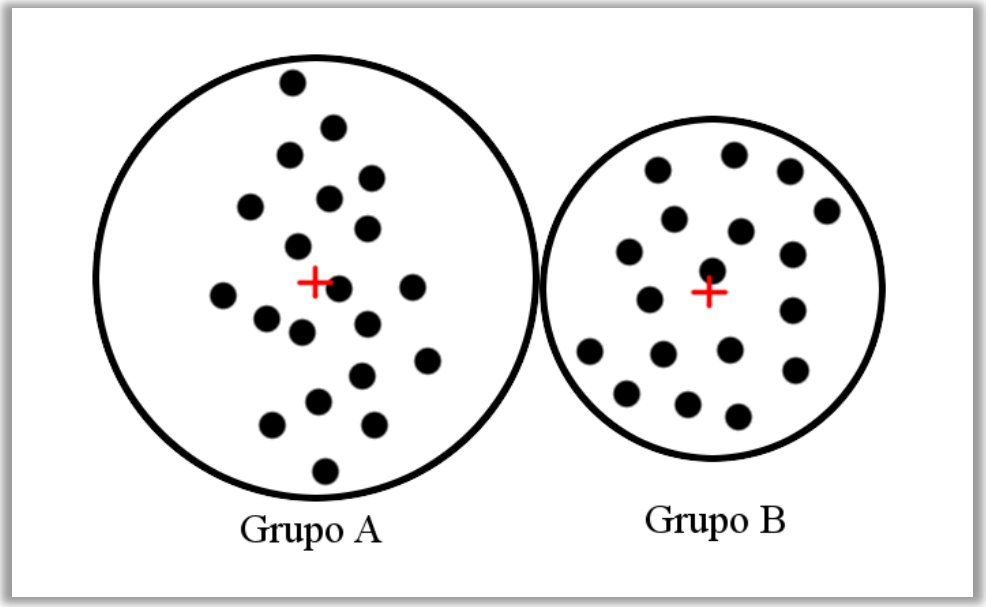


Fonte: Produzido pelo próprio autor.

2.2.2.2 Baseado em Protótipos

Um grupo é um conjunto de objetos no qual cada objeto está mais próximo do protótipo que define o grupo do que do protótipo de qualquer outro grupo. Para dados com atributos contínuos, o protótipo de um grupo é muitas vezes um centroide. Quando um centróide não é significativo, como quando os dados possuem atributos categorizados, o protótipo é muitas vezes um medóide (objeto representativo, mas que está restrito a fazer parte do grupo que representa). Em diversas ocasiões o protótipo pode ser considerado como o ponto mais central do grupo, então o grupo baseado em protótipo também pode ser chamado de grupo baseado em centro. Por força da própria definição, estes grupos tendem a ser globulares. (TAN, 2006). A figura 7 ilustra um agrupamento baseado em protótipos, onde a cruz vermelha no centro de cada grupo representa o protótipo do grupo.

Figura 7 – Exemplo de grupos baseados em protótipos



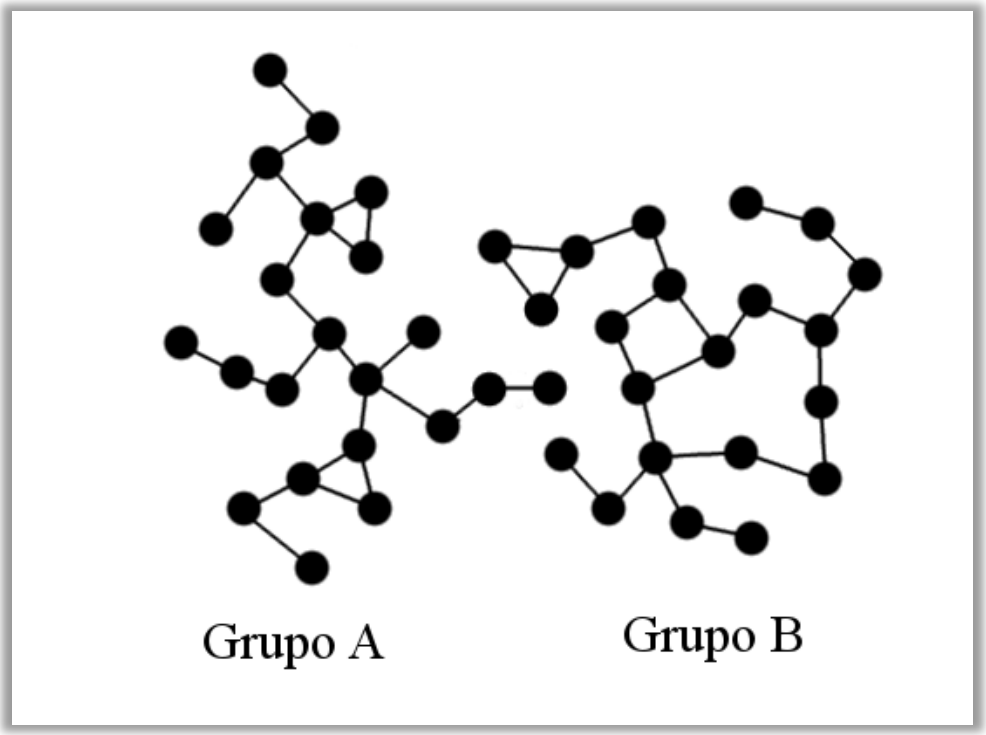
Fonte: Produzido pelo próprio autor.

2.2.2.3 Baseado em Grafos

Se os dados forem representados como um grafo, onde os nodos são objetos e existem conexões entre os objetos, então um grupo pode ser definido como um componente conectado, isto é, objetos que tenham conexão entre si, mas que não tenham conexão com objetos de outros grupos.

Um exemplo importante de grupos baseados em grafos são os grupos formados por contiguidade, onde cada objeto de um grupo está mais próximo de outro objeto do mesmo grupo do que de um objeto de outro grupo. A figura 8 mostra uma representação de agrupamento baseado em grafos, onde o grupo A e o grupo B estão entrelaçados.

Figura 8 – Exemplo de grupos baseados em grafos



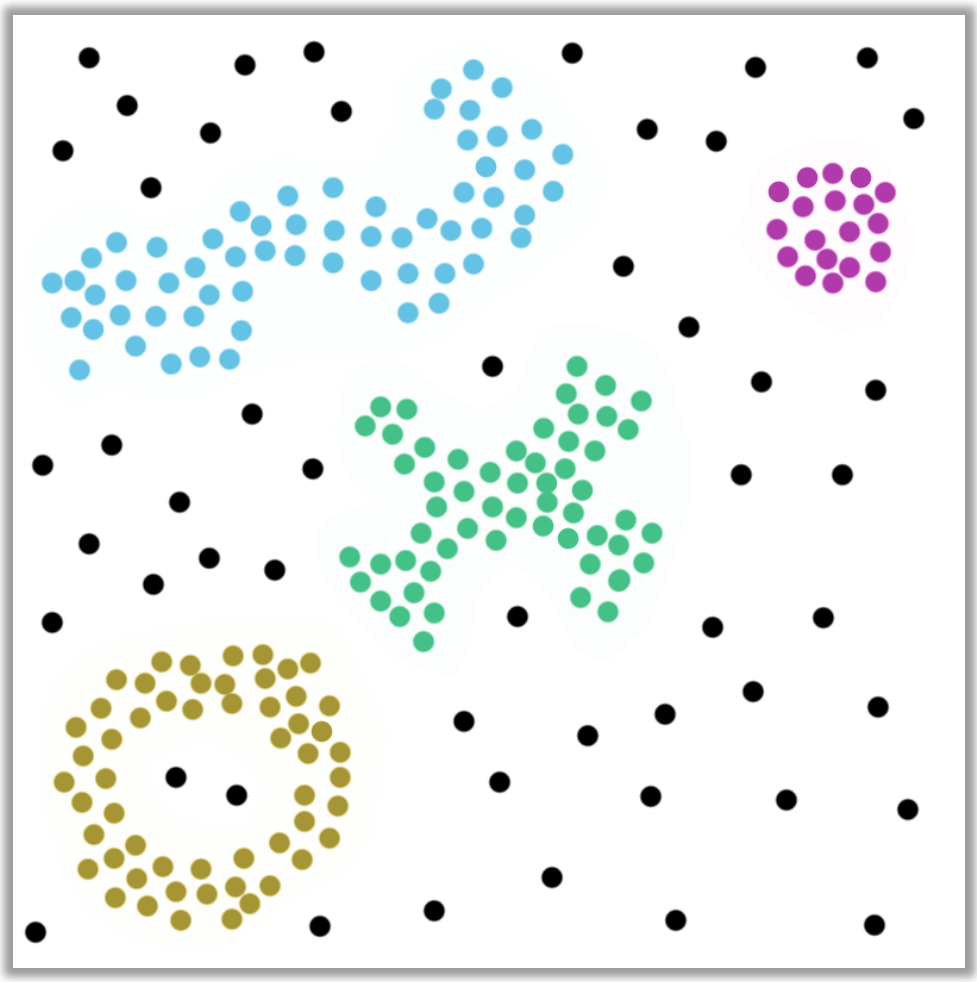
Fonte: Produzido pelo próprio autor

Grupos baseados em grafo podem ser muito úteis em casos de grupos irregulares ou entrelaçados, mas são especialmente sensíveis à ruído, o que pode causar fusão entre dois grupos distintos (TAN, 2006).

2.2.2.4 Baseado em Densidade

Um grupo é uma região densa de objetos que seja rodeada por uma região de baixa densidade. A definição de grupos baseados em densidade muitas vezes é empregada quando os grupos são irregulares ou entrelaçados e existe ruídos ou elementos externos. (TAN, 2006). A figura 9 representa um conjunto de dados que um agrupamento baseado em densidade poderia encontrar grupos com diferentes formatos e tamanhos (grupos de pontos azuis, amarelos, verdes e roxos) e descartar ruídos (pontos pretos).

Figura 9 – Exemplo de grupos baseados em densidade



Fonte: Produzido pelo próprio autor

Apesar de identificar grupos de diferentes formatos e tamanhos e de possuir resistência a ruídos, este tipo de grupo traz também inconveniências. A descrição de um grupo denso com áreas adjacentes com uma significativa diferença de densidade não é muito informativo. Grupos baseados em protótipos por exemplo, indicam que os elementos do grupo têm medidas similares ao protótipo, inferência que não pode ser feita entre os elementos de um grupo baseado em densidade.

2.2.2.4 Grupos Conceituais

Um grupo pode ser definido de modo mais geral como um conjunto de objetos que compartilham alguma propriedade. Esta definição engloba as definições anteriores, mas inclui novos tipos de grupo (TAN, 2006).

No mais, grupos conceituais podem ser utilizados com objetos que são representados por descritores não numéricos ou simbólicos. O objetivo é que os objetos sejam reunidos em classes conceitualmente simples. Por exemplo, agrupar trens utilizando os atributos: número de vagões, cor dos vagões, número de rodas e número de itens carregados. O conceito é definido utilizando atributos, e agrupar trens com dois vagões vermelhos pode ser um conceito de agrupamento para o caso dos trens (JAIN E DUBES, 1988).

2.3 Algoritmos de Agrupamento

Theodoridis e Koutroumbas (2009) e Xu e Wunch (2009) destacam diversos algoritmos de agrupamento, fundamentado em redes neurais, baseado em núcleo, de dados sequenciais, de abordagem evolucionária, entre outros. Sendo que, cada um destes algoritmos possuem particularidades, pontos fortes e fracos. Nesta seção serão abordados os algoritmos particionais K-Means e DBSCAN que serão no capítulo três utilizados no contexto de dois estudos de caso, um no domínio da mineração de dados educacionais, e o outro para identificar pontos de interesse baseado em fotografias geo-referenciadas.

O K-Means e o DBSCAN, apresentados respectivamente nas seções 2.3.1 e 2.3.2, são dois dos principais algoritmos de agrupamento de dados. Ainda que simples, são algoritmos representativos e úteis para o entendimento da análise de grupos e muito utilizados em diversos domínios, como por exemplo nas duas aplicações que serão apresentadas no capítulo três.

O K-Means e o DBSCAN são algoritmos de agrupamento particionais e exclusivos, porém, enquanto o K-Means produz agrupamento completo com grupos baseados em protótipos, o DBSCAN produz agrupamento parcial com grupos baseados em densidade. Estas diferenças entre a forma do agrupamento e tipo de

grupo produzido, como visto no capítulo anterior, trazem consigo características específicas para cada algoritmo.

O autor Berkhin (2006) lista propriedades relevantes de um algoritmo de agrupamento para a mineração de dados. Estas propriedades incluem:

1. O tipo de atributo que o algoritmo consegue lidar.
2. A escalabilidade para conjuntos de dados grandes.
3. A habilidade para encontrar grupos com formas irregulares.
4. Capacidade de lidar com ruídos.
5. Complexidade.
6. Forma de atribuição e organização dos grupos.
7. Dependência de um conhecimento prévio e parâmetros definidos pelo usuário.

De acordo com estas propriedades, posteriormente, será apresentado na seção 2.3.3 uma discussão comparativa entre os algoritmos K-Means e DBCAN.

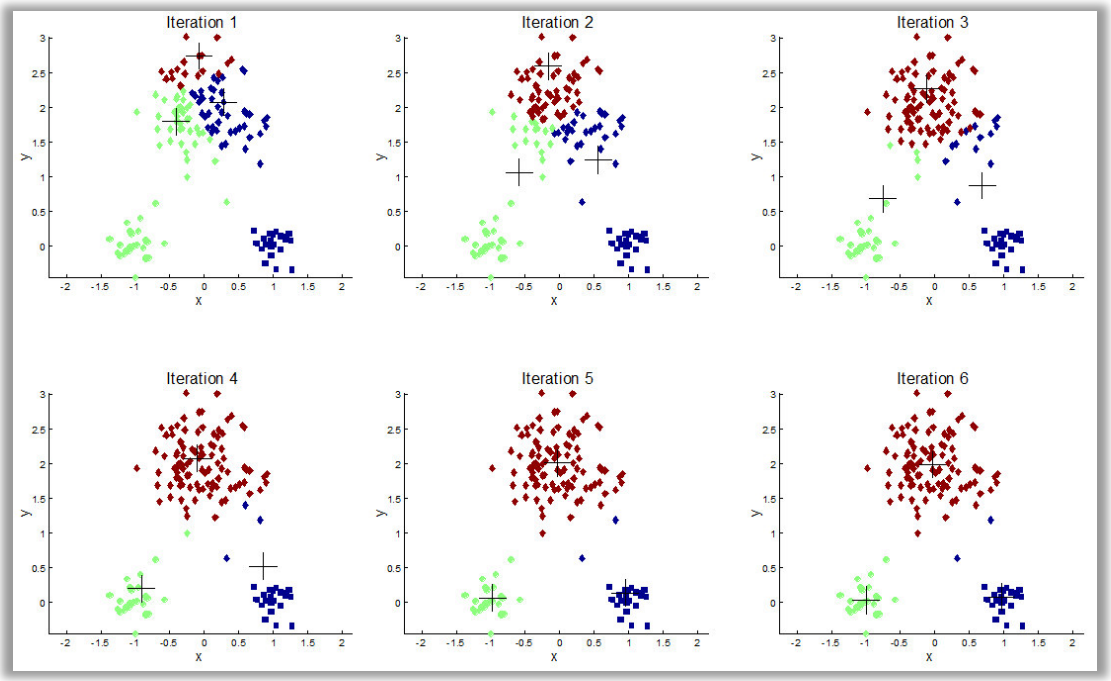
2.3.1 K-Means

O algoritmo K-Means é um dos algoritmos de agrupamento mais conhecidos e populares (BERKHIN, 2006; DUDA et al., 2001; THEODORIDIS e KOUTROUMBAS, 2008; XU e WUNCH, 2009), isto ocorre por ser um algoritmo simples e amplamente utilizado em diversas áreas.

O K-Means básico produz um agrupamento particional, baseado em centróide, que agrupa todos os objetos e utiliza como medida de proximidade a distância euclidiana. Descrevendo o algoritmo, primeiro escolhemos *K* centróides iniciais, sendo *K* o número de grupos desejado que foi dado de entrada pelo usuário. Então todos os outros pontos são atribuídos ao centróide mais próximo, e cada coleção é um grupo. Cada grupo tem seu centróide recalculado e, em seguida, repetimos os passos de atribuição de pontos e atualização de centróide até que nenhum ponto mude de grupo, ou até que os centróides não sejam alterados (TAN, 2006).

A figura 10 representa as iterações do algoritmo K-Means para *K* = 3, para cada iteração o centróide é recalculado, se adaptando ao conjunto de dados, até que não ocorra mudanças da quinta para sexta iteração, e a execução termine.

Figura 10 - Convergência de um agrupamento utilizando o algoritmo K-Means.



Fonte: DATA VISUALIZATION, Cluster Analysis: see it 1st. Online. Disponível em: https://goo.gl/z1cU6J. Acesso em 15.02.2018

O algoritmo pode ser estruturado em cinco etapas:

1. Selecione *K* centróides iniciais.
2. **faça:**
3. Atribua todos os outros pontos ao centróide mais próximo.
4. Recalcule o centróide de cada grupo.
5. **até que:** Os centróides não mudem **e/ou** nenhum ponto mude de grupo.

Para algumas funções de proximidade e tipos de centróide, K-means sempre converge para uma solução, e como a maioria da convergência ocorre nas primeiras iterações, a condição na linha 5 do algoritmo acima é normalmente mais fraca.

É necessário definir também a proximidade entre os pontos, frequentemente a distância euclidiana é usada para pontos de dados no espaço euclidiano, mas o K-Means também pode ser utilizado para outros tipos de dados, como para documentos, que a semelhança do coseno é mais apropriado (TAN, 2006). Geralmente as medidas de distância são relativamente simples, já que o algoritmo tem que calcular a semelhança de cada ponto com cada centróide repetidamente.

Após executar o algoritmo, precisamos de alguma forma de avaliar se a execução produziu uma boa partição do conjunto de dados. Como o objetivo do agrupamento usualmente é expresso por uma função objetiva que depende das proximidades dos pontos de um grupo, podemos medir a qualidade de um agrupamento em um espaço euclidiano utilizando a soma do erro quadrado (SSE) (TAN, 2006; THEODORIDIS e KOUTROUMBAS, 2008; JAIN e DUBES, 1988).

Para isso, calculamos o erro de cada ponto de dados (distância entre o centróide mais próximo) e depois a soma total de todos os erros dos pontos de um grupo. Ao compararmos várias execuções do algoritmo, preferimos a com o menor SSE, já que isto significa que os centróides deste agrupamento representam melhor os pontos do seu grupo. A SSE é definida formalmente da seguinte forma:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist\,(c_i, x)^2$$

Onde x é um objeto, C_i é o grupo de índice i, c_i é o centróide do grupo C_i, K é o número de grupos e *dist* é a distância euclidiana entre dois objetos no espaço Euclidiano.

Escolher os centróides iniciais é uma tarefa crucial para o K-Means, uma vez que este é um algoritmo heurístico de subida da colina, e desta forma, suas iterações tendem a levar para um ótimo, mas não necessariamente para um ótimo global (JAIN e DUBES, 1988). A forma mais simples de se escolher os centróides iniciais é aleatóriamente, mas os grupos resultantes são frequentemente pobres, uma técnica comum para se contornar este problema é executar várias vezes e selecionar o resultado com a menor SSE (TAN, 2006; THEODORIDIS e KOUTROUMBAS, 2008).

Sendo assim, esta estratégia pode não funcionar muito bem dependendo do número de grupos procurados e dos dados, sendo assim, outras formas de escolher os centróides iniciais são pegar uma amostra de pontos de um agrupamento hierárquico feito préviamente, mas que é prática apenas se a amostra for relativamente pequena e se K for relativamente pequeno comparado com o tamanho da amostra, ou ainda, gerar aleatóriamente o primeiro centróide e os seguintes são selecionados calculando o ponto de maior distância dos anteriores, o que pode selecionar elementos externos em vez de pontos em regiões densas além de ser custoso calcular estes pontos iniciais.

De fato, K-Means tem limitações para encontrar diferentes tipos de grupos, possuindo dificuldade para detectar grupos que não sejam esféricos ou com tamanhos e densidades muito diferentes (TAN, 2006), tendendo a dividir grupos muito grandes e pouco densos e fundindo grupos pequenos e compactos.

Destaca-se que ruídos e elementos externos interferem no sucesso do K-Means. Os elementos externos, como são pontos do conjunto de dados, são agrupados. Desta forma eles influenciam no cálculo da SSE, e por consequência no agrupamento final (THEODORIDIS e KOUTROUMBAS, 2008). Levando em consideração que em geral grupos pequenos provavelmente são formados por elementos externos, uma versão do algoritmo feita por BALL e HALL (1967) trata destes elementos externos simplesmente descartando grupos considerados pequenos.

O K-Means é naturalmente bem aplicável em dados de atributos contínuos e a princípio não é adequado para conjuntos de dados com coordenadas nominais, embora existam variantes deste algoritmo para tratar destes casos (THEODORIDIS e KOUTROUMBAS, 2008). Os algoritmos K-Medoides, em troca de custo computacional, são uma possibilidade.

Entre as suas desvantagens, o K-Means assume que o usuário já sabe o número de grupos *K* distintos no conjunto de dados, o que na prática quase nunca é verdade, como na situação da inicialização dos centróides, não há uma forma global e eficiente para selecionar o número de grupos (XU e WUNCH, 2009). Assim sendo, indentificar préviamente *K* se torna um tópico muito importante para a validação dos grupos (DUBES, 1993).

A maior vantagem do K-Means é a sua simplicidade computacional, o que o faz um candidato atraente para uma variedade de aplicações. Sua complexidade de tempo é *O(KmI)*, onde *m* é o número de pontos de dados e *I* o número de iterações necessárias para a convergência. O número de iterações pode ser controlado, uma vez que a convergência maior ocorre nas primeiras iterações, *I* e *K* são significativamente menores que *m*, fazendo com que este algoritmo seja aproximadamente linear em *m*, e assim elegível para processar grandes conjuntos de dados (TAN, 2006; THEODORIDIS e KOUTROUMBAS, 2008; XU e WUNCH, 2009).

Trabalhos como o de Dhillon e Modha (2002) mostram que o K-Means aceita paralelização direta, podendo ter sua execução dividida em vários processadores, aumentando sua capacidade em lidar com grandes conjuntos de dados.

2.3.2 DBSCAN

O algoritmo DBSCAN (Density Based Spatial Clustering of Applications with Noise) introduzido por Ester et al. (1996) é o maior representante dos algoritmos com abordagem de agrupamento baseado em densidade (BERKHIN, 2006). Por ser um algoritmo baseado em densidade simples e eficaz, e por apresentar diversos conceitos importantes para qualquer algoritmo baseado em densidade, seu estudo é pertinente para o entendimento do assunto.

Embora não existam tantas formas para definir densidade quanto como para definir similaridade, existem diversos métodos distintos. O algoritmo DBSCAN tradicional utiliza a abordagem de densidade baseada em centro. Nesta abordagem a densidade é avaliada para um determinado ponto no conjunto de dados contando-se o número de pontos dentro de um determinado raio de vizinhança (*Eps*) do ponto em questão (TAN, 2006).

Este método é simples de ser implementado, mas a densidade de qualquer ponto dependerá do raio especificado, e para um raio muito grande, todos os pontos terão a densidade com o número de pontos no conjunto de dados (*m*), e ainda, para um raio muito pequeno, todos os pontos terão densidade de 1.

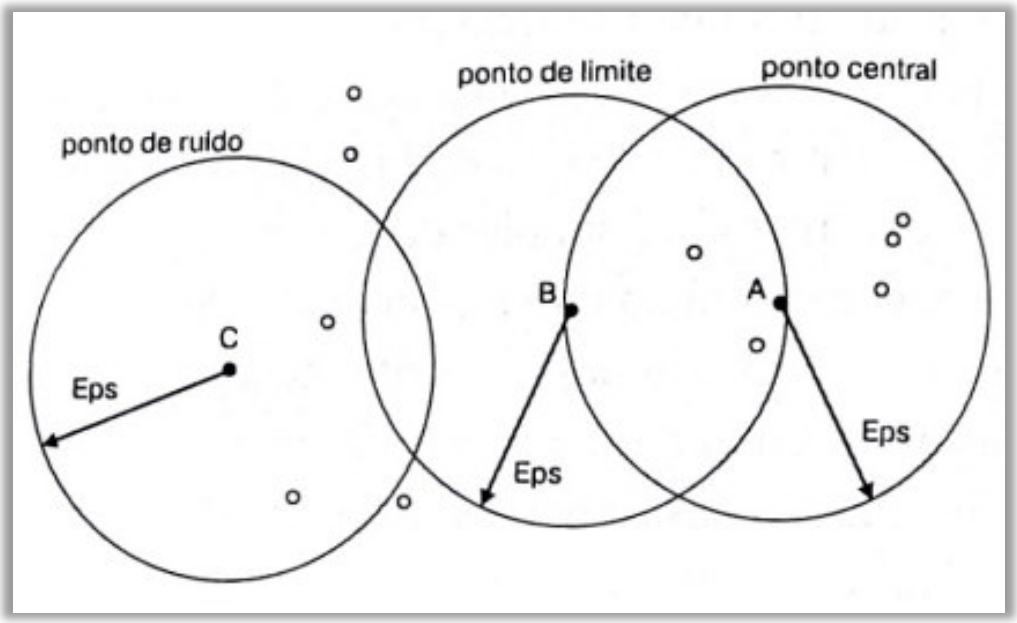
A abordagem da densidade baseada em centro permite classificar um ponto das seguintes formas:

1. Pontos Centrais: São pontos que estão no interior de um grupo baseado em densidade. Um ponto é classificado como central se o número de pontos dentro de uma determinada vizinhança em torno do ponto, conforme determinado pela função de distância e *Eps* dado, exceder um determinado limite, *MinPts*, que também é um parâmetro dado pelo usuário. A exemplo de um ponto central, veja o ponto A da figura 7 para *MinPts* = 7.
2. Pontos de Limite: Um ponto de limite é um ponto que não é central, mas que fica dentro da vizinhança de um ponto central, e pode estar simultaneamente dentro da vizinhança de vários pontos centrais. Na figura

11, o ponto B é um ponto de limite, pois não cumpre o requerimento de *MinPts* = 7, mas está no *Eps* do ponto A.

3. Pontos de Ruído: É qualquer ponto que não seja nem um ponto central nem um ponto de limite, por exemplo o ponto C na figura 11.

Figura 11 - Exemplo de Ponto de ruído, limite e central



Fonte: Adaptado de (TAN, 2006)

Como é possível verificar destas classificações de pontos, o algoritmo DBSCAN implementa o conceito de densidade e alcance para definir um grupo, traduzido nos dados de entrada *MinPts* e *Eps*, respectivamente.

Podemos descrever o DBSCAN de maneira informal da seguinte forma. Todos os pontos de centro que estejam próximos o suficiente entre si, são colocados no mesmo grupo. Cada ponto de limite é colocado no grupo do ponto central que esteja suficientemente próximo, nesta etapa pode ser necessário resolver conexões que podem ocorrer entre grupos. Os pontos de ruído são descartados. (TAN, 2006)

O algoritmo pode ser estruturado em cinco etapas:

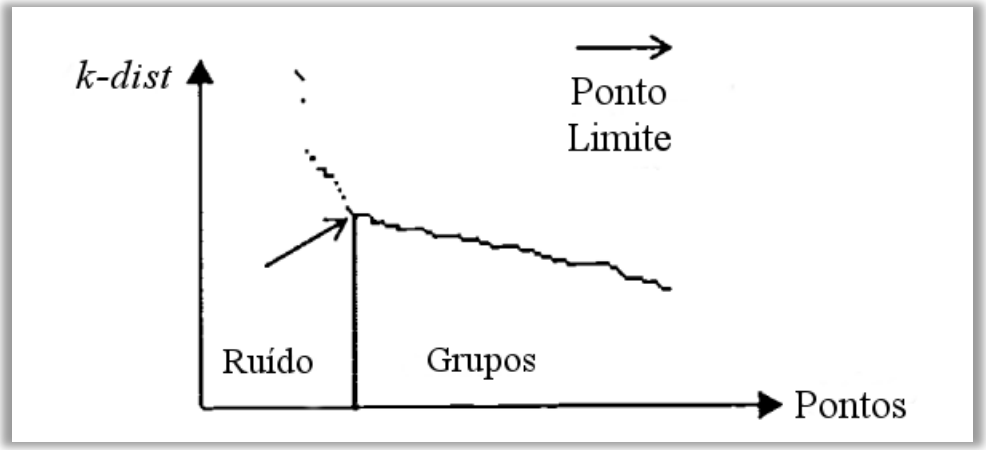
1. Classificar todos os pontos do conjunto como ponto de centro, limite ou ruído.

2. Descartar os pontos de ruído.
3. Conectar todos os pontos de centro que estejam dentro da *Eps* uns dos outros.
4. Tornar cada grupo de pontos de centro conectados um grupo separado.
5. Atribuir cada ponto limite ao grupo de algum dos seus pontos centrais.

A seleção dos parâmetros de entrada para o DBSCAN não é um problema trivial, depende do conjunto de dados, e tem um caráter decisivo para o sucesso do resultado final do algoritmo. Ester et al. (1996), no trabalho que apresenta o DBSCAN, sugere uma heurística simples mas eficiente para determinar os parâmetros *Eps* e *MinPts* do grupo de menor densidade no conjunto de dados. Sendo *d* a distância entre um ponto *p* e o seu k-ésimo vizinho mais próximo, a vizinhança de *p* dentro da distância *d* terá *k* + 1 elementos apenas se dois ou mais pontos tiverem exatamente a mesma distância de *p*, o que é improvável. Se mudarmos o ponto *k* para um ponto do mesmo grupo, a distância *d* não sofrerá grandes alterações, isto só acontecerá se os elementos *k=1, 2, 3, ...* do grupo estiverem mais ou menos em linha, o que em geral não é verdade.

Dado um *k*, podemos definir a função *k-dist*, que mapeia a distância de cada ponto para o seu *k*-ésimo vizinho mais próximo. A figura 12 mostra um gráfico da relação entre *k-dist* e o número de pontos, se escolhermos um ponto limite *p, Eps* será igual a *k-dist(p)* e *MinPts* igual a *k*. O ponto limite *p* é o primeiro ponto do "vale" no gráfico da figura 11, e embora seja difícil identificar automaticamente este ponto, esta é uma tarefa relativamente simples para o usuário ao visualizar este gráfico.

Figura 11 - Gráfico de pontos e k-dist



Fonte: Adaptado de (ESTER el al., 1996)

A complexidade de tempo básica do DBSCAN é *O(m x tempo para encontrar pontos na vizinhança)*, onde *m* é o número de pontos no conjunto de dados. No pior caso, esta complexidade é $O(m^2)$. DBSCAN utiliza árvores R (GUTTMAN, 1984) para realizar a busca de pontos na vizinhança, o que implica, para *m* pequenos, em complexidade de $O(m \log_2 m)$ (TAN, 2006; THEODORIDIS e KOUTROUMBAS, 2008; XU e WUNCH, 2009). Experimentos confirmam esta complexidade log-linear (BERKHIN, 2006). Com relação ao requisito de espaço, o DBSCAN só guarda uma pequena quantidade de dados, como o rótulo do grupo e identificação de cada ponto.

Entre as suas desvantagens, o DBSCAN pode ter problemas se a densidade de grupos variar muito (TAN, 2006), ocasionando na fusão dos grupos muito densos ou classificação dos grupos pouco densos como ruído.

Entre as vantagens do DBSCAN estão o processamento independente da ordem dos dados, a relativa imunidade à ruidos, e a capacidade de lidar com grupos de tamanhos e formas arbitrárias. Em compensação, pode ser custoso quando calcular os vizinhos mais próximos (TAN, 2006).

2.3.3 Discutindo o K-Means e o DBSCAN

A partir dos fudamentos teóricos que foram apresentados nas seções 2.1, 2.2, 2.3.1 e 2.3.2, e de trabalhos como Tan (2006), Berkhin (2006), Jain et al.(1999) e

Theodoridis e Koutroumbas (2008) que apontam diferenças entre o K-Means e o DBSCAN, é possível realizar uma breve discussão comparativa.

Embora ambos os algoritmos K-Means e DBSCAN sejam algoritmos particionais e produzam agrupamentos exclusivos, a semelhança entre eles praticamente acaba neste ponto. O K-Means geralmente agrupa todos os objetos, o DBSCAN descarta os objetos que classifica como ruído. No que diz respeito ao tipo de grupo que eles produzem, o K-Means agrupa os objetos do conjunto utilizando a noção de protótipo de um grupo, já o DBSCAN identifica os grupos baseado na densidade de pontos em uma região.

O DBSCAN pode lidar com grupos de diferentes tamanhos e formatos e não é fortemente afetado por ruído e elementos externos. Já o K-Means tem dificuldade com grupos não globulares e de tamanhos diferentes. Mas apesar da diferença entre a capacidade de detectar grupos, ambos os algoritmos podem ter o desempenho comprometido quando os grupos possuírem densidades muito variadas. O K-Means só pode ser utilizado quando os dados possuírem um centróide bem definido, como uma média ou mediana. O DBSCAN requer que a sua definição de densidade, baseada na noção Euclidiana tradicional de densidade, tenha significado para os dados. O K-Means pode ser aplicado a dados esparsos e de alta dimensionalidade, como documentos. O DBSCAN geralmente tem um desempenho ruim para dados de alta dimensionalidade, uma vez que a definição Euclidiana tradicional de densidade não funciona bem para eles. K-Means pode encontrar grupos que não estejam bem separados, mesmo que exista intersecção entre eles, o DBSCAN tende a fundir grupos que possuam intersecção entre si. O algoritmo K-Means possui uma complexidade de tempo de $O(m)$, enquanto que o DBSCAN se aproxima de $O(m^2)$ com exceção de alguns casos, e esta é uma diferença bastante significativa, que pode tornar o DBSCAN impraticável para grandes conjuntos de dados.

DBSCAN produz o mesmo conjunto de grupos a cada execução, enquanto que o K-Means, que geralmente utiliza uma inicialização aleatória, não. O DBSCAN determina o número de grupos automaticamente, mas precisa de dois parâmetros a serem especificados pelo usuário, *Eps* (raio de vizinhança) *e MinPts* (mínimo de pontos no *Eps* para a formação de um grupo). K-Means precisa do número de grupos na sua entrada. O agrupamento K-Means pode ser visto como um problema de otimização e como um caso específico de abordagem de agrupamento estatístico, já

o DBSCAN não é baseado em algum modelo formal. A tabela 1 foi construída, inspirada na lista de Berkhin (2006) de propriedades relevantes de um algoritmo de agrupamento para mineração de dados, com o objetivo de facilitar a visualização das semelhanças e diferenças entre o K-Means e o DBSCAN.

Tabela 1 – Comparação entre as características do K-Means e DBSCAN

Características	**K-Means**	**DBSCAN**
Agrupamento particional	Sim	Sim
Agrupamento exclusivo	Sim	Sim
Agrupamento completo	Sim	Não
Agrupamento parcial	Não	Sim
Tipo de grupo	Baseado em protótipo	Baseado em densidade
Resistência a ruídos e elementos externos	Não	Sim
Capacidade de lidar com grupos de diferentes tamanhos e formatos	Não, tem dificuldades com grupos não globulares	Sim
Capacidade de lidar com grupos com densidades muito variadas	Não	Não
Utilização quanto às características dos dados	Quando os dados possuírem um centróide bem definido	Quando a noção de densidade tenha significado para os dados
Aplicação para dados com alta dimensionalidade	Pode ser aplicado	Geralmente tem um desempenho ruim
Capacidade de encontrar grupos interseccionados	Sim	Não, geralmente funde os grupos
Complexidade de tempo	$O(m)$	$O(m^2)$
Aplicação em grandes conjuntos de dados	Sim	Pode ser impraticável
Consistência de resultados	Não, pode produzir grupos diferentes para cada execução	Sim
Dados de entrada	Número de grupos	*Eps* e *MinPts*
Baseado em modelo formal	Sim, pode ser visto como um problema de otimização	Não

Fonte: Produzido pelo autor

3 APLICAÇÕES DE ANÁLISE DE GRUPOS

Este capítulo ilustra a aplicação dos conceitos sobre a análise de grupos apresentados no capítulo 2, apresentando duas aplicações de agrupamentos de dados. Na seção 3.1, uma aplicação no domínio da mineração de dados educacionais, com o objetivo de agrupar usuários conforme o perfil de utilização de um ambiente virtual de aprendizado. Na seção 3.2, é discutida uma aplicação de informações geográficas, para a identificação de locais de interesse a partir de fotografias geo-referenciadas.

Vale ressaltar que apenas estas duas aplicações não são capazes de sumarizar a utilidade da análise de grupos, mas a diferença entre os domínios das aplicações dão uma noção da amplitude da área de atuação desta tarefa de mineração de dados.

Para entender as escolhas das técnicas de agrupamento utilizadas nas aplicações nestes dois domínios, diversos fatores precisam ser considerados. Para algumas aplicações, como a criação de uma taxonomia lógica, é preferido utilizar uma técnica hierarquica, já quando o objetivo é realizar um agrupamento por resumo, tipicamente um agrupamento particional é escolhido. Em outras aplicações ambos podem ser úteis.

No que tange aos tipos de grupo, se os grupos são globulares, como quando precisamos resumir os dados para reduzir seu tamanho, uma técnica particional pode ser mais apropriada. Se o agrupamento for utilizado para segmentar grupos com formatos diversos, a técnica de agrupamento por densidade pode gerar melhores resultados.

As próprias características dos conjuntos de dados e atributos podem sugerir o tipo de algoritmo a ser utilizado, por exemplo o K-Means só pode ser usando quando uma medida de proximidade esteja disponível e que permita o cálculo significativo de um centróide de grupo. Ruídos e elementos externos podem ser interessantes ou não para a aplicação, às vezes o que é ruído para uma, é exatamente o que se procura em outra.

A quantidade de objetos de dados também pode ser determinante, para um número muito grande, o custo computacional pode ser alto de mais ou até inviável

(TAN, 2006). Na apresentação de cada aplicação de exemplo, estas discussões serão levantadas, e com base nos conceitos sobre técnicas de agrupamento apresentados no capítulo dois, e particularmente os algoritmos K-Means e DBSCAN, será indicada a técnica de agrupamento mais promissora.

3.1 Agrupamento em Ambiente Virtual de Aprendizado

Nesta seção discutiremos especificamente o trabalho realizado por Pinheiro et al. (2014) na identificação de grupos de alunos em ambiente virtual de apendizado (AVA).

Os AVA armazenam grande quantidade de dados dos usuários na plataforma, acessos, ações realizadas, erros, interações com o fórum e chat, entre outros. Estes dados são muito valiosos para analisar o comportamento dos estudantes (Mostowand e Beck, 2006) e as informações podem ser obtidas através de Mineração de Dados.

A Mineração de Dados têm sido usada na área da educação com a intenção de investigar questões, como os fatores que afetam a aprendizagem, como desenvolver sistemas educacionais mais eficazes, entender a relação da abordagem pedagógica e aprendizado. No contexto da área da educação, surgiu a Mineração de Dados Educacionais (do inglês, *Educational Data Mining* – EDM), definida como a área de pesquisa que tem como foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais (PINHEIRO et al.,2014).

Com o intuito de favorecer a inclusão digital, nações latino-americanas têm investido na criação e utilização de centros tecnológicos comunitários, ou telecentros, onde o acesso às Tecnologias de Informação e Comunicação (TIC) é disponibilizado para as comunidades menos privilegiadas a um custo mínimo ou isento de custos (SILVA et al.,2013). A implantação destas políticas públicas tem como um dos aspectos críticos a necessidade de formar os agentes para inclusão digital. E esse é o contexto do Programa Telecentros.BR (BRASIL, 2009), que é uma ação do Governo Federal que tem a finalidade de apoiar a implantação de novos telecentros públicos e fortalecer unidades já existentes no País.

A Rede Telecentros BR até o ano de 2011 contava com a participação de cinco polos regionais, dois polos estaduais e um polo nacional. Sob responsabilidade dos polos regionais estava a formação dos agentes de inclusão digital (monitores de

telecentro), gestores de telecentros (administradores do telecentro), tutores (que atuavam na formação dos monitores) e supervisores de tutoria (responsáveis pela supervisão e acompanhamento do trabalho dos tutores) (SILVA et al., 2013).

No período de fevereiro de 2010 a dezembro de 2010, os membros dos polos de formação se articularam para construir e aplicar o Curso de Formação de Monitores dos Telecentros e a ativação das redes sociais de agentes de inclusão social atuantes nas comunidades. O projeto de formação dos agentes de inclusão social ofertou um curso de 480 horas, disponibilizado na plataforma Moodle. Os agentes de inclusão social percorreram os temas oferecidos pelo curso sem um percurso pré-definido, e durante o decorrer do curso contaram com o apoio de tutores e supervisores de tutores dos diversos polos regionais (SILVA et al., 2013).

O Moodle foi o AVA selecionado para o desenvolvimento do curso de formação de agentes de inclusão social do Telecentros.BR, é uma plataforma de aprendizado projetada para fornecer educadores, administradores e alunos um sistema único, robusto, seguro e integrado para criar e personalizar ambientes de aprendizado (MOODLE, 2018). Também foi necessário realizar um curso para formação de tutores que atuaram auxiliando os monitores durante a formação do Telecentros.BR, que tiveram o acompanhamento e auxilio dos supervisores de tutoria no processo de formação.

A intenção desta aplicação é agrupar os agentes de inclusão digital conforme seus perfis de utilização do AVA, desta forma, entregando ao usuário da aplicação uma partição do conjunto de dados que possa ajudar a classificar qualitativamente os alunos conforme a sua interação com o curso.

Para este caso, é interessante que todos os pontos do conjunto de dados sejam classificados, uma vez que informações de uso, mesmo que entre agentes de inclusão sociais e tutores, podem ser de grande valia para a compreensão e apoio de decisões futuras.

Os próprios níveis de avaliação qualitativa do curso: Excelente, Bom, Regular e Insuficiente (PINHEIRO et al., 2014), nos dá uma idéia da partição desejada do conjunto de dados.

Dadas as características de agrupamento particional, a classificação de todos os pontos, e uma noção da quantidade de grupos existentes, o algoritmo K-Means se torna a primeira escolha. Para um algoritmo baseado em técnicas de densidade como

o DBSCAN, a possibilidade de não agrupar alguns pontos e a quantidade de dados acumulados em um log de eventos podem ser fatores que dificultem a chegada a um resultado satisfatório. O tamanho do conjunto de dados também pode ser um fator crítico para um algoritmo hierárquico aglomerativo, somado ao fato de que uma taxonomia lógica não é o objeto final deste trabalho.

Intuitivamente, é possível formular a hipótese de que dado um determinado conjunto de alunos, classificados em uma mesma avaliação qualitativa, para o mesmo curso, e com as mesmas tarefas a serem cumpridas, provavelmente não apresentarão um comportamento extremamente distintos entre si, e que os grupos tendem a ser globulares. O que nos direciona novamente para o algoritmo K-Means.

O Moodle armazena os logs em uma base de dados relacional, e para esta aplicação foram utilizados os registros da tabela *mdl_log*, responsável pelo armazenamento de eventos do sistema. O pré-processamento realizado nestes dados consistiu na limpeza das informações de demais tabelas e eliminação de alunos excluídos e administradores.

O algoritmo escolhido por Pinheiro et al. (2014) foi o K-Means, a entrada *k* setada em cinco, um para cada nível de avaliação qualitativa e mais um para agrupar valores fora do esperado. O processo de agrupamento pelo K-Means utilizou duas dimensões, as colunas "course" e "userid" da tabela *mdl_log*, foram filtradas também somente as linhas correspondentes aos eventos dentro da categoria "course".

Para o K-Means, os 5 grupos encontrados (denominados K1, K2, K3, K4, K5) possuem as seguintes caraterísticas, como mostra a tabela 2:

Tabela 2 - Grupos encontrados para o K-Means

	K1	K2	K3	K4	K5
Número de Usuários	5	32	183	825	3251
Acessos	228728	598867	1197017	2024758	1152300
Cursos	50	50	50	49	50
Acesso Médio por Usuário	45745,6	18714,59375	6541,076502732	2454,252121212	354,44478622
Acesso Médio por Usuário e Por Curso	914,912	374,291875	130,821530055	50,086777984	7,088895724
Ação Mais Utilizada	*View* 93930 execuções	*View* 266186 execuções	*View* 614599 execuções	*View* 1183165 execuções	*View* 93930 execuções
Recurso Mais Utilizado	*ErrorLogin* 59430 acessos	*Course* 153197 acessos	*Forum* 294337 acessos	*Course* 513217 acessos	*Course* 310376 acessos
Quantidade Caminhos Médios	145	163	167	137	133

Fonte: (PINHEIRO et al., 2014)

A tabela 2 demonstra que foram encontrados grupos distintos de usuários. Resumidamente, o grupo K1, o pequeno número de usuários e o grande número de erro de acesso indica que este é o grupo formado por usuários do AVA com o comportamento fora do esperado. No grupo K2, dado o número de usuários e a grande quantidade de acesso, podemos deduzir que se trata do grupo de usuários responsáveis pela capacitação dos agentes de inclusão social. No grupo K3 estão os agentes de inclusão social mais participativos, com uma média de acessos por curso maior que K4 e K5, e com o fórum como recurso mais utilizado. O grupo K4 representa os agentes com participação mediana, em seguida no grupo K5 estão os agentes menos participativos (PINHEIRO et al., 2014).

Os resultados obtidos demonstram algum grau de sucesso na análise de grupos realizada para este conjunto de dados, uma vez que foi capaz de inferir conhecimento da partição feita pelo algoritmo K-Means.

3.2 Agrupamento em identificação de locais de interesse utilizando fotografias geo-referenciadas

Nesta sessão abordaremos o trabalho de Ponciano (2016), que apresentou uma estratégia *online* que utiliza fotografias geo-referenciadas e seus metadados para

identificar locais de interesse pertencentes a uma dada região geográfica e recuperar informações relevantes relacionadas.

Segundo a Organização Mundial do Turismo (OMT), o turismo mundial têm crescido nos últimos anos e tende a crescer nos próximos. Em 2013 o turismo internacional alcançou 1,087 bilhão de chegadas, em 2014 foram 1,1 bilhão de chegadas e a perspectiva de crescimento é de 3,3% ao ano entre 2010 e 2030 (UNWTO, 2015; UNWTO, 2014).

Parte das viagens turísticas internacionais, ou nacionais, são realizadas por pessoas que visitam seu destino pela primeira vez, o que exige um certo esforço para obter informações potencialmente interessantes e quais locais visitar. Geralmente se trata de um planejamento complexo e demorado, que pode envolver pesquisa em diversos locais como *websites*, blogs, mapas turísticos, redes sociais, sugestões de conhecidos, entre outros (PONCIANO, 2016).

A rede social Flickr[1] possui cerca de 6,47 bilhões de imagens, com cerca de 597 milhões postadas apenas no ano de 2017 (a uma média de 1.63 milhões por dia), e foi o repositório *online* escolhido para a obtenção de imagens.

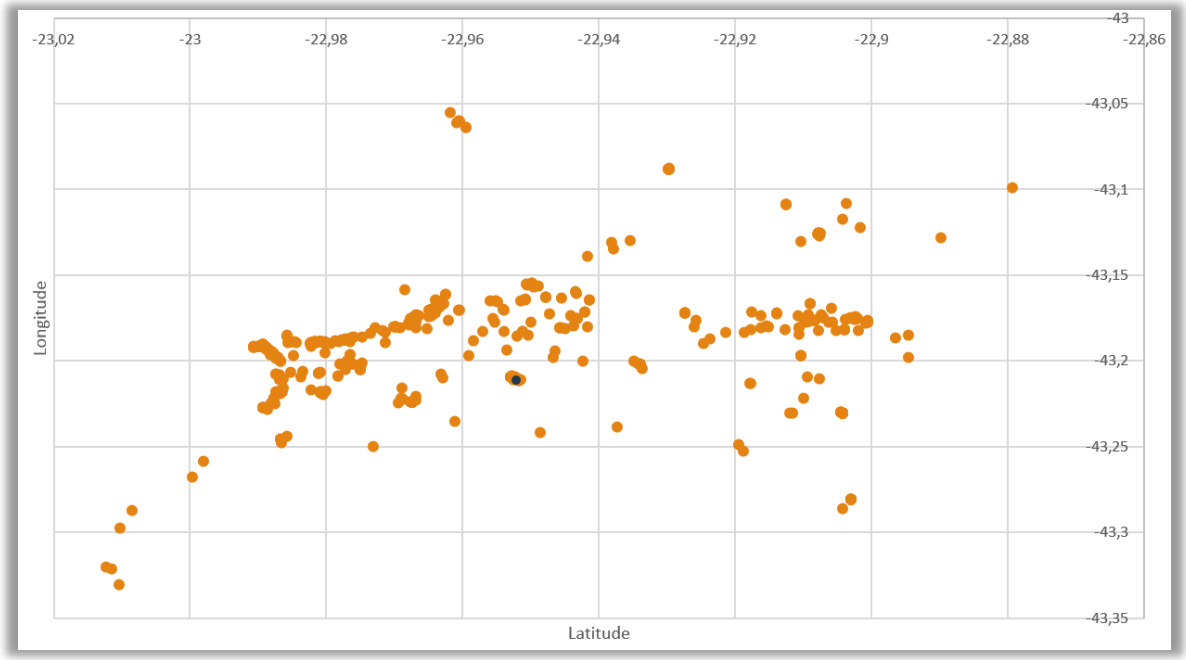
Ponciano (2016) apresenta um método, denominado *Tourist Place Identification* (ToPI), que visa identificar pontos de interesse, baseado nas imagens obtidas no repositório online escolhido, e em contexto dinâmico, no qual as informações disponíveis sobre pontos de interesse podem sofrer alterações com o passar do tempo.

Parte do processo ToPI consiste em agrupar as fotografias de uma dada região geográfica, utilizando para isso as suas coordenadas GPS (latitude e longitude). A concentração de fotos tiradas em um mesmo local indica que ele é um lugar potencialmente turístico. Cada grupo encontrado é tratado como um possível ponto de interesse para as etapas posteriores (PONCIANO, 2016).

A figura 13 mostra um conjunto de 500 imagens selecionadas na região próxima ao Cristo Redentor (Rio de Janeiro, RJ, Brasil).

[1] https://www.flickr.com/

Figura 9 - Gráfico de dispersão com 500 fotografias recuperadas nas proximidades do Cristo Redentor (representado pelo ponto azul no centro da imagem).



Fonte: (PONCIANO, 2016)

Os grupos deste domínio de aplicação podem apresentar formatos distintos, que variam conforme a disposição física do próprio ponto de interesse, como por exemplo uma região litorânea muito frequentada pode ser representada por um conjunto de fotografias tiradas em uma extensa faixa. Esta característica não é ideal para um algoritmo como o K-Means, em contrapartida um algoritmo baseado em densidade pode lidar com esse tipo de situação.

Em um ponto de interesse é comum que existam diversos pontos de vista onde se tirem fotos, o que pode ocasionar vários adensamentos de pontos. Esta característica pode se tornar um problema de identificação de grupos diferentes onde deveria existir apenas um, algoritmos baseados em densidade tendem a fundir grupos interseccionados, o que se torna uma característica desejável para situações como esta.

No conjunto de fotografias retornadas em uma região geográfica não teremos apenas imagens de pontos de interesse, por exemplo, uma pessoa pode ter postado fotos do seu próprio jardim, outras pessoas podem ter postado fotos do seu dia a dia em outras áreas que não são potenciais pontos turísticos. Esta situação cria ruído no

conjunto de dados que podem influenciar negativamente o resultado final caso sejam colocados em um grupo.

Dada uma região geográfica, não é possivel estimar previamente quantos pontos de interesse existem. Ponciano (2016) aponta que esta característica da base de dados foi determinante para escolher técnicas baseadas em densidade para realizar o agrupamento das fotografias, uma vez que, diferente de algoritmos como o K-Means, não necessitam da entrada prévia do número de grupos a serem encontrados.

A necessidade de ser capaz de lidar com grupos de diferentes formatos, menor sensibilidade na detecção de grupos interseccionados, resistência a ruídos e a determinação de grupos de forma dinâmica direciona o uso de algoritmos baseados em densidade para o agrupamento dos dados desta aplicação.

Ponciano (2016) desenvolveu o sistema ToPI *Trip*, uma aplicação Web com a finalidade de avaliar o método proposto, e avaliou o método em relação à relevância dos pontos de interesse identificados, os pontos de interesse indentificados pelo TripAdvisor[2]. Além disso também foram considerados quão consistentes os pontos de interesse encontrados são de acordo com o Google Maps[3], em termos de localização geográfica.

Após experimentos e análises dos resultados, nas conclusões de Ponciano (2016) o método foi positivamente avaliado, de forma que os locais identificados estão de acordo com os exibidos por outras plataformas já consolidadas nos cenários analisados.

No próximo capítulo serão apresentadas algumas conclusões acerca da pesquisa desenvolvida nesta monografia, bem como apresentados sugestões de trabalhos futuros.

[2] http://www.tripadvisor.com
[3] https://www.google.com/maps

4 CONCLUSÃO

Uma grande quantidade de dados é gerada todos os dias, e técnicas tradicionais de análise de dados não são suficientes para extrair, de forma eficiente, informações relevantes destes dados. Neste contexto, surge a mineração de dados e as suas tarefas, uma tecnologia para processar e extrair informação de grandes volumes de dados.

O presente trabalho apresentou a análise de grupos, uma das tarefas da mineração de dados, que tem como característica a extração de conhecimento ao gerar agrupamentos, revelando estruturas de relacionamento entre os dados, antes desconhecidas, em conjuntos de dados sobre os quais não se tem, ou se tem poucas informações a priori.

Com o intuito de entender como pode ser realizado um agrupamento de dados, e consequentemente indentificar tais estruturas, foram introduzidos os conceitos dos diferentes tipos de agrupamento, entre eles, na seção 2.2.1.1, o agrupamento particional, que divide o conjunto de dados em grupos que não se interseccionam, e na seção 2.2.1.2, o agrupamento hierárquico, que divide os dados em grupos aninhados. Foram definidos também alguns conceitos de diferentes tipos de grupos, entre eles, na seção 2.2.2.2, os grupos baseados em protótipo , em que um grupo é um conjunto de objetos no qual cada objeto está mais próximo do protótipo que define o seu grupo do que do protótipo de outro grupo, e na seção 2.2.2.4, os grupos baseados em densidade, em que um grupo é uma região densa de objetos que seja rodeada por uma região de baixa densidade.

Os conceitos de tipos e agrupamento e tipos de grupo são fundamentais para a compreensão dos dois algoritmos de agrupamento de dados apresentados e discutidos neste trabalho, o K-Means e o DBSCAN, bem como as suas semelhanças e diferenças.

O K-Means e o DBSCAN, apresentados respectivamente nas seções 2.3.1 e 2.3.2, são considerados dois dos principais algoritmos de agrupamento de dados por serem simples, representativos, úteis para o entendimento da análise de grupos e amplamente utilizados. Foi realizada também uma discussão, na seção 2.3.3, entre o K-Means e o DBSCAN, afim de tornar claro o contexto de utilização de cada um dos

algoritmos, contextos estes, ilustrados por meio de dois estudos de caso, evidenciando os elementos significativos para apoiar a tomada de decisão de se utilizar o K-Means ou o DBSCAN, uma vez que cada um possui suas particularidades, pontos fortes e fracos. Como por exemplo, o K-Means produz um agrupamento particional e completo, divide o conjunto de dados em um número K dado pelo usuário, é baseado em protótipo e agrupa os objetos utilizando como medida de proximidade a distância euclidiana. Já o DBSCAN, que é maior representante dos algoritmos com abordagem de agrupamento baseados em densidade, produz um agrupamento particional e parcial, e possui resistência a ruído no conjunto de dados. Ficou evidente na discussão entre o K-Means e o DBSCAN, que embora ambos sejam algoritmos de agrupamento particionais que produzem agrupamentos exclusivos, a semelhança entre eles praticamente acaba neste ponto, que o K-Means é geralmente mais indicado quando se têm um conjunto de dados muito grande, grupos globulares e há a necessidade de se agrupar todos os objetos, enquanto o DBSCAN é mais promissor que o K-Means em conjuntos de dados que apresentem ruído e grupos com diferentes formatos e tamanhos.

Ainda que simples, o K-Means e o DBSCAN são muito utilizados em diversos domínios, como por exemplo, nas duas aplicações que foram apresentadas no capítulo três, que tiveram como finalidade ilustrar a utilização do agrupamento de dados.

A primeira aplicação (seção 3.1) no domínio da mineração de dados educacionais, teve por objetivo agrupar os agentes de inclusão digital conforme seus perfis de utilização do AVA, e desta forma, entregar ao usuário da aplicação uma partição do conjunto de dados que possa ajudar a classificar qualitativamente os alunos conforme a sua interação. Para este caso, foi verificado que o K-Means é a técnica mais promissora entre as aqui apresentadas, uma vez que, entre outros fatores, existe uma indicação do número de grupos a ser encontrado neste conjunto de dados e todos os elementos devem ser agrupados.

A segunda aplicação (seção 3.2) teve por objetivo identificar pontos de interesse baseado em fotografias geo-referenciadas. Para este caso, o DBSCAN apresenta características mais apropriadas que o K-Means, uma vez que, entre outros fatores, o conjunto de dados possui ruído, os grupos desta aplicação podem ter

tamanhos e formatos distintos, além de não ser possível determinar previamente a quantidade de grupos diferentes em uma dada região.

A diferença entre os domínios destas aplicações dão uma noção da amplitude da área de atuação da análise de grupos, e também deixa evidente a complexidade inerente desta tarefa, que depende de uma observação cuidadosa dos objetivos e do conjunto de dados da aplicação, para em seguida, definir a técnica de agrupamento mais apropriada.

Ao ilustrar conceitos e disponibilizar uma discussão a respeito da análise de grupos e suas aplicações, este trabalho contribui para servir como material de referência para a área acadêmica ou de mercado que necessitem fazer uso de técnicas de agrupamento de dados.

Como limitações deste trabalho podemos listar:

- A discussão entre os algoritmos de agrupamento se restringe ao K-Means e ao DBSCAN.
- No desenvolvimento da discussão para se definir qual técnica utilizar em cada aplicação de exemplo, não foram realizados testes comparativos entre os algoritmos, então a real diferença da partição resultante não pôde ser explicitada.
- Os casos de uso deste trabalho estão limitados a dois domínios de aplicação da análise de grupos, existem diversos outros, por exemplo, técnicas de agrupamento são amplamente utilizados em dados genéticos.
- Não foi mostrado um caso de uso em que tanto o K-Means quanto o DBSCAN não fossem apropriados, como em uma aplicação em que seja necessário utilizar uma técnica de agrupamento hierárquico.

Como complemento e trabalhos futuros, podemos listar:

- Seria relevante abordar outros métodos de agrupamento de dados e suas técnicas, entre elas, apresentar variações do K-Means e DBSCAN, uma vez que estas variações minimizam ou resolvem alguns dos problemas das versões aqui apresentadas, como por exemplo, o K-Medoid elimina o problema do K-Means de não ser capaz de lidar com atributos categóricos.

- Seria interessante introduzir a descrição do algoritmo hierárquico aglomerativo e incluir aplicações de exemplo, que em especial são muito signficativas nas áreas biológica e médica.
- Propor uma discussão incluindo técnicas que implementem abordagens evolutivas, em que candidatos para a solução do problema de agrupamento são codificados como cromossomos, e operadores evolucionários como a seleção, recombinação e mutação são utilizados com o intuito de encontrar uma partição ótima global.

Referências Bibliográficas

ANDERBERG, Michael R. **Cluster analysis for applications**. Office of the Assistant for Study Support Kirtland AFB N MEX, 1973.

BARBARA, D. **An introduction to cluster analysis for data mining.** Retrieved March, v. 13, p. 2006, 2000.

BALL, Geoffrey H.; HALL, David J. **A clustering technique for summarizing multivariate data.** Systems Research and Behavioral Science, v. 12, n. 2, 1967.

BERKHIN, Pavel. **A survey of clustering data mining techniques.** Grouping multidimensional data, 2006.

Brasil. Decreto n.° 6991, de 27 de outubro de 2009. **Institui o Programa Nacional de Apoio à Inclusão Digital nas Comunidades Telecentros.BR, no âmbito da política de inclusão digital do Governo Federal, e dá outras providências.** Diário Oficial [da] Republica Federativa do Brasil, Brasília, DF, n. 206, Seção 1, pág. 3. 2009.

DHILLON, Inderjit S.; MODHA, Dharmendra S. **A data-clustering algorithm on distributed memory multiprocessors.** In: Large-Scale Parallel Data Mining. Springer, Berlin, Heidelberg, 2002. p. 245-260.

DUBES, Richard C. **Cluster analysis and related issues.** In: Handbook of pattern recognition and computer vision. 1993. p. 3-32.

DUDA, Richard O. et al. **Pattern classification. 2nd**. **Edition.** New York, 2001.

ESTER, Martin et al. **A density-based algorithm for discovering clusters in large spatial databases with noise.** In: Kdd. 1996. p. 226-231.

EVERITT, Brian. **Cluster analysis 122**. 1974.

EVERITT, Brian S. et al. **Applied multivariate data analysis**. London: Arnold, 2001.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases.** AI magazine. 1996.

GUTTMAN, Antonin. **R-trees: A dynamic index structure for spatial searching**. ACM, 1984.

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

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. **Data clustering: a review.** ACM computing surveys (CSUR). 1999.

KAUFMAN, L. Rousseeuw; ROUSSEEUW, P. PJ. **Finding groups in data: An introduction to cluster analysis.** Hoboken NJ John Wiley & Sons Inc. 1990.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining.** John Wiley and Sons, Inc, 2005.

PINHEIRO, Márcia F. et al. **Identificação de Grupos de Alunos em Ambiente Virtual de Aprendizagem: Uma Estratégia de Análise de Log Baseada em Clusterização.** In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2014.

PONCIANO, Jean Roberto. **ToPI-uma abordagem online para identificar locais de interesse utilizando fotografias geo-referenciadas.** 2016.

MICHEL. **How many public photos are uploaded to Flickr every day, month, year?** 2018.<https://www.flickr.com/photos/franckmichel/6855169886>. Acessado em: 02/02/2018.

MITCHELL, Tom M. et al. **Machine learning.** 1997. Burr Ridge, IL: McGraw Hill, v. 45, n. 37, 1997.

Moodle (2018). Disponível: http://www.moodle.org. Acesso em 01/02/2018.

DATA VISUALIZATION, **Cluster Analysis: see it** 1st. Online. Disponível em: https://goo.gl/z1cU6J. Acesso em 15.02.2018

SILVA, Aleksandra do Socorro et al. **Análise de Redes Sociais para avaliação e monitoramento de programas de treinamento em larga escala baseados no uso de ambientes de aprendizagem e redes sociais online**. In:II Brazilian Workshop on Social Network Analysis and Mining. 2013.

TAN, Pang-Ning. **Introduction to data mining**. 2006.

THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. **Pattern Recognition**, 2008.

UNWTO. **UNWTO tourism highlights.** [S.l.]: United Nations World Tourism Organization Madrid, 2014.

_____. Over 1.1 billion tourists travelled abroad in 2014. 2015. <http://media.unwto.org/press-release/2015-01-27/over-11-billion-tourists-travelled-abroad-2014>. Acesso em: 01/02/2018.

XU, Rui; WUNSCH, Don. **Clustering**. John Wiley & Sons, 2009.