Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Distância adaptativa e algoritmo do tipo nuvens dinâmicas com dados simbólicos poligonais

Pedro José Carneiro de Souza

Trabalho de Graduação

Recife
25 de abril de 2023

Universidade Federal de Pernambuco
Centro de Informática

Pedro José Carneiro de Souza

Distância adaptativa e algoritmo do tipo nuvens dinâmicas com dados simbólicos poligonais

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: *Renata Maria Cardoso Rodrigues de Souza*

Recife
25 de abril de 2023

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Souza, Pedro José Carneiro de.
Distância adaptativa e algoritmo do tipo nuvens dinâmicas com dados simbólicos poligonais / Pedro José Carneiro de Souza. - Recife, 2023.
29 : il., tab.

Orientador(a): Renata Maria Cardoso Rodrigues de Souza
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Ciências da Computação - Bacharelado, 2023.

1. Agrupamento. 2. Aprendizado não-supervisionado. 3. Análise de dados simbólicos. 4. Distância adaptativa. 5. Dados poligonais. I. Souza, Renata Maria Cardoso Rodrigues de. (Orientação). II. Título.

000 CDD (22.ed.)

Agradecimentos

Aos meus pais, Climário e Solange, agradeço primeira e infinitamente. Por vocês foi o meu passado, é o meu presente e será o meu futuro. Com vocês aprendi o que é ser justo, respeitoso e honesto. Pude entender o que é a vida em sua essência: dores e felicidades, tropeços e conquistas, amores e decepções. Me mostraram que a vida nem sempre será justa, mas que sempre há por quem ou pelo o que tentar novamente.

Aos meus irmãos Felipe e Natália, onde encontrei segurança para dividir várias preocupações e inúmeros ótimos momentos. Obrigado por todo o apoio e por sempre me fazer enxergar o quão eu sou capaz.

Agradeço a minha esposa Gabriella, por me mostrar que há uma vida fora da minha mente metódica. Sempre me mostrou como nossa existência pode ser mais completa quando vivida sem tantos planejamentos e mais improvisos. Obrigado de ser a minha melhor companhia, pelos incentivos e confiança, e me mostrar a vida pode nos levar para qualquer lugar, basta querermos.

Aos meus companheiros de PROAES, agradeço todo o apoio nessa longa caminhada. Nesse grupo que há 9 anos sempre lutou por um serviço público de qualidade, preciso agradecer especialmente ao meu camarada Emanoel, pessoa que me ajudou demais na caminhada acadêmica, profissional e pessoal. Aprendi muito vendo a dedicação e amor que ele aplica em todas as camadas da vida. Não menos importante, agradeço imensamente ao meu camarada Adriano, com quem compartilho curiosidade científica, debates profundos sobre a vida e boas risadas das coisas mais simples da internet, como dublagens feitas por desenhos animados.

Aos colegas e amigos do CIn, que me adotaram como "vovô", mesmo tendo na maioria do percorrer do curso menos que 30 anos de idade. Obrigado por todos os momentos de brincadeira e seriedade, algazarras e concentração pré-prova. Com certeza seria muito mais difícil sem vocês ao meu lado. Carregarei vocês no coração! Viva o *podcast* da área 2!

À minha orientadora Renata agradeço pelo conhecimento, apoio e compromentimento que me fez ter mais certeza sobre a escolha de seguir na carreira acadêmica.

Agradeço a todas as outras pessoas, sendo conhecidos ou não, que contribuiram de alguma forma por conseguir dar meu primeiro passo no que acredito ser minha grande jornada do sonho de Pedrinho: torna-se Professor Pedro.

O homem que diz "sou" não é, porque quem é mesmo é,"não sou"
—BADEN POWELL (Canto de Ossanha)

Resumo

Com o constante crescimento da quantidade de dados que são produzidas pela humanidade, temos cada vez mais a necessidade de traçarmos estratégias que nos auxilie na extração de valores utilizando, por exemplo, mineração de dados. Tendo em vista esta problemática, esse trabalho tem como objetivo apresentar um método de agrupamento para dados simbólicos poligonais utilizando um algoritmo de nuvens dinâmicas com distância adaptativa de Hausdorff. Algoritmos de nuvens dinâmicas têm como objetivo agrupar uma partição em classes, identificando seus representantes baseados na minimização um critério que mede a adequação entre os dados e os protótipos. Acrescido da abordagem adaptativa na distância, estes algoritmos recebem a capacidade de reconhecer grupos com diferentes formas e tamanhos. Para avaliar e demonstrar a utilidade da abordagem proposta, experimentos com dados poligonais sintéticos foram realizados e foram avaliados com base no índice de Rand ajustado.

Palavras-chave: Agrupamento; Aprendizado não-supervisionado; Análise de dados simbólicos; Distância adaptativa; Dados poligonais

Abstract

The constant growth of quantity of data produced by humanity, we need even more to draw strategies to auxiliary us in value extraction using, for example, data mining. Owing to this problem, this paper has as objective presents a dynamic clustering method based on adaptive Hausdorff distance for symbolic polygonal data. Dynamic clustering algorithms aims to group a set of data in clusters, identifying their prototypes based on criterion minimization that measures the adequacy between the data and their prototypes. Adding adaptive distance, dynamic clustering methods turns capable to recognize clusters with different sizes and shapes.To evaluate and show the usefulness of the proposed method, experiment with synthetic polygonal data were perfomed and they were evaluated based on corrected Rand Index.

Keywords: Symbolic data analysis; Polygonal data; Clustering; Unsupervisioned learning; Adaptative distance

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Desde os tempos mais remotos, os povos registravam suas memórias graficamente, fosse nas paredes de uma caverna ou em papiros e pergaminhos. O poder de tornar o conhecimento algo físico nos deu o poder de perpetuar informações através das gerações, habilitando o acúmulo dos registros, dando suporte à continuidade do conhecimento histórico. De posse de dados antigos, redigidos das formas mais diversas possíveis, surge o cuidado em guardá-los e preservá-los por se tornarem cada vez mais necessários para crescimento técnico-científico ao longo do tempo.

Chegando ao final do século 19, temos a concepção dos primeiros sistemas de informação, e sendo um conceito com várias décadas de idade, bancos de dados por muito tempo foram tratados apenas com enfoque operacional e transacional, mas atualmente, com mais poder computacional e de armazenamento, podemos aplicar estratégias mais sofisticadas aos dados, nos auxiliando também em objetivos analíticos e informacionais [11].

Métodos estatísticos, análises exploratórias e procura por padrões em tabelas são exemplos de como atuamos para encontrar valor nos grandes volumes de informações armazenadas. Com isso, como forma a utilizar os dados em objetivos mais complexos, surge a descoberta de conhecimento em bases de dados (Knowledge Discovery in Databases - KDD), uma área de pesquisa que buscar resolver o problema que surge com a posse de bancos de dados e a necessidade de extrair valores deles de forma automática. Para metodificar o KDD, podemos organizar sua lógica em três etapas: o pré-processamento dos dados aparece como primeira etapa do KDD para que tenhamos segurança na qualidade dos dados; já a segunda etapa, a descoberta do conhecimento propriamente dito, toma forma com a escolha de uma estratégia ou algoritmo de mineração de dados para nos fornecer as informações de valor da massa de dados; e por último e não menos importante, avaliamos o conteúdo extraído para apoiar-nos em alguma tomada de decisão [17].

São muitas as técnicas e estratégias que utilizam dessa linha de processo, mas ainda surgem desafios que constantemente encaramos no que se diz respeito à análise de grande base de dados. O enorme volume das bases é combustível para várias técnicas de inteligência artificial, como aprendizagem de máquina, onde o sistema aprende com exemplos para reconhecer padrões ou extrair valor que possa não ser possível de ser visto por um humano, mas a alta dimensionalidade às vezes é um problema até para métodos que se alimentam desses grande conjunto de informações.

A partir desse problema, percebe-se a necessidade de representar os dados de formas mais complexas, produto de agregação de dados tradicionais em estruturas mais complexas, como por exemplo de dados representados por um intervalo, chamado de dados intervalares, são representados por uma estrutura bivariada composta pelo mínimo e máximo daqueles dados que

estão sendo representados. Tendo essas representações formatos e estruturas diferentes dos dados tradicionais, visualiza-se vários problemas para os métodos tradicionais de extração de valores dos dados, já que estes métodos não estão preparados para lidar com essas representações. Com isso, surge a análise de dados simbólicos (*Symbolic Data Analysis* - SDA) [5] para generalizar as técnicas tradicionais para essas representações mais complexas. Métodos de agrupamento, regressões e outras metodologias estatísticas são propostas para os diversos tipos de dados simbólicos presentes na literatura de SDA. Do mesmo modo que há a possibilidade de surgirem novos métodos para extração de valor dos dados, também é factível novas formas de representar os dados como estruturas mais complexas.

Como nova concepção de representação de dados, Silva et al. [15] propõe representar os dados em polígonos, estes constituídos por um centro e um raio, que representam, respectivamente, a média e duas vezes o desvio padrão dos dados agregados representados. Como nova forma de dados simbólicos, o poligonal também sofre da necessidade da construção de métodos de extração de valor assim como os dados tradicionais e com isso, surge o método de nuvens dinâmicas para dados simbólicos poligonais [18], onde por meio de aprendizagem não supervisionada, conseguimos agrupar em k clusters um conjunto p-dimensional de dados poligonais.

Tendo em vista que a construção da literatura em torno de dados simbólicos e, nesse caso, dados simbólicos poligonais, esse trabalho vem acrescentar uma estratégia já conhecida em *SDA* ao método exposto anteriormente. Propõe-se um algoritmo de nuvens dinâmicas com distância adaptativa para dados simbólicos poligonais. Este método adiciona ao algoritmo de aprendizagem não-supervisionada proposto por Silva et al. a estratégia de distâncias adaptativas, que tem como vantagem permitir que o algoritmo reconheça diferentes tamanhos e formatos de clusters. [1]

Este trabalho está organizado da seguinte forma: o capítulo 2 apresenta uma visão geral sobre aprendizado não supervisionado e o seu método mais comum para dados tradicionais, o k-médias; no capítulo 3 encontramos uma breve introdução à análise simbólica de dados e ao método de nuvens dinâmicas com distância adaptativa aplicado a dados intervalares; no capítulo 4 apresenta-se dados simbólicos poligonais e o método de nuvens dinâmicas para esse tipo de dado; seguindo para o capítulo 5, temos a proposta de algoritmo de nuvens dinâmicas com distância adaptativa para dados simbólicos poligonais; o capítulo 6 traz os resultados do método anteriormente proposto; e conclui-se o trabalho com o capítulo 7, onde levanta-se as considerações finais e possíveis trabalhos futuros.

CAPÍTULO 2

Fundamentação Teórica

O conhecimento, independente da área, parte da vontade de alguém tentar responder alguma pergunta, resolver algum problema ou em construir herança teórica e prática sobre a natureza dos seres e das coisas. Contudo, podemos partir de conceitos e teorias já estabelecidas ou até mesmo questioná-las, mas nunca poderemos ignorar o empenho de quem dedicou seu tempo para contribuir com a evolução intelectual.

Portanto, precisamos nos apropriar pelo menos um pouco de alguns conceitos que nos auxiliam na formação da proposta desse trabalho.

2.1 Análise de Dados Simbólicos

Com o objetivo de lidar com o crescimento das bases de dados e suas enormes dimensionalidades, surge como alternativa representar esses dados de modo mais compacto. Dados simbólicos são obtidos pelo resultado de agregação de dados tradicionais individuais [5], formando estruturas mais complexas que trazem informação do grupo representado. A construção dessas representações devem observar a conservação máxima de informações, ao mesmo tempo que a dimensionalidade da tabela de dados seja reduzida, enfrentando assim o problema que dados simbólicos se propõe a combater.

Dados intervalares, por exemplo, representa um conjunto de dados por uma forma bivariada contendo o mínimo e o máximo do conjunto para aquela variável, já dados poligonais dar-se por centro e raio, sendo obtidos respectivamente pela média e duas vezes o desvio padrão dos dados tradicionais agregados.

Após aplicação dessa estratégia de mudança de representação, a tabela de dados tradicionais dá lugar à tabela de dados simbólicos, onde encontramos dados simbólicos nas células ao invés de dados clássicos. E as diferenças entre dados tradicionais e dados simbólicos vão além do formato que se apresentam nas tabelas, mostrnando-se presente também na forma que aplicamos metodologias para extração de valor desses dados. Portanto tornando-se as tabelas de dados simbólicos a fonte de informações que alimentarão os métodos de mineração de dados e inteligência artificial, teremos que analisar como adaptar os métodos para dados tradicionais para essa nova representação. Não podemos apenas aplicar as metodologias construídas para dados clássicos, já que estamos tratando de estruturas mais complexas, mas também temos que nos atentar se o método apresentado para outro tipo de dado simbólico se aplica para o que estamos lidando. Com isso, percebe-se que a construção de metodologias estatísticas faz-se necessária pelo crescimento de dados de natureza simbólica.

Com isso, a análise de dados simbólicos vem apresentar diversos métodos para extrair va-

lor das tabelas de dados símbolicos, e com isso possibilitando a utilização dessa estratégia representativa de agregação da dados como matéria para análises estatísticas, observando a particularidade de cada tipo de representação. Podemos nos guiar pelo entendimento geral e as proposições do método e suas devidas estratégias de extrair conhecimento dos dados, mas como já vimos que os dados simbólicos têm particularidades entre cada um deles, precisa-se estudar se a metodologia escolhida e seu ferramental cabe para a tal representação.

2.2 Dados simbólicos Poligonais

Muitas representações simbólicas de dados compõe, a literatura, contudo outras estratégias são propostas para reforçar o arcabouço de alternativas para problemas já conhecidos e também para novos desafios que surgem a todo tempo. Dados simbólicos poligonais, proposto em [15], é uma nova forma de representar simbolicamente dados num formato bivariado.

Segundo o autor, dados poligonais têm uma estrutura mais complexa do que dados intervalares, trazendo principalmente mais acurácia no cálculo de medidas descritivas. Juntamente dessa vantagem sobre dados intervalares, dados poligonais carregam mais informação em sua estrutura, melhorando a extração de valor com aplicação de mineração de dados.

Para os métodos apresentados neste trabalho, iremos utilizar de vetores de polígonos para representar os objetos, por isso devemos nos apropriar de como dar-se a construção dessa abordagem poligonal. Seja Ω um conjunto de m objetos indexados por i e descritos por p variáveis indexadas por j. Seja Z uma variável aleatória onde $Z : \Omega \to \mathbb{R}^2$ [15]. Para cada $i \in \Omega$, $Z(i) = \{(a_{i1}, b_{i1}), \ldots, (a_{i\ell}, b_{i\ell})\}$, onde $\ell = 1, \ldots, L$ for $L \geq 3 \in \mathbb{N}$ é o número de vértices. Adicionalmente, cada objeto i é representado como um vetor de polígonos $\mathbf{z}_{i1} = (z_{i1}, \ldots, z_{ip})$. Deste modo, dados poligonais podem ser construídos como

$$z_{ij\ell} = \begin{bmatrix} c_{ij} \\ c_{ij} \end{bmatrix} + r_{ij} \begin{bmatrix} \cos(2\pi\ell/L) \\ \sin(2\pi\ell/L) \end{bmatrix}, \tag{2.1}$$

onde $z_{ij\ell}$ indica os vétices do polígono i na variável j, c_{ij} e r_{ij} são o centro e o raio do polígono i para a variável j, respectivamente. O par (c_{ij}, r_{ij}) corresponde à média e à duas vezes o desvio padrão da descrição associada à classe i para a variável contínua X_j [15].

Ainda em [15], o autor contribui com a literatura dos dados poligonais, com um modelo de regressão linear, o que inicia o trabalho que será incrementado com o primeiro algoritmo de nuvens dinâmicas para dados poligonais apresentando em [18].

2.3 Métodos de agrupamento

Como estratégia para extrair valores de conjuntos de dados, métodos de agrupamento são amplamente utilizados para reconhecimento de padrões nos objetos analisados, nos permitindo visualizar informações que nos auxilie na resolução de problemas. Podemos usar dessa metodologia não apenas com dados do tipo clássico, mas precisamos adaptar os métodos para dados simbólicos por conta das diferenças presentes em comparação aos dados tradicionais.

Tendo como objetivo a construção de programas que melhorem seu desempenho por meio de exemplos [14], técnicas de aprendizagem de máquina são orientadas a dados, isto é, aprendem automaticamente a partir de grandes volumes de dados [12], volumes esses definidores do problema capital.

Métodos de agrupamento são estratégias de aprendizagem não supervisionada, aprendizagem essa que trabalha com conjuntos de dados não rotulados, cabendo ao método agrupar os dados de acordo com a similaridade dos seus atributos. Em resumo, podemos afirmar que esses métodos buscam encontrar grupos que os elementos tenham um alto grau de similaridade com elementos do seu mesmo grupo e os grupos tenham um alto grau de dissimilaridade entre eles.

Como forma de organizar esses métodos de agrupamento, define-se dois grandes grupos: hierárquicos e de partição [8]. A classificação hierárquica consiste numa sequência de partições, partindo de n classes unitárias, finalizando em uma única classe com todos os elementos, geralmente representada por um dendograma. Já a classificação de partição apresenta um número predefinido de grupos disjuntos k para um conjunto de n elementos, onde $k \leq n$ grupos disjuntos [17].

Os métodos de partição possuem duas grandes linhas: agrupamento *hard* e agrupamento *fuzzy*. Agrupamento *hard* consiste que cada elemento do conjunto que está sendo particionado pertença a um e somente um grupo. Já em métodos com estratégia *fuzzy*, os elementos possuem um grau de pertencimento para cada grupo, deixando de ser uma atribuição rígida (método *hard*) para ser uma atribuição difusa (método *fuzzy*).

2.4 Algoritmos do tipo nuvens dinâmicas

Como exemplo de método direcionado para agrupamento não hierárquico, algoritmos do tipo nuvens dinâmicas têm como objetivo obter uma partição de um conjunto de n elementos em um número predefinido k de classes, ao mesmo tempo que identifica um conjunto de protótipos ou representantes das classes minimizando um critério para mensurar a adequação entre as classes e os protótipos [4]. Mesmo uma das vantagens sendo exatamente a formulação de um problema de classificação em termos de uma otimização de um critério de adequação entre as classes e os protótipos, os mesmos enfrentam o problema de que a convergência nessa metodologia depende muito da configuração inicial dos protótipos como da escolha da função de representação de acordo com a distância entre um grupo e seu representante [17].

Essa estratégia busca a convergência repetindo duas etapas até que alcance um valor estacionário. Iniciando com um conjunto de protótipos aleatório, aplica-se iterativamente a etapa de alocação que atribui a cada elemento uma classe na qual a proximidade entre o elemento e o protótipo do grupo é mínima. Logo após, inicia-se a etapa de representação, na qual os protótipos são atualizados de acordo com o resultado da etapa de alocação. Como a qualidade do algoritmo pode ser muito afetada pela configuração inicial, tem-se como estratégia para vencer esse problema executar o algoritmo diversas vezes e escolher dentre elas a melhor configuração.

SDA apresenta diversos algoritmos de nuvens dinâmicas para dados intervalares. O cálculo de pesos nas distâncias é um ponto importante e bem explorado nos trabalhos, formando o entendimento de distância adaptativa, que vem atacar problemas que possuem clusters de tamanhos e formatos distintos, fazendo com que a estratégia de computar pesos para cada variável

dentro do grupo seja de suma importância.

Como trabalhos relacionados com o que foi exposto anteriormente, podemos citar alguns como:

1. Algoritmos do tipo nuvens dinâmicas para dados intervalares com e sem distâncias adaptativas baseados na distância City-Block são apresentados por Souza e De Carvalho. [16]

2. Algoritmos do tipo nuvens dinâmicas baseados na distância L_2 são propostos por De Carvalho, Brito e Bock [2].

3. Algoritmos do tipo nuvens dinâmicas baseados em distância de Hausdorff adaptativa e não-adaptativa, proposto por De Carvalho et al.[1].

2.5 Distância Adaptativa

Tendo como exemplo algoritmos de agrupamento como o apresentado na seção 2.4 desse capítulo, as iterações se dão, de modo geral, em duas etapas que constroem nesse momento novos agrupamentos e seus respectivos representantes através da busca pela otimização de uma função de adequação. No K-médias, por exemplo, temos a distância euclidiana sendo usada para o papel de minimizar o critério de adequação. Contudo, viu-se que distâncias determinadas de forma absoluta tendem a não ter a devida sensibilidade para que o algoritmo reconheça as diversas formas e tamanhos que um grupo pode tomar.

Tomando como inspiração o cálculo de pesos apresentado em [7], onde aparece presente no algoritmo dinâmico baseado em distâncias adaptativas, podemos seguir a estratégia proposta e aplicar para o método que pretendemos estudar. No caso desse trabalho, aplicamos a metodologia apresentada em outras publicações como [1] e [17], alcançando resultados que demonstram o ganho de performance diante do método que não utiliza da distância adaptativa.

Da mesma forma, o algoritmo de nuvens dinâmicas com distância adaptativa tem como objetivo obter uma partição dos dados e um conjunto de representantes para as classes otimizando um critério de adequação, mas diferentemente do método sem distâncias adaptativas, consegue reconhecer formatos e tamanhos diferentes.

CAPÍTULO 3

Algoritmo de nuvens dinâmicas para dados poligonais

Continuando o trabalho de construção de literatura para dados poligonais, Silva et al propõe em [18] um algoritmo de nuvens dinâmicas para dados poligonais.

Para isso, precisamos definir a métrica de similaridade entre esses dados, pois é onde reside toda a aplicação do algoritmo proposto.

3.1 Distância de Hausdorff para dados poligonais

Dado dois objetos i e v de S descritos por p variáveis poligonais $z_i = (z_{i1}, \ldots, z_{ip})$ e $z_v = (z_{v1}, \ldots, z_{vp})$, sendo z_{ij} ez_{vj} representados pelos pares (c_{ij}, r_{ij}) e (c_{ij}, r_{ij}), respectivamente, para $j = 1, \ldots, p$. A distância entre z_i e z_v é definida como:

$$d(\mathbf{z}_i, \mathbf{z}_v) = \sum_{j=1}^{p} \Phi(z_{ij}, z_{vj}), \tag{3.1}$$

onde Φ representa a distância entre os objectos i e v para a variável j, sendo Φ como a distância de Hausdorff entre dois polígonos indenpendentemente do número de lados como:

$$\Phi(z_{ij}, z_{vj}) = \sqrt{2} \mid c_{ij} - c_{vj} \mid + \mid r_{ij} - r_{vj} \mid, \tag{3.2}$$

3.2 O algoritmo

Seja $P = \{C_1, \ldots, C_k, \ldots, C_K\}$ uma partição de S em K grupos. Cada grupo $C_k \in P$ é descrito por um vetor de variáveis simbólicas poligonais $w_k = (w_{k1}, \ldots, w_{kp})$ chamadas de protótipo.

O objetivo do método é encontrar uma partição $\mathscr{S}$ em K grupos e um conjunto de protótipos $\{w_k\}$ $(k = 1, \ldots, K)$ minimizando o seguinte critério:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d(z_i, w_k) \tag{3.3}$$

com

$$d(z_i, w_k) = \sum_{j=1}^{p} \Phi(z_{ij}, w_{kj}) = \sqrt{2} \sum_{j=1}^{p} \mid c_{ij} - \beta_{kj} \mid + \sum_{j=1}^{p} \mid r_{ij} - \gamma_{kj} \mid, \tag{3.4}$$

onde β_{kj} e γ_{kj} é o centro e o raio do k-ésimo polígono, respectivamente, representando o k-ésimo protótipo do grupo C_k e j-ésima variável simbólica poligonal.

O método inicia com uma partição aleatória e alterna entre dois passos (identificação dos protótipos e construção dos grupos) até a convergência, quando o critério J alcança um valor estacionário que representa do mínimo local.

3.3 Definição dos melhores protótipos

O critéiro J pode ser reescrito como

$$J = \sum_{k=1}^{K} J(k) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \Phi(z_{ij}, w_{kj}), \tag{3.5}$$

o critério $J(k)$ por ser aditivo, o problema torna-se encontrar um polígono para a variável j $(j = 1, \ldots, p)$ que minimiza

$$J_{kj} = \sqrt{2} \sum_{i \in C_k} \mid c_{ij} - \beta_{kj} \mid + \sum_{i \in C_k} \mid r_{ij} - \gamma_{kj} \mid . \tag{3.6}$$

O par $(\beta_{kj}, \gamma_{kj})$ é atualizado por β_{kj} e γ_{kj}, sendo eles a mediana dos conjuntos $\{c_{ij}\}$ e $\{r_{ij}\}$, respectivamente, para $\forall i \in C_k$.

Demonstração. Desde que o critério J_{kj} na Equação (3.6) seja aditivo, ele produz dois problemas de minimização já conhecidos em L_1: encontrar β_{kj} e γ_{kj} que minimiza, respectivamente, $\sqrt{2}\sum_{i \in C_k} \mid c_{ij} - \beta_{kj} \mid$ e $\sum_{i \in C_k} \mid r_{ij} - \gamma_{kj} \mid$.

A prova para cada um desses problemas de minimização pode ser encontrado em [7]. Com isso, o critério J_{kj} é minimizado. □

3.4 Definição da melhor partição

Assuma que os protótipos $(\beta_{kj}, \gamma_{kj})$ para $k = 1, \ldots, K$ e $j = 1, \ldots, p$ são fixos. Os *clusters* C_k $(k = 1, \ldots, K)$ que minimizam o critério J são atualizados de acordo com a seguinte regra de atribuição:

$$C_k = \{i \in S : d(z_i, w_k) \leq d(z_i, w_h) \forall h \neq k (h = 1, \ldots, K)\} \tag{3.7}$$

3.5 O algoritmo

O algoritmo 1 é descrito em quatro passos: inicialização, representação, atribuição e finalização.

Algorithm 1 Algoritmo *hard* para dados poligonais

Data: Número de *clusters* K, $2 \leq K < m$; *Dataset* $\mathscr{S}$;
Resultado: Os *clusters* C_k, Critério J
 # *Inicialização*
Escolha aleatoriamente $(w_1, w_2, \ldots, w_K)$ K prototypes dentre os elementos de $\mathscr{S}$
for $k = 1 : K$ **do**
  $C_k \leftarrow \emptyset$
end for
for $i = 1 : m$ **do**
  # Atribuia i-ésimo objeto ao cluster C_l *tal que:*
  $l = argmin_{k=1,\ldots,K} d(z_i, w_k)$
  $C_l = C_l \cup \{i\}$
end for
 # *Representação*
Compute o melhor protótipo w_k.
 # *Alocação*
$mudou \leftarrow 0$
for $i = 1 : m$ **do**
  $l = argmin_{k=1,\ldots,K} d(z_i, w_k)$
  if $i \in C_k$ e $l \neq k$ **then**
    $mudou \leftarrow 1$
    $C_l = C_l \cup \{i\}$
    $C_k = C_k - \{i\}$
  end if
end for
 # *Finalização*
if $mudou = 0$ **then**
  Parar
else
  Ir para Representação
end if

CAPÍTULO 4

Algoritmo de nuvem dinâmica para dados poligonais com distâncias adaptativas

A principal ideia de algoritmo do tipo nuvens dinâmicas com distâncias adaptativas é associar uma distância d_k a cada *cluster* C_k e seu protótipo y_k de modo que a soma das distâncias $d_k(z_i, w_k)$ entre os elementos $i \in C_k$ e o protótipo w_k seja o melhor possível [1]. O critério de agrupamento é

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d_k(z_i, w_k), \tag{4.1}$$

mas assumindo a estratégia adaptativa da distância, temos que a d_k é a soma ponderada das distâncias d^j, onde d^j compara um par de objetos de acordo com a variável j

$$d_k(z_i, z_{i'}) = \sum_{j=1}^{p} d^j(z_i^j, z_{i'}^j) = \sum_{j=1}^{p} \lambda_k^j d(z_i^j, z_{i'}^j), \tag{4.2}$$

e de posse da equação 3.4 temos

$$d_k(z_i, w_k) = \sum_{j=1}^{p} \lambda_k^j \Phi(z_{ij}, w_{kj}) = \sqrt{2} \sum_{j=1}^{p} \lambda_k^j \mid c_{ij} - \beta_{kj} \mid + \sum_{j=1}^{p} \lambda_k^j \mid r_{ij} - \gamma_{kj} \mid, \tag{4.3}$$

sendo $d^j(x_i^j, x_{i'}^j) = \lambda_k^j d(x_i^j, x_{i'}^j)$, $\lambda_k^j > 0$ e $\prod_{j=1}^{p} \lambda_k^j = 1$.

4.1 Definição dos melhores protótipos

Tendo como base De Carvalho et al. [1], partimos do entendimento que a etapa de representação possui dois estágios, correspondendo a dois problemas de minimização. Partindo de uma partição P e o vetor de pesoas λ fixos, o primeiro problema é achar para $k = 1, \ldots, K$ o protótipo y_k que minimiza o critério de adequação $\sum_{i \in C_k} d_k(x_i, y_k)$. Pela definição de d_k em 4.3 e com $x_i = (x_i{}^1, \ldots, x_i{}^p)$ e $y_i = (y_i{}^1, \ldots, y_i{}^p)$, o critério de adequação é:

$$\sum_{i \in C_k} d_k(x_i, y_k) = \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j \Phi(x_{ij}, y_{kj}) = \sum_{j=1}^{p} \lambda_k^j \sum_{i \in C_k} \Phi(x_{ij}, y_{kj}) \tag{4.4}$$

Dado como fixo o vetor de pesos, o problema torna-se encontrar para $j = 1, \ldots, p$ o polígono y_{kj} que minimiza

$$\sum_{i\in C_k} d_k(x_i,y_k) = \sqrt{2}\sum_{i\in C_k} \mid c_{ij} - \beta_{kj} \mid + \sum_{i\in C_k} \mid r_{ij} - \gamma_{kj} \mid . \tag{4.5}$$

O problema de minização em 4.5 é o já demonstrado em [18] na equação 3.6.

4.2 Definição das melhores distâncias

O segundo estágio do passo de representação do algoritmo de nuvens dinâmicas com distâncias adaptativas dar-se quando temos a partição P e o conjunto de protótipos G fixos e buscamos para $k = 1,\ldots,K$ o vetor de pesos λ_k que minimiza o critério de adequação apresentado em 4.4

$$\sum_{i\in C_k} d_k(x_i,y_k) = \sum_{j=1}^{p} \lambda_k^j \Phi_j \quad \text{onde } \Phi_j = \sum_{i\in C_k} d(x_{ij},y_{kj}) \tag{4.6}$$

Assim como levantado por De Carvalho et al [1], segue-se o entendimento de Diday e Govaert [6], os pesos λ_k^j são calculados pelo método dos multiplicadores de Lagrange

$$\frac{\partial}{\partial \lambda_k^j}\left(\sum_{j=1}^{p} \lambda_k^j \Phi_j - \mu \prod_{h=1}^{p} \lambda_k^h \right) = 0 \quad \text{for} j = 1,\ldots,p \tag{4.7}$$

A partir da equação 4.7, temos o seguinte resultado:

$$\Phi_j - \mu \frac{\prod_{h=1}^{p} \lambda_k^h}{\lambda_k^j} = 0 \Rightarrow \lambda_k^j = \frac{\mu}{\Phi_j}\left(\prod_{h=1}^{p} \lambda_k^h \right) \tag{4.8}$$

Relembrando que $\prod_{h=1}^{p} \lambda_k^h = 1$, o parâmetro λ_k^j na equação 4.8 é dado por

$$\lambda_k^j = \frac{\mu}{\Phi_j} \tag{4.9}$$

A restrição $\prod_{h=1}^{p} \lambda_k^h = 1$ pode ser escrita como

$$1 = \prod_{h=1}^{p} \frac{\mu}{\Phi_j} = \frac{\mu^p}{\prod_{h=1}^{p} \Phi_h} \quad \text{logo } \mu = \left(\prod_{h=1}^{p} \Phi_h \right)^{\frac{1}{p}} \tag{4.10}$$

Chegando na solução $\hat{\lambda_k^j}$ para o parâmetro λ_k^j é

$$\hat{\lambda_k^j} = \frac{\mu}{\Phi_j} = \frac{\left[\prod_{h=1}^{p} \left(\sqrt{2}\sum_{i\in C_k} \mid c_{ih} - \beta_{kh} \mid + \sum_{i\in C_k} \mid r_{ih} - \gamma_{kh} \mid \right) \right]^{\frac{1}{p}}}{\sqrt{2}\sum_{i\in C_k} \mid c_{ij} - \beta_{kj} \mid + \sum_{i\in C_k} \mid r_{ij} - \gamma_{kj} \mid} \tag{4.11}$$

4.3 O algoritmo

O algoritmo de nuvens dinâmicas com distâncias adaptativas para dados simbólicos poligonais é definido de acordo com o pseudocódigo 2.

Por conta da possibilidade de modificar os passos de de inicialização e finalização nos métodos tradicionais de algoritmos de nuvens dinâmicas, podemos traçar alguma estratégia de escolher os objetos iniciais de forma que sejam os menos similares possível entre eles, assim como determinar os pesos λ_k^j como 1, temos o algoritmo equivalente ao sem distância adaptativa [1].

Algorithm 2 Algoritmo de nuvens dinâmicas com distâncias adaptativas para dados poligonais

Data: Número de *clusters* K, $2 \leq K < m$; *Dataset* $\mathscr{S}$;
Resultado: Os *clusters* C_k, Critério J
 # *Inicialização*
Escolha aleatoriamente $(w_1, w_2, \ldots, w_K)$ K prototypes dentre os elementos de $\mathscr{S}$
for $k = 1 : K$ **do**
  $C_k \leftarrow \emptyset$
end for
for $i = 1 : m$ **do**
  # *Atribuia i-ésimo objeto ao cluster* C_l *tal que:*
  $l = argmin_{k=1,\ldots,K} d(z_i, w_k)$
  $C_l = C_l \cup \{i\}$
end for
 # *Representação*
# Compute o melhor protótipo w_k.
for $j = 1 : p$ **do**
  for $k = 1 : K$ **do**
    # Compute $\hat{\lambda}_k^j$.
  end for
  # *Atribuia i-ésimo objeto ao cluster* C_l *tal que:*
  $l = argmin_{k=1,\ldots,K} d(z_i, w_k)$
  $C_l = C_l \cup \{i\}$
end for
 # *Alocação*
$mudou \leftarrow 0$
for $i = 1 : m$ **do**
  $l = argmin_{k=1,\ldots,K} \hat{\lambda}_k^j d(z_i, w_k)$
  if $i \in C_k$ e $l \neq k$ **then**
    $mudou \leftarrow 1$
    $C_l = C_l \cup \{i\}$
    $C_k = C_k - \{i\}$
  end if
end for
 # *Finalização*
if $mudou = 0$ **then**
  Parar
else
  Ir para Representação
end if

CAPÍTULO 5

Experimentos e resultados

Para demonstrar a utilidade da estratégia proposta por esse trabalho, utilizaremos duas bases de dados simulados com *clusters* de diferentes tamanhos e formatos [9]. Como objetivo, temos a comparação entre os métodos com distância adaptativa e com distância não adaptativa.

Para avaliar a diferença entre as estratégias sobre o método de agrupamento utilizaremos um índice de validação externo [10]. Essa validação externa é dada pela comparação entre a partição original dos dados e a partição encontrada pelo algoritmo. Para isso, utilizaremos o índice de Rand ajustado (ARI) [10] para comparar as partições. Para cada conjunto de dados simulados, é avaliado o ARI através do experimento Monte Carlo, sendo 100 repetições do algoritmo para cada caso. Como métrica final, considera-se a média e desvio padrão dos índices das 100 repetições.

5.1 Dados simulados

Como objeto para avaliarmos o método proposto por esse trabalho, utilizaremos de dados organizados em dois conjuntos e simulados a partir de quatro parâmetros para obtermos os dados simbólicos poligonais. A metodologia organiza-se nos três seguintes passos:

1. **Dados simulados**: Os dados que darão início à montagem da tabela de dados simbólicos são simulados no $\mathbb{R}^2$ de acordo com diferentes distribuições normais bivariadas de componentes independentes. Cada conjunto de dados simulados tem 350 elementos, sendo distribuído em três clusters de tamanhos diferentes: dois clusters com 150 elementos e um cluster com 50 elementos.
   O conjunto 1, ilustrado na Figura 5.1, é composto de *clusters* bem separados, seguindo as seguintes distribuições normais bivariadas:
   a) Classe 1: $\mu = (28,22)^T$, $\sigma_1^2 = 100, \sigma_2^2 = 9$, e $\sigma_{12} = 0$;
   b) Classe 2: $\mu = (60,30)^T$, $\sigma_1^2 = 9, \sigma_2^2 = 144$, e $\sigma_{12} = 0$.
   c) Classe 3: $\mu = (45,38)^T$, $\sigma_1^2 = 9, \sigma_2^2 = 9$, e $\sigma_{12} = 0$.
   O conjunto 2, ilustrado na Figura 5.2, é composto de *clusters* com sobreposição, seguindo as seguintes distribuições normais bivariadas:
   a) Classe 1: $\mu = (45,22)^T$, $\sigma_1^2 = 100, \sigma_2^2 = 9$, e $\sigma_{12} = 0$;
   b) Classe 2: $\mu = (60,30)^T$, $\sigma_1^2 = 9, \sigma_2^2 = 144$, e $\sigma_{12} = 0$.
   c) Classe 3: $\mu = (52,38)^T$, $\sigma_1^2 = 9, \sigma_2^2 = 9$, e $\sigma_{12} = 0$.
2. **Dados de classe**: A partir de cada semente bivariada $(s_1, s_2)^T$ são gerados dados bivariados de classe. O tamanho n de uma classe é definido pela distribuição uniforme U[15, 20]. Os elementos de cada classe $\{u_1, \ldots, u_n\}$ são gerados a partir de uma distribuição

de probabilidade bivariada com componentes independentes. Dado um $n\sim$U[15,20], o vetor bivariado (u_1, u_2) pode ser definido como:

- Normal: os componentes u_1 e u_2 seguem $N(s_1, \delta)$ e $N(s_2, \delta)$, respectivamente.
- Uniforme: ambos componentes u_1 e u_2 seguem $U[\delta_1, \delta_2]$. respectivamente.

A tabela 5.1 apresenta os parâmetros usados para gerar as classes de acordo com cada distribuição de probabilidade, que valem para os *datasets* simulados 1 e 2.

3. **Dados simbólicos poligonais**: Cada classe é um subconjunto da agregação dos elementos que podem ser descritos por dados poligonais, representados pelo seu centro e seu raio. Dessa forma, os *datasets* são construídos.

Distribuição	
Uniform	Normal
$\delta_1 = 1, \delta_2 = 8$	$\delta = 1$
$\delta_1 = 1, \delta_2 = 16$	$\delta = 3$
$\delta_1 = 1, \delta_2 = 32$	$\delta = 5$
$\delta_1 = 1, \delta_2 = 40$	$\delta = 7$
-	$\delta = 16$
-	$\delta = 25$
-	$\delta = 36$
-	$\delta = 49$

Tabela 5.1 Parâmetros usados para gerar os dados para as tabelas de dados simbólicos 1 e 2

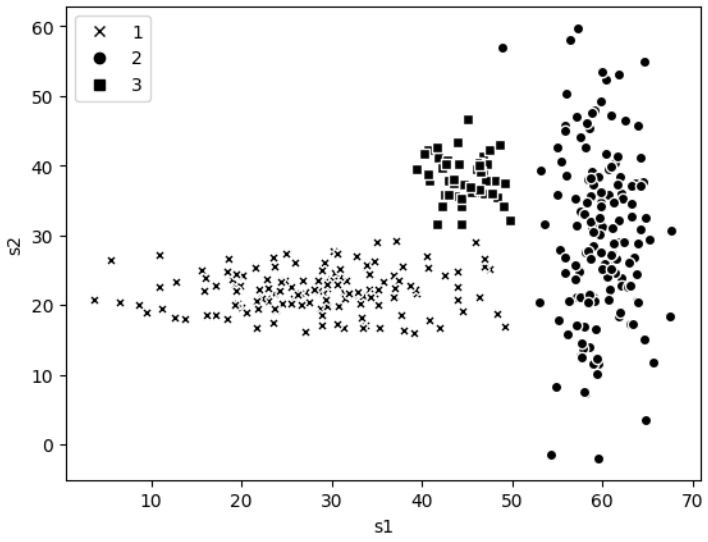

Figura 5.1 Conjunto de dados (sementes) simulados com classes bem separadas - *dataset* 1

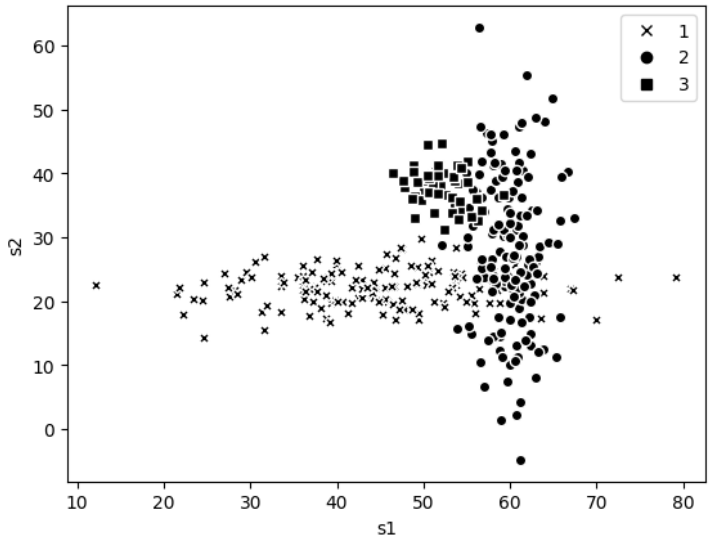

Figura 5.2 Conjunto de dados (sementes) simulados com classes sobrepostas - *dataset* 2

5.2 Resultados

A Tabela 5.2 mostra a média e o desvio padrão (esse último entre parênteses) do índice de Rand ajustado para os algoritmos do tipo nuvens dinâmicas para dados poligonais com e sem distância adaptativa para os *datasets* 1 e 2.

Considerando diferentes parâmetros para as distribuições normal e uniforme, podemos observar que à medida que a variabilidade interna de cada classe aumenta, o valor do índice de Rand diminui. Esse resultado é esperado uma vez que o aumento de variabilidade torna mais difícil a separação linear dos grupos. Além disso, podemos notar que:

1. **Uniforme**: Em todos os cenários, percebe-se um ganho de desempenho com a aplicação da estratégia adaptativa para a distância.

2. **Normal**: Como na distribuição uniforme, temos ganhos de desempenho também na distribuição normal.

De acordo com os resultados dos experimentos, podemos afirmar que seguir com o uso de distâncias adaptativas para calcular a medida de adequeção do algoritmo de nuvens dinâmicas para dados poligonais melhora a qualidade do agrupamento e também ver-se a melhora da capacidade de lidar com conjuntos de dados com formatos e tamanhos diferentes.

		Conjunto de Dados 1		Conjunto de Dados 2	
	Parâmetros	Não adaptativa	Adaptativa	Não adaptativa	Adaptativa
Uniforme	$[\delta_1, \delta_2] = [1, 8]$	0.675 (0.120)	0.747 (0.168)	0.380 (0.035)	0.427 (0.066)
	$[\delta_1, \delta_2] = [1, 16]$	0.687 (0.119)	0.765 (0.157)	0.378 (0.039)	0.420 (0.072)
	$[\delta_1, \delta_2] = [1, 32]$	0.641 (0.108)	0.712 (0.142)	0.364 (0.036)	0.393 (0.055)
	$[\delta_1, \delta_2] = [1, 40]$	0.640 (0.104)	0.667 (0.123)	0.358 (0.033)	0.383 (0.051)
Normal	$\delta = 1$	0.683 (0.126)	0.758 (0.164)	0.387 (0.041)	0.440 (0.078)
	$\delta = 3$	0.680 (0.121)	0.749 (0.170)	0.381 (0.032)	0.433 (0.077)
	$\delta = 5$	0.687 (0.128)	0.733 (0.165)	0.377 (0.037)	0.434 (0.078)
	$\delta = 7$	0.680 (0.132)	0.734 (0.158)	0.379 (0.032)	0.430 (0.063)
	$\delta = 16$	0.677 (0.128)	0.726 (0.159)	0.385 (0.039)	0.433 (0.071)
	$\delta = 25$	0.676 (0.123)	0.740 (0.159)	0.376 (0.035)	0.415 (0.054)
	$\delta = 36$	0.653 (0.111)	0.745 (0.152)	0.373 (0.034)	0.410 (0.065)
	$\delta = 49$	0.658 (0.110)	0.711 (0.138)	0.376 (0.038)	0.407 (0.056)

Tabela 5.2 Índice de Rand Ajustado para algoritmos do tipo nuvens dinâmicas baseados em distâncias adaptativas e não adaptativas para as tabelas de dados poligonais 1 e 2.

CAPÍTULO 6

Conclusões e Trabalhos Futuros

Esse trabalho propôs um algoritmo do tipo nuvens dinâmicas com distância adaptativa para dados poligonais foi proposto. O método busca a otimização local de um critério de adequação medido entre as classes e seus representantes. A estratégia de distância adaptativa foi apresentada como uma versão ponderada da distância usada em [18]. Para avaliar a performance do método de agrupamento proposto, experimentos foram realizados baseados em dados simulados. Tamanhos de grupos diferentes foram considerados para reforçar a aplicabilidade de distâncias adaptativas. A performance do método foi medida por um índice externo, baseando-se no método Monte Carlo. Comparou-se o método proposto com o algoritmo de nuvens dinâmicas para dados poligonais com distâncias fixas.

A partir dos resultados, vimos que há um significativo ganho de performance quando aplicamos a estratégia adaptativa na distância usada no método abordado neste trabalho. O problema presente nos algoritmos de nuvens dinâmicas com distâncias fixas sobre como lidar com *clusters* de diferentes tamanhos e formatos é enfrentado e mostra que a estratégia proposta por esse trabalho têm resultados melhores nessas situações.

Podemos afirmar que nosso trabalho contribui de alguma forma para a construção dos métodos de inteligência artificial para dados simbólicos, somando mais forças para tornar essa estratégia ainda mais possível para enfrentar o problema das altas dimensionalidades das bases de dados.

Como trabalhos futuros, podemos listar as seguintes ideias:

1. Experimentos em dados reais para reforçar a sua utilidade no mundo real
2. Comparar com métodos de dados intervalares com a mesma estratégia adaptativa para as distâncias
3. Aplicar configuração inicial inteligente
4. Método *Fuzzy* para dados poligonais
5. Disponibilizar algoritmo em *Python* para possibilitar agilidade na pesquisa de outros pesquisadores

Referências Bibliográficas

[1] Francisco de A.T. de Carvalho, Renata M.C.R. de Souza, Marie Chavent, and Yves Lechevallier. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, 27(3):167–179, 2006.

[2] F A T De Carvalho, P Brito, and H H Bock. Dynamic clustering for interval data based on l2 distance. *Computational Statistics*, 21(2):231–250, 2006.

[3] F A T De Carvalho, R M C R Souza, M Chavent, and Y Lechevallier. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, 27(3):167–179, 2006.

[4] E. Diday and J. C. Simon. *Clustering Analysis*, pages 47–94. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

[5] Edwin Diday and Lynne Billard. Symbolic data analysis: Conceptual statistics and data mining. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, 12 2006.

[6] Edwin Diday and Gerard Govaert. Classification avec distances adaptives. *Comptes Rendues Acad. Sci. Paris, série A*, 278:993, 1974.

[7] Edwin Diday and Gerard Govaert. Classification automatique avec distances adaptatives. *R.A.I.R.O. Inform. Comput. Sci*, 11(4):329–349, 1977.

[8] Allan David Gordon. *Classification*. CRC Press, 1999.

[9] K Chidananda Gowda and G Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112, 1978.

[10] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[11] W. H. Inmon. *Building the Data Warehouse,3rd Edition*. John Wiley Sons, Inc., USA, 3rd edition, 2002.

[12] Teresa Bernarda Ludermir. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, 35(Estud. av., 2021 35(101)):85–94, Jan 2021.

[13] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.

[14] Tom Michael Mitchell et al. *Machine learning*, volume 1. McGraw-hill New York, 2007.

[15] Wagner J.F. Silva, Renata M.C.R. Souza, and Francisco José A. Cysneiros. Polygonal data analysis: A new framework in symbolic data analysis. *Knowledge-Based Systems*, 163:26–35, 2019.

[16] Renata M C R Souza and F A T De Carvalho. Clustering of interval data based on city–block distances. *Pattern Recognition Letters*, 25(3):353 – 365, 2004.

[17] Renata M.C.R. Souza. *Métodos de cluster para intervalos usando algoritmos do tipo nuvens dinâmicas*. Tese (doutorado), Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, 2003.

[18] Renata M.C.R. Souza Wagner J. F. Silva, Pedro J.C. Souza and Francisco José A. Cysneiros. A clustering algorithm for polygonal data applied to scientific journal profiles. Em fase de revisão, 2023.