UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Fernanda Teixeira Dos Santos

Espaços de dissimilaridade simbólicos usando distâncias intervalares para classificação e regressão

Recife

2022

Fernanda Teixeira Dos Santos

Espaços de dissimilaridade simbólicos usando distâncias intervalares para classificação e regressão

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Renata Maria Cardoso Rodrigues de Souza

Coorientador (a): Telmo de M. Silva Filho

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S237e Santos, Fernanda Teixeira dos
Espaço de dissimilaridade simbólicos usando distâncias intervalares para classificação e regressão / Fernanda Teixeira dos Santos. – 2022.
51 f.:il., fig, tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.

Inclui referências.

1. Inteligência computacional. 2. Regressão. 3. Distâncias intervalares. I. Souza, Renata Maria Cardoso Rodrigues de (orientadora). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2022-198

Fernanda Teixeira dos Santos

“Espaço de dissimilaridade simbólicos usando distâncias intervalares para classificação e regressão”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 12 de agosto de 2022.

BANCA EXAMINADORA

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática / UFPE

_____

Prof. Dr. Marcus Costa de Araujo
Departamento de Engenharia Mecânica / UFPE

_____

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE
(**Orientadora**)

À minha avó Josefa Teixeira dos Santos (in memoriam) por todo imenso amor, carinho, dedicação e amizade. Principalmente por acreditar em mim por toda a sua vida. Por ser meu espelho de mulher, minha pilastra no mundo e a quem eu sempre serei eternamente grata.

AGRADECIMENTOS

Primeiramente a Deus por todo amor, zelo e cuidado em minha vida. Ao Senhor agradeço por sempre estar ao meu lado cultivando a minha fé. À Universidade Federal de Pernambuco por me conceder a honra de fazer parte de sua grife seleta de estudantes, por todas as lutas e bênçãos alcançadas neste lugar.

À professora Dra. Renata Maria Cardoso Rodrigues de Souza, que aceitou fazer parte deste trabalho, sendo minha orientadora, acreditando em mim, passando-me confiança e me ajudando a ultrapassar todos os percalços encontrados. E, também, ao meu co-orientador Telmo de M. Silva Filho por todo suporte e ajuda no desenvolvimento deste trabalho.

Agradeço também à banca pela disponibilidade e apreciação deste trabalho. À minha família por todo apoio incondicional, amor e carinho com o qual estiveram ao meu lado durante estes 30 anos, até aqui, em especial a Francisco Inácio dos Santos, meu avô, por ser o homem que me ensina sobre a vida, e os melhores sentimentos que se possa ter no mundo. Aos meus amigos por todo apoio e momentos de felicidade proporcionados, com vocês sei que sou forte, obrigada por sempre estarem ao meu lado quando precisei.

À Elizabeth Regina, por acreditar sempre em mim e ser calmaria e paz, sempre me apoiando e estando ao meu lado em todos os momentos. Por fim, não existem palavras que mensuram o tamanho da minha gratidão a Deus por ter cruzado meu caminho com cada um de vocês. Espero que nestas poucas, porém sinceras palavras vocês percebam o quanto são importantes na minha vida. Aqui fica a minha imensa e eterna GRATIDÃO!

RESUMO

Apresenta como objetivo a aplicação de algoritmos clássicos de aprendizagem de máquina na transformação de dados simbólicos em dados clássicos, na busca de alcançar resultados tão bons quanto utilizando dados clássicos. Tendo em vista que, o estudo de dados simbólicos originou a área de análise de dados simbólicos, que está diretamente ligada a uma abordagem na área de descoberta automática de máquinas para os dados clássicos. Para tanto, utilizar-se-á da matriz de dissimilaridade já que vem sendo amplamente explorada na literatura de aprendizagem de máquina, por obter bons resultados em variadas tarefas. Esta pesquisa apresenta-se sob a forma de dissertação e está dividida em seções, sendo a primeira apresentando os aspectos da introdução; a segunda a revisão de literatura; a terceira denota acerca dos procedimentos metodológicos, a quarta apresenta os resultados que foram compreendidos, a quinta e última demonstra as conclusões e considerações finais, além de prospectar projetos futuros, a partir desta pesquisa. Com base nos resultados, foi possível identificar a construção de espaços de dissimilaridade usando distâncias do tipo intervalar, transformando os dados em um novo ambiente permitindo assim, o treinamento de modelos de aprendizagem de máquina clássicos.

Palavras-chaves: dados simbólicos; distâncias intervalares; classificação; regressão.

ABSTRACT

This research aims to apply classical machine learning algorithms in the transformation of symbolic data into classical data, in the search to achieve results as good as using classical data. Considering that, the study of symbolic data originated the area of analysis of symbolic data, which is directly linked to an approach in the area of automatic machine discovery for classical data. To do so, the dissimilarity matrix will be used, as it has been widely explored in the machine learning literature, as it obtains good results in various tasks. This research is presented in the form of a dissertation and is divided into sections, the first presenting aspects of the introduction; the second, the literature review; the third denotes about the methodological procedures, the fourth presents the results that were understood, the fifth and last demonstrates the conclusions and final considerations, in addition to prospecting future projects, from this research. Based on the results, it was possible to identify the construction of dissimilarity spaces using interval-type distances, transforming the data into a new environment, thus allowing the training of classical machine learning models.

Keywords: symbolic data; interval distances; classification; regression.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

γ	Letra grega Gama
$\in$	Pertence
δ	Delta
θ	Teta
σ	Sigma
μ	Mi

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O mundo da tecnologia mudou de forma considerável ao longo da história humana e o avanço da ciência figura como um dos fatores que mais contribuíram para essa mudança. O surgimento da internet transformou o paradigma informacional vigente trazendo à tona a sociedade da informação moldada pela ciência de dados, como resultado direto da modernização. Conforme Sampaio Neto (2021) atualmente, muitos serviços estão sendo executados na internet, como e-commerce, plataformas de entretenimento, internet banking, e-learning e redes sociais.

Assim, existe uma infinidade de dados que passam por esses serviços e levam a uma sobrecarga de informações. A era do Big Data viabilizou os processos que nos levam a assimilar novas realidades e conceitos, e representá-los com algumas teorias que irão trabalhar com a ideia de que o conhecimento está conosco a priori e são frutos de ideias pré-moldadas, outras se destacam por apontar que o conhecimento é produzido a partir da nossa interação com o meio em que estamos inseridos, mas uma ideia comum entre essas teorias é a de que o mundo é representado por modelos e essa modelização funciona para que possamos entender quantitativa e qualitativamente os seus fenômenos complexos.

Neste ínterim, pelas palavras de Reyes (2022) ultimamente, existe uma enorme capacidade de armazenamento de dados, em decorrência da modernidade, que disparou e fez com que as bases de dados tenham se tornado excessivamente grandes e complexas. Assim, os conjuntos de dados, que são extraídos, podem facilmente ter centenas de variáveis e milhares de registros, o que os torna difíceis de processar com as técnicas tradicionais de análise de dados. Para Souza (2016) a representação desta infinidade de dados pode ser contornada de forma condensada e compactada.

Onde a condensação de várias informações em um mesmo dado pode ser feita com a utilização de dados simbólicos. E continua dizendo que, "assim, a análise de dados simbólicos (ADS) propõe estratégias para a representação compacta de dados, com o uso de elementos complexos que guardam as informações de várias instâncias. Estes elementos são exemplificados por intervalos, histogramas e distribuições de probabilidade". Portanto, pode-se perceber que a representação viabilizada pelos dados simbólicos se tornou uma alternativa para a manipulação destas bases de dados, que otimiza o tempo necessário, encontra soluções para

o problema de organização de grandes volumes de dados e ainda viabiliza a categorização, quando necessário.

Segundo Reyes (2022) a ADS tem suas raízes na Análise Exploratória de Dados (BEATON; TUKEY, 1974; BOCK, 1974; SAPORTA, 1990), Inteligência Artificial (BEATON; TUKEY, 1974; RUSSELL et al., 2003; LUGER, 2005) e Taxonomia Numérica (SNEATH; SOKAL, 1962).

De acordo com Souza (2016) a ADS direciona seus estudos para métodos de obtenção de conhecimento em dados simbólicos. E, desta maneira, apresenta duas grandes áreas que se relacionam com a obtenção de conhecimento sendo estas: a análise de agrupamento; e a regressão linear. Desta forma, a análise de agrupamento envolve técnicas para a separação de um conjunto de dados em grupos cujos elementos apresentam similaridade entre si (SOUZA, 2016). A análise de regressão envolve a modelagem da dependência linear do valor esperado de uma variável resposta em relação a outras variáveis (chamadas regressoras). A partir do modelo construído, é possível encontrar estimativas para a resposta utilizando valores diversos e não-observados para a variável regressora. Os métodos de agrupamento e regressão para dados simbólicos são muito variados por causa da complexidade que esses dados apresentam (BILLARD; DIDAY, 2006).

Neste estudo há o desenvolvimento da metodologia pautada nas seguintes etapas:

a) Busca em bases de dados acerca da literatura de dados simbólicos;

b) Elaboração do referencial teórico sobre dados simbólicos, classificadores e regressores;

c) aplicação dos algoritmos de classificação e regressão.

Espera-se com esta pesquisa aplicar algoritmos clássicos de aprendizagem de máquina na transformação de dados simbólicos em dados clássicos, na busca de alcançar resultados tão bons quanto utilizando dados clássicos.

1.2 OBJETIVO

Esta pesquisa apresenta como objetivo a aplicação de algoritmos clássicos de aprendizagem de máquina na transformação de dados simbólicos em dados clássicos, na busca de alcançar resultados tão bons quanto utilizando dados clássicos. Tendo em vista que, o estudo de dados simbólicos originou a área de análise de dados simbólicos, que está diretamente ligada a uma abordagem na área de aprendizagem de máquinas. Para tanto,utilizar-se-á da matriz de dissimilaridade já que vem sendo amplamente explorada na literatura de aprendizagem de

máquina,por obter bons resultados em variadas tarefas.

1.3 ESTRUTURA DO DOCUMENTO

Esta pesquisa apresenta-se sob a forma de dissertação e está dividida em seções, sendo a primeira apresentando os aspectos da introdução; a segunda a revisão de literatura; a terceira denota acerca dos procedimentos metodológicos, a quarta apresenta os resultados que foram obtidos, a quinta e última demonstra as conclusões e considerações finais, além de prospectar projetos futuros, a partir desta pesquisa.

2 REVISÃO DE LITERATURA

2.1 DADOS SIMBÓLICOS

Na área de Ciência de Dados, a análise de dados simbólicos (Symbolic Data Analysis) é um novo campo na descoberta de conhecimento e gerenciamento de dados, onde muitos estudos vêm sendo desenvolvidos com temas centrais de análise multivariada, reconhecimento de padrões e inteligência artificial. Nesta perspectiva a análise de dados simbólicos oferece uma estrutura de dados chamada de dados simbólicos (DIDAY, 2016), de forma que a informação fundida considera variabilidade entre os membros de cada classe, por meio de intervalos, distribuições e conjunto de categorias que podem ser obtidas, a partir de um processo de fusão (SILVA, 2017).

No universo de grandes e desordenados volumes de dados, os bancos de dados são amplamente utilizados nas empresas, no mercado de trabalho do mundo corporativo, nas administrações públicas e privadas, onde seu crescimento exponencial faz com que se tenha uma conjuntura complexa para o gerenciamento de seus sistemas (SILVA, 2017). De tal maneira, ao longo dos anos, várias abordagens foram surgindo para identificar, selecionar, visualizar e estruturar ou até mesmo extrair informações desses grandes volumes de dados.

A ADS teve a sua origem pela influência de três principais áreas e seus principais trabalhos foram surgindo na década de 80 (BILLARD; DIDAY, 2006):

1. análise exploratória de dados;
2. inteligência artificial;
3. taxonomia numérica.

Dentro deste contexto de grandes conglomerados de dados a análise de dados simbólicos é utilizada em diferentes grupos de dados, sendo eles grandes ou pequenos. Desta maneira, a ADS consiste, em construir de forma automática, grupos homogêneos de observações, a partir de grandes conjuntos de dados, definindo assim novas unidades, os dados simbólicos, que descrevem estes grupos. O que resulta em tabelas de dados, denominadas de tabelas de dados simbólicos, de estrutura mais ampla e complexa, em vista que, em cada uma destas células não há notoriamente apenas um único valor quantitativo ou um único valor qualitativo, porém também pode conter informações complexas, como subconjuntos, intervalos, funções de

diferentes semânticas (probabilista, possibilista, credibilista) ligadas por taxonomias ou dependências. Pode-se considerar, portanto, que o principal objetivo da análise de dados simbólicos é generalizar a mineração de dados e estatísticas para unidades de alto nível descritas por dados simbólicos (ASSIS, 2011).

Pelas palavras de Silva Filho (2013) tem-se um cenário no qual seja necessário descrever peixes de uma determinada espécie. Assim, afirma-se que "o comprimento é de 30 a 60 cm, mas não passa de 40 cm nas fêmeas". Como não se pode demonstrar este tipo de informação em tabelas de dados clássicos, por serem constituídas de células com simples valores de pontos. Também não é fácil representar regras em tabelas clássicas. Logo estes peixes são descritos corretamente da seguinte forma: [Comprimento = [30, 60]], [Sexo = macho, fêmea] e [Se Sexo = fêmea Então Comprimento <= 40] o objetivo da ADS é obter ferramentas que possam lidar com esses tipos complexos de dados.

Pelo exemplo acima é possível perceber que a análise de dados é ampla e pode ser utilizada em qualquer tipo de cenário, em qualquer tempo. Nesta conjuntura, faz-se necessário que as variáveis de um objeto assumam informações sumariamente complexas, no que concernem histogramas, distribuição de probabilidade, intervalos e conjuntos, todavia para dados mais complexos são utilizados os dados simbólicos, na tentativa de conseguir a extração de mais informações. Por meio da literatura na ciência de dados, tem-se que esta tipologia de dados é especificamente a mais indicada para representar as classes de indivíduos (objetos agregados) (SILVA, 2006).

De maneira simples e objetiva o campo da análise de dados simbólicos e as estatísticas envolvidas na manipulação de dados foram introduzidas pelos autores Bock e Diday (2000). Tais ferramentas atuam na aplicação de estatística descritiva que representa um conjunto de dados, tais como: média, variância, correlação, distribuição de probabilidades e histogramas (SOUZA, 2016).

A análise de dados simbólicos busca métodos para a obtenção de conhecimento em dados simbólicos. Neste sentido, duas grandes áreas que estão relacionadas com a obtenção de conhecimento são a análise de agrupamento e a regressão linear (SOUZA, 2016). A análise de agrupamento envolve técnicas para a separação de um conjunto de dados em grupos cujos elementos apresentam similaridade entre si, ou seja, separa os elementos de um conjunto de acordo com as suas similaridades, ou qualidades específicas que cada objeto possui em comum.

Já a análise de regressão envolve a modelagem da dependência linear do valor esperado de uma variável resposta em relação a outras variáveis (chamadas regressoras). A partir do

Tabela 1 – Tabela de dados simbólicos

ID	Peso	Modelo de Celular	Avaliação de uso
C1	[65,3:72,1]	Nokia	Regular
C2	[55,3:62,4]	Motorola	Bom
C3	[46,3:49,2]	Iphone	Ótimo
C4	[57,7:60,5]	Samsung	Excelente
C5	[66,4:74,5]	Xiaomi	Ótimo

Fonte: Elaborado pela Autora (2022)

modelo construído, é possível encontrar estimativas para a resposta utilizando valores diversos e não-observados para a variável regressora. Os métodos de agrupamento e regressão para dados simbólicos são muito variados por causa da complexidade que esses dados apresentam (BILLARD; DIDAY, 2006).

No campo da ciência de dados, os dados simbólicos são obtidos de diferentes maneiras, sendo estas:

I. a aplicação de um algoritmo de classificação não supervisionada para simplificar grandes grupos de conjuntos de dados, identificar e representar de forma clara e autoexplicativa as classes associadas aos grupos obtidos;

II. através da descrição pelos especialistas dos conceitos;

III. por meio de bases de dados relacionais para estudar os conjuntos de unidades cuja descrição envolve variadas relações (ASSIS, 2011).

Por conseguinte, os dados simbólicos podem descrever os indivíduos, tendo como base ou não as características complexas que o condicionam nos grupos de indivíduos. Assim, ainda de acordo com Assis (2011) os objetos de um conjunto de dados simbólicos podem ou não ser definidos pelas mesmas variáveis; cada variável pode ter mais que um valor ou mesmo um intervalo de valores; podem incluir um ou mais objetos elementares; a descrição pode vir a depender das relações existentes com os outros dados. Assim, em uma tabela de dados simbólicos, as linhas correspondem aos indivíduos ou classes de indivíduos, onde as variáveis simbólicas são representadas pelas colunas, que descrevem as características dos indivíduos. À fim de ilustrar, segue a tabela 1 que denota uma tabela de dados simbólicos:

Neste sentido, os dados simbólicos podem surgir de forma natural, de situações do cotidiano como o intervalo de variação de temperatura de uma cidade, ou uma lista de categorias da

Tabela 2 – Variáveis e dados simbólicos

ID	Modelo de celular	Cidades
C1	Nokia	{Recife,Rio de Janeiro,Manaus,Fortaleza,Vitória,São Paulo,Belém}
C2	Motorola	{Fortaleza,Vitória,São Paulo,Belém,Porto Alegre}
C3	Iphone	{Recife,Vitória,São Paulo,Belém}
C4	Samsung	{Manaus,Fortaleza,Vitória}
C5	Xiaomi	{Porto Alegre}

Fonte: Elaborado pela Autora (2022)

Tabela 3 – Representação simbólica

ID	Peso	Modelo de Celular	Avaliação de uso
C1	[65,3:72,1]	Nokia	Regular
⋮	⋮	⋮	⋮
C5	[66,4:74,5]	Xiaomi	Ótimo

Fonte: Elaborado pela Autora (2022)

população, e frequências respectivas, ou ainda a lista de avaliações de usuários de celulares smartphones, como bem mostra a tabela 1 acima. Logo, também podem ser o resultado de agregação da base de dados que contém estes dados. À exemplo disso, a agregação de uma base de dados que se tratava inicialmente dos moradores de um país pode gerar uma nova base, cujas instâncias são as cidades do país. Essa agregação também pode ser efetuada porque o conjunto original era tão grande que tornava difícil sua análise. (SILVA FILHO, 2013).

Como é possível perceber a tabela 1, visualizada anteriormente, viabilizou ser transformada em uma tabela simbólica, isto é, quando os dados clássicos são modificados para dados de nível mais elevado, e porventura acontece a representação em uma tabela simbólica.

2.2 CLASSIFICADORES INTERVALARES

Neste sentido, de acordo com estudos matemáticos de Grigoletti, et al (2006) um sistema linear intervalar do tipo $Ax = b$ possui a matriz de coeficientes A e os vetores x e b como dados intervalares. Como na aritmética pontual, existem também vários métodos para a resolução de problemas intervalares. Os estudos sobre versões intervalares para métodos indiretos ou iterativos foram iniciados por E. Hansen e posteriormente continuados por um grande número de pesquisadores, inclusive em abordagens de programação paralela.

Por conseguinte, Souza (2016) denota que os dados intervalares, em sua complexidade,

apresentam dificuldades em sua construção, o autor afirma que os intervalos podem ser vistos como uma agregação de pontos, as propostas que existem na literatura sugerem a escolha de alguns deles para a construção dos modelos de regressão.

A partir das pesquisa de Souza Filho (2013) existem atualmente diferentes tipos de algoritmos clássicos que são usados para trabalhar com os dados simbólicos, sendo estes: redes neurais, classificadores orientados a regiões, classificadores que tem como base as abordagens de regressão e k-vizinhos mais próximos.

O autor disserta que, no que concerne a redes neurais, as variações foram propostas levando em consideração as entradas, saídas ou pesos intervalares, e desta maneira defende, que historicamente inúmeras propostas com esta metodologia foram lançadas nos estudos de análises de dados intervales, tais como: Beheshti que propôs que tanto os valores de entrada quanto os pesos, biases e saídas são intervalares e para além disso, mostra como obter pesos e biases ótimos para um dado conjunto de treinamento; Rossi e Conan-guez e Roque que para lidar com entradas e saídas intervalares, introduziram abordagens de MLP utilizando pesos e bias clássicos, apresentando como vantagem direta a simplicidade de sua arquitetura, se for comparada com redes neurais que utilizam pesos e bias intervalares (SOUZA FILHO, 2013).

2.3 REGRESSORES INTERVALARES

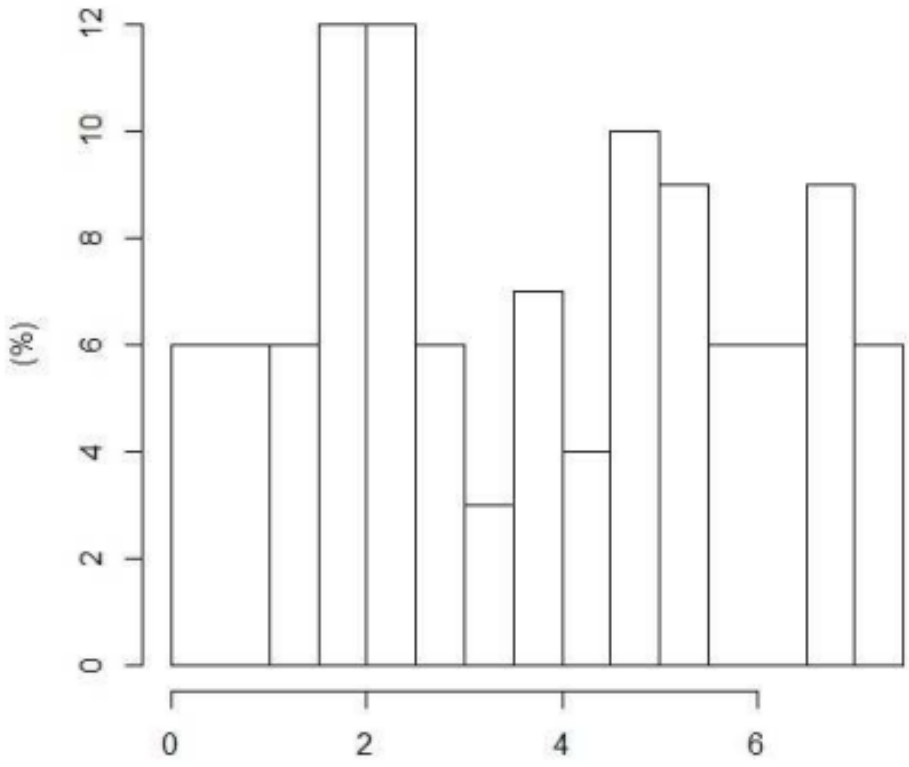
Souza (2016) defende que os dados intervalares multivariados podem ser interpretados na condição de hipercubos em um espaço multidimensional, o que acontece é que os hipercubos apresentam uma região interna que denota uma variação que os intervalos representam (relacionada com uma incerteza).

Segundo as palavras de Hoffman (2006) há diversas maneiras de uso de equações de regressão, em situações envolvendo duas variáveis. À exemplo disso, um pesquisador em EDM pode tentar explicar as variações do desempenho de alunos em função do aumento da carga horária diária de estudos. Desta maneira, a lógica de uma relação causal deve advir de fenômenos externos ao âmbito da estatística. A análise estatística de regressão apenas modela qual relacionamento matemático pode existir, se existir algum (RODRIGUES, 2013).

Pelas preposições de Malaquias (2017) atualmente são consideradas as variáveis intervalares, no contexto da Análise Simbólica de Dados, os modelos de regressão linear propostos, que se consideram mais relevantes sendo estes: o Método do Centro (Billard e Diday 2000); o Método do Mínimo e Máximo (Billard e Diday 2002); o Método do Centro e Raio (LIMA

NETO; DE CARVALHO, 2008); o Método do Centro e Raio com Restrição (LIMA NETO; DE CARVALHO, 2010); o Lasso - Constrained Regression Model (Giordani 2014); e o Interval Distribution (ID) Regression Model (DIAS; BRITO, 2017). Desta maneira, entende-se que com qualquer um destes métodos e possível, em uma análise de dados simbólicos estimar uma variável intervalar resposta, por meio de variáveis intervalares explicativas. Dias e Brito (2017) também consideram que ultimamente foram propostos inúmeros métodos de regressão linear para variáveis intervalares.

Figura 1 – Histograma para dados intervalares



Fonte: Fagundes (2013)

Assim, a figura 1 acima conforme Fagundes (2013) exemplifica a representação do histograma intervalar de classes definidas para cada elemento do vetor. Com isso, as oito classes definidas contém os intervalos obtidos dentro de cada faixa definida. Portanto, quanto mais intervalos existirem, maiores serão as frequências das classes.

3 ESTADO DA ARTE

3.1 CLASSIFICADORES INTERVALARES

Pela percepção de Sampaio Neto (2021) no mundo contemporâneo as técnicas baseadas em modelos abordam o problema como um problema de regressão ou classificação, dependendo do sistema de classificação, por exemplo, o modelo deve estimar pontuações ou classificações para produtos, dado o usuário. Para esta categoria, métodos de agrupamento, redes Bayesianas, redes neurais artificiais ou regressões lineares são os métodos mais utilizados.

Assim, Reyes (2022) defende que, em grande parte dos métodos que existem na literatura para a escolha de protótipos foi elaborada para a resolução de problemas de classificação, afirmando que enquanto apenas alguns artigos trazem o enfoque para o problema de seleção de protótipos para a regressão. Desta forma entende-se que a seleção de protótipos para regressão é muito mais complexa. A razão é que nos problemas de classificação apenas os limites entre as classes devem ser determinados com precisão, enquanto nos problemas de regressão o valor de saída deve ser calculado adequadamente em cada ponto do espaço de entrada (REYES, 2022).

Neste viés, a decisão em problemas de classificação é frequentemente bidirecional ou há no máximo várias classes diferentes, enquanto que em problemas de regressão, a saída do sistema é contínua, portanto há um número ilimitado de valores possíveis previstos pelo sistema (REYES, 2022). Isso faz com que a compactação do conjunto de dados obtida pela seleção de protótipos possa ser muito maior em problemas de classificação do que em problemas de regressão. Além disso, a decisão sobre a rejeição de um determinado vetor nos problemas de classificação pode ser tomada com base na classificação correta ou incorreta do vetor por algum algoritmo (REYES, 2022).

De tal maneira, Silva Filho, Sousa e Prudêncio (2016) denotam que os classificadores que são baseados em protótipos, logo aprendem posicionando iterativamente os protótipos, já que eles estão mais próximos dos dados de suas respectivas classes. Normalmente, diz-se que um teste em postura pertence a uma classe Cg quando seu protótipo mais próximo, de acordo com uma distância predeterminada, pertence a Cg.

Para tanto, os classificadores que são baseados em protótipos podem ser compreendidos otimizando um critério, como a soma das distâncias entre protótipos e suas instâncias afetadas. Isso permite o uso de algoritmos de otimização baseados em conjuntos para treinar protótipos

(SILVA FILHO; SOUSA; PRUDÊNCIO, 2016). Portanto, algumas pesquisas foram dedicadas à criação de protótipos treinados por enxames de classificadores, com o objetivo de se beneficiar da capacidade de escapar de mínimos locais demonstrados por métodos baseados em enxames, para mitigar o problema de inicialização de protótipo ruim. Infelizmente, todos esses métodos lidam com dados pontuais e uma abordagem semelhante para dados de intervalo ainda é um assunto em aberto (SILVA FILHO; SOUSA; PRUDÊNCIO, 2016). Neste sentido, os autores descrevem um novo algoritmo para dados intervalares que tem a capacidade de selecionar melhor as variáveis em um conjunto de dados de forma significativa. Tal método é chamado de classificador intervalar de abelhas artificiais baseado em velocidade (IVABC).

3.1.1 Método IVABC

O método IVABC utiliza otimização de enxames para encontrar os melhores protótipos e para fazer seleção de variáveis, assumindo que pode existir diversas soluções para o problema e vai tentar encontrar a melhor solução possível.

Para resolver o problema de classificação de dados intervalares utiliza diversas heurísticas para tentar encontrar a melhor solução possível, por meio de otimização baseada em enxames, mais especificamente, colônia de abelhas, ele faz seleção de variáveis, seleção de protótipos e encontra o melhor conjunto de protótipos e por isso demora muito tempo para processar.

De acordo com Silva Filho (2017) O IVABC procede da seguinte maneira: novas partículas são selecionadas aleatoriamente e suas velocidades e valores de qualidade são inicializados (passos 1 a 3), a seguir, o método itera pelas fases de abelhas empregadas (passos 5 a 8), observadoras (passos 9 a 13) e exploradoras (passos 14 a 16), até um critério de convergência ser atingido.

O Algoritmo 1 abaixo apresenta os passos detalhados do IVABC.

Algoritmo 1: IVABC

Inicialize $\mathcal{SN}$ fontes de alimento e suas velocidades;
Atribua valores para os pesos N_1 e N_2 do critério e para o limite de tentativas;
Inicialize w_{max} e w_{min}
Calcule os pesos da distância para todas as fontes de alimento;
Calcule o valor do critério para cada fonte de alimento, usando Equação (a);
Inicialize $pbest_l$ para cada fonte de alimento ($l = 1, ..., SN$) e $gbest$ para a população;
Enquanto o critério de parada não for atingido Faça:
Envie as abelhas empregadas:
    PARA $l = 1 \rightarrow SN$ **FAÇA:**
        Selecione uma fonte de alimentos aleatória $\mathbf{W}_k$, $k \neq l$;
        Selecione um índice aleatório $0 \leq m \leq M$;
        Selecione uma variável aleatória j;
        Use a Equação (b) ou a Equação (c) para gerar uma nova fonte de alimento $\mathbf{W}'l$;
        Calcule os pesos da distância e o novo valor do critério $\psi'_{IVABC}(l)$;
        Se $\psi'_{IVABC}(l) < \psi_{IVABC}(l)$ **ENTÃO**
            $\mathbf{W}_l \leftarrow \mathbf{W}'_l$
            $pbest_l \leftarrow \mathbf{W}'_l$
            $tentativas_l \leftarrow 0$
        SENÃO
            $tentativas_l \leftarrow tentativas_l + 1$
        FIM SE
    FIM PARA
    Atualize $gbest = \mathbf{W}_c$, onde $c = argmin\ \psi_{IVABC}(l), \forall l \in (1, ..., SN)$.
Envie as abelhas observadoras:
    Calcule as probabilidades das fontes de alimento, usando a Equação (d);
    Calcule ω, usando a fórmula empregada por Imanian et al. (2014);
    PARA $l = 1 \rightarrow SN$ **FAÇA:**
        Selecione aleatoriamente uma fonte de alimento $\mathbf{W}_k$, $k \neq l$,baseado nas probabilidades;
        Selecione aleatoriamente r_1 e r_2 do intervalo $[0, 1]$ e c_1 e c_2 do intervalo $[0.5, 1]$;
        Atualize a velocidade $\mathbf{V}_k$, usando a Equação (e);
        Calcule a nova posição $\mathbf{W}'_k$, usando a Equação (f);
        Calcule os pesos da distância e o novo valor do critério $\psi_{IVABC}(k)$;
        SE $\psi'_{IVABC}(k) < \psi_{IVABC}(k)$ **ENTÃO**
            $\mathbf{W}_k \leftarrow \mathbf{W}'_k$
            $pbest_l \leftarrow \mathbf{W}'_k$
            $tentativas_l \leftarrow 0$
        SENÃO
            $tentativas_l \leftarrow tentativas_l + 1$
        FIM SE
    FIM PARA
Envie as abelhas exploradoras:
    PARA $l = 1 \rightarrow SN$ **FAÇA:**
        SE $tentativas_l > limit$ **ENTÃO**
            Reinicialize $tentativas_l$, $\mathbf{W}_l$ e sua velocidade;
            Calcule os pesos da distância e o novo valor do critério $\psi_{IVABC}(l)$;
        FIM SE
    FIM PARA
    $gbest = \mathbf{W}_c$, onde $c = argmin\ \psi_{IVABC}(l), \forall l \in (1, ..., SN)$.
    FIM ENQUANTO
    RETORNE $gbest$ e seus pesos λ
Fonte: (SILVA FILHO, 2017)

Equação (a) usada na etapa 5

$$\psi_{IVABC}(l) = n_1\psi_1(l) + n_2 j_I VABC(l)$$

onde n_1 e n_2 são pesos escolhidos do intervalo [0;1], tal que $n_1+n_2 = 1$ e $\psi(l) = N_{erros,l}/N$ é a taxa de erro de classificação da $l-ésima$ partícula. Os valores de $\psi_1(l)$ situam-se no intervalo [0;1]. A função $j_{IVABC}(l)$ é calculada pela equação abaixo:

$$j_{IVABC}(l) = \frac{\sum_{m=1}^{M}\sum_{i=1}^{N}\frac{d\pi_{m,l}(\vec{x}_i, \vec{w}_{m,l,\vec{\lambda}_{m,l}})\delta(m,l,k,i)}{k_i}}{N_{acertos,l}}$$

Onde $\delta(m,l,k,i) = 1$ se o $m-ésimo$ protótipo da $l-ésima$ partícula for um dos k protótipos mais próximos da $i-ésima$ instância e $y_i = y_{m,l}$, caso contrário $\delta(m,l,k,i) = 0$. $d\pi_{m,l}$ é uma distância.(SILVA FILHO, 2017)

Equação (b) usada na etapa 13

$$w'_{0,l,j} = w_{0,l,j} + \phi * (w_{0,l,j} - w_{0,k,j})$$

Equação (c) usada na etapa 13

$$w'_{m,l,j,L} = w_{m,l,j,L} + \phi * (w_{m,l,j,L} - w_{m,k,j,L})$$
$$w'_{m,l,j,U} = w_{m,l,j,U} + \phi * (w_{m,l,j,U} - w_{m,k,j,U})$$

Equação (d) usada na etapa 25

$$\text{setlength3cm}pl = \frac{f(l)}{\sum_{k=1}^{S} Nf(k)}$$

Equação (e) usada na etapa 30

$$\vec{V}_k(t+1) = \omega\vec{V}_k(t) + c_1\vec{r}_1 \cdot (pbest_k(t) - \vec{w}_k(t)) + c_2\vec{r}_2 \cdot (gbest(t) - \vec{w}_k(t))$$

Equação (f) usada na etapa 31

$$\vec{w}_k(t+1) = \vec{w}_k(t) + \vec{v}_k(t)$$

Onde $\vec{w}_k(t)$ e $\vec{v}_k(t)$ são, respectivamente, a posição e a velocidade da $k-ésima$ partícula ou fonte de alimento no instante t, ω é a inércia, c_1 e c_2 são coeficientes de aceleração, $pbest_k(t)$ é a melhor posição obtida pela $k-ésima$ partícula até o instante t, $gbest(t)$ é a melhor posição obtida pelo enxame até o instante t r $\vec{r}_1$ e $\vec{r}_2$ são vetores, cujos valores são escolhidos aleatoriamente no intervalo [0; 1].(SILVA FILHO, 2017)

3.2 REGRESSORES INTERVALARES

Na preposição de Souza (2016) existe uma certa complexidade em razão da natureza dos dados intervalares, e desta maneira, a tarefa de construir um modelo que relacione variáveis intervalares não é algo fácil e simples. Os intervalos são vistos como uma agregação de pontos e desta forma, alguns autores na literatura denotam de formas diferentes, sem que haja um consenso quanto ao conjunto de pontos utilizados na representação dos intervalos. Para tanto, Souza (2016) defende que um método de regressão linear para intervalos deve oferecer, de forma simultânea, duas estimativas:

1. os limites inferiores e superiores dos intervalos.
2. os limites devem manter a coerência matemática dos intervalos, em que o limite superior é maior ou igual ao inferior.

De acordo com Reyes (2017) dentre as técnicas mais utilizadas para a construção de modelos na descrição de comportamento de uma variável dependente, a regressão linear segue sendo uma das mais utilizadas, assim sua metodologia vem sendo objeto de estudo para as mais variadas áreas do conhecimento, tais como: a financeira, a epidemiológica, a médica, a econômica, a geográfica, dentre outras. Na literatura, conforme os estudos de Souza (2016), atualmente usa-se a seguinte notação para as variáveis que trabalham com a regressão:
Y é a variável resposta (dependente) intervalar com n observações. Neste caso, $Y = (y_1, y_2, ..., y_n)^t$, em que $y_i = [\underline{Y}_i, \bar{y}_1](i = 1, ..., n)$. São consideradas p variáveis regressoras (independentes ou explicativas) intervalares $X_1, X_2, ..., X_p$. Cada variável regressora possui n observações intervalares. Desta forma, $X_j = (x_{j1}, x_{j2}, ..., x_{jn})^t(j = 1, ..., p)$, em que $x_{ji} = [\underline{x}_{ji}, \bar{x}_{ji}]$. São definidos, também, um vetor $p - dimensional$ de intervalos x_ϕ, com $x_\phi = (x_{\phi 1} x_{\phi 2} ... x_{\phi p})$ e sua respectiva estimativa intervalar obtida através de um modelo de regressão, $\hat{y}_\phi = [\underline{\hat{y}}_\phi, \bar{\hat{y}}_\phi]$. Os vetores linha (pontuais) $x_\phi^\infty = (1\underline{x}_{\phi 1}\underline{x}_{\phi 2}...\underline{x}_{\phi p})$ e $x_\phi^{sup} = (1\bar{x}_{\phi 1}\bar{x}_{\phi 2}...\bar{x}_{\phi p})$, baseados em x_ϕ,

são usados pelos modelos para estimar $\hat{y}_\phi$. Fagundes (2013) enfatiza que na literatura uma grande quantidade de autores propuseram modelos de regressão para dados simbólicos do tipo intervalo e ilustra que Lima Neto e De Carvalho (2008) propuseram o Método do centro e da amplitude para ajustar o Modelo de regressão linear básico (MRLC) com um melhor desempenho do que antes fora apresentado em Billard e Diday (2000). Desta forma em Maia, et al (2008) há a apresentação de abordagens para a previsão de séries temporais para dados simbólicos do tipo intervalar. Em 2010 Neto e De Carvalho, (FAGUNDES, 2013) propuseram uma nova abordagem para melhorar o modelo de previsão de séries temporais de regressão linear com restrição no centro e na amplitude dos intervalos, com o intuito de assegurar a coerência matemática entre as variáveis previstas nos limites inferior e superior.

3.2.1 Método do Centro

De acordo com Reyes (2022), Billart e Diday propuseram um método que ajusta o modelo de regressão linear ao centro dos intervalos assumidos pelas variáveis simbólicas e os aplica aos limites inferior (inf) e superior (sup) dos intervalos das variáveis preditoras para verificar a previsão do limite inferior e da variável resposta. Neste sentido, o método de centro diz que, conforme Souza (2016) os coeficientes do modelo construído e os limites dos regressores são usados na predição dos limites da resposta. O autor diz que os limites inferiores dos regressores determinam o limite inferior da resposta. O mesmo ocorre com os limites superiores.

O método do centro para variáveis simbólicas do tipo intervalo pode ser formalmente definido do seguinte modo: Seja $E = \{e_1, e_2, ..., e_n\}$ um conjunto de exemplos descritos por $p + 1$ variáveis intervalares Y e $X_1, X_2, ..., X_n$; e seja cada exemplo $e_i \in E(i = 1, ..., n)$ representado por um vetor de intervalos $Z_i = (X_i, y_i)$, onde $X_i = (X_{i1}, X_{i2}, ..., X_{ij}, ..., X_{ip})$, com $X_{ij} = \xi_{ij} = [a_{ij}; b_{ij}] \in \Omega = \{[a; b] : a \leq b \text{ e } \{a, b\} \in R^1\}(j = 1, ..., p)$ e $y_i = [y_i^{inf}; y_i^{sup}] \in \Omega$, caracterizando os valores observados de X_j e Y. Considere Y como variável resposta e $X_1, X_2, ..., X_p$ como variáveis regressoras relacionadas por (REYES, 2017):

$$y_i^{inf} = \beta_0 + \beta_{1ai1} + ... + \beta_{paip} + \varepsilon_i^{inf}$$

$$y_i^{sup} = \beta_0 + \beta_{1bi1} + ... + \beta_{pbip} + \varepsilon_i^{sup}$$

3.2.2 Método do Mínimo e do Máximo

Neste ínterim, Souza (2016) discorre também sobre outro método proposto por Billart e Diday, sendo chamado de método do mínimo e máximo que ajusta dois modelos independentes de regressão linear para os limites inferiores e superiores das variáveis simbólicas.

Considere o conjunto de variáveis $X_1, X_2, ..., X_p$ como variáveis regressoras relacionadas linearmente com uma variável resposta Y através do modelo:

$$y_i^{inf} = \beta_0^{inf} + \beta_1^{inf} ai1 + ... + \beta_p^{inf} aip + \varepsilon_i^{inf},$$
$$y_i^{sup} = \beta_0^{sup} + \beta_1^{sup} bi1 + ... + \beta_p^{sup} bip + \varepsilon_i^{sup} \quad (3.1)$$

A partir da equação (3.1), pode-se deduzir a soma dos quadrados dos erros no método dos limites mínimo e máximo, que são:

$$\sum_{i=1}^{n} (\varepsilon_i^{inf})^2 + \sum_{i=1}^{n} (\varepsilon_i^{sup})^2 = \sum_{i=1}^{n} (y_i^{inf} - \beta_0^{inf} - \beta_1^{inf} ai1 - ... - \beta_p^{inf} aip)^2 + \sum_{i=1}^{n} (y_i^{sup} - \beta_0^{sup} - \beta_1^{sup} bi1 - ... - \beta_p^{sup} bip)^2 \quad (3.2)$$

Essa equação representa a soma dos quadrados dos resíduos dos limites inferiores e dos limites superiores de forma independente, considerando também independentes os vetores de parâmetros β utilizados para predição dos limites da variável resposta $\hat{Y}$.

Os estimadores de mínimos quadrados de $\beta_0^{inf}, \beta_1^{inf}, ..., \beta_p^{inf}$ e $\beta_0^{sup}, \beta_1^{sup}, ..., \beta_p^{sup}$ que minimizam a equação (3.2) podem ser escritas na notação matricial por:

$$\hat{\beta} = (\hat{\beta}_0^{inf}, \hat{\beta}_1^{inf}, ..., \hat{\beta_p^{inf}}, \beta_0^{\hat{sup}},, \beta_1^{\hat{s}}up, ..., \beta_p^{\hat{sup}})^T.$$

(3.3)

Onde A é uma matriz $2(p+1) \times 2(p+1)$ e b é um vetor $2(p+1) \times 1$.

3.2.3 Método do Centro e da Amplitude

Neste cenário, Fagundes (2013) afirma que foi proposto um novo método de regressão simbólica levando em consideração o centro e a amplitude das variáveis intervalares. Neste método existe um critério de minimização para estimação dos parâmetros e considera a soma dos quadrados dos erros relativos do centro e da amplitude dos intervalos de forma independente. A expectativa é de que com a inclusão de informações da amplitude dos intervalos haja uma melhoria na predição do modelo. O ajuste dos limites inferiores e superiores da variável resposta é realizado por meio da aplicação do vetor de parâmetros $\hat{\beta}$ ao centro e amplitude das variáveis regressoras.(FAGUNDES,2013).

Sejam y^c e x_j^c com $(j = 1, 2, ..., p)$ variáveis quantitativas relativas ao centro dos intervalos das variáveis simbólicas y e x_j com $(j = 1, 2, ..., p)$. Além de que, considere y^r e $x_j^r (j = 1, 2, ..., p)$ variáveis quantitativas que assumem como valores a metade da amplitude (ou meia-amplitude) dos intervalos das variáveis simbólicas y e $x_j (j = 1, 2, ..., p)$. Considere y^c e y^r como variáveis resposta e x_j^c e $x_j^r (j = 1, 2, ..., p)$ um conjunto de variáveis regressoras relacionadas por:

$$y_i^c = \beta_0^c + \beta_1^c x_{i1}^c + ...\beta_p^c x_{ip}^c + \varepsilon_i^c,$$
$$y_i^r = \beta_0^r + \beta_1^r x_{i1}^r + ...\beta_p^r x_{ip}^r + \varepsilon_i^r$$

(3.4)

Neste método, os vetores de parâmetros $\hat{\beta} = ((\hat{\beta}^c)^T, (\hat{\beta}^r)^T)^T$ são estimados de forma independente para o centro e a amplitude dos intervalos. Sendo assim, a soma dos quadrados dos erros é dada por:

$$\sum_{i=1}^{n}[(\varepsilon_i^c)^2+\varepsilon_i^r)^2] = \sum_{i=1}^{n}(y_i^c-\beta_0^c-\beta_1^c x_{i1}^c-...-\beta_p^c x_{ip}^c)^2+\sum_{i=1}^{n}(y_i^r-\beta_0^r-\beta_1^r x_{i1}^r-...-\beta_p^r x_{ip}^r)^2. \quad (3.5)$$

Os estimadores de mínimos quadrados de $\beta_0^c, \beta_1^c, ..., \beta_p^c$ e $\beta_0^r, \beta 1_r, ..., \beta_p^r$ que minimizam a equação (3.5) podem ser escritas em notação matricial por:

$$\hat{\beta} = (\hat{\beta}_0^c, \hat{\beta}_1^c, ..., \hat{\beta}_p^c, \hat{\beta}_0^r, \hat{\beta}_1^r, ..., \hat{\beta}_p^r)^T = (A)^-1b, \quad (3.6)$$

Em que A é uma matriz $2(p+1) \times 2(p+1)$ e b é um vetor $2(p+1) \times 1$.

3.2.4 Métodos com restrições

A partir das análises realizadas e dos estudos bibliográficos desenvolvidos, faz-se importante ressaltar uma síntese acerca de dados simbólicos, que fora apresentado anteriormente neste trabalho. Deste modo, este tipo de conotação para dado simbólico é definido a partir da realização simbólica $\xi = [a : b]$, com $\{a, b\} \in \Re^1$, a e b, variáveis quantitativas, representando, respectivamente, o limite inferior e o limite superior de um intervalo, onde necessariamente a condição $(a \leq b)$ deve ser atendida.(FAGUNDES,2013).

Para além disso, conforme Fagundes (2013) ainda afirma que pode-se demonstrar que, em alguns cenários, não se tem a garantia de que os intervalos preditos pelos métodos apresentados nesta seção contemplem a definição dos dados simbólicos do tipo intervalo, como por exemplo, mantendo a estimativa do limite inferior do intervalo predito menor do que a estimativa do limite superior deste intervalo para qualquer observação intervalar x_i.

A partir disso, foram propostos métodos na resolução deste problema, e assim impor restrições quanto aos valores estimados dos parâmetros dos modelos do método do centro, método do centro e da amplitude, e do método dos mínimos e máximos. Neste sentido,

o modelo estabelece uma relação linear entre a variável resposta e as variáveis regressoras, impondo restrições aos parâmetros do vetor β, da seguinte forma:

$$
\begin{aligned}
&\mathsf{y}_i^{inf} = \beta_0 + \beta_1 a_{i1} + ... + \beta_p a_{ip} + \varepsilon_i^{inf}, \\
&y_i^{sup} = \beta_0 + \beta_1 b_{i1} + ... + \beta_p b_{ip} + \varepsilon_i^{sup}, \\
&\text{restritos a } \beta_j \geq 0, j = 0, 1, ..., p. \quad (3.7)
\end{aligned}
$$

A estimação dos parâmetros β do modelo com restrições segue os mesmos passos dos métodos simbólicos descritos anteriormente. No entanto, o uso de restrições no vetor de parâmetros β restringe o espaço de prováveis soluções que minimiza a soma dos quadrados dos erros, podendo assim, ocasionar uma perda de desempenho de predição se comparado com os métodos sem restrição descritos nas seções 3.2.1, 3.2.1, 3.2.3. Isto posto, sugere-se utilizar, a princípio, o modelo sem restrição, a fim de alcançar as estimativas dos parâmetros que minimizam a soma dos quadrados dos erros.Entretanto, caso sejam identificados observações onde os valores estimados para os limites inferior e superior estejam incoerentes, é recomendado a abordagem corespondente com restrições para re-estimar apenas aquelas observações que apresentam problemas.

4 ABORDAGEM PROPOSTA

Para dar prosseguimento às etapas desta pesquisa, levando em consideração que não existe uma facilidade na adoção de um modelo de dados, em se tratando de dados intervalares, para a construção de um modelo de regressão linear que relaciona tais variáveis. Desta forma, Souza (2016) afirma que os intervalos podem ser vistos como uma agregação de pontos, as propostas que existem na literatura sugerem a escolha de alguns deles para a construção dos modelos de regressão. Os métodos presentes na literatura diferem quanto ao conjunto de pontos utilizados na representação dos intervalos. E desta maneira, continua dizendo que um método de regressão linear para intervalos deve oferecer, de forma simultânea, duas estimativas: os limites inferiores e superiores dos intervalos. Além disso, as estimativas dos limites devem manter a coerência matemática dos intervalos, em que o limite superior é maior ou igual ao inferior.

Tendo em vista que aqui, há como objetivo principal a aplicação de algoritmos clássicos de aprendizagem de máquina na transformação de dados simbólicos em dados clássicos, buscando alcançar resultados tão bons quanto utilizando dados clássicos, aplicando as etapas de encontrar os protótipos, as matrizes de distância e aplicando os modelos preditivos.

Para tanto, o estudo de dados simbólicos originou a área de análise de dados simbólicos, que está diretamente ligada a uma abordagem na área de descoberta automática de máquinas para os dados clássicos. Assim,utilizar-se-á da matriz de dissimilaridade já que vem sendo amplamente explorada na literatura de aprendizagem de máquina, por obter bons resultados em variadas tarefas.

Neste sentido, de acordo com o andamento da pesquisa, pretende-se melhorar e contribuir com os resultados de uma outra pesquisa de Silva Filho e Pontes (2020), tendo como base o estudo da média e desvio padrão das acurácias e dos tempos de treinamento de cada método para todos os conjuntos de dados em modelos de classificação.

Os resultados dos autores mostraram que as variações SDSA tendem a ser mais rápidas e que pelo menos um dos métodos preditivos aplicados teve um melhor resultado que o algoritmo IVABC. Tendo em vista a construção desta pesquisa usando a matriz de dissimilaridade é possível observar a transformação dos dados para a busca de melhores resultados, o que viabilizará atualização e o treinamento de novos modelos, permitindo que haja a existência de resultados de acurácia em um tempo menor. Assim, esta metodologia aponta como resultados a serem obtidos a construção de novos espaços de dissimilaridade usando distâncias do tipo

intervalar e a visualização para a transformação de dados simbólicos em dados clássicos. Esta pesquisa destina-se aos estudos de classificação e regressão, em que, dado um conjunto de treinamento $T = \{(\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\}$ deve-se encontrar uma função $f : X \to Y$, onde Y é o espaço de saída e X é o espaço de entrada, sendo este último um espaço simbólico intervalar. De tal maneira, o objetivo do método proposto, chamado de Análise de Espaços de Dissimilaridade Simbólicos (SDSA – do inglês Symbolic Dissimilarity Spaces Analysis) é encontrar uma transformação $g : X \to R^M$, em que R^M é um espaço real com M dimensões, de forma que a função f pode ser encontrada por qualquer algoritmo de aprendizagem de máquina clássica.

De acordo com Souza (2016) os métodos existentes na literatura para agrupamento de dados intervalares utilizam os limites dos intervalos como pontos representativos e, a partir disso com eles, são propostas as mais variadas distâncias para medir a dissimilaridade entre os intervalos.

Dando continuidade aos procedimentos metodológicos, o algoritmo SDSA primeiro encontra um conjunto W com M centróides a partir do conjunto de treinamento simbólico T. A seguir, dados T, W e uma distância escolhida d, o SDSA calcula a matriz de distâncias $D \in R^M$, usando a equação abaixo:

$$\mathsf{D}_N \times M = \begin{bmatrix} d(\vec{x}_1, \vec{w}_1) & d(\vec{x}_1, \vec{w}_2) & \cdots & d(\vec{x}_1, \vec{w}_M) \\ d(\vec{x}_2, \vec{w}_1) & d(\vec{x}_2, \vec{w}_2) & \cdots & d(\vec{x}_2, \vec{w}_M) \\ \vdots & \vdots & \ddots & \vdots \\ d(\vec{x}_N, \vec{w}_1) & d(\vec{x}_N, \vec{w}_2) & \cdots & d(\vec{x}_N, \vec{w}_M) \end{bmatrix} \quad (4.1)$$

A partir disso, o SDSA ajusta um classificador ou regressor $\hat{f}(D, \vec{y}) : R^M \to Y$, onde $\vec{y} = \{y_1, ..., y_N\}$. O SDSA pode ser ajustado usando um classificador C ou Regressor R qualquer e esta pesquisa fundamentou-se nos seguintes algoritmos para a classificação: Árvore de Decisão $(SDSA_{DT})$, Floresta Aleatória$(SDSA_{RF})$, Máquina de Vetores de Suporte $(SDSA_{SVM})$ e, por fim, a Regressão Logística$(SDSA_{LR})$(SILVA FILHO; PONTES, 2020). Para regressão: Regressão Linear $(SDSR_{linear})$, Regressão Multivariada$(SDSR_{ols})$, Random forest Regressor$(SDSR_{rf})$, Support Vector Regression$(SDSR_{svr})$ e Árvore de decisão$(SDSR_{dt})$.

Este trabalho apresentou modelos de classificação e regressão, os classificadores e regressores foram todos avaliados usando as bibliotecas scikit-learn e statsmodels, com seus

valores pré-definidos de hiperparâmetros. Assim, para encontrar os centróides, o SDSA aplica um algoritmo de agrupamento intervalar qualquer G aos dados pertencentes a cada classe $k \in \{1, ..., K\}$, retornando um conjunto W_k de centroides da classe, tal que $W = U_{k=1}^{K} W_k$.

Nesta pesquisa, o método de agrupamento escolhido foi o k-médias++ (Arthur e Vassilvitskii 2007) e a distância escolhida d foi a Euclidiana quadrática intervalar, tendo em vista que uma variável aleatória simbólica intervalar X, toma valores em um intervalo, $X = [l; u] \subset R^1$, onde $l \leq u$ e $l, u \in R^1$ são os limites inferior e superior do intervalo, respectivamente. A partir disso, nessa definição, a distância euclidiana padrão não pode ser diretamente aplicada a vetores de variáveis intervalares (SILVA FILHO; PONTES, 2020). Uma forma de adaptar a distância euclidiana quadrática para dados intervalares é dada pela Equação abaixo:

$$d_{lu}(\vec{x}, \vec{w}) = d_l(\vec{x}, \vec{w}) + d_u(\vec{x}, \vec{w}) \quad (4.2)$$

Pela equação euclidiana acima d_l e d_u são respectivamente a distância euclidiana quadrática dos limites inferiores e a distância euclidiana quadrática dos limites superiores e $\vec{x}$ e $\vec{w}$ são vetores de variáveis intervalares.

Além do k-médias++, também foi testado o SDSA apenas selecionando os centróides (usando o passo de inicialização do k-médias), mas sem atualizá-los. Assim, o conjunto de centróides W fica composto de instâncias obtidas do conjunto de treinamento e chamamos os modelos resultantes de $SDSA^*_{DT}$, $SDSA^*_{RF}$, $SDSA^*_{SVM}$ e $SDSA^*_{LR}$ e $SDSR^*_{linear}$, $SDSR^*_{ols}$,$SDSR^*_{rf}$ e $SDSR^*_{svr}$. Desta maneira, abaixo apresenta-se os Algoritmos 2 e 3, respectivamente com os passos para ajustar um modelo SDSA e para testá-lo (SILVA FILHO; PONTES, 2020).

Algoritmo 2: Passos para treinamento de um modelo de SDSA

Input: Conjunto de treinamento T, número de centróides por classe $\{M_1, M_2, ..., M_K\}$, onde $\sum_{k=1}^{K} M_k = M$, algoritmo de agrupamento G e algoritmo de classificação C
Output: conjunto W com M centroides e classificador $\hat{f} : R^M \rightarrow Y$
begin
    Inicialize o conjunto de centroides $W = \phi$;
    for $k = 1 : K$ **do**
        Obtenha o conjunto de observações da classe $k : T_k = \{(\vec{x}_i, y_i) | (\vec{x}_i, y_i) \in T$ e $y_i = K\}$
        Obtenha o conjunto de centroides $W_k = G(T_k, M_k)$;
        faça $W \cup W_k$;
        end Calcule a matriz de distâncias D, usando a equação(4.1)
        Ajuste o classificador $\hat{f} = C(D, \vec{y})$;
        returnconjunto de centroides W e classificador $\hat{f}$;
end

Algoritmo 3: Passos para teste de um modelo de SDSA

Input: Conjunto de teste T, conjunto de centroides W e classificador $\hat{f}$
Output: Predições para as instâncias T
begin
        calcule a matriz de distância D, usando a equação(4.1);
        return predições$\hat{f}(D)$
end

O modelo SDA de classificação proposto foi comparado com o IVABC (SILVA FILHO; SOUZA; PRUDÊNCIO, 2016), método que representa o estado da arte em classificação de dados intervalares. O IVABC usa um método de otimização baseado em enxames para encontrar os protótipos ótimos e fazer seleção de variáveis, além de uma distância Euclidiana intervalar ponderada, capaz de modelar regiões de variadas formas e tamanhos nos dados. Os métodos foram avaliados usando quatro conjuntos de dados intervalares reais: cogumelos, climas, climas secos e climas europeus (Silva Filho, Souza e Prudêncio 2016).

Os experimentos foram realizados por meio de dez repetições de validações cruzadas com dez partições , tendo como resultado 100 taxas de acurácia para cada par de método e conjunto. Os valores dos hiperparâmetros do IVABC foram mantidos em seus valores ótimos para cada conjunto, segundo o artigo original (Silva Filho, Souza e Prudêncio 2016). Além disso, as quantidades de centróides usadas pelas variações do SDSA foram as mesmas usadas pelo IVABC. Além das atividades ligadas ao desenvolvimento do método proposto. Já o modelo SDA para regressão proposto foi comparado com o método do centro e da amplitude (LIMA

NETO; CARVALHO, 2008), que neste trabalho está sendo exibido nos resultados como CA, que pode ser compreendido como um método de regressão simbólica que leva em consideração o centro e a amplitude das variáveis intervalares.

Foi proposto na literatura por Neto e de Carvalho (2008), que concerne como sendo um novo método da regressão simbólica que leva em consideração o centro e a amplitude das variáveis intervalares. Desta maneira, neste método de acordo com Souza (2016) a modelagem da regressão intervalar é feita através de dois modelos independentes: um baseado nos centros e outro baseado na amplitude dos intervalos. Para os centros, é utilizada a mesma estratégia aplicada ao Método de Centro (SOUZA, 2016).

Conforme Souza (2016) Wang, Guan e Wu (2012) consideram o Método do centro e da amplitude um avanço com relação ao Método do centro. Ele pode melhorar o desempenho para a predição dos intervalos quando existe uma dependência linear entre as amplitudes dos regressores e da resposta. A coerência matemática na predição dos limites não é garantida por este método.

5 RESULTADOS

Realizados os passos descritos na Seção 4, obtive-se a média das acurácias e do tempo de treinamento, os resultados obtidos para cada conjunto de dados e os algoritmos de classificação podem ser visualizados abaixo:

Tabela 4 – Resultados de Classificação para o dataset climates

dataset	classifier	Acurácia		Tempo de Execução	
		mean	std	mean	std
climates	ivabc	0.816612227	0.029674174	6187.47569	8380.567485
climates	sdsa_dt	0.694693837	0.034446855	1.725049489	0.316631517
climates	**sdsa_lr**	**0.844351713**	**0.030171214**	**165.7748546**	**64.69536384**
climates	sdsa*lr	0.842801918	0.034747005	102.1051543	41.84330035
climates	sdsa*dt	0.69888223	0.037271893	1.201718194	0.424185134
climates	sdsa_rf	0.765786635	0.035955997	4.452389119	1.210107315
climates	sdsa*rf	0.76703676	0.035018191	3.381959441	0.981558292
climates	sdsa_svm	0.655473979	0.038542685	3.143751638	0.841205342
climates	sdsa*svm	0.655688742	0.037744265	0.618288281	0.143090494

Fonte: elaborado pela autora

Tabela 5 – Resultados de classificação para o dataset Dry-climates

dataset	classifier	Acurácia		Tempo de Execução	
		mean	std	mean	std
dry-climates	ivabc	0.943080552	0.031173802	2904.91557	8619.503338
dry-climates	sdsa_dt	0.926255443	0.037230528	0.797075696	0.24
dry-climates	sdsa_lr	0.95	0.03182834	15.83493268	6.819594678
dry-climates	**sdsa*lr**	**0.951738026**	**0.031910607**	**5.361375728**	**1.883687519**
dry-climates	sdsa*dt	0.92738389	0.036421165	0.104988456	0.028505968
dry-climates	sdsa_rf	0.947122642	0.029292493	1.653425231	0.413918828
dry-climates	sdsa*rf	0.946923077	0.028794367	0.585601799	0.168833104
dry-climates	sdsa_svm	0.93004717	0.03647949	1.322152054	0.340091207
dry-climates	sdsa*svn	0.930050798	0.036981675	0.076554077	0.013374975

Fonte: elaborado pela autora

Tabela 6 – Resultados de classificação para o dataset European-Climates

dataset	classifier	Acurácia		Tempo de Execução	
		mean	std	mean	std
european-climates	ivabc	0.955814	0.034165	216.6892272	96.47757
european-climates	sdsa_dt	0.924025	0.039408	0.346861215	0.065136
european-climates	**sdsa_lr**	**0.962244**	**0.038767**	**3.724894056**	**1.458465**
european-climates	sdsa*lr	0.961638	0.036275	1.3677022	0.489421
european-climates	sdsa*dt	0.91875	0.047037	0.043602273	0.010218
european-climates	sdsa_rf	0.929886	0.043296	1.264159784	0.246582
european-climates	sdsa*rf	0.930492	0.045219	0.397822094	0.123514
european-climates	sdsa_svm	0.941051	0.040122	0.780743027	0.208622
european-climates	sdsa*svm	0.940739	0.040272	0.040408401	0.008203

Fonte: elaborado pela autora

Tabela 7 – Resultados de Classificação para o dataset Mushroom

dataset	classifier	Acurácia		Tempo de Execução	
		mean	std	mean	std
mushroom	ivabc	0,718541667	0,057827215	16,18089485	6,767719702
mushroom	sdsa_dt	0,613833333	0,113734621	0,043276808	0,009395465
mushroom	sdsa_lr	0,725208333	0,037676702	0,191923614	0,039938353
mushroom	sdsa*lr	0,730208333	0,032552809	0,055095918	0,014884755
mushroom	sdsa*dt	0,610041667	0,111933658	0,00567003	0,002513716
mushroom	sdsa_rf	0,665541667	0,083880474	0,508492069	0,097149781
mushroom	sdsa*rf	0,68325	0,086649061	0,34764087	0,102066268
mushroom	**sdsa_svm**	**0,74**	**0,008206099**	**0,094385512**	**0,019053248**
mushroom	**sdsa*svm**	**0,74**	**0,008206099**	**0,007999225**	**0,002378547**

Fonte: elaborado pela autora

Também foi realizado o teste de hipótese não paramétrico de Wilcoxon a um nível de significância de 5% , já que as médias das acurácias não seguem uma distribuição normal. Para avaliar a performance foram formuladas as seguintes hipóteses:

Hipótese nula: média da acurácia melhor classificador = média acurácia dos outros classificadores.

Hipotese alternativa: média da acurácia melhor classificador $>$ média acurácia dos outros classificadores.

Os testes foram realizados comparando a acurácia do melhor classificador para cada dataset contra todos os outros classificadores, os resultados para cada dataset podem ser visualizados nas tabelas abaixo:

Tabela 8 – Resultados de classificação para o dataset climates

Dataset	Melhor classificador	classificador comparado	P-Valor
Climates	$SDSA_{LR}$	$SDSA^{*}_{LR}$	2,15E-01
Climates	$SDSA_{LR}$	IVABC	**1,66E-11**
Climates	$SDSA_{LR}$	$SDSA_{RF}$	**2,90E-18**
Climates	$SDSA_{LR}$	$SDSA^{*}_{RF}$	**2,80E-18**
Climates	$SDSA_{LR}$	$SDSA_{DT}$	**1,92E-18**
Climates	$SDSA_{LR}$	$SDSA^{*}_{DT}$	**1,92E-18**
Climates	$SDSA_{LR}$	$SDSA_{SVM}$	**1,93E-18**
Climates	$SDSA_{LR}$	$SDSA^{*}_{SVM}$	**1,93E-18**

Fonte: elaborado pela autora

Tabela 9 – Resultados de classificação para o dataset dry-climates

Dataset	Melhor classificador	classificador comparado	P-Valor
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA_{LR}$	1,80E-01
Dry-Climates	$SDSA^{*}_{LR}$	IVABC	**5,60E-03**
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA_{RF}$	**3,20E-02**
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA^{*}_{RF}$	**1,50E-02**
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA_{SVM}$	**2,24E-07**
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA^{*}_{SVM}$	**1,84E-07**
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA_{DT}$	**2,53E-10**
Dry-Climates	$SDSA^{*}_{LR}$	$SDSA^{*}_{DT}$	**3,17E-09**

Fonte: elaborado pela autora

Tabela 10 – Resultados de classificação para o dataset european-climates

Dataset	Melhor classificador	classificador comparado	P-Valor
European-Climates	$SDSA_{LR}$	$SDSA^{*}_{LR}$	3,80E-01
European-Climates	$SDSA_{LR}$	IVABC	**2,73E-02**
European-Climates	$SDSA_{LR}$	$SDSA_{SVM}$	**1,94E-05**
European-Climates	$SDSA_{LR}$	$SDSA^{*}_{SVM}$	**1,47E-05**
European-Climates	$SDSA_{LR}$	$SDSA^{*}_{RF}$	**5,11E-10**
European-Climates	$SDSA_{LR}$	$SDSA_{RF}$	**2,84E-10**
European-Climates	$SDSA_{LR}$	$SDSA_{DT}$	**9,92E-12**
European-Climates	$SDSA_{LR}$	$SDSA^{*}_{DT}$	**2,98E-13**

Fonte: elaborado pela autora

Tabela 11 – Resultados de classificação para o dataset Mushroom

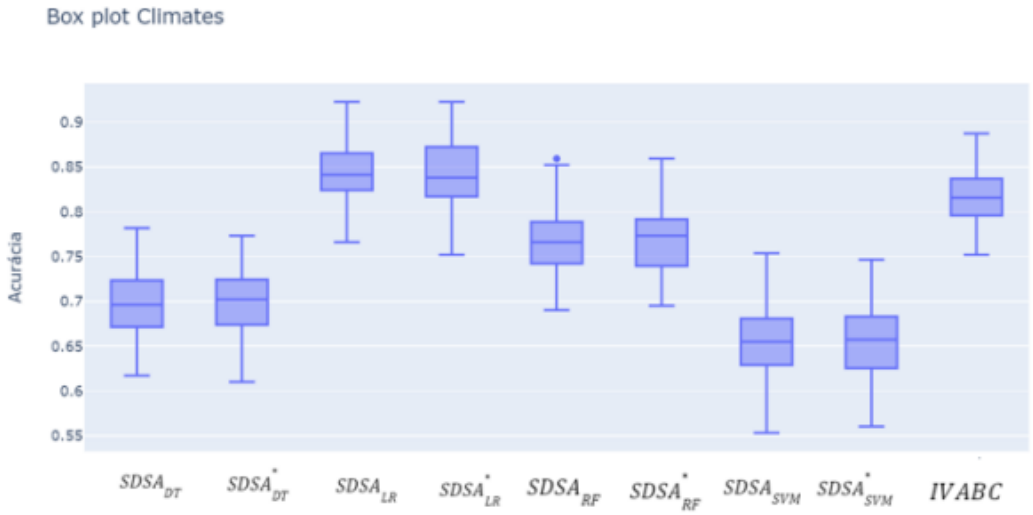
Dataset	Melhor classificador	classificador comparado	P-Valor
Mushroom	$SDSA_{SVM}$	$SDSA^{*}_{SVM}$	Valores iguais entre os métodos
Mushroom	$SDSA_{SVM}$	$SDSA^{*}_{LR}$	**5,00E-03**
Mushroom	$SDSA_{SVM}$	$SDSA_{LR}$	**1,20E-03**
Mushroom	$SDSA_{SVM}$	IVABC	**9,24E-04**
Mushroom	$SDSA_{SVM}$	$SDSA^{*}_{RF}$	**3,99E-07**
Mushroom	$SDSA_{SVM}$	$SDSA_{RF}$	**8,95E-11**
Mushroom	$SDSA_{SVM}$	$SDSA_{DT}$	**3,42E-13**
Mushroom	$SDSA_{SVM}$	$SDSA^{*}_{DT}$	**2,31E-14**

Fonte: elaborado pela autora

Sendo assim, pode-se observar que pelo menos uma versão do SDA foi melhor que o IVABC em cada um dos conjuntos de dados, e o tempo de execução foi consideravelmente menor em todas as versões do SDA, também é possível observar que o tempo de execução para as versões do SDA apenas selecionando os centróides(usando o passo de inicialização do k-médias++), mas sem atualizá-los foram ainda menores, já com relação as acurácias, pode-se observar que não há diferença entre as versões com e sem atualização dos centróides. As duas versões do SDA que usam regressão logística tiveram as maiores médias, sendo maiores do que o IVABC, o que torna um resultado interessante, já que os classificadores clássicos foram treinados com valores pré-definidos de hiperparâmetros não sendo otimizados para estes conjuntos como é o caso do IVABC. Assim, como é possível observar nos testes de hipótese que, em sua grande maioria, com exceção dos casos em que são comparados os mesmos metódos com e sem atualização dos centróides, rejeitamos a hipótese nula.

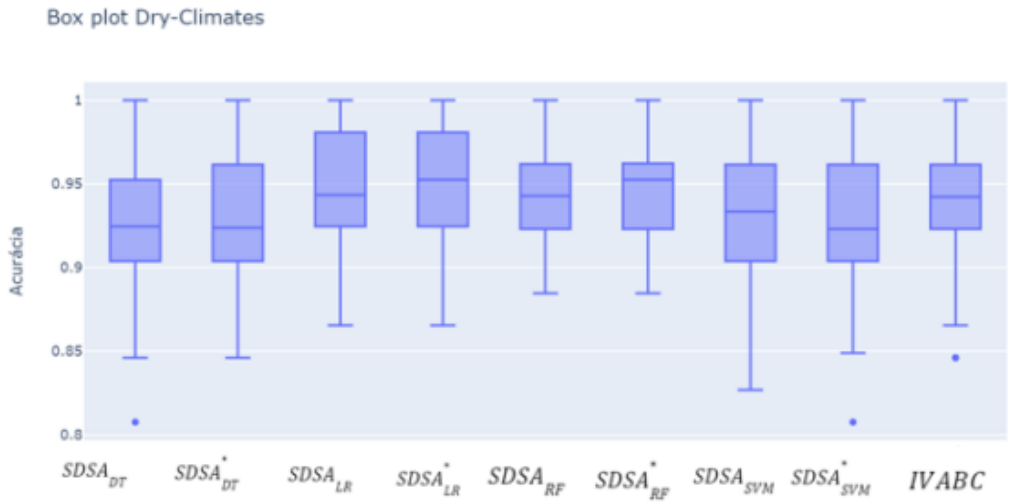
As figuras abaixo mostram os gráficos de Boxplot com os resultados da acurácia para cada um dos classificadores em cada um dos conjuntos de dados:

Figura 2 – Box plot Climates



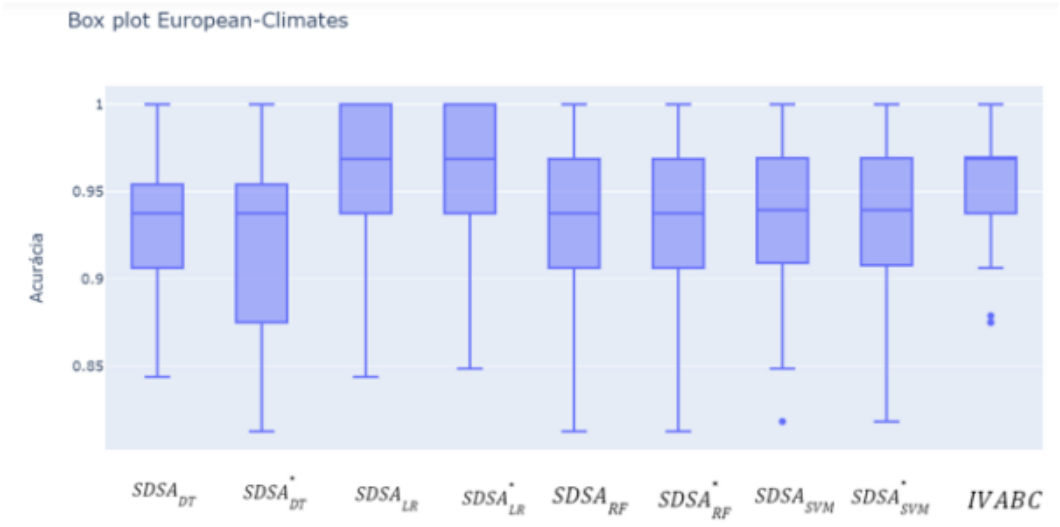
Fonte: elaborado pela autora

Figura 3 – Box plot Dry-Climates



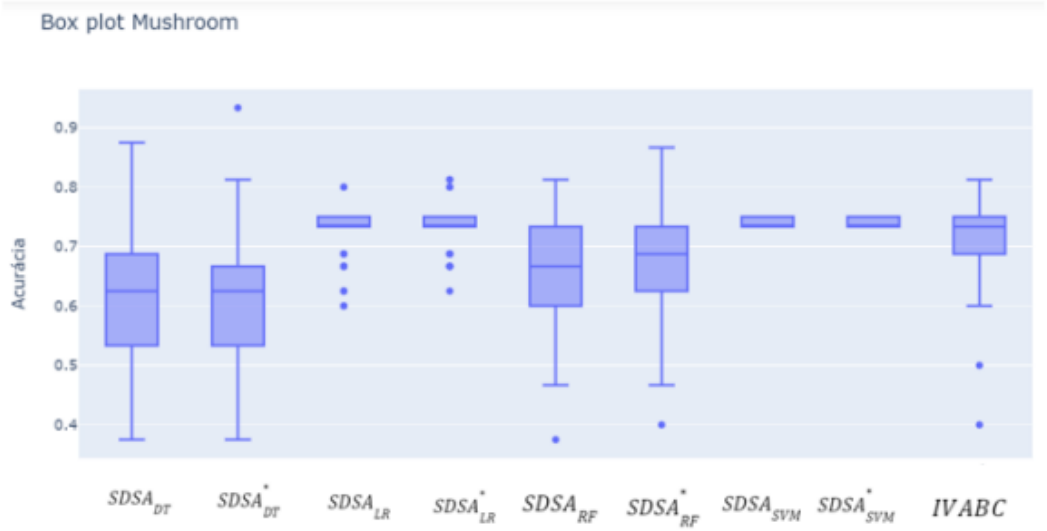
Fonte: elaborado pela autora

Figura 4 – Box plot European-Climates



Fonte: elaborado pela autora

Figura 5 – Box plot Mushroom



Fonte: elaborado pela autora

Os resultados obtidos para os algoritmos de regressão para cada dataset podem ser visualizados nas tabelas abaixo:

Tabela 12 – Resultados de regressão para o dataset akc-data

dataset	regression_model	MMRE		Tempo de Execução	
		mean	std	mean	std
akc-data	sdsr_dt	8,840196	5,146827	0,012167	0,002272
akc-data	sdsr*dt	9,204592	5,534139	0,005377	0,00113
akc-data	**sdsr_linear**	**4,289295**	**3,236581**	**0,02856**	**0,011359**
akc-data	sdsr*linear	4,332344	3,296841	0,009341	0,004741
akc-data	sdsr_ols	4,585102	3,287012	0,041578	0,0124
akc-data	sdsr*ols	4,616782	3,312911	0,030087	0,009876
akc-data	sdsr_rf	5,377162	3,648564	0,079202	0,013012
akc-data	sdsr*rf	5,530037	3,611569	0,041985	0,006729
akc-data	sdsr_svr	4,305679	3,369003	0,029627	0,006202
akc-data	sdsr*svr	4,305453	3,368383	0,011502	0,002427
akc-data	CA	4,294721	3,252893	0,027391	0,010273

Fonte: elaborado pela autora

Tabela 13 – Resultados de regressão para o dataset Climates

dataset	regression_model	MMRE		Tempo de Execução	
		mean	std	mean	std
climates	sdsr*dt	10,06273	3,082844	0,125647	0,021644
climates	sdsr_linear	11,35653	2,361217	0,321978	0,100577
climates	sdsr*linear	10,63326	2,163873	0,023157	0,007923
climates	sdsr_ols	12,49896	3,202104	0,625178	0,175804
climates	sdsr*ols	12,86494	3,402374	0,454257	0,135479
climates	**sdsr_rf**	**5,722489**	**1,456756**	**1,19977**	**0,126277**
climates	sdsr*rf	6,036764	2,054476	0,969114	0,079059
climates	sdsr_svr	14,04706	4,260744	5,427838	0,52273
climates	sdsr*svr	13,91148	3,954808	4,910331	0,452699
climates	CA	4,389859	0,267491	0,03098	0,011503

Fonte: elaborado pela autora

Tabela 14 – Resultados de regressão para o dataset Mushroom

dataset	regression_model	MMRE		Tempo de Execução	
		mean	std	mean	std
mushroom	sdsr*dt	1,221011	0,680675	0,005285	0,000918
mushroom	**sdsr_linear**	**0,716585**	**0,437681**	**0,027705**	**0,008855**
mushroom	sdsr*linear	0,715456	0,433186	0,009189	0,004885
mushroom	sdsr_ols	0,745569	0,446509	0,047193	0,014321
mushroom	sdsr*ols	0,756616	0,440914	0,038771	0,011698
mushroom	sdsr_rf	0,798443	0,447039	0,056698	0,012152
mushroom	sdsr*rf	0,787839	0,430709	0,054154	0,012136
mushroom	sdsr_svr	0,729824	0,471818	0,016217	0,003286
mushroom	sdsr*svr	0,729036	0,470359	0,013156	0,002921
mushroom	CA	0,74925	0,409546	0,026536	0,0144

Fonte: elaborado pela autora

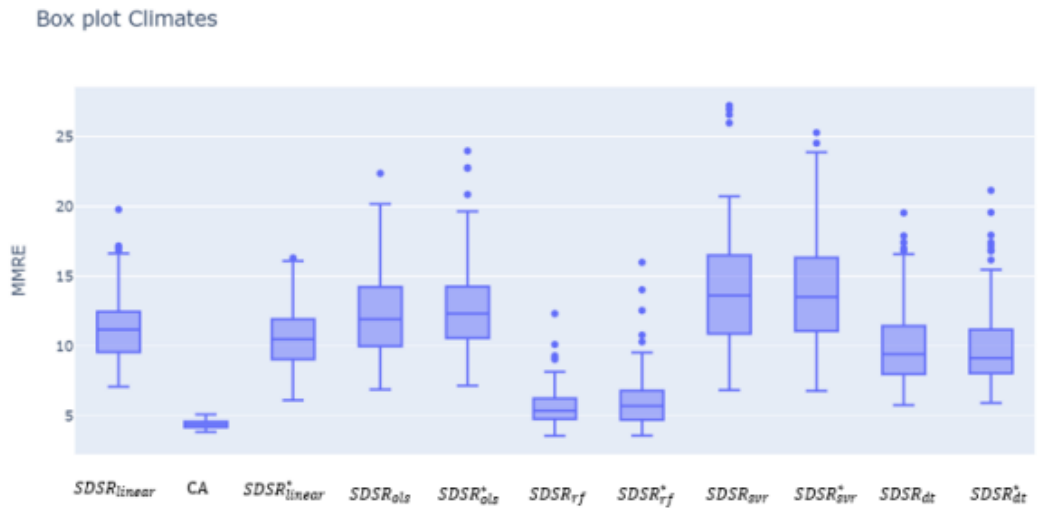
A precisão dos modelos de regressão foi calculada a partir da Magnitude Média do Erro Relativo(MMRE), o MMRE para dados simbólicos de tipo intervalar é dado por:

$$MMRE = \frac{1}{2_n} \sum_{i=1}^{n} \{ \left| \frac{y_{Li} - \hat{y}_{Li}}{y_{Li}} \right| + \left| \frac{y_{Ui} - \hat{y}_{Ui}}{y_{Ui}} \right| \} \quad (5.1)$$

Assim como em classificação, é possível observar que o tempo de execução para os modelos clássicos foram ligeiramente menores em alguns casos se comparado com o método do centro e da amplitude(CA), com relação a precisão não há uma diferença muito grande com relação aos modelos, a versão do $SDSR_linear$ chegou a apresentar um resultado relativamente menor que o método do centro e da amplitude para os conjuntos de dados akc-data e mushroom.

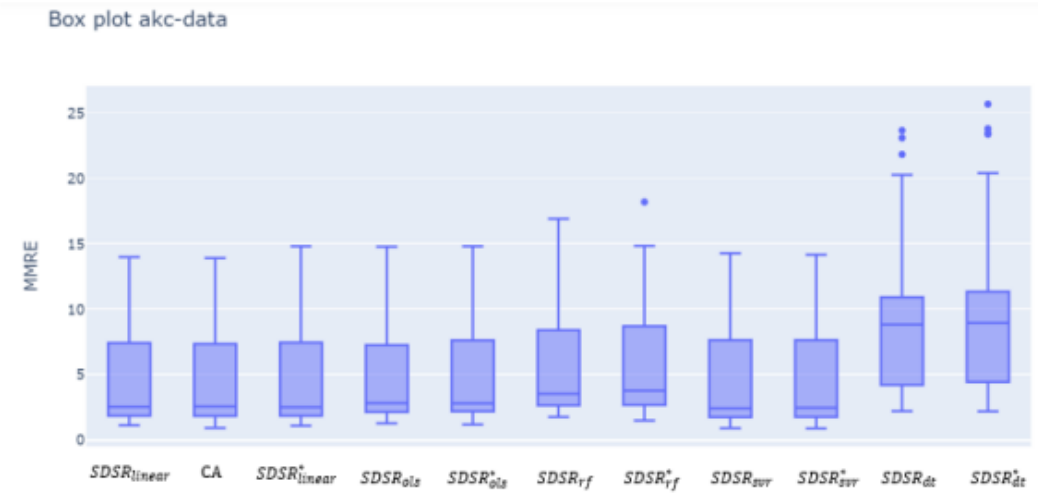
As figuras abaixo mostram os gráficos de Boxplot com os resultados do MMRE para cada dataset e cada um dos Regressores:

Figura 6 – Box plot Climates



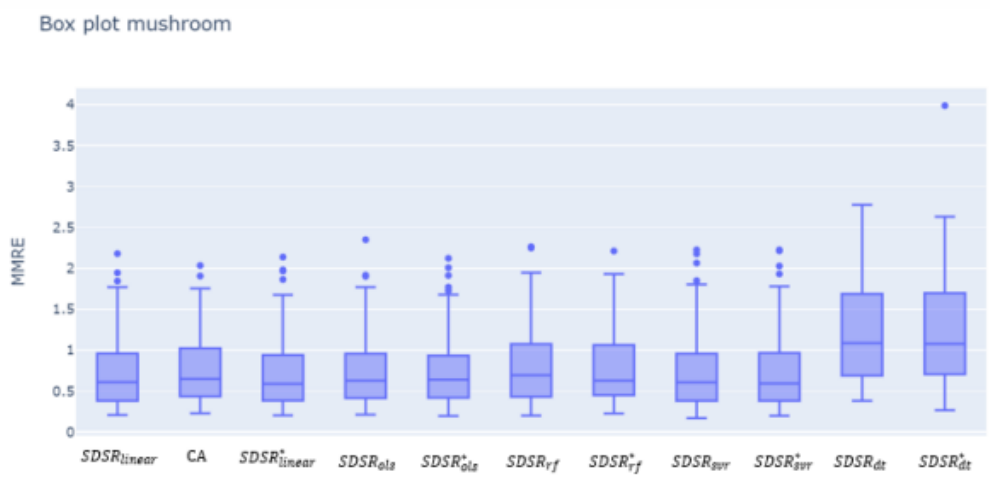
Fonte: elaborado pela autora

Figura 7 – Box plot Akc-data



Fonte: elaborado pela autora

Figura 8 – Box plot Mushroom



Fonte: elaborado pela autora

Também foi realizado o teste de hipótese não paramétrico teste de Wilcoxon a um nível de significância de 5% , já que as médias das acurácias não seguem uma distribuição normal, para avaliar a performance foram formuladas as seguintes hipóteses:

Hipótese nula: média MMRE do melhor regressor = média MMRE dos outros regressores.

Hipotese alternativa: média MMRE do melhor regressor $<$ média MMRE dos outros regressores.

Os testes foram realizados comparando os MMREs do melhor regressor para cada dataset contra todos os outros regressores, os resultados para cada dataset podem ser visualizados nas tabelas abaixo:

Tabela 15 – Resultados de regressão para o dataset Akc-data

Dataset	Melhor Regressor	regressor comparado	P-Valor
akc-data	$SDSR_{linear}$	CA	2,09E-01
akc-data	$\boldsymbol{SDSR_{linear}}$	$SDSR^{*}_{linear}$	**4,90E-02**
akc-data	$SDSR_{linear}$	$SDSR^{*}_{svr}$	3,98E-01
akc-data	$SDSR_{linear}$	$SDSR_{svr}$	4,28E-01
akc-data	$SDSR_{linear}$	$SDSR_{ols}$	**6,29E-08**
akc-data	$SDSR_{linear}$	$SDSR^{*}_{ols}$	**2,10E-09**
akc-data	$SDSR_{linear}$	$SDSR_{rf}$	**3,83E-18**
akc-data	$SDSR_{linear}$	$SDSR^{*}_{rf}$	**5,91E-18**
akc-data	$SDSR_{linear}$	$SDSR_{dt}$	**1,94E-18**
akc-data	$SDSR_{linear}$	$SDSR^{*}_{dt}$	**4,38E-18**

Fonte: elaborado pela autora

Tabela 16 – Resultados de regressão para o dataset Climates

Dataset	Melhor Regressor	regressor comparado	P-Valor
Climates	CA	$SDSR_{rf}$	2,44E-16
Climates	**CA**	$SDSR^{*}_{rf}$	**1,67E-15**
Climates	CA	$SDSR^{*}_{dt}$	1,94E-18
Climates	CA	$SDSR_{dt}$	1,94E-18
Climates	**CA**	$SDSR^{*}_{linear}$	**1,94E-18**
Climates	**CA**	$SDSR_{linear}$	**1,94E-18**
Climates	**CA**	$SDSR_{ols}$	**1,94E-18**
Climates	**CA**	$SDSR^{*}_{ols}$	**1,94E-18**
Climates	**CA**	$SDSR^{*}_{svr}$	**1,94E-18**
Climates	**CA**	$SDSR_{svr}$	**1,94E-18**

Fonte: elaborado pela autora

Tabela 17 – Resultados de regressão para o dataset Mushroom

Dataset	Melhor Regressor	regressor comparado	P-Valor
Mushroom	$SDSR_{linear}$	$SDSR^{*}_{linear}$	3,42E-01
Mushroom	$SDSR_{linear}$	$SDSR_{svr}$	**7,38E-02**
Mushroom	$SDSR_{linear}$	$SDSR^{*}_{svr}$	8,72E-02
Mushroom	$SDSR_{linear}$	CA	9,05E-02
Mushroom	$SDSR_{linear}$	$SDSR_{ols}$	**5,49E-02**
Mushroom	$SDSR_{linear}$	$SDSR^{*}_{ols}$	**4,06E-05**
Mushroom	$SDSR_{linear}$	$SDSR^{*}_{rf}$	**0,9998**
Mushroom	$SDSR_{linear}$	$SDSR_{rf}$	**8,75E-06**
Mushroom	$SDSR_{linear}$	$SDSR^{*}_{dt}$	**1,73E-16**
Mushroom	$SDSR_{linear}$	$SDSR_{dt}$	**5,41E-18**

Fonte: elaborado pela autora

Assim como em classificação, em regressão também é possível observar que em todos os conjuntos de dados tem-se testes que rejeitam a hipótese nula.

6 CONCLUSÕES

Assim,cita-se, aqui como resultados obtidos a construção de novos espaços de dissimilaridade usando distâncias do tipo intervalar e a visualização para a transformação de dados simbólicos em dados clássicos. De tal maneira, os resultados motivarão pesquisas futuras e o desenvolvimento de artigos científicos em periódicos internacionais. Diante do exposto acima, Neste sentido, de acordo com o andamento da pesquisa foi possível melhorar os resultados de uma outra pesquisa de Silva Filho e Pontes (2020), tendo como base o estudo e a média e desvio padrão das acurácias e dos tempos de treinamento de cada método para todos os conjuntos de dados, acrescentando os algoritmos de regressão que foram comparados com o método do Centro e da Amplitude.

Neste sentido, os resultados deste trabalho motivarão trabalhos futuros que incluem:

1. a otimização dos hiperparâmetros tanto dos classificadores como dos regressores utilizados como parte do SDA;
2. o incremento de mais conjuntos de dados;
3. desenvolvimento de artigos científicos que serão submetidos em revistas científicas e eventos internacionais.

Pretende-se ainda que as hipóteses deste estudo, possam vir a ser utilizadas, no futuro próximo, em uma possível tese de Doutorado, com maior amplitude de universos estudados.

REFERÊNCIAS

BEATON, A.E. : TUKEY, J.W. **The fitting of power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data.** *Technometrics 16*,147-185, 1974.

BILLIARD, L. ; DIDAY, E. **Regression Analysis for Interval-Valued Data.** *Data Analysis, Classification, and Related Methods, eds. Henk A L Kiers, Jean-Paul Rasson, Patrick J F Groenen, and Martin Schader. Berlin, Heidelberg: Springer Berlin Heidelberg 369–74, 2000.* https://doi.org/10.1007/978-3-642-59789-3_58

BILLIARD, L. ; DIDAY, E. **Symbolic Regression Analysis.** *Classification, Clustering, and Data Analysis: Recent Advances and Applications, eds. Krzysztof Jajuga, Andrzej Sokołowski, and Hans-Hermann Bock. Berlin, Heidelberg: Springer Berlin Heidelberg 281–88.* https://doi.org/10.1007/978-3- 642-56181-8_31.

BILLIARD, L. ; DIDAY, E. **Symbolic Data Analysis**: Conceptual Statistics and Data Mining, *Wiley, West Sussex, England 2006*.

BOCK, H.H., ; DIDAY, E. **Analysis of Symbolic Data.** *Study in Classification, Data Analysis and Knowledge Organization. Springer Verlag* 2000.

CHAVENT, M. An Hausdorff distance between hyper-rectangles for clustering interval data. In: AL, D. B. et (Ed.). ***Classification, Clustering an Data Mining Application, Proceedings of the IFCS'04.*** $[S.1]$*: Springer, 2004.* p.333–340.

DA SILVA, W.J.F. **Análise de dados poligonais: uma nova abordagem para dados simbólicos.** *Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,* 2017.

DIDAY, E.. **Thinking by classes in data science: the symbolic data analysis paradigm.** *Wiley Interdisciplinary Reviews: Computational Statistics 8.5, pp. 172–205. DOI: 10.1002/ wics.1384. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1384.*
url: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1384.

FAGUNDES, R. A. A.; SOUZA, R. M. C. R. d.; CYSNEIROS, F. J. A. **Interval kernel regression. Neurocomputing**, $[S.1]$, *v.128, p.371–388,* 2013.

GRIGOLETTI, P. S. Et al. **Análise Intervalar de Circuitos Elétricos**. *Rio Grande do Sul: Universidade Católica de Pelotas,*2006.

HOFFMANN, R. **Análise de regressão: uma introdução à econometria** [recurso eletrônico] / Rodolfo Hoffmann. - - *5. ed. Piracicaba: O Autor,*2006.

MALAQUIAS, P. J. **Modelos de Regressão Linear para Variáveis Intervalares**. *Dissertação. 2017 (Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão) - Faculdade de Economia do Porto,*2017.

REYES, D.M.A. **Predição para Dados Simbólicos Multi-valorados de Tipo Quartis: Caso Especial Dados Representados por Boxplots. 2022.** *Tese (Doutorado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,*2022.

REYES, D.M.A. **Ensaios de modelos de regressão linear e não-linear para dados simbólicos de tipo intervalo. 2017.** *Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,*2017.

RUSSEL, S. J. e NORVIG, P. **Artificial Inteligence:** A modern Approach. *Pearson Education,*2003.

RODRIGUES, S.C.A. **Modelo de Regressão Linear e suas Aplicações.** *Dissertação (Mestrado em Matemática)Universidade da Beira Interior, Portugal,*2013.

RODRIGUES, S.C.A. Histogram-Based User and Item Profiles for Recommendation Systems. *Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,* Pernambuco,2013.

SANTOS, D.S. **Aprendizado de máquina: estatística bayesiana em método de regressão linear simples com aplicação em magnitudes de quasares.** *Trabalho de Conclusão de Curso (Bacharelado em Física - Pesquisa Básica).* Universidade Federal do Rio Grande do Sul, 2018.

SILVA, A. ; BRITO, P. **Linear Discriminant Analysis for Interval Data.** *Computational Statistics, 21, 289–308,* 2006.

SILVA FILHO, T.M.; SOUZA, M.C.R; PRUDÊNCIO, R B.C. **A swarm-trained knearest prototypes adaptive classifier with automatic feature selection for interval data.** *Em: Neural Networks 80, pp. 19–33. issn: 0893-6080. doi: http://dx.doi.org/10.1016/j.neunet.2016.04.006.* http://www.sciencedirect.com/science/article/pii/S0893608016300363.

SILVA FILHO, T.M.; PONTES, W. **Espaços de dissimilaridade simbólicos usando distâncias intervalares.** *Programa de Iniciação Científica da Paraíba - PIBIC,* 2020.

SILVA FILHO, T.M. **Uma abordagem adaptativa de learning vector quantization para classificação de dados intervalares.** *Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,* Pernambuco, 2013.

SILVA FILHO, T.M. **Classificação baseada em protótipos de decisão mais próximos e distâncias adaptativas.** *Tese (Doutorado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,* Pernambuco, 2017.

SILVA FILHO, T.M. learning vector quantization approaches for interval data. *FUZZ-IEEE 2013, IEEE INTERNATIONAL CONFERENCEON FUZZY SYSTEMS PROCEEDINGS, Hyderabad, India.* **Anais**. . . $[S.l.:s.n.]$, 2013. p.1–8.

SILVA FILHO, T.M.; SOUZA, M.C.R. **Bivariate elliptical regression for modeling. Computational Statistics,2022** *Disponível em:<https://link.springer.com/article/10.1007/s00180-021-01189-x>.* Acesso em 10 jul. 2022.

SOUZA, L.C. **Agrupamento e regressão linear de dados simbólicos intervalares em novas apresentações.** *2016. Tese (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco,* Pernambuco, 2016.

SOUZA, L.C. **Numerical Taxonomy.** *The Principle and Pratice of Numerical Classification. Freeman,* Pernambuco, 1973.