Pós-Graduação em Ciência da Computação

Avyner Henrique Bezerra da Fonseca Lucena

Investigação do Uso de Computação Evolucionária para Descoberta de Subgrupos em Conjuntos de Dados Numéricos de Alta Dimensionalidade

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

Avyner Henrique Bezerra da Fonseca Lucena

Investigação do Uso de Computação Evolucionária para Descoberta de Subgrupos em Conjuntos de Dados Numéricos de Alta Dimensionalidade

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência computacional
Orientador: Renato Vimieiro

Recife
2019

Catalogação na fonte
Bibliotecária Arabelly Ascoli CRB4-2068

L935i Lucena, Avyner Henrique Bezerra da Fonseca
Investigação do uso de computação evolucionária para descoberta de subgrupos em conjuntos de dados numéricos de alta dimensionalidade / Avyner Henrique Bezerra da Fonseca Lucena. – 2019.
63 f.: il., fig., tab.

Orientador: Renato Vimieiro
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação. Recife, 2019.
Inclui referências e apêndice.

1. Inteligência computacional. 2. Subgroup discovery. 3. Conjunto de dados numéricos. 4. Alta dimensionalidade. I. Vimieiro, Renato (orientador). II. Título.

006.31 CDD (22. ed.) UFPE-MEI 2019-173

Avyner Henrique Bezerra da Fonseca Lucena

“Investigação do uso de computação evolucionária para descoberta de subgrupos em conjunto de dados numéricos de alta dimensionalidade”

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.

Aprovado em: 09 de agosto de 2019.

BANCA EXAMINADORA

_____

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

_____

Prof. Dr. Miguel Couceiro
Université de Lorraine, França/Computer Science – Inria Nancy

_____

Prof. Dr. Renato Vimieiro
Departamento de Ciência da Computação / UFMG
(**Orientador**)

Dedico este trabalho à minha família, que sempre me apoiou e deu suporte.

AGRADECIMENTOS

Agradecimentos a todos os envolvidos na produção deste trabalho de pesquisa, em especial à FACEPE, que financiou o projeto.

RESUMO

O trabalho proposto nesta dissertação, se trata de um novo algoritmo evolucionário para a área de Subgroup Discovery "SD", com foco na mineração de padrões em conjuntos de dados numéricos de alta dimensionalidade. Subgroup Discovery é uma técnica descritiva para mineração de dados, cujo objetivo é encontrar e descrever subgrupos em conjuntos de dados, a partir de propriedades de interesse previamente definidas. A área em questão possui uma ampla gama de aplicações e casos de uso, porém poucas dessas técnicas são capazes de atuar adequadamente sobre atributos numéricos contínuos. O que pode ser considerado um problema, tendo em vista que conjuntos de dados provenientes do mundo real recorrentemente possuem atributos de diferentes tipos. Para poder trabalhar com dados contínuos, algumas técnicas de SD demandam discretização prévia de tais atributos. Porém, esse tipo de solução tende a trazer perda de informações e resultados imprecisos. Devido ao rápido desenvolvimento das tecnologias de coleta e armazenamento de dados, problemas cada vez mais complexos tendem a surgir. Um bom exemplo, são os conjuntos de dados de alta dimensionalidade, que, por sua vez, podem possuir centenas de milhares de atributos, tornando ainda mais desafiadora a tarefa de mineração de padrões e consequentemente, a descoberta de subgrupos. Até o momento, não existem trabalhos publicados na área com foco em conjuntos numéricos de alta dimensionalidade. Então, o trabalho aqui proposto visa otimizar o processo de descoberta de subgrupos por meio de dois aspectos principais, que são: i) trabalhar adequadamente com dados contínuos sem deixar de abranger categóricos, e ii) propor uma estratégia evolucionária capaz de lidar com conjuntos de dados de alta dimensionalidade. Após a realização de um amplo estudo experimental, o algoritmo proposto se demonstrou competitivo e, muitas vezes, superior em relação a outras técnicas do estado da arte e trabalhos recém-publicados na área.

Palavras chave: Subgroup Discovery. Conjuntos de dados numéricos. Alta Dimensionalidade.

ABSTRACT

The work proposed in this dissertation, is a new evolutionary algorithm for the Subgroup Discovery "SD" area, and its main focus is the mineration of patterns in high dimensionality numerical datasets. Subgroup Discovery is a descriptive technique for data mining, which aims to find and describe subgroups in data, accordingly to predefined interest properties. The SD area have a wide range of applications and use cases, but few are able to perform properly over continuous numerical attributes, what can be seen as a problem, given that real word related datasets, usually have mixed data types. To work with continuous data, some SD techniques require a previous discretization, but this kind of solution may lead to information loss and imprecise results. The fast development of data collecting and storage technologies is leading to the emergence of new complex problems, such as high dimensional data, which can easily have hundreds of thousands attributes, making the pattern mining task and consequently subgroup discovery, even harder in those scenarios. To the moment this work is being presented, there are no published techniques in the area, that address the problem of high dimensional numerical datasets; so the proposed algorithm aims to optimize the subgroup discovery process by means of: Properly representation of numerical data without neglecting categorical data; together with an evolutionary strategy able to deal with high dimensionality data. After an extensive experimental study, the proposed algorithm showed itself competitive and many times superior when compared with other state of the art techniques and recently published works.

Keywords: Subgroup Discovery. Evolutionary Algorithm. Data Mining.

LISTA DE FIGURAS

Figura 1 – EASD - Visão Geral de Funcionamento 24
Figura 2 – Breast Cancer - Média dos Melhores Valores GARSD+ 44
Figura 3 – Breast Cancer - Evolução Média EASD 46
Figura 4 – Gordon - Evolução Média EASD 47
Figura 5 – Gravier - Evolução Média EASD 48

LISTA DE TABELAS

Tabela 1 – Exemplo de Padrões de Consumo em Supermercados 17
Tabela 2 – Exemplo de Conjunto de Dados sobre Aprovação de Crédito 25
Tabela 3 – Descrição dos Conjuntos de Dados Categóricos e Mistos 35
Tabela 4 – Descrição dos Conjuntos de Dados Numéricos 36
Tabela 5 – Comparação de WRACC da Replicação dos Experimentos do GARSD+ 37
Tabela 6 – Configurações Para as Três Melhores Avaliações de WRACC 38
Tabela 7 – Métricas Completas Para as Três Melhores Avaliações de WRACC . . 38
Tabela 8 – Media das Métricas Conjuntos de Baixa Dimensionalidade. 39
Tabela 9 – Teste Ranqueado de Friedman sobre WRACC Observado 40
Tabela 10 – Teste Wilcoxon Signed para o Wracc Observado 40
Tabela 11 – Comparação de WRACC EASD X SSDP+ 41
Tabela 12 – Comparação de WRACC EASD X RefineAndMine 42
Tabela 13 – Métricas Replicação dos Experimentos do GARSD+ 54
Tabela 14 – Resultados Experimentos RefineAndMine 54
Tabela 15 – Comparação de WRACC EASD X RefineAndMine 55
Tabela 16 – Comparação de WRACC EASD X SSDP+ 55
Tabela 17 – Comparação de WRACC Experimentos de Baixa Dimensionalidade . . 56
Tabela 18 – Resultados Experimentos GARSD+ Dados Categoricos 57
Tabela 19 – Resultados Experimentos GARSD+ Dados Mistos 58
Tabela 20 – Resultados Experimentos GARSD Dados Numéricos 59
Tabela 21 – Resultados Médios Experimentos EASD Alta Dimensionalidade 60
Tabela 22 – Resultados Experimentos EASD Dados Numéricos 61
Tabela 23 – Resultados Experimentos EASD Dados Mistos 62
Tabela 24 – Resultados Experimentos EASD Dados Categoricos 63

SUMÁRIO

1 INTRODUÇÃO 12

2 REVISÃO DA LITERATURA 15

2.1 MINERAÇÃO DE DADOS 15

2.1.1 Desafios Motivadores 15

2.1.2 Conceitos 16

2.2 SUBGROUP DISCOVERY 18

2.2.1 Tipos de Variável Alvo 18

2.2.2 Estratégias de Busca 19

2.2.3 Linguagem de Descrição 20

2.2.4 Métricas 20

2.3 ALGORITMOS ESCOLHIDOS 21

3 ALGORITMO EVOLUCIONÁRIO PARA DESCOBERTA DE SUBGRUPOS EM CONJUNTOS DE DADOS NUMÉRICOS DE ALTA DIMENSIONALIDADE 24

3.1 DEFINIÇÃO E OBJETIVOS 24

3.2 REPRESENTAÇÃO E ESTRUTURAÇÃO DE REGRAS 25

3.3 RESTRIÇÕES DE INICIALIZAÇÃO E ALTA DIMENSIONALIDADE 26

3.4 OTIMIZANDO BUSCAS E GARANTINDO DIVERSIDADE 27

3.4.1 Função de Aptidão 27

3.4.2 Critérios de Parada 27

3.4.3 Operadores Genéticos 29

3.5 EASD PASSO A PASSO 32

4 ESTUDO EXPERIMENTAL 34

4.1 METODOLOGIA 34

4.2 DESCRIÇÃO DOS CONJUNTOS DE DADOS 34

4.3 IMPLEMENTAÇÃO E REPLICAÇÃO DO GARSD+ 37

4.4 SELEÇÃO E VARIAÇÃO DE PARÂMETROS 37

4.5 RESULTADOS DOS EXPERIMENTOS 39

4.5.1 Algoritmos Evolucionários e Baixa Dimensionalidade 39

4.5.2 Algoritmos Evolucionários e Alta Dimensionalidade 40

4.5.3 Considerações acerca do RefineAndMine 41

4.6 ANÁLISES DOS RESULTADOS 42

4.7 EVOLUÇÃO POPULACIONAL 43

5 CONCLUSÃO 49
5.1 CONCLUSÕES 49
5.2 PONTOS DE MELHORIAS E TRABALHOS FUTUROS 49

REFERÊNCIAS 51

APÊNDICE A – TABELAS 54

1 INTRODUÇÃO

A geração e utilização massiva de dados é uma característica expressiva da tão aclamada era da informação. Ela é caracterizada pelo crescimento exponencial no volume de dados gerados pelos mais diversos setores da sociedade, em, praticamente, todos os aspectos da vida moderna. Tal contexto vem se tornando cada vez mais evidente devido ao incentivo advindo de empresas e governos em escala global. Nas últimas décadas, ocorreu o rápido desenvolvimento de uma infraestrutura cada vez mais robusta para a coleta, armazenamento e transferência de dados. Sendo assim, *petabytes* são gerados diariamente por empresas, mídias sociais e uma infinidade de outras fontes, incitando a constante necessidade de novas técnicas de mineração de dados capazes de percorrer os mais variados espaços de busca e extrair informações úteis a seus utilizadores.

O problema investigado nessa dissertação, é de como o uso de uma abordagem evolucionária composta por estratégias adequadas, pode permitir a de descoberta de subgrupos em um cenário desafiador de *datasets* com milhares de características contínuas.

A pesquisa apresentada se encaixa no atual contexto da área de mineração de dados, em que o principal objetivo se dá a partir do estudo e desenvolvimento de um novo método heurístico para a área. Mais especificamente, para a tarefa de *Subgroup Discovery*, com foco de atuação sobre conjuntos de dados numéricos de alta dimensionalidade.Devido à seu alto potencial para extração de informações relevantes, a heurística proposta serve como mais um acessório a auxiliar a transição entre uma era de dados para uma real era da informação.

O processo de mineração de dados pode ser tratado por alguns autores como um sinônimo do conhecido KDD (*Knowledge Discovery in Databases*), que é definido por (FAYYAD et al., 1996), como: "O processo não trivial de identificação de padrões em conjuntos de dados, que sejam válidos, novos, potencialmente úteis e por fim compreensíveis." A visão mais aceita sobre mineração de dados é de que se trata de uma importante etapa do processo de KDD. Tal processo por sua vez possui etapas adicionais, tais como: seleção e limpeza dos dados, incorporação de conhecimento prévio e a interpretação e apresentação de maneira adequada, dos resultados obtidos pelas etapas anteriores. Enquanto a mineração de dados se trata de um processo pelo qual são aplicados algoritmos de análise e descoberta de dados, sob níveis aceitáveis de eficiência computacional, produzindo, assim, um subconjunto enumerado de padrões, baseados nos dados utilizados.

Existem diferentes abordagens para a mineração e descoberta de padrões e técnicas específicas para as mesmas, tais como abordagens preditivas ou descritivas, respectivamente. Abordagens preditivas têm como objetivo a construção de modelos baseados em dados que sejam capazes de atribuir valores de uma variável alvo a instâncias ainda não conhecidas pelo modelo em questão. Também englobam algoritmos de aprendizagem de máquina utilizados para classificação e regressão. Por outro lado, na mineração descritiva de dados, os valores

de atributos alvo já são conhecidos para todas as instâncias. Então, o objetivo das técnicas baseadas nessa abordagem é de descrever relações entre atributos e a variável alvo, possibilitando uma melhor compreensão acerca de determinado conjunto de dados. Dentre as técnicas desta abordagem, estão *Emerging Patterns* (DONG; LI, 1999), *Contrast Sets* (BAY; PAZZANI, 2001) e *Subgroup Discovery* (HERRERA et al., 2011), a área a qual o algoritmo proposto neste trabalho pertence.

A tarefa de subgroup discovery pode ser descrita como uma técnica de mineração de dados amplamente aplicável para a mineração descritiva e exploratória de dados. Seu objetivo é extrair conhecimento a partir de subgrupos encontrados em conjuntos de dados, de acordo com propriedades de interesse previamente definidas. Nelas, o conhecimento é descrito por meio de regras, que são compostas por conjuntos baseados em atributo e valor, representados no contexto da área por condições e propriedades de interesse.

Técnicas de *Subgroup Discovery* podem ser aplicadas em uma ampla gama de problemas reais, tais como: *Marketing* (JESUS et al., 2007), Bioinformática (GAMBERGER et al., 2004) e Medicina (LAVRAČ, 2005). Uma característica usual de conjuntos de dados provenientes do mundo real é possuir diferentes tipos de dados, porém são poucas as técnicas de subgroup discovery capazes de trabalhar diretamente com atributos numéricos contínuos. Em (HERRERA et al., 2011), são apresentados os tipos de dados que 16 (dezesseis) algoritmos da área de SD podem atuar sobre, dentre eles apenas 2 (dois) lidavam com atributos numéricos contínuos. Para poder lidar com dados contínuos, algumas técnicas demandam uma discretização prévia. Esse tipo de abordagem, porém, pode levar à perda de informação e a resultados imprecisos. Devido a evidente necessidade de novos algoritmos capazes de atuar sobre conjuntos de dados numéricos, o algoritmo proposto neste trabalho é capaz de lidar diretamente com tais atributos sem prévias discretizações. Utilizando uma abordagem baseada em intervalos numéricos visando representar os atributos contínuos de uma maneira mais apropriada, permitindo ao algoritmo proposto encontrar descrições de subgrupos com uma maior qualidade em relação a outras técnicas da área.

Um segundo ponto explorado pela técnica proposta está no desafio de atuar sobre conjuntos de dados de alta dimensionalidade. Tais conjuntos de dados são encontrados em algumas áreas de atuação, como, por exemplo, a área de bioinformática (WANG et al., 2005). O que caracteriza um conjunto de dados como sendo de alta dimensionalidade é a grande quantidade de colunas, que podem chegar a centenas de milhares.

Existem diferentes abordagens para a tarefa de descoberta de subgrupos, podendo ser divididas entre Exaustivas e Heurísticas (HERRERA et al., 2011). Devido a restrições computacionais e de tempo, abordagens de busca exaustivas demonstram algumas limitações. Pois, mesmo garantindo encontrar a melhor solução possível para um dado problema, acabam se tornando inviáveis em espaços de busca extremamente amplos como os encontrados em conjuntos de dados de alta dimensionalidade. Enquanto abordagens exaustivas exploram todo o espaço de busca disponível por soluções, abordagens heurísticas fazem uso de estratégias específicas para

explorar o espaço de busca de maneira mais inteligente. Se tornando, assim, mais viáveis em espaços de buscas mais extensos e complexos. Dentre as abordagens heurísticas, algoritmos baseados em computação evolucionária (CARMONA et al., 2014) vêm recebendo grande atenção na comunidade de *Subgroup Discovery*, dada a sua bem conhecida capacidade de lidar com problemas de busca e otimização.

O Algoritmo proposto, "*Evolutionary Algorithm for Subgroup Discovery*" (EASD), visa encontrar regras capazes de descrever com precisão subgrupos encontrados em conjuntos de dados, de acordo com uma métrica estatística de não usualidade (CARMONA; JESUS; HERRERA, 2018). Algumas funcionalidades aplicadas na estratégia evolucionária proposta, são: controles de inicialização, reinicialização de populações e punições baseadas na amplitude das regras geradas. Elas demonstram bons resultados nos experimentos realizados otimizando a busca por soluções durante o processo evolucionário; permitindo ao algoritmo tratar problemas de alta dimensionalidade e também ajudando a evitar áreas de ótimos locais aumentando suas chances de convergir para melhores soluções, de acordo com o tempo disponibilizado para a busca.

2 REVISÃO DA LITERATURA

2.1 MINERAÇÃO DE DADOS

Em um contexto em que técnicas tradicionais de armazenamento e análise de dados se tornam rapidamente obsoletas e inaplicáveis devido ao aumento na complexidade de problemas emergentes, novos *frameworks*, como SD *Subgroup Discovery*, têm grande importância na manutenção da tarefa de mineração de dados. A seguir, são apresentados alguns desafios motivadores, tal como conceitos básicos sobre mineração de dados, permitindo um melhor entendimento da área onde a tarefa de SD se encaixa.

2.1.1 Desafios Motivadores

Escalabilidade

A geração massiva de dados e, consequentemente, a evolução de tecnologias para sua coleta, transferência e armazenamento, cria um contexto no qual se torna cada vez mais comum que técnicas de análise de dados sejam capazes de atuar sobre *giga*, *tera* ou até *petabytes*. Trazendo a necessidade do desenvolvimento de algoritmos de mineração de dados e estruturas, focados no tratamento de dados em larga escala. Temos por exemplo, a utilização de computação paralela, algoritmos "*out of core*" e estratégias de busca específicas para espaços com tamanho exponencial.

Análise não-Tradicional

Boa parte do trabalho de análise de dados gira em torno da aplicação de avaliações estatísticas por meio de testes de hipóteses. Porém, devido à complexidade das tarefas atuais de análise, muitas vezes se faz necessária a geração de milhares de hipóteses, motivando a criação de novos algoritmos de mineração capazes de automatizar o processo de geração e avaliação de tais hipóteses.

Alta Dimensionalidade

Em algumas áreas de atuação, é possível encontrarmos conjuntos de dados com centenas de milhares de atributos. Como, por exemplo, as tarefas de processamento de linguagem natural (CHEN et al., 2018), em que os dados podem ser representados por meio de uma vetorização, utilizando a frequência dos termos encontrados em todo um corpo de texto. Ou conjuntos de dados genômicos (WANG et al., 2005), advindos dos avanços das tecnologias de *microarray*, na área de bioinformática. A maioria dos algoritmos de mineração e descoberta de padrões desenvolvidos até o momento possuem o foco em conjuntos de baixa dimensionalidade, não sendo capazes de atuar sobre alta dimensionalidade. Isso evidencia a necessidade de novas técnicas capazes de atuar de maneira satisfatória sobre tais espaços de busca extremamente amplos.

2.1.2 Conceitos

A área de Mineração de Dados engloba conceitos de áreas como a computação, matemática e estatística, visando à resolução de diversos desafios, incluindo os citados acima. Segundo (FAYYAD et al., 1996), pode ser definida como a etapa do processo de KDD, em que são aplicados algoritmos de análise e descoberta de dados, de maneira repetitiva e iterativa, visando recuperar um conjunto de padrões compreensíveis. De acordo com (TAN; STEINBACH; KUMAR, 2013), tarefas de mineração de dados são, geralmente, divididas entre duas categorias mais amplas, abordagens preditivas e descritivas. Abordagens preditivas têm como objetivo prever o valor de um determinado atributo alvo, cujos padrões são desconhecidos, tomando por base exemplos previamente observados. No caso das abordagens descritivas, o objetivo é definido pela descoberta de padrões interessantes que representam correlações entre atributos, descrevendo, assim, suas características.

Para o funcionamento das abordagens preditivas, se faz necessária a construção de modelos capazes de prever um determinado atributo alvo em função dos demais atributos de um conjunto de dados. Existem, basicamente, duas tarefas principais focadas na construção de modelos preditivos: a classificação e a regressão; com relação a primeira, os modelos classificadores são utilizados para prever atributos alvo discretos; já em relação a segunda, nos problemas de regressão, a classe alvo é contínua. Em muitos casos, o objetivo da tarefa se dá na previsão em função de um determinado período de tempo. Um exemplo é a previsão de séries temporais advindas do mercado financeiro.

De acordo com (TAN; STEINBACH; KUMAR, 2013), o objetivo da tarefa de mineração descritiva de dados é a obtenção de padrões (correlações, tendências, anomalias etc). Tais padrões devem ser capazes de descrever relações em conjunto de dados. Não tendo o seu foco voltado para a classificação de novas instâncias, mas no entendimento de um determinado fenômeno, no qual os padrões extraídos devem ser humanamente compreensíveis. Várias técnicas e *frameworks* compõem o arcabouço da mineração descritiva. Um conceito relevante para a área é o de "*Association Analysis*", que consiste na descoberta de relações estatísticas entre instâncias de conjuntos de dados, sendo originalmente proposta como uma metodologia para identificação de tendências em padrões de consumo de clientes em redes de supermercados. Temos um exemplo na Tabela 1, em que tais relações seriam posteriormente apresentadas por meio de regras de associação ou conjuntos de itens frequentes. De forma geral o conceito de "*Association Analysis*" teve um viés muito forte na construção de aplicações e novas técnicas de mineração descritiva de dados. Um exemplo é o trabalho de (VIMIEIRO et al., 2012), com a mineração de padrões nos complexos conjuntos de dados biomédicos.

Um exemplo de regra baseado na Tabela 1 pode ser a famosa relação: $\{Fraldas\} \rightarrow \{Cerveja\}$. Essa regra sugere que existe uma relação forte entre a venda de fraldas e cervejas. Com base no cadastro de compras de clientes, se um cliente compra fraldas, implicaria numa grande possibilidade dele também comprar cervejas, devido a possibilidade de se extrair esse tipo de padrões a partir da relação entre itens em cadastros de transações. O conceito de

Tabela 1 – Exemplo de Padrões de Consumo em Supermercados

ID	**Itens**
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Coca}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Coca}

itens frequentes age como um filtro para a composição de regras. Estas são compostas apenas por conjuntos de itens com um nível mínimo de ocorrência. Sendo assim, estatisticamente interessantes levando a descoberta de padrões com fortes correlações.

Além de descritivas e preditivas, técnicas de mineração de dados também se dividem entre globais e locais (MANNILA, 2002). Abordagens globais realizam análises e verificações utilizando todos os padrões disponíveis. Enquanto o foco de métodos locais se dá pela verificação do quão localmente excepcional, relevante e interessante pode ser um padrão identificado. Essa detecção, como o próprio nome sugere, é bem diferente de detecções globais, focando em padrões localmente distribuídos. Abordagens de busca local podem ser delimitadas por uma série de restrições, tais como: marcadores de uma classe, determinado espaço de tempo etc.

A já consolidada literatura das áreas de "Regras de Associação" e "Mineração de Padrões Frequentes" serve de inspiração. Muitas técnicas descritivas tomam por base e generalizam seus conceitos e ideias. Um grupo, em especial, vem ocupando seu espaço na área. Devido a seu desenvolvimento em paralelo, porém, com diferentes nomenclaturas, em (NOVAK; LAVRAČ; WEBB, 2009) foi introduzida a ideia de se apresentar tal grupo em um *framework* chamado de "*Supervised Descriptive Rule Discovery*", composto por um conjunto de abordagens de mineração descritiva supervisionadas. Tendo o objetivo de realizar uma exploração e descrição de conjuntos de dados, por meio da geração de regras, englobando as técnicas de: *Contrast Sets* (BAY; PAZZANI, 2001), *Emerging Patterns* (DONG; LI, 1999) e *Subgroup Discovery* (HERRERA et al., 2011). Uma característica interessante sobre o *framework* é a presença do termo supervisionado, já que se tratam de técnicas de mineração descritiva de dados, e essas, geralmente, são não-supervisionadas. O que não é o caso das três técnicas citadas, já que estas utilizam os marcadores das classes durante o processo de descoberta de regras.

De maneira geral, a ideia da técnica *Contrast Sets* é de se extrair conjunções de atributo e valor que possuam uma distribuição estatística significantemente diferente entre grupos ou classes em conjuntos de dados. A ideia de se analisar e compreender diferenças entre classes com valores opostos é um problema real explorado em diferentes domínios. Em (BAY; PAZZANI, 2001), pesquisas de ciências sociais e análises de censo são utilizados como exemplos de domínio de aplicação. O objetivo, então, gira em torno de explorar e apresentar diferenças entre os grupos de maneira automática e completa.

Uma revisão ampla sobre a técnica de *Emerging Patterns*, no contexto de mineração

descritiva de dados, pode ser encontrada em (GARCÍA-VICO et al., 2018). A técnica em questão possui muitos casos de uso tanto para a mineração descritiva quanto para a preditiva. Tal característica é possível, pois os padrões emergentes minerados de maneira supervisionada pelas técnicas da área, geralmente, possuem um grande indicativo discriminativo entre classes. O objetivo da técnica é capturar tendências emergentes em conjuntos de dados relacionados a marcadores de tempo ou características úteis para diferenciar conjuntos de instâncias entre classes.

Subgroup Discovery tem como objetivo a descoberta de subgrupos de interesse relacionados com o valor de um dado marcador de classe, possuindo alta adaptabilidade quanto à seleção e utilização de métricas para definir o nível de interesse em padrões descobertos. Tal característica é um dos fatores que mais a diferencia das demais técnicas de mineração descritiva citadas nessa sessão. Em SD, se busca extrair padrões interessantes sobre cada classe de um conjunto de dados. Na técnica de *Contrast Sets*, se busca por padrões que apresentem grande diferença de cobertura entre classes comparadas. Por fim, *Emerging Patterns* foca em encontrar padrões que demonstrem aumento de cobertura de uma classe para outra.

2.2 SUBGROUP DISCOVERY

O problema de subgroup discovery foi primeiramente introduzido por (KLÖSGEN, 1996) e (WROBEL, 1997), sendo definido em (WROBEL, 2001) como: "Em subgroup discovery, assumimos que nos é dada uma população de indivíduos (objetos, clientes,...) e uma propriedade de interesse acerca desses indivíduos à qual estamos interessados. A tarefa de descoberta de subgrupos é descobrir os subgrupos nessa população que sejam estatisticamente mais interessantes, em relação a uma propriedade de interesse.". Ainda de acordo com (WROBEL, 2001), o que torna um subgrupo interessante é a existência de uma distribuição estatística não usual, em que a não usualidade se relaciona com uma propriedade alvo predefinida. Tais afirmações, quanto à não usualidade, se embasam em métricas específicas e em suas análises.

De acordo com os trabalhos de (HERRERA et al., 2011), (HELAL, 2016) e (CARMONA et al., 2014), os principais elementos de um algoritmo da área de Subgroup Discovery são: o tipo de variável alvo, a estratégia de busca, a linguagem de descrição e medidas de qualidade.

2.2.1 Tipos de Variável Alvo

Diferentes análises se fazem necessárias devido aos diferentes tipos de dados encontrados em atributos. Por esse motivo, muitas técnicas possuem restrições acerca dos tipos de dados sobre os quais poderão atuar. Na literatura da área os três tipos de dados citados são: Binários ou Boleanos, nos quais basicamente pode ter apenas dois tipos de valores, verdadeiro ou falso, nominais ou categóricos. Estes acabam tendo sua análise um pouco semelhante à dos boleanos, em que valores de domínio por atributo são considerados de maneira individual. Por fim, existem os atributos numéricos ou contínuos que são considerados os mais complexos

de se tratar. Por seus valores de domínio não serem discretos, tais atributos exigem abordagens diferentes em sua análise. Muitas das técnicas da área não foram projetadas para atuar diretamente sobre conjuntos de dados numéricos. Discretização de todo e qualquer atributo numérico é, muitas vezes, uma opção, o que pode levar à perda de informação e resultados imprecisos, como é demonstrado no trabalho de (GROSSKREUTZ; RÜPING, 2009).

2.2.2 Estratégias de Busca

Diferentes abordagens de busca podem ser utilizadas para a tarefa de descoberta de subgrupos. Como um todo, as técnicas se dividem entre abordagens exaustivas ou heurísticas. As técnicas baseadas em busca exaustiva percorrem todo o espaço de busca de um determinado conjunto de dados, garantindo encontrar soluções ótimas. Para problemas mais simples e conjuntos de dados com poucas dimensões, a busca exaustiva se mostra eficaz e completa, o que não ocorre em conjuntos de dados com espaços de busca mais amplos, em que, devido a amplitude e complexidade, abordagens exaustivas se tornam inviáveis devido à seu custo computacional. Muitas técnicas exaustivas fazem uso de estratégias de poda, visando reduzir o espaço de hipóteses, mesmo assim esse tipo de busca continua sendo extremamente custosa.

Dentre as abordagens heurísticas, técnicas baseadas em *beam search* e algoritmos evolucionários têm grande relevância na área, tendo capacidade de atuar sobre conjuntos de dados com espaços de busca maiores e mais complexos. Um ponto negativo sobre essas abordagens está no fato de em muitos casos não se encontrar a melhor solução possível, uma vez que essas técnicas geralmente não percorrem todo o espaço de busca disponível, sendo esse um *trade-off* a ser considerado.

Assim como algoritmos evolucionários, *Beam search* é uma heurística de busca e otimização. Seu funcionamento tem em foco reduzir gastos com memória e otimizar o tempo, abstraindo o espaço de hipóteses como um grande grafo e ordenando possíveis soluções de acordo com algum critério predefinido. Além disso, permite a heurística de apenas expandir nós mais promissores.

O sucesso de técnicas baseadas em *beam search* sobre espaços de buscas extensos é muito relacionado à maneira como tais algoritmos inicializam a busca por soluções, limitando o tamanho e a quantidade de padrões gerados utilizando um parâmetro chamado largura de banda. Assim, ele refina a sua busca no espaço de soluções, a partir da qualidade dos padrões encontrados em cada iteração.

Trabalhos como (CARMONA et al., 2014) e (FERNÁNDEZ et al., 2010) demonstram estudos aprofundados sobre a relevância de algoritmos evolucionários para o processo de indução de regras. Algoritmos evolucionários têm como principal inspiração o processo de evolução natural, incorporando conceitos da área de ciências biológicas, como: mutação, seleção e cruzamento. Tais algoritmos são utilizados amplamente como mecanismos de busca e otimização de funções.

2.2.3 Linguagem de Descrição

Existe um grande foco na interpretabilidade dos resultados gerados por um processo de descoberta, ressaltando a importância de uma representação clara dos resultados obtidos. O conhecimento obtido é representado por conjuntos de regras, que são compostas da seguinte maneira: *"Condição ⇒ Valor Alvo"*. A parte antecedente da regra, 'Condição', é representada por uma conjunção de atributos de um objeto, sendo organizada como pares de atributos e seus respectivos valores. Já o 'Valor Alvo', na maioria dos casos, representa o indicador de uma classe, e é apresentado na parte consecutiva da regra.

De maneira um pouco mais formal, podemos assumir que ΩA é o conjunto de todos os atributos de um conjunto de dados. Para cada atributo $a \in \Omega A$, uma amplitude de valores é definida em $dom(a)$. A seguir, assumimos que VA é o conjunto universal de valores de atributos na forma de $(a = v)$ para atributos categóricos ou $(v \leq a \leq v)$ para numéricos, em que $(a \in \Omega A)$ é um atributo, e $(v \in dom(a))$ é um valor definido. Digamos que Ex seja um conjunto de exemplos de padrões, $e \in Ex$ é definido pelo conjunto de n-tuplas $e = ((a_1 = v_1), (v_2 \geq a_2 \leq v_2)...(an = vn))$ no qual n é a dada quantidade de atributos, para $vi \in dom(a_i)$.

O tipo das variáveis ou atributos alvo influenciam em como são realizadas análises sobre conjuntos de dados. Como pode ser visto na simples definição formal acima, a representação de dados numéricos pode divergir da representação de dados categóricos. Uma abordagem que vem se demonstrando muito eficaz para o tratamento de atributos numéricos é a representação de padrões por meios de intervalos numéricos.

2.2.4 Métricas

Medidas ou funções de qualidade estão entre os principais componentes de um algoritmo de *Subgroup Discovery*. Elas são responsáveis por filtrar os subgrupos estatisticamente menos usuais e mais interessantes, guiando o processo de descoberta de subgrupos como um todo. Algumas das métricas mais comuns da área são apresentadas a seguir.

Para uma melhor compreensão das equações descritas abaixo, Cond. e Val. são abreviações respectivas à Condição e Valor Alvo.

- A Taxa de *Recall* Eq.2.1 retorna a proporção dos exemplos corretamente classificados cobertos por uma regra, na qual ***n(Cond.Val)*** é o número de exemplos que satisfazem a parte antecedente e consequente da regra. Já ***n(class)*** representa o número total de objetos na classe observada.

$$Recall = n(Cond.Val)/n(class) \tag{2.1}$$

- Confiança de uma regra Eq.2.2 é a frequência relativa de exemplos que satisfazem a regra completa, dentre aqueles que satisfazem apenas a parte antecedente, em que ***n(Cond)***

representa todos os padrões cobertos que satisfazem a parte antecedente da regra.

$$Conf = n(Cond.Val)/n(Cond) \tag{2.2}$$

- A medida de *WRACC* (Weighted Relative Accuracy of a Rule) Eq.2.3, também conhecida na literatura como *"unusualness"* (medida de não usualidade) ou *"gain accuracy"* (ganho de Acurácia), representa um *trade-off* entre generalidade e acurácia relativa. Além disso, é uma métrica muito popular e utilizada por diferentes algoritmos da área. Em (CARMONA; JESUS; HERRERA, 2018), são demonstrados diferentes casos de estudo, reforçando a relevância da métrica. A princípio, a métrica é voltada para problemas binários, mas pode ser facilmente generalizada para problemas com múltiplas classes. Para se fazer isso, aplicamos uma abordagem "Um contra todos", que consiste em selecionar uma classe por vez como positiva e tratar as demais como negativas.

$$WRACC = \frac{n(Cond)}{N} * \left(\frac{(n(Cond.Val)}{n(Cond)} - \frac{n(class)}{N} \right) \tag{2.3}$$

2.3 ALGORITMOS ESCOLHIDOS

Nessa seção, são apresentados alguns algoritmos recentemente publicados e o estado da arte da área de *Subgroup Discovery*. A seleção de cada um dos algoritmos utilizados foi baseada em dois critérios básicos: sua capacidade de atuar sobre conjuntos numéricos e a capacidade de atuar sobre conjuntos de alta dimensionalidade, servindo, assim, como inspiração e para futuras comparações com os objetivos do algoritmo proposto.

Existem diferentes abordagens para a tarefa de descoberta de subgrupos. Técnicas baseadas em busca puramente exaustiva tendem a ser inviáveis em espaços de busca amplos. Mesmo aplicando estratégias para limitar o espaço de busca, esse tipo de abordagem continua sendo extremamente custosa, não sendo uma boa opção para tratar de conjuntos de dados de alta dimensionalidade.

Proposto em (BELFODIL; BELFODIL; KAYTOUE, 2018), o algoritmo *RefineAndMine* é uma nova técnica de busca exaustiva focada na descoberta de subgrupos em conjuntos de dados numéricos. Suas principais características são listadas a seguir. Primeiramente, existe a sua representação do conhecimento, que encontra padrões intervalares por sua vez refinados e otimizados no decorrer do tempo. Como o nome do *Anytime SD* sugere, ele permite parar a busca por soluções a qualquer momento. Por fim, o algoritmo fornece uma garantia teórica do quão longe o processo se encontra de realizar uma busca exaustiva. O formalismo por trás de todo o trabalho apresentado por essa técnica é muito interessante. Para poder atuar sobre conjuntos de dados numéricos, a abordagem os representa por meio de intervalos. Cada intervalo é tratado como um hiper-retângulo, sendo discretizado e refinado a cada iteração.

Técnicas baseadas em Computação Evolucionária vêm tomando cada vez mais espaço na área de subgroup discovery. Alguns algoritmos baseados em Computação Evolucionária, que

são no momento, soluções de estado da arte para a área, estão listados a seguir: SDIGA (JESUS et al., 2007), MESDIF (JESUS; GONZÁLEZ; HERRERA, 2007), NMEEF (CARMONA et al., 2010), GARSD+ (PACHÓN; MATA; DOMÍNGUEZ, 2017) e SSDP+ (LUCAS; VIMIEIRO; LUDERMIR, 2018).

SDIGA (*Subgroup Discovery Iterative Genetic Algorithm*) é um algoritmo mono-objetivo que utiliza regras *fuzzy* e várias medidas de qualidade que são adaptações da área de regras de associação. No final, ele avalia o *WRACC* para definir a qualidade de suas regras. Assim como outros *fuzzy systems*, faz uso de regras linguísticas para a descrição dos subgrupos. Ele utiliza um algoritmo genético no começo do seu processo de busca utilizando busca global, evoluindo, consequentemente, para uma busca local e utilizando o conhecido *Hill Clibing* no refinamento de soluções.

O MESDIF (*Multi-objective Evolutionary Approach for Subgroup Discovery*) é um algoritmo multiobjetivo baseado no SPEA2. O conceito de algoritmos de otimização multiobjetivos, de uma maneira simplificada, significa que serão otimizadas diferentes funções ao mesmo tempo durante o processo evolucionário, em que normalmente se espera algum tipo de concorrência entre as funções otimizadas, o que nem sempre ocorre. Para o seu funcionamento, o MESDIF aplica uma estratégia elitista baseada no conceito de dominância de regiões de Pareto. A inicialização populacional utiliza todos os atributos de um dado conjunto de dados em cada regra, fazendo uso de valores aleatoriamente gerados para criar pares de atributo e valor. A técnica cria diferentes populações durante suas iterações: uma intermediária, na qual são aplicados operadores genéticos básicos, e outra, em que são armazenados os padrões de elite.

O NMEEF-SD (*Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery*) se trata de um algoritmo multiobjetivo baseado no algoritmo NSGA-II, sendo um algoritmo elitista que utiliza as regiões de Pareto durante o processo evolucionário. Porém, com uma inicialização populacional muito mais suave que a do previamente citado MESDIF, aproveitando apenas um quarto dos atributos de um conjunto de dados para 75% população e apenas 25% semelhantemente ao MESDIF. Também tem algumas features de reinicialização populacional baseado em critérios de convergência.

Os três últimos algoritmos citados, SDIGA, MESDIF e NMEEF, utilizam indução por lógica *fuzzy*. Tais técnicas, baseadas em lógica *fuzzy*, tratam atributos numéricos de uma maneira interessante, em que os atributos numéricos são simplesmente representados por termos linguísticos [(JESUS et al., 2007). Mesmo que essas técnicas baseadas em lógica *fuzzy* não demandem uma discretização prévia para as características contínuas, a maneira como dados contínuos são representados pode não ser muito adequada. Com isso aparentam uma visível pseudodiscretização, dado que atributos contínuos são substituídos e representados internamente por labels nominais.

O Algoritmo SSDP+ (*Simple Search Discriminative Patterns*) é um dos mais recentemente desenvolvidos dentre os algoritmos baseados em computação evolucionária, podendo facilmente ser aplicado à tarefa de descoberta de subgrupos. O foco da técnica é lidar com

conjuntos de dados de alta dimensionalidade, porém também demonstra bons resultados em conjuntos de dados de baixa dimensionalidade. Apesar de grandes e inegáveis contribuições para a área, o SSDP+ também necessita de uma discretização prévia para conjuntos de dados numéricos.

O GARSD+ (*Genetic Algorithm for Subgroup Discovery*) é um algoritmo genético multiobjetivo adaptado da área de Regras de Associação para a tarefa de SD. Sua principal característica é de atuar sobre conjuntos de dados numéricos sem discretização prévia. Tal característica é possibilitada pela maneira como atributos numéricos são representados por meio de intervalos numéricos.

Em (LUCAS et al., 2017), os algoritmos SDIGA, MESDIF e NMEEF foram testados sobre conjuntos de dados de alta dimensionalidade. O resultado dos experimentos demonstrou que ambas as técnicas não são capazes de atuar apropriadamente sobre tais conjuntos, apresentando problemas graves de convergência. No mesmo trabalho, o autor também levanta discussões sobre pontos chaves para se atuar de maneira satisfatória em alta dimensionalidade. O autor cita a possibilidade de se criar restrições de inicialização populacional, que no caso do SSDP+, se demonstraram muito satisfatórias.

O trabalho publicado em (TORREÃO; VIMIEIRO, 2018) aponta, de maneira aprofundada, os efeitos do controle na inicialização populacional para alta dimensionalidade, em que o autor utiliza os algoritmos MESDIF e NMEEF em seus experimentos. Assim aplica diferentes distribuições estatísticas, que forçavam uma certa porcentagem da população gerada a ter um tamanho reduzido. Como resultado, essas técnicas, que, previamente, não convergiam, se tornaram capazes de atuar apropriadamente em alta dimensionalidade.

Como citado há pouco, os algoritmos baseados em regras *fuzzy* não necessitam de discretização prévia para lidar com conjuntos de dados numéricos, devido à sua pseudo discretização, na qual internamente dados numéricos são representados por termos linguísticos. Nos experimentos realizados por (PACHÓN; MATA; DOMÍNGUEZ, 2017), diferentes configurações dos algoritmos SDIGA, MESDIF e NMEEF foram testadas. Nesse teste, se demonstraram inferiores ao GARSD+ em algumas das métricas, porém se mantiveram competitivos em outras.

Inspirado pela necessidade de novas técnicas capazes de trabalhar diretamente com dados numéricos não discretizados e conjuntos de dados de alta dimensionalidade, o trabalho proposto neste projeto de mestrado surge como um novo e competitivo algoritmo para a tarefa de mineração de subgrupos.

3 ALGORITMO EVOLUCIONÁRIO PARA DESCOBERTA DE SUBGRUPOS EM CONJUNTOS DE DADOS NUMÉRICOS DE ALTA DIMENSIONALIDADE

3.1 DEFINIÇÃO E OBJETIVOS

(EASD)*Evolutionary Algorithm for Subgroup Discovery* se trata de uma abordagem evolucionária para a tarefa de Subgroup Discovery. A técnica proposta tem como objetivo apresentar maneiras de como tratar dois grandes desafios para a área, os conjuntos de dados de alta dimensionalidade e como, propriamente, representar atributos numéricos sem prévia discretização. A dificuldade de trabalhar com conjuntos de dados de alta dimensionalidade se dá devido ao grande número de características, que podem chegar a centenas de milhares, tornando o espaço de busca extremamente complexo e extenso. Quanto aos conjuntos de dados numéricos, sua discretização pode levar à perda de informação e resultados imprecisos, já que essa não seria a maneira mais apropriada de se representar tais tipos de dados. Durante essa seção, será apresentado o funcionamento do algoritmo e outros demais operadores genéticos, além de estratégias que visam garantir pontos como convergência e diversidade populacional.

O funcionamento do EASD pode ser observado na Figura 1, o qual é resumidamente descrito a seguir: o algoritmo é executado uma vez para cada classe disponível. Populações de indivíduos são geradas, avaliadas, sofrem mutações, cruzamentos e até reinicializações ao não conseguirem avançar para melhores soluções. Quando uma população atinge critérios de convergência, o melhor indivíduo segue para o conjunto de soluções da classe. O algoritmo encerra o processo evolucionário e segue para a próxima classe, assim que o conjunto de soluções atinge um critério mínimo de qualidade. No final da execução, são retornados os conjuntos de regras para cada uma das classes.

Durante os próximos capítulos, serão introduzidos e explanados os parâmetros do EASD. Em caso de dúvida, todos estão listados no Pseudocódigo 1.

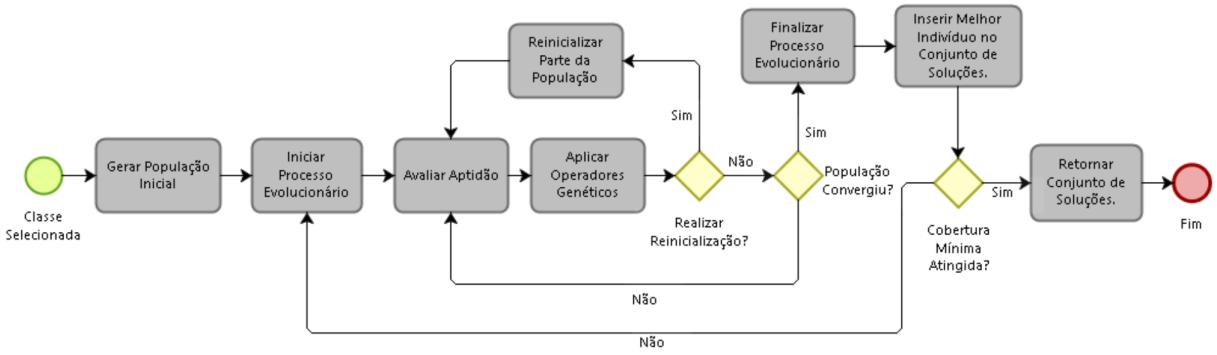

Figura 1 – EASD - Visão Geral de Funcionamento

3.2 REPRESENTAÇÃO E ESTRUTURAÇÃO DE REGRAS

No EASD, cada indivíduo de uma dada população é representado por uma regra, seguindo a estrutura já apresentada anteriormente e possuindo duas partes, antecedente (conjuntos de restrições) e posterior (indicador de classe). A representação interna das regras varia de acordo com os respectivos tipos dos dados que as formam. Atributos categóricos são representados como conjuntos de seus possíveis valores de domínio. Atributos numéricos são, por sua vez, representados por intervalos numéricos compostos por dois limites de valores, um inferior e um superior, respeitando seus valores de domínio. Ainda que o foco da técnica esteja em tratar conjuntos de dados numéricos, a mesmo não negligencia variáveis categóricas, trabalhando com ambos os tipos de dados ao mesmo tempo. Para ilustrar melhor a composição e estrutura de regras geradas, utilizaremos um conjunto de dados fictício, representado pela Tabela **??** a qual apresenta dados sobre aprovação de crédito. Alguns exemplos de regras que poderiam ser encontradas pelo EASD são demonstradas a seguir.

Tabela 2 – Exemplo de Conjunto de Dados sobre Aprovação de Crédito

Educação	**Estado Civil**	**Sexo**	**Filhos**	**Salário**	**Aprovação**
Fundamental	Solteiro	Masculino	Não	1200	Não
Fundamental	Solteiro	Masculino	Sim	1200	Não
Fundamental	Casado	Masculino	Não	2000	Sim
Superior	Divorciado	Feminino	Não	5000	Sim
Superior	Casado	Feminino	Sim	3500	Sim
Ensino Médio	Solteiro	Masculino	Não	1800	Não
Superior	Solteiro	Masculino	Não	4500	Sim
Ensino Médio	Divorciado	Feminino	Não	1800	Sim
Ensino Médio	Solteiro	Feminino	Sim	1800	Não
Ensino Médio	Divorciado	Masculino	Sim	1800	Não

1. Regra com apenas atributos categóricos:
   $(\mathbf{Educação} = (\mathbf{Superior}) \wedge \mathbf{Sexo} = (\mathbf{Feminino})) \rightarrow (\mathbf{Aprovação} = \mathbf{Sim})$

2. Regra com apenas atributos numéricos:
   $((\mathbf{4500} \leq \mathbf{Salário} \leq \mathbf{5000})) \rightarrow (\mathbf{Aprovação} = \mathbf{Sim})$

3. Regra com ambos os tipos de dados:
   $((\mathbf{1200} \leq \mathbf{Salário} \leq \mathbf{1800}) \wedge \mathbf{Filhos} = (\mathbf{Sim})) \rightarrow (\mathbf{Aprovação} = \mathbf{Não})$

O processo de descoberta de subgrupos pelo EASD, se inicia com a geração de uma população inicial de indivíduos, no caso, um conjunto de regras. O tamanho da população é definido pelo parâmetro "Tamanho-População". Para se ter uma boa convergência e seguir uma ideia implementada em várias outras soluções, todos os indivíduos gerados são baseados

em instâncias selecionadas aleatoriamente de suas classes. Isso permite que o algoritmo sempre se encontre no espaço de busca do problema.

No caso de atributos categóricos, seus valores de domínios são representados por um conjunto com, no mínimo, um valor de domínio do atributo selecionado. Atributos numéricos são representados por um intervalo numérico delimitado por um limite inferior e um superior. Os limites são selecionados de maneira aleatória, porém uniforme, sendo, no máximo, 10% menores ou maiores que um dado valor. Este é estocasticamente escolhido a partir do domínio de valores de um atributo. O processo de seleção de atributos para compor uma determinada regra é estocástico e sem repetição de variáveis na mesma regra. Por outro lado, a mesma instância pode ser selecionada mais de uma vez para gerar novos indivíduos.

3.3 RESTRIÇÕES DE INICIALIZAÇÃO E ALTA DIMENSIONALIDADE

Um dos principais fatores para o sucesso, quando se pretende explorar conjuntos de dados de alta dimensionalidade, é a maneira como se inicia a busca por padrões. No caso de algoritmos evolucionários, pode-se aplicar uma série de restrições a esse objetivo. No trabalho de (LUCAS et al., 2017), que se trata da técnica mais relevante publicada até o momento para a descoberta de padrões em conjuntos de alta dimensionalidade, o autor comenta sobre como técnicas baseadas em *beam search* limitam o tamanho de suas populações iniciais, visando encontrar regras mais gerais. Além disso, o autor aplica uma limitação de inicialização, em que toda a população do *SSDP+* se inicia com tamanho um, cobrindo apenas um atributo de um conjunto de dados. Para reforçar a importância da inicialização populacional nesse contexto, o trabalho publicado por (TORREÃO; VIMIEIRO, 2018) demonstra todo um estudo aprofundado nesse tema.

A restrição de inicialização utilizada na abordagem proposta foi de se limitar o tamanho dos indivíduos gerados na população inicial. O tamanho de um indivíduo, ou regra, se dá pela quantidade de atributos que compõem sua parte antecedente. No caso do EASD, esse tamanho é determinado de maneira estocástica e uniforme, variando entre tamanho um e o valor do parâmetro "Tamanho-Indivíduo", que define um limiar máximo para o tamanho das regras geradas.

O uso dessa estratégia, que visa limitar o tamanho máximo de um indivíduo gerado, se demonstrou extremamente positiva, visto que a geração de regras cada vez mais gerais e no caso, mais interessantes. Como já citado anteriormente, quanto maior uma regra, mais específica ela se torna, dificultando assim, a identificação de características estatisticamente interessantes. Desse modo, regras que se tornam muito específicas acabam por corresponder apenas a um número ínfimo de indivíduos. Foram realizados experimentos com técnicas que não realizam nenhum controle de inicialização. Os resultados apontam que essa ideia, ainda que simples, seria um dos principais fatores do sucesso do algoritmo proposto, servindo para lidar com conjuntos de dados de alta dimensionalidade e influenciando uma convergência mais rápida dos processos evolucionários, dado que encontra rapidamente boas soluções.

3.4 OTIMIZANDO BUSCAS E GARANTINDO DIVERSIDADE

Algoritmos evolucionários são notoriamente reconhecidos por sua flexibilidade de uso pelas mais diversas tarefas e áreas de aplicação. No caso do algoritmo aqui proposto, todas as estratégias voltadas para controle, como convergência, manutenção de diversidade e demais operadores genéticos, foram projetados para otimizar sua atuação na tarefa específica de descoberta de subgrupos em conjuntos de dados numéricos de alta dimensionalidade.

Nesse trabalho, a diversidade de um conjunto de soluções não é verificada por meio de métricas. A diversidade é atestada de maneira empírica, com a ideia de que uma população diversa é aquela que possui diferentes indivíduos, explorando diferentes pontos no espaço de soluções.

3.4.1 Função de Aptidão

Todo algoritmo evolucionário tem, em sua essência, uma função a ser otimizada, a qual se denominada função de aptidão ou, pelo termo em inglês, *"Fitness Function"*. Para a tarefa de *subgroup discovery*, utilizando o algoritmo proposto, a função a ser otimizada é a métrica de qualidade utilizada para determinar quão interessantes são determinados conjuntos de subgrupos. No caso do EASD, a função é o já explanado *WRACC*, que foi escolhido devido à sua grande aderência por muitas outras técnicas, facilitando, assim, futuras comparações. Além disso, o *WRACC* foi selecionado em função de sua capacidade de equilibrar métricas, como *recall* e confiança.

Um conceito comum para a área de computação evolucionária são as estratégias de punição. Como o nome sugere, essas estratégias funcionam como um reforço por meio de penalizações de fitness, que buscam impedir possíveis comportamentos indesejáveis advindos da natureza estocástica da área. A maneira pela qual o algoritmo proposto representa os valores de domínio de seus atributos - na forma de conjuntos de possíveis soluções para os dados categóricos e intervalos numéricos para os dados contínuos - pode levar a um crescimento descontrolado na amplitude dos valores representados. Isso é em grande parte um efeito colateral à aplicação do operador genético de mutação. De maneira simples, a punição no fitness de soluções é aplicada caso se atinja um dado limiar de amplitude que varia em decorrência dos tipos de dados. A análise sobre essas amplitudes é realizada sobre cada um dos cromossomos, de acordo com a diferença entre o limiar e a amplitude coberta pelo cromossomo. Então, a punição aplicada no fitness é ponderada pela quantidade de cromossomos a serem punidos e os excessos de amplitude.

3.4.2 Critérios de Parada

Anteriormente, abordamos a importância de uma inicialização adequada para o processo descoberta de subgrupos. Entretanto apenas uma boa inicialização não garante o sucesso de um processo de busca e otimização. O EASD utiliza uma série de outras estratégias para garantir

critérios como convergência e exploração ótimas do espaço de busca, equilibrando tanto a questão da exploração como o refinamento de soluções.

Como o nome sugere, critérios de parada determinam até que ponto uma dada busca por soluções deve continuar. O EASD possui três possíveis critérios de parada, dois internos ao processo evolucionário e um mais externo, que determina quando a busca em uma classe foi concluída.

O critério de parada mais externo é controlado pelo parâmetro percentual de Suporte Mínimo, que é a adaptação de um conceito da área de mineração de regras de associação, funcionando da seguinte maneira: enquanto o conjunto de regras de uma classe não for capaz de atingir um valor mínimo de suporte ou cobertura, processos evolucionários devem ser performados, para que sejam geradas novas regras, e por fim, o critério de cobertura seja atingido. Essa abordagem se demonstrou bastante efetiva, porém um pouco perigosa, podendo levar uma população a nunca convergir, caso o conjunto de regras encontradas não consiga alcançar um padrão de cobertura mínima. Então, esse parâmetro deve ser configurado de maneira consciente e correta, evitando valores muito altos, que significariam a cobertura da classe inteira e também valores muito baixos, o que pode gerar regras ruins. Vale salientar que instâncias cobertas mais de uma vez por diferentes indivíduos contam apenas como um ponto de cobertura. O valor mínimo de cobertura é definido na Equação 3.1, em que N é o número total de padrões em uma classe, e SupMin é a porcentagem definida pelo parâmetro de Suporte Mínimo.

$$CoberturaMínima = N * SupMin \tag{3.1}$$

Existem mais duas condições de término no algoritmo, ambas sendo internas ao processo evolucionário. A primeira condição de término é padrão e ocorre quando é atingido o valor relativo ao parâmetro Número-Gerações, que determina a quantidade máxima de iterações ou gerações que podem ser percorridas durante a execução do processo evolucionário.

A última condição de término se relaciona com os parâmetros: "Número Máximo Reinicializações", Checagem-Reinicialização e Porcentagem-Geracional-Máxima. O "Número Máximo Reinicializações" define quantas vezes uma dada população pode ser reinicializada. Já a Checagem-Reinicialização define uma porcentagem que checa quando uma população deve ser reiniciada, com base na falta de melhoria do melhor indivíduo da população. A Porcentagem-Geracional-Máxima determina uma porcentagem que, ao ser atingida, finaliza o processo evolucionário, apenas se já tiverem ocorrido uma quantidade de reinicializações populacionais igual ao valor encontrado no parâmetro Número-Máximo-Reinicializações. Em seguida, é solicitada uma nova reinicialização da população com base no parâmetro Checagem-Reinicialização.

Essa solução é uma saída para situações em que ocorre um aumento no valor do fitness da melhor solução de uma população, por um dado período de tempo, mesmo após a aplicação de reinicializações e demais operadores genéticos para diversidade. Consideramos, então, que essa população já convergiu; evitando assim gasto de recursos apenas com tempo de *takeover*

dos indivíduos mais aptos sobre os demais valores da população. por exemplo, digamos que o Número-Máximo-Reinicializações é igual a 8, a Checagem-Reinicialização igual a 10% e a Porcentagem-Geracional-Máxima é igual a 80%. A partir do momento em que 80% do número de gerações já tenha se passado e a população já tenha sido inicializada por 8 vezes, na próxima Checagem-Reinicialização sobre a evolução do melhor indivíduo, se não ocorrer nenhuma melhoria, o processo é finalizado. Assim, a finalização desse processo garante a adição da melhor solução ao conjunto de regras de uma determinada classe.

3.4.3 Operadores Genéticos

Um problema a ser considerado em qualquer estratégia de busca são os "Ótimos Locais", que são inerentes à exploração de basicamente qualquer espaço de hipóteses, em que um algoritmo de abordagem heurística poderia facilmente ficar preso no ponto de ótimo local. Assim, podem nunca encontrar um ótimo global nesse espaço de busca.

Para evitar esse tipo de problema, manter uma população diversa em um conjunto de soluções serve para garantir uma maior exploração do espaço de busca. A estratégia de reinicializações populacionais segue essa ideia e tenta mitigar a falta de diversidade por meio da inserção de novos indivíduos em uma população. Desse modo, o parâmetro Checagem-Reinicialização é utilizado para definir uma porcentagem de gerações, que é utilizada para avaliar quando ocorreu o último aumento de fitness da melhor solução da população. Se essa porcentagem é alcançada e o Número-Máximo-Reinicializações ainda não, a população é reinicializada. O parâmetro Taxa-Reinicialização define a porcentagem de indivíduos que serão substituídos durante o processo. Dessa forma, a população de indivíduos é ordenada com base em sua aptidão, e os padrões substituídos pela reinicialização são os menos aptos da população, trazendo, assim, diversidade e oportunidade de se escapar ótimos locais. Todavia nunca perdendo os melhores indivíduos. Mantendo o balanceamento entre exploração e refinamento de soluções.

A cada geração do processo evolucionário, a população sofre mutações e cruzamentos, com base nos em seus respectivos parâmetros e taxas pre definidas, com o objetivo de melhorar a diversidade da população e a exploração e refinamento de soluções no espaço de busca, respectivamente.

Cruzamento

O operador de cruzamento se inicia no EASD com a seleção de dois pais, que são escolhidos aleatoriamente dentro da população. A seguir, se escolhe uma das duas possíveis versões do operador, as quais são aplicadas de acordo com o tamanho dos pais selecionados.

Primeira Opção

Devido à estratégia de controle de tamanho de indivíduos, existe uma grande ocorrência de regras de tamanho um, representando o valor de apenas um atributo em sua parte antecedente. Dado que pelo menos um dos pais selecionados para o cruzamento tenha tamanho igual a um, a primeira opção de cruzamento é selecionada.

- Caso 1: Apenas um dos pais selecionados tem tamanho igual um.
  Condição 1: Se o atributo coberto pelo pai de tamanho um não se encontra entre os atributos cobertos pelo pai de tamanho maior, o filho retornado, nesse caso, é uma junção dos dois pais.
  Exemplo:
  Pai 1: $((4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.
  Pai 2: $(Educação = (Superior) \wedge Sexo = (Feminino)) \rightarrow (Aprovação = Sim)$.
  Resultado: $(Educação = (Superior) \wedge Sexo = (Feminino)) \wedge (4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.

  Condição 2: Se o atributo coberto pelo pai de tamanho um se encontra entre os atributos cobertos pelo pai de tamanho superior, o valor de atributo do pai menor substitui o valor de atributo do segundo pai.
  Exemplo:
  Pai 1: $((4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$
  Pai 2: $(Educação = (Superior) \wedge 4200 \leq Salário \leq 4500)) \rightarrow (Aprovação = Sim)$.
  Resultado: $(Educação = (Superior) \wedge (4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.

- Caso 2: Os dois pais selecionados têm tamanho igual um.
  Condição 1: Caso eles não representem o mesmo atributo, o filho retornado, nesse caso, é uma junção dos dois pais. Exemplo:
  Pai 1: $(Educação = (Superior) \rightarrow (Aprovação = Sim)$.
  Pai 2: $((4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.
  Resultado: $(Educação = (Superior) \wedge (4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.

  Condição 2: Se ambos os pais selecionados cobrirem o mesmo atributo, outro par de pais é selecionado.

Segunda Opção

A segunda opção de crossover é utilizada quando ambos os pais são maiores que um. Então, é aplicada uma adaptação do simples *Single Point Crossover*, retornando dois novos indivíduos a partir da mistura entre os valores dos pais selecionados, sem repetição de atributos. Em ambas as opções de *crossover*, os filhos gerados substituem os pais, caso possuam maior fitness.
Pai 1: $((4500 \leq Salário \leq 5000) \wedge Sexo = (Feminino)) \rightarrow (Aprovação = Sim)$
Pai 2: $(Educação = (Superior) \wedge 4200 \leq Salário \leq 4500)) \rightarrow (Aprovação = Sim)$. Resultado:
Filho 1:$(Educação = (Superior) \wedge (4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.
Filho 2: $((4200 \leq Salário \leq 4500) \wedge Sexo = (Feminino)) \rightarrow (Aprovação = Sim)$

Mutação

No EASD, o operador de mutação consiste na seleção aleatória de ao menos um atributo ou cromossomo de um dado indivíduo, em que são realizadas transformações aleatórias baseadas no tipo de dados do atributo. Caso o melhor indivíduo da população seja selecionado durante o processo de mutação, ele só poderá ser substituído caso sua versão modificada possua um fitness superior ao atual. Como a ideia desse operador é trazer novas soluções e explorar o espaço de busca, os demais indivíduos têm seus valores alterados, mesmo que isso acarrete numa diminuição em sua função de aptidão.

Para atributos numéricos, a mutação, consiste em mudanças nos seus intervalos, obedecendo a um parâmetro percentual interno à técnica e definindo, assim, a amplitude até onde esses valores podem ser alterados. Os novos limites de até quanto os intervalos serão alterados são selecionados aleatoriamente dentro da amplitude permitida, de oito maneiras diferentes:

1. Aumentar ambos os limites.
2. Diminuir ambos os limites.
3. Aumentar o limite à direita.
4. Aumentar o limite à esquerda.
5. Diminuir o limite à direita.
6. Diminuir o limite à esquerda.
7. Diminuir o limite à esquerda e aumentar o da direita.
8. Aumentar o limite à esquerda e diminuir o limite à direita.
9. Remover o atributo da regra, (apenas válido para indivíduos com mais de um atributo).

Exemplo para mutação, Opção 1:
Antes: $((4500 \leq Salário \leq 5000)) \rightarrow (Aprovação = Sim)$.
Depois: $((4700 \leq Salário \leq 5200)) \rightarrow (Aprovação = Sim)$.

Para atributos categóricos, existem três opções de mutação:

1. Remoção do valor da lista atributos de um atributo selecionado, (apenas disponível para atributos com mais de um valor de domínio).
2. Adição de um novo valor aleatório do mesmo conjunto de valores de domínio do atributo selecionado.
3. A substituição de um valor de atributo selecionado aleatoriamente por um valor de domínio do mesmo atributo selecionado.

Exemplo para mutação, Opção 2:

Antes: $(Educação = (Superior) \wedge Sexo = (Feminino)) \rightarrow (Aprovação = Sim)$

Depois: $(Educação = (Superior, EnsinoMédio) \wedge Sexo = (Feminino)) \rightarrow (Aprovação = Sim)$

3.5 EASD PASSO A PASSO

Algorithm 1 EASD Pseudocódigo

Require: Tamanho-População, Tamanho-Indivíduo, Número-Gerações, Taxa-Mutação, Taxa-Crossover, Porcentagem-Geracional-Máxima, Checagem-Reinicialização, Taxa-Reinicialização, Número-Máximo-Reinicializações.

for each Classe **do**
  $regras \leftarrow []$
  while $LinhasNãoCobertas > SuporteMínimo$ **do**
    Gerar População Inicial
    Avaliar Fitness da População
    $nGen \leftarrow 0$
    $Terminate \leftarrow False$
    $ReiniciarPopulação \leftarrow False$
    while $nGen < Número - Gerações$ **do**
      Aplicar Crossover e Mutação
      Avaliar Fitness da População
      if $Checagem - Reinicialização$ is True **then**
        $ReiniciarPopulação$ is True
      if $ReiniciarPopulação$ is True **then**
        Reiniciar População
      if $Prct - Gen - Máxima$ is True and $Num - Máx - Reini$ is True **then**
        Terminate is True
      if $Terminate$ is True **then**
        Encerrar Processo Evolucionário
      else
        $nGen++$
    $regras \leftarrow MelhorIndivíduo$
    Avaliar Regras
    Atualizar Linhas Não Cobertas
  $Regras \leftarrow regras$
  $Classe++$
return $Regras$

O EASD é executado para cada uma das classes de um conjunto de dados, em que para cada uma dessas classes, o processo evolucionário é repetido até que se encontre um ponto mínimo de suporte. Assim, retorna o conjunto de regras no final dos processos.

Ao final de um processo evolucionário, os indivíduos mais aptos formam, então, o conjunto de regras que descrevem os subgrupos encontrados. Em vez de utilizar uma função complexa

de fitness com conceitos mais rebuscados de computação evolucionária, a métrica WRACC, já definida e explanada, é utilizada nesse caso, como função de fitness ou medida de interesse.

Após a primeira geração de indivíduos e uma avaliação de fitness, se inicia um *loop* interno que segue até que seja atingido um valor máximo de gerações predefinidas ou alguma outra condição de término.

A cada geração, todas as avaliações de fitness realizadas são armazenadas. Com base nessas avaliações, são aplicadas estratégias de reinicialização populacional e encerramento do processo evolucionário.

Os operadores genéticos de mutação e crossover são aplicados sobre a população a cada geração ou iteração, respeitando os valores definidos em seus referidos parâmetros de porcentagem.

Após a conclusão do processo evolucionário, o melhor indivíduo da população final é adicionado ao conjunto de regras de uma determinada classe, e parte para a próxima classe assim que um critério de cobertura mínima é atingido.

4 ESTUDO EXPERIMENTAL

4.1 METODOLOGIA

O estudo experimental seguiu uma série de etapas, que são demonstradas a seguir. A primeira etapa dos experimentos foi a seleção e preparação de técnicas da área de SD, sendo endossada pelo trabalho de revisão da literatura. Nela, foram listados alguns dos algoritmos mais relevantes da área, tanto para servirem como inspiração com seus pontos fortes e fracos, mas também por serem competitivos em relação a pontos específicos explorados pelo algoritmo proposto. Foi possível encontrar implementações oficiais de todos os algoritmos utilizados, com exceção do GARSD+, que foi implementado ao melhor das habilidades do autor. Quanto aos demais, o MESDIF, NMEEF e SDIGA foram utilizadas suas implementações do pacote de software Keel. Tanto no SSDP+ quanto no RefineAndMine, os autores originais disponibilizaram implementações em suas respectivas páginas.

Após a seleção e preparação das técnicas, partimos para a seleção dos parâmetros a serem utilizados. No caso do algoritmo proposto, foi realizada uma busca exaustiva, ou *Grid Search*, sobre um conjunto limitado de parâmetros. Para as demais técnicas, foram utilizadas as configurações apontadas como ótimas por seus autores.

Por fim, para permitir uma amostragem populacional suficiente para uma análise estatística, os experimentos foram repetidos por trinta vezes, tendo seus valores médios apresentados no decorrer da seção.

4.2 DESCRIÇÃO DOS CONJUNTOS DE DADOS

A seleção dos conjuntos de dados utilizados levou em conta alguns fatores, tais como tipo e quantidade de atributos. Nas Tabelas 3 e 4, são apresentadas informações básicas acerca dos conjuntos de dados utilizados. No total, foram testados 29 conjuntos de dados encontrados no site *UCI Machine Learning*, dos quais: 10 conjuntos possuem apenas atributos numéricos, 9 deles contém ambos atributos numéricos como categóricos, ou seja, possuem dados mistos, e os demais, 10, possuem apenas atributos categóricos, sendo todos os 29 conjuntos de dados de baixa dimensionalidade. O motivo de se utilizar conjuntos de dados de baixa dimensionalidade. Isso se deu para poder comparar e observar o comportamento algoritmo proposto, juntamente com as demais técnicas da área, dado que a maioria delas não consegue atuar sobre conjuntos de dados de alta dimensionalidade.

Duas das principais funcionalidades do algoritmo proposto é lidar com conjuntos de dados numéricos e de alta dimensionalidade. Foram utilizados 10 conjuntos de alta dimensionalidade, advindos da área de bioinformática. Cada conjunto de dados possui milhares de atributos numéricos.

Tabela 3 – Descrição dos Conjuntos de Dados Categóricos e Mistos

Conjuntos de Dados	**N° Instâncias**	**N° Atributos**	**N° Classes**	**Tipos**
Balanced Scale	625	4	3	Categórico
Lymphography	148	18	4	Categórico
Monk	432	6	2	Categórico
Primary Tumor	339	17	21	Categórico
Solarflare	1066	11	6	Categórico
TicTacToe	958	9	2	Categórico
Bridges Version2	106	11	7	Categórico
Housevotes84	435	17	2	Categórico
Spect	187	24	2	Categórico
Hayes Roth	132	4	3	Categórico
Credit Approval	690	15	2	Mistos
German	1000	20	2	Mistos
Hypothyroid	151	23	2	Mistos
Kidney Disease	400	24	2	Mistos
Hepatitis	155	20	2	Mistos
StatlogHeart	270	13	2	Mistos
AcuteInflamations	120	7	2	Mistos
Saheart	462	9	2	Mistos
Australian Credit	690	14	2	Mistos

Tabela 4 – Descrição dos Conjuntos de Dados Numéricos

Conjuntos de Dados	**N° Instâncias**	**N° Atributos**	**N° Classes**	**Tipos**
Appendicitis	106	7	2	Numérico
Breast Cancer	569	30	2	Numérico
Diabetes	768	8	2	Numérico
Ionosphere	351	33	2	Numérico
Iris	150	4	3	Numérico
Vehicle	846	18	4	Numérico
Wine	178	13	3	Numérico
Glass	214	9	6	Numérico
Ecoli	336	6	8	Numérico
Breast Tissue	106	9	6	Numérico
Chin	118	22215	2	Numérico
Alon	62	2000	2	Numérico
Burczynski	127	22283	3	Numérico
Chiaretti	111	12625	4	Numérico
Christensen	217	1413	3	Numérico
Gravier	168	2905	2	Numérico
Nakayama	105	22283	10	Numérico
Tian	173	12625	2	Numérico
Yeoh	248	12625	6	Numérico
Gordon	181	12533	2	Numérico

4.3 IMPLEMENTAÇÃO E REPLICAÇÃO DO GARSD+

Realizamos a implementação do algoritmo, reproduzindo os resultados apresentados no seu artigo original (PACHÓN; MATA; DOMÍNGUEZ, 2017). Para realizar a comparação, mantivemos os mesmos parâmetros, conjuntos de dados e quantidades de execuções usados no artigo original. Os valores de WRACC, tanto para a implementação dos autores como do artigo original, são apresentados na tabela 5. As demais métricas podem ser consultadas nos apêndices. Utilizando o teste estatístico de *Wilcoxon*, é possível atestar que a hipótese nula não foi rejeitada, com $pvalue = 0.483839851394$. Logo, a implementação dos autores é válida e estatisticamente equivalente, quando comparada aos valores encontrados no artigo original.

Tabela 5 – Comparação de WRACC da Replicação dos Experimentos do GARSD+

Datasets	**WRACC Implementação**	**WRACC Artigo Original**
Vehicle	0.0755	0.046000
Ionosphere	0.05202379	0.126000
Wine	0.13973765	0.148000
Iris	0.18218519	0.195000
German	0.05057521	0.031000
Hypothyroid	0.03501	0.015000
Diabetes	0.0666823	0.055000
Appendicitis	0.04259172	0.033000

4.4 SELEÇÃO E VARIAÇÃO DE PARÂMETROS

Devido a necessidade de se verificar configurações ótimas em relação aos parâmetros da técnica proposta, assim como se faz necessária uma melhor compreensão acerca das correlações entre parâmetros, foi realizada uma busca exaustiva por um conjunto limitado de possíveis valores. Ela resultou na configuração utilizada pelo EASD em todos os experimentos.

Abaixo, são demonstrados os valores dos parâmetros utilizados no *Grid Search*, que geram 5120 combinações de parâmetros, que foram executadas em quatro conjuntos de dados distintos, totalizando 20480 execuções. Após a execução, foi retirada a média das métricas alcançadas, sendo ordenadas pelo maior valor de WRACC, por ser a métrica a ser otimizada. Nas Tabelas 6 e 7, estão listadas as três melhores configurações para a média das 5120 configurações e seus valores de métricas.

Configurações Testadas:

1. Tamanho-População (P.Size) = [100,200,300,400,500]
2. Número-Gerações (Max.Gen) = [100,250,500,1000]
3. Taxa-Mutação (TaxMu) = [10,25,50,75]

4. Taxa-Crossover (TaxCr) = [50,60,70,80]
5. Checagem-Reinicialização (RestChk) = [10,25,50,100]
6. Taxa-Reinicialização (TaxRest) = [10,25,50,75]

Tabela 6 – Configurações Para as Três Melhores Avaliações de WRACC

Index	**P.Size**	**Max.Gen**	**TaxMu**	**TaxCr**	**RestChk**	**TaxRest**
864	500	500	50	50	10	10
545	500	250	50	50	10	25
3456	200	500	70	70	10	10

Tabela 7 – Métricas Completas Para as Três Melhores Avaliações de WRACC

Index	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**
864	0.8231	0.8780	0.7250	1	1.5
545	0.8161	0.8845	0.7227	1	1.5
3456	0.8139	0.8845	0.7205	1	1.5

Os parâmetros a seguir não foram variados durante o *Grid Search*, devido a sua sensibilidade e forte correlação: Tamanho-Indivíduo, Porcentagem-Geracional-Máxima, Número-Máximo-Reinicializações e o limiar de punição.

Configuração Final Utilizada:

1. Tamanho-População = 500
2. Número-Gerações = 500
3. Taxa-Mutação = 50
4. Taxa-Crossover = 50
5. Checagem-Reinicialização = 10
6. Taxa-Reinicialização = 10
7. Tamanho-Indivíduo = 4
8. Porcentagem-Geracional-Máxima = 80
9. Número-Máximo-Reinicializações = 8
10. Limiar de punição = 50

Todos os resultados dos experimentos realizados podem ser facilmente atestados e reproduzidos, e estão disponíveis em (LUCENA, 2019).

4.5 RESULTADOS DOS EXPERIMENTOS

O principal objetivo do estudo experimental realizado foi observar e analisar o comportamento do novo algoritmo proposto, ao lidar com os desafios impostos por diferentes conjuntos de dados, provando, assim, a eficiência das estratégias implementadas. Para uma visualização mais intuitiva dos resultados obtidos nos experimentos, a métrica *WRACC* apresentada nas seguintes tabelas desta seção foi ajustada para que seus valores variem entre 0 e 1. Por fim, uma visualização mais ampla dos resultados, como, por exemplo, desvios padrões e outras métricas além do *WRACC*, pode ser obtida nas tabelas inseridas nos apêndices.

4.5.1 Algoritmos Evolucionários e Baixa Dimensionalidade

Na Tabela 8, são apresentados os resultados médios obtidos pela aplicação dos algoritmos listados, sobre as bases de baixa dimensionalidade, em que os valores destacados demonstram os melhores resultados. As métricas de *Recall*, Confiança e *WRACC* evidenciam que a técnica proposta possui em média valores superiores às demais. O número médio de regras geradas e seus respectivos tamanhos também são apresentados na tabela. Como já explanado anteriormente, quanto menor uma regra, mais geral, e portanto, mais interessante a mesma se torna. Sobre a quantidade de regras, um número baixo de regras tende a indicar uma maior qualidade das mesmas. Por serem boas regras, alcançam mais facilmente os critérios de parada em suas respectivas técnicas, não havendo necessidade da geração de mais regras para atingir restrições mínimas de qualidade.

O algoritmo proposto tem, em média, as menores regras geradas, porém fica em segundo lugar como o algoritmo que gera menos regras, perdendo apenas para o GARSD+. Tal fato ocorre devido à ineficiência da estratégia de controle de amplitude de cobertura e controle no tamanho das regras implementado no GARSD+, fazendo com que, em média, suas regras sejam três vezes maiores do que as geradas pelo EASD. Por esse motivo, sua aplicação em conjuntos de dados com uma maior quantidade de atributos também é impedida, como é o caso de dados de alta dimensionalidade.

Tabela 8 – Media das Métricas Conjuntos de Baixa Dimensionalidade.

Algoritmos	**Recall**	**Confiança**	**WRACC**	**N° Regras**	**Tamanho-Regras**
EASD	**0.8076**	**0.7630**	**0.4811**	3.8958	**1.8279**
GARSD+	0.6924	0.7320	0.3953	**2.4651**	5.8218
MESDIF	0.6054	0.4899	0.1311	10.7586	6.4507
NMEEF	0.6104	0.7329	0.3166	9.0072	3.2668
SDIGA	0.6916	0.5783	0.2337	6.7931	3.4610

As Tabelas 9 e 10 apresentam os resultados da aplicação de dois testes estatísticos não paramétricos, visando atestar que existe diferença estatística entre os algoritmos comparados.

Então dentre os algoritmos evolucionários aplicados sobre os dados de baixa dimensionalidade, tomando como base os *p-values* obtidos nos testes estatísticos aplicados, a hipótese nula de que não existe diferença entre o EASD e os demais algoritmos é rejeitada. Logo, além de possuir em média métricas mais interessantes, também é atestada a sua diferença estatística para com os demais algoritmos da área.

A escolha dos testes estatísticos não paramétricos se dá pela não assunção da ocorrência de uma distribuição normal, nas amostras observadas. O teste de Friedman é utilizado para comparar mais de dois algoritmos ao mesmo tempo. Enquanto seu ranqueamento é baseado na média dos valores apresentados pelas amostras obtidas por cada algoritmo. O teste de Wilcoxon foi utilizado, no caso, para verificar diferenças estatísticas entre apenas duas amostras por vez, permitindo uma visão mais detalhada para cada algoritmo.

Tabela 9 – Teste Ranqueado de Friedman sobre WRACC Observado

Algorithms	**Avg. WRACC**	**Avg. Ranks**
EASD	**0.4811**	**1.5517**
GARSD+	0.3953	2.5172
NMEEF	0.3166	2.8275
SDIGA	0.2337	3.8275
MESDIF	0.1311	4.2758
Friedman Ranked Test p-value		7.8856E−11

Tabela 10 – Teste Wilcoxon Signed para o Wracc Observado

Algorithms	**p-values**
EASD x GARSD	8.35E−04
EASD x MESDIF	7.99E−06
EASD x NMEEF	1.91E−03
EASD x SDIGA	3.46E−05

4.5.2 Algoritmos Evolucionários e Alta Dimensionalidade

Como citado na revisão da literatura, o SSDP+ é, no momento, o algoritmo mais relevante para a área de Subgroup Discovery, no que tange à aplicação de tais técnicas sobre conjuntos de alta dimensionalidade. Nos experimentos propostos por seu autor, a classe majoritária de um conjunto de dados é considerada como a classe alvo ou positiva, enquanto as demais classes são tratadas como negativas. Tal padrão não se repete nos experimentos apresentados nessa dissertação, então cada uma das classes é selecionada em uma vez como alvo, enquanto as demais são tratadas como negativas e, no final, se retira uma média da métrica de otimização; no caso, o *WRACC*. O mesmo processo se segue com o EASD.

Os resultados do EASD foram comparados com o algoritmo SSDP+ sobre 10 conjuntos de dados de alta dimensionalidade, no qual foi observado o *WRACC*, dado que ele foi utilizado como função de otimização para ambas as técnicas, obtendo-se uma média da métrica para todas as classes encontradas nos *datasets*. Utilizando o teste estatístico Wilcoxon, os resultados demonstram diferença estatísticas entre as duas técnicas superior a 99% de confiança, com *p-Value*:$5.06\mathrm{E}{-03}$. A média dos valores obtidos pelo EASD:0.418504294 foi superior à obtida pelo SSDP+: 0.2069, como pode ser atestado na Tabela 16.

Em relação à quantidade média de regras geradas, como o SSDP+ é uma estratégia top-k, com $k = 5$ para todos os experimentos realizados nesse trabalho. Não é uma comparação justa dizer que o EASD é superior nesse ponto, por gerar conjuntos de regras menores e mais concisos. Em relação ao tamanho médio das regras, o EASD tem tamanho médio $= 1.771$, enquanto o SSDP+ tem tamanho médio $= 2.779$.

Tabela 11 – Comparação de WRACC EASD X SSDP+

Datasets	**EASD**	**SSDP+**
Alon	0.47806417	0.2580
burczynski	0.48835167	0.2744
chiaretti	0.25168194	0.1263
chin	0.61861222	0.2584
christensen	0.737535	0.2106
gordon	0.41543167	0.1810
gravier	0.28016833	0.2122
nakayama	0.261671	0.1390
tian	0.25619333	0.2168
yeoh	0.39733361	0.1696

4.5.3 Considerações acerca do RefineAndMine

O RefineAndMine é o único algoritmo da comparação que não utiliza uma abordagem evolucionária, mas uma abordagem exaustiva. O algoritmo é bem limitado quanto aos tipos de dados que podem ser utilizados por ele, então, nesse caso, apenas foram utilizados conjuntos de dados numéricos de baixa dimensionalidade. De um ponto de vista mais geral, o algoritmo é bem diferente dos demais utilizados, ficando a cargo do usuário, por exemplo, definir quais serão os atributos utilizados na geração de regras. Por fim, na utilização, como pode ser visto na Tabela 15, em que seu *WRACC* é comparado com o do EASD, a técnica tem o típico resultado de algoritmos de busca exaustiva, oferecendo resultados ótimos em espaços de busca pequenos. Porém, mesmo oferecendo a opção de se parar a busca a qualquer momento, acaba não obtendo resultado algum, mesmo para alguns conjuntos de dados numéricos com poucas dezenas de atributos.

Acreditamos que pela pequena quantidade de valores disponíveis para comparação, o teste estatístico de *Wilcoxon* não rejeitou a hipótese nula com *p-value*: 0.2026 de que não existe diferença entre os valores obtidos pelo RefineAndMine e o EASD, mesmo o EASD possuindo, em média, um WRACC:0.5072 bem superior ao RefineAndMine:0.2706, quando comparados sobre conjuntos de dados numéricos de baixa dimensionalidade.

Tabela 12 – Comparação de WRACC EASD X RefineAndMine

Datasets	**EASD**	**RefineAndMine**
Vehicle	0.3843	0.0
Ionosphere	0.4784	0.0
Diabetes	0.3899	0.3956
Wine	0.7601	0.0
Iris	0.8367	0.8503
Appendicitis	0.3875	0.4587
Glass	0.3096	0.3459
BreastCancer	0.7611	0.0
Ecoli	0.2860	0.3187
BreastTissue	0.4785	0.3372

4.6 ANÁLISES DOS RESULTADOS

Mesmo tendo seu foco voltado para conjuntos de dados numéricos de alta dimensionalidade, os resultados apresentados na Tabela 8 demonstram que o EASD também foi capaz de atuar em conjuntos de dados de baixa dimensionalidade. Não apenas numéricos, mas também conjuntos categóricos e com ambos, categóricos e numéricos, ao mesmo tempo. Tamanha versatilidade se dá pela excelente representação dos padrões extraídos de diferentes tipos de dados, aplicada pelo EASD. Com taxas de Recall e Acurácia, pelo menos 10% superiores aos algoritmos com a segunda melhor avaliação para essas medidas, o EASD demonstra sua capacidade de cobertura de padrões superior aos demais, ao mesmo tempo que consegue manter o maior nível de acurácia dentre as demais observações. Diferentemente do caso do algoritmo SDIGA, por exemplo, que, ao possuir um alto nível de suporte, acaba por ter uma baixa confiança e consequentemente, tendo sua acurácia prejudicada.

Um outro ponto é a quantidade e tamanho das regras geradas pelos algoritmos, tendo sua importância relacionada a necessidade de interpretabilidade das regras geradas. Então, quanto menores e menos regras, melhor para os resultados. O EASD fica atrás apenas do GARSD+ quando o quesito é quantidade de regras geradas, porém obtém o melhor valor dentre os algoritmos testados em relação ao tamanho das regras. Tendo seu tamanho, em média, pelo menos 45% inferior ao segundo colocado, demonstra, assim, mais um ponto positivo do algoritmo.

Os algoritmos utilizados no estudo experimental dessa dissertação são algumas das técnicas mais relevantes no que tange às abordagens evolucionárias para a área de descoberta de subgrupos. Porém, apresentam uma grande deficiência quanto à sua atuação sobre conjuntos de dados de alta dimensionalidade, com exceção do SSDP+. O EASD realiza um controle sobre a inicialização dos indivíduos de suas populações, limitando o tamanho máximo do tamanho das regras geradas, com intuito de mitigar a dificuldade de se atuar sobre conjuntos de alta dimensionalidade. Tal estratégia se demonstrou bem sucedida à medida em que o algoritmo se demonstrou capaz de atuar sobre conjuntos de dados de alta dimensionalidade, também sendo capaz de obter em média uma acurácia 50% superior em comparação aos resultados obtidos ao ser comparado com o SSDP+. Este é, até o momento, uma referência para a área de mineração de padrões discriminativos em conjuntos de dados de alta dimensionalidade. Muito provavelmente, a disparidade dos resultados entre as duas técnicas se dá devido à maneira como o EASD não necessita de discretização prévia de dados numéricos, os representando de forma mais adequada e, consequentemente, permitindo a descoberta de regras com melhor acurácia.

O algoritmo proposto possui um número razoável de parâmetros e suas interações foram testadas durante o *Grid Search*. A melhor combinação foi utilizada como configuração para os experimentos realizados. Como comentado na seção anterior, alguns parâmetros foram mantidos inalterados, dada sua importância para o processo evolucionário. O tamanho máximo de indivíduo foi mantido como 4, pois o menor conjunto de dados do trabalho possui apenas quatro atributos. Variar esse parâmetro poderia causar problemas em alguns conjuntos de dados. O limiar de punição por amplitude de cobertura foi mantido a 50% tanto para dados numéricos como para categóricos, funcionando como mais um fator para a geração de regras menores e mais gerais. Por fim, os últimos dois parâmetros que não foram alterados são os de porcentagem geracional máxima e o número máximo de reinicializações, que fazem parte da estratégia de reinicialização populacional e finalização do processo no momento de convergência na população.

4.7 EVOLUÇÃO POPULACIONAL

Algoritmos evolucionários são conhecidos por sua boa capacidade de busca e otimização de funções e fácil adaptação a diferentes domínios de atuação. No caso do EASD, a analogia de uma população de indivíduos representa, na realidade, um conjunto de regras geradas aleatoriamente, com base em restrições e padrões. Para avaliar a qualidade dessas regras, utilizamos medidas. No caso desse estudo, a principal medida avaliada é o *WRACC*. Sendo usada como função a ser otimizada ou função de aptidão do EASD.

A maneira pela qual se avalia a evolução de uma dada população durante um processo evolucionário é por meio da verificação do fitness ou função de adaptação dessa população. Por meio de elementos gráficos, como, por exemplo, a Figura 2, é possível se ter uma ideia mais clara. Avaliações semelhantes à da figura serviram como inspiração para as estratégias

de controle de convergência e reinicializações do EASD.

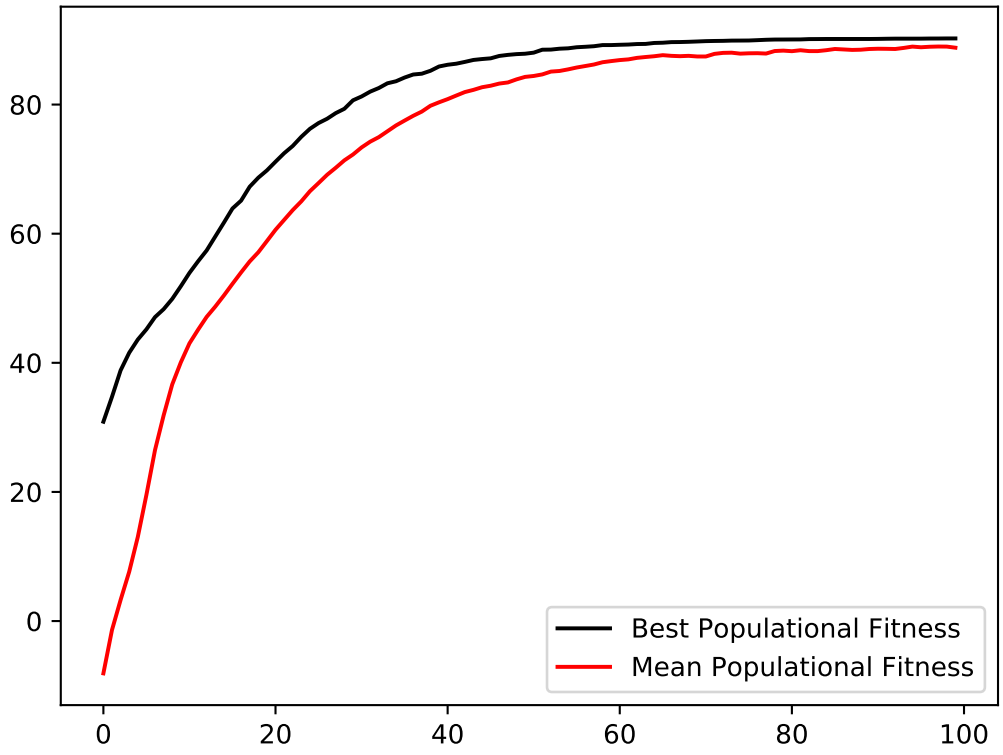

Figura 2 – Breast Cancer - Média dos Melhores Valores GARSD+

O primeiro ponto a ser observado, ainda levando em conta a figura demonstrada, é que a partir 60% do número máximo de iterações, basicamente não houve melhora nos valores de fitness dessa população. Esse tipo de situação é problemática porque esse tipo de comportamento pode demonstrar que a busca ficou presa em um mínimo local e não consegue escapar. Ou, simplesmente, o número de iterações é muito superior ao que a técnica realmente precisa para convergir, levando apenas a desperdício de recursos computacionais.

A solução aplicada no EASD é uma ideia simples. Primeiro, se mantêm registros de fitness da melhor solução da população a cada 10% da quantidade de iterações ou gerações definidas. Após isso, são criados alertas para que depois de 10% das gerações atingidas, se a melhor solução não tem nenhum avanço, se aplica uma reinicialização populacional, trazendo novos indivíduos para explorar mais o espaço e reforçando a ideia de fuga de ótimos locais.

A estratégia de reinicializações é muito interessante, todavia não resolve, por si só, o problema de se ter demasiadas iterações, se tornando inefetiva quando uma dada população pode já ter convergido para seus ótimos. Então, para mitigar essa situação, usamos dois parâmetros, a quantidade máxima de reinicializações e a porcentagem máxima de gerações a serem cumpridas. O conceito é simples de entender, caso já se tenha reiniciado a população

a quantidade máxima. Então no momento que se chega à porcentagem máxima de geração e se atinge novamente o marcador 10% sem melhoras, se considera que a população convergiu e o processo de busca é finalizado. No caso dos experimentos performados, 80% é o máximo da população que se segue sem encerrar o processo, até que a quantidade de reinicializações chegue a 8, que foi o valor utilizado.

É muito importante ter cautela com esses valores de parâmetros, podemos tomar como exemplo que o parâmetro de Checagem de Reinicialização tenha sido configurado para um valor muito baixo, isso poderia levar a muitas reinicializações e, no final do processo, encontrar apenas uma população sub ótima.

Apesar do uso de diversas estratégias para uma diversificação na população por meio de seus operadores genéticos, o EASD ainda se trata de uma abordagem elitista, tendo, assim, a premissa de que a solução com melhor valor nunca é perdida, podendo ser substituída apenas em caso de melhoria de fitness. As Figuras 4 e 5 demonstram a evolução média do fitness em dois conjuntos de dados numéricos de alta dimensionalidade, acompanhando o crescimento do valor da melhor solução para cada geração, tal como o valor para a média do fitness da população. Nelas as linhas pretas representam seus valores médios máximos; enquanto as linhas tracejadas vermelhas, os valores médios de suas populações, evidenciando por meio de sua evolução, que as estratégias aplicadas pelo EASD funcionaram como se esperava para os conjuntos de dados numéricos de alta dimensionalidade.

Ambos os gráficos demonstram uma evidente evolução do Fitness (Eixo Y) em relação as gerações (Eixo X). A maneira como não ocorrem gerações demasiadas sem a melhoria da população ótima demonstra a melhoria no funcionamento quando se compara a Figura 3, que são os resultados do EASD para o conjunto de dados *Breast Cancer*, com a Figura 2, que são as avaliações de fitness realizados pelo GARSD+ para o mesmo *dataset*.

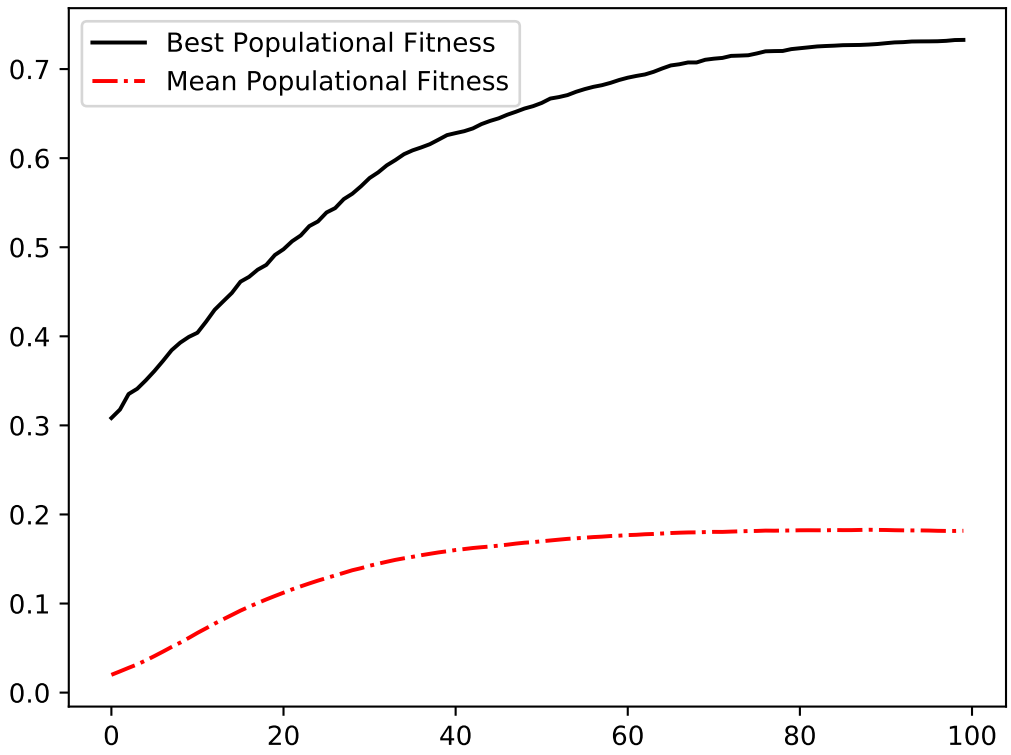

Figura 3 – Breast Cancer - Evolução Média EASD

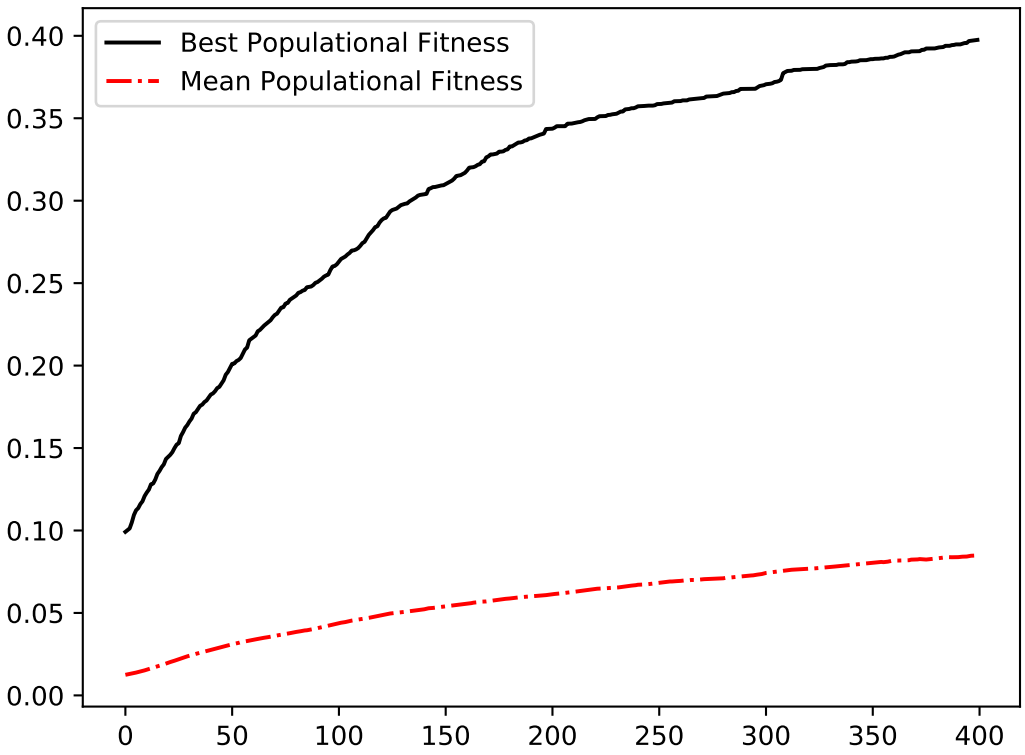

Figura 4 – Gordon - Evolução Média EASD

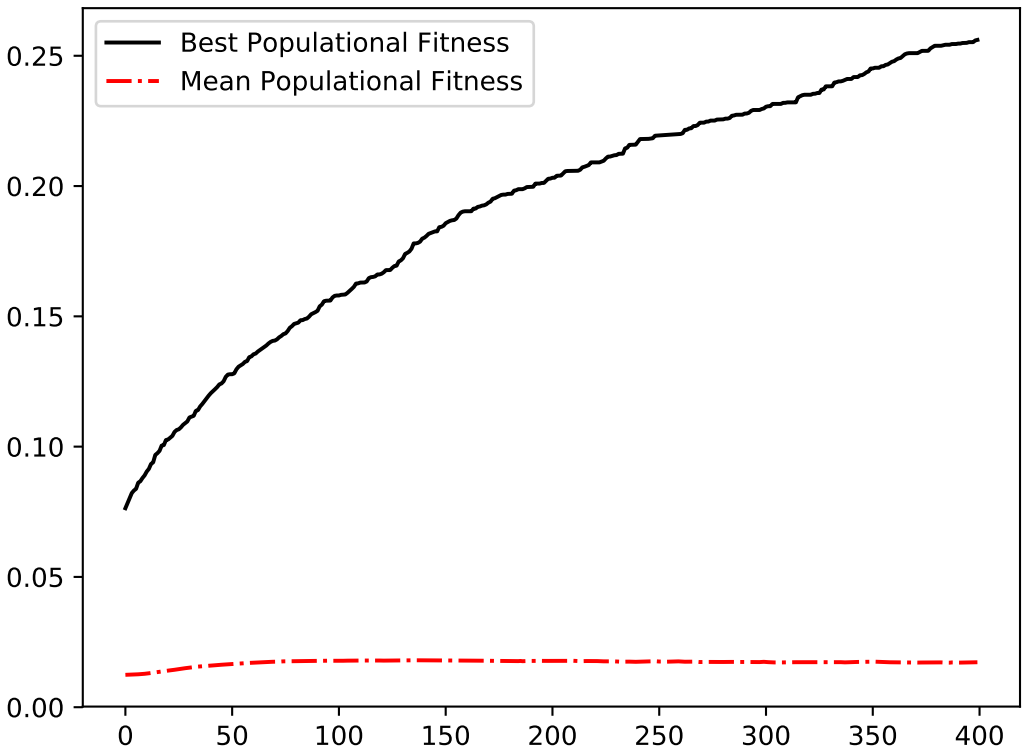

Figura 5 – Gravier - Evolução Média EASD

5 CONCLUSÃO

5.1 CONCLUSÕES

A pesquisa apresentada nesse trabalho propôs uma nova técnica para a área de Subgroup Discovery, capaz de atuar sobre conjuntos de dados numéricos de alta dimensionalidade.

Após uma revisão da literatura da área de mineração de dados, mais especificamente da área de Subgroup Discovery, pudemos atestar a falta de técnicas capazes de lidar com conjuntos de dados numéricos de maneira adequada. O problema apenas se tornava maior quando se pensava em conjuntos de dados de alta dimensionalidade. Estes são, por sua vez, um desafio cada vez mais presente na sociedade atual, devido ao crescimento da complexidade de dados gerados.

Com a apresentação da literatura e ideias que inspiraram o trabalho, foram demonstrados de maneira cuidadosa, cada um dos elementos principais da técnica proposta, fazendo uso de exemplos e analogias. Assim, concluindo com um passo a passo de um pseudocódigo, que pode ser usado como guia do processo seguido pelo algoritmo.

Com todas as características do algoritmo devidamente explanadas, apresentamos um longo estudo experimental, no qual foram realizadas implementações e adequações de algoritmos do estado da arte da área, permitindo uma comparação justa com a técnica proposta. Além disso, foi seguido por um estudo exploratório acerca de interações e correlações entre os parâmetros da técnica proposta, permitindo, assim, se encontrar uma configuração ótima para seus parâmetros. Também foram descritos os 39 conjuntos de dados utilizados nos experimentos, assim como sua importância devido a suas variações de tipos de dados, dimensão e quantidade de registros. Por fim, foram apresentadas as métricas advindas dos experimentos e testes estatísticos, que reforçam as hipóteses levantadas durante o estudo experimental.

Também foram destrinchados detalhes acerca dos resultados obtidos, reforçando argumentos com o auxílio de gráficos, contendo observações coletadas durante o estudo experimental.

A contribuição do trabalho proposto para a área de Subgroup Discovery pode ser verificada, uma vez que o EASD é o único algoritmo até o corrente momento capaz de lidar com conjuntos de dados numéricos de alta dimensionalidade. Ele se demonstra competitivo e superior quando comparado com o estado da arte e técnicas recém publicadas na área.

5.2 PONTOS DE MELHORIAS E TRABALHOS FUTUROS

A pesquisa e técnica aqui proposta ainda deixa em aberto diversos desafios. Não apenas para a área de Subgroup Discovery, mas para a área de Mineração de Dados de maneira mais geral.

A implementação atual do EASD não faz uso de paralelismo. Logo, incorrendo em dificuldades para atuar sobre conjuntos de dados com uma quantidade muito grande de registros, tornando o algoritmo lento e inviável para aplicações de Big Data. Uma solução para esse pro-

blema seria uma adaptação, utilizando tecnologias como: *Cython*, *Dask* e *Spark*, que tem, em seu foco, a solução de problemas de computação concorrente e Big data. Existe dependência na ordem de execução das iterações do EASD, por esse motivo as paralelizações devem ser realizadas nas avaliações de fitness e aplicação de alguns dos operadores genéticos.

Por mais que as estratégias de reinicialização e os operadores genéticos sejam uma tentativa de se atingir uma boa diversidade quanto as regras obtidas, em alguns casos, o algoritmo pode gerar regras utilizando o mesmo atributo dado a suas características discriminativas, variando apenas intervalos que não se sobrepõem, para atingir o critério de suporte mínimo. Este fato não é interessante para a diversidade. Então, um tratamento desse problema poderia ser uma melhoria relevante, seja por meio de uma abordagem top-k ou até mesmo com um simples pós-processamento, removendo regras muito semelhantes após o processo evolucionário.

O algoritmo se demonstrou muito eficiente em encontrar atributos com potencial discriminativo alto em conjuntos de dados. Então uma possível melhoria poderia ser proposta por meio da aplicação de técnicas exaustivas sobre atributos com potencial alto. Atingindo, assim, os melhores valores possíveis para um dado conjunto de regras.

A utilização de meta-heurísticas para ajustar os parâmetros da técnica de acordo com os dados apresentados também pode ser algo a se pensar. Além de se utilizar taxas variáveis para os operadores de crossover, mutação e taxa da população a ser iniciada, os define de acordo com o tempo disponível de exploração e as fases de exploração e refinamento de soluções.

REFERÊNCIAS

BAY, S. D.; PAZZANI, M. J. Detecting group differences: Mining contrast sets. *Data mining and knowledge discovery*, Springer, v. 5, n. 3, p. 213–246, 2001.

BELFODIL, A.; BELFODIL, A.; KAYTOUE, M. Anytime subgroup discovery in numerical domains with guarantees. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2018. p. 500–516.

CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J. del; HERRERA, F. Nmeef-sd: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 18, n. 5, p. 958–970, 2010.

CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J. del; HERRERA, F. Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 4, n. 2, p. 87–103, 2014.

CARMONA, C. J.; JESUS, M. del; HERRERA, F. A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy. *Knowledge-Based Systems*, Elsevier, v. 139, p. 89–100, 2018.

CHEN, P.-H.; ZAFAR, H.; GALPERIN-AIZENBERG, M.; COOK, T. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *Journal of digital imaging*, Springer, p. 1–7, 2018.

DONG, G.; LI, J. Efficient mining of emerging patterns: Discovering trends and differences. In: CITESEER. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 1999. p. 43–52.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. et al. Knowledge discovery and data mining: Towards a unifying framework. In: *KDD*. [S.l.: s.n.], 1996. v. 96, p. 82–88.

FERNÁNDEZ, A.; GARCÍA, S.; LUENGO, J.; BERNADÓ-MANSILLA, E.; HERRERA, F. Genetics-based machine learning for rule induction: state of the art, taxonomy, and comparative study. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 14, n. 6, p. 913–941, 2010.

GAMBERGER, D.; LAVRAČ, N.; ŽELEZNỲ, F.; TOLAR, J. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of biomedical informatics*, Elsevier, v. 37, n. 4, p. 269–284, 2004.

GARCÍA-VICO, A.; CARMONA, C. J.; MARTÍN, D.; GARCÍA-BORROTO, M.; JESUS, M. del. An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 1, p. e1231, 2018.

GROSSKREUTZ, H.; RÜPING, S. On subgroup discovery in numerical domains. *Data mining and knowledge discovery*, Springer, v. 19, n. 2, p. 210–226, 2009.

HELAL, S. Subgroup discovery algorithms: A survey and empirical evaluation. *Journal of Computer Science and Technology*, Springer, v. 31, n. 3, p. 561–576, 2016.

HERRERA, F.; CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J. D. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, Springer, v. 29, n. 3, p. 495–525, 2011.

JESUS, M. J. D.; GONZÁLEZ, P.; HERRERA, F.; MESONERO, M. Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 15, n. 4, p. 578–592, 2007.

JESUS, M. J. del; GONZÁLEZ, P.; HERRERA, F. Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In: IEEE. *2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*. [S.l.], 2007. p. 50–57.

KLÖSGEN, W. Explora: A multipattern and multistrategy discovery assistant. In: AMERICAN ASSOCIATION FOR ARTIFICIAL INTELLIGENCE. *Advances in knowledge discovery and data mining*. [S.l.], 1996. p. 249–271.

LAVRAČ, N. Subgroup discovery techniques and applications. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2005. p. 2–14.

LUCAS, T.; SILVA, T. C.; VIMIEIRO, R.; LUDERMIR, T. B. A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data. *Applied Soft Computing*, Elsevier, v. 59, p. 487–499, 2017.

LUCAS, T.; VIMIEIRO, R.; LUDERMIR, T. Ssdp+: A diverse and more informative subgroup discovery approach for high dimensional data. In: IEEE. *2018 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.], 2018. p. 1–8.

LUCENA, A. *Resultados Experimentos*. 2019. Disponível em: <https://github.com/AvynerLucena/MSCC-Experimentos>.

MANNILA, H. Local and global methods in data mining: Basic techniques and open problems. In: SPRINGER. *International Colloquium on Automata, Languages, and Programming*. [S.l.], 2002. p. 57–68.

NOVAK, P. K.; LAVRAČ, N.; WEBB, G. I. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, v. 10, n. Feb, p. 377–403, 2009.

PACHÓN, V.; MATA, J.; DOMÍNGUEZ, J. L. Searching for the most significant rules: an evolutionary approach for subgroup discovery. *Soft Computing*, Springer, v. 21, n. 10, p. 2609–2618, 2017.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining: pearson new international edition*. [S.l.]: Pearson Higher Ed, 2013.

TORREÃO, V. de A.; VIMIEIRO, R. Effects of population initialization on evolutionary techniques for subgroup discovery in high dimensional datasets. In: IEEE. *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2018. p. 25–30.

VIMIEIRO, R. et al. Mining disjunctive patterns in biomedical data sets. 2012.

WANG, J. T.; ZAKI, M. J.; TOIVONEN, H. T.; SHASHA, D. Introduction to data mining in bioinformatics. In: *Data Mining in Bioinformatics*. [S.l.]: Springer, 2005. p. 3–8.

WROBEL, S. An algorithm for multi-relational discovery of subgroups. In: SPRINGER. *European Symposium on Principles of Data Mining and Knowledge Discovery*. [S.l.], 1997. p. 78–87.

WROBEL, S. Inductive logic programming for knowledge discovery in databases. In: *Relational data mining*. [S.l.]: Springer, 2001. p. 74–101.

APÊNDICE A – TABELAS

Tabela 13 – Métricas Replicação dos Experimentos do GARSD+

Datasets	**Recall**	**Conf**	WRACC	Recall-std	Conf-std	WRACC-std
Vehicle	0.5826	0.5517	0.0755	2.5509e-02	5.0023e-02	4.8220e-03
Ionosphere	0.30013072	0.82675538	0.05202379	0.08810771	0.10543304	0.01931273
Wine	0.70731922	0.84346749	0.13973765	0.06018181	0.04323944	0.01099273
Iris	0.85779915	0.93167551	0.18218519	0.02386029	0.00880523	0.00557945
German	0.53629156	0.62313496	0.05057521	0.11930824	0.04325141	0.02210538
Hypothyroid	0.5463	0.7843	0.03501	1.3638e-01	0.03191917	9.4975e-03
Diabetes	0.52169578	0.70251242	0.0666823	0.08913823	0.0452942	0.01178442

Tabela 14 – Resultados Experimentos RefineAndMine

Datasets	**Recall**	**Confidence**	**Wracc**	**N-Regras**	**Tamanho-Regras**
Vehicle	0.00	0.0	0.0	0.0	0.0
Ionosphere	0.00	0.0	0.0	0.0	0.0
Diabetes	0.7368	0.7065	0.3956	1.0	8.0
Wine	0.00	0.0	0.0	0.0	0.0
Iris	0.9866	0.9444	0.8503	1.0	4.0
Appendicitis	0.9232	0.7632	0.4587	1.0	7.0
Glass	0.8936	0.7466	0.3459	1.0	6.0
BreastCancer	0.00	0.0	0.0	0.0	0.0
Ecoli	0.9745	0.7398	0.3187	1.0	6.0
BreastTissue	0.6805	0.7424	0.3372	1.0	9.0

Tabela 15 – Comparação de WRACC EASD X RefineAndMine

Datasets	**EASD**	**RefineAndMine**
Vehicle	0.3843	0.0
Ionosphere	0.4784	0.0
Diabetes	0.3899	0.3956
Wine	0.7601	0.0
Iris	0.8367	0.8503
Appendicitis	0.3875	0.4587
Glass	0.3096	0.3459
BreastCancer	0.7611	0.0
Ecoli	0.2860	0.3187
BreastTissue	0.4785	0.3372

Tabela 16 – Comparação de WRACC EASD X SSDP+

Datasets	**EASD**	**SSDP+**
Alon	0.47806417	0.2580
burczynski	0.48835167	0.2744
chiaretti	0.25168194	0.1263
chin	0.61861222	0.2584
christensen	0.737535	0.2106
gordon	0.41543167	0.1810
gravier	0.28016833	0.2122
nakayama	0.261671	0.1390
tian	0.25619333	0.2168
yeoh	0.39733361	0.1696

Tabela 17 – Comparação de WRACC Experimentos de Baixa Dimensionalidade

Datasets	**EASD**	**GARSD**	**MESDIF**	**N-NMEEF**	**SDIGA**
solarflare	0.3744	0.3540	0.1702	0.5373	0.1445
bridgesVersion2	0.3625	0.2666	0.0937	0.0824	0.1971
balancescale	0.3174	0.2502	0.0958	0.1498	0.1717
housevotes84	0.8645	0.7914	0.2570	0.9378	0.5641
lymphography	0.3567	0.2996	0.0325	0.0	0.1111
Spect	0.1676	0.1291	0.0810	0.0062	0.0042
TicTacToe	0.2941	0.3028	0.1315	0.2604	0.1209
Monk2	0.7685	0.6697	0.0058	0.0785	0.0032
hayes$_r oth$	0.2998	0.2800	0.1250	0.2852	0.2813
primarytumor	0.1078	0.0880	0.0277	0.03145	0.0194
German	0.4143	0.3489	0.0480	0.1504	0.0293
AcuteInflamations	0.3184	0.2386	0.5782	0.6865	0.6291
CreditApproval	0.5243	0.4895	0.1328	0.5301	0.0042
Hypothyroid	0.8667	0.8045	0.0333	0.0678	0.0143
KidneyDisease	0.7326	0.5814	0.0513	0.3109	0.35
saheart	0.2084	0.0408	0.0623	0.2149	0.1834
StatlogHeart	0.9008	0.7525	0.2	0.4592	0.3637
Hepatite	0.2849	0.2591	0.0	0.3055	0.2631
australianCrx	0.7153	0.6090	0.1423	0.4566	0.7117
Vehicle	0.3843	0.3289	0.0665	0.0	0.0161
Ionosphere	0.4784	0.2833	0.0	0.4351	0.2726
Diabetes	0.3899	0.6270	0.0276	0.2735	0.1909
Wine	0.7601	0.8124	0.3566	0.7098	0.3209
Iris	0.8367	0.2616	0.4148	0.7822	0.6518
Appendicitis	0.3875	0.5409	0.2258	0.4297	0.4958
Glass	0.3096	0.2616	0.0751	0.0	0.0413
BreastCancer	0.7611	0.2204	0.1713	0.8776	0.6808
Ecoli	0.2860	0.3025	0.0658	0.0839	0.0663
BreastTissue	0.4785	0.2682	0.1308	0.2093	0.1039

Tabela 18 – Resultados Experimentos GARSD+ Dados Categoricos

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
solarflare	0.8449	0.5485	0.3540	6.9	2.5833	4.28953730e-02	2.23836884e-02	4.15863609e-03
bridgesVersion2	0.7987	0.5883	0.2666	11.0333	3.1285	0.05581154	0.05271437	0.0029381
balancescale	0.4740	0.5155	0.2502	8.5333	3.3666	2.36562727e-02	2.14788547e-02	3.59959363e-03
housevotes84	0.9062	0.9205	0.7914	2.0	1.9333	6.02588026e-02	3.67288517e-02	1.91820459e-02
lymphography	0.8225	0.7169	0.2996	4.7	2.6250	0.05585477	0.13304129	0.00573416
Spect	0.5576	0.6443	0.1291	4.4666	4.35	0.06519806	0.0386054	0.00271575
TicTacToe	0.6226	0.6823	0.3028	2.1666	1.7166	5.27192006e-02	2.42583715e-02	2.42013690e-03
Monk2	0.7804	0.9075	0.6697	2.0	1.4	2.60138182e-02	6.71799868e-02	2.24177462e-02
hayes$_r$*oth*	0.4177	0.7940	0.2800	9.6	3.8777	0.02455227	0.02661773	0.00185682
primarytumor	0.7960	0.2720	0.0880	32.4333	5.0968	3.07711974e-02	3.67957808e-02	7.87759030e-04

Tabela 19 – Resultados Experimentos GARSD+ Dados Mistos

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
hepatitis	0.7340	0.7568	0.3489	2.3333	1.9833	0.0922	0.0743	0.0127
German	0.5870	0.6398	0.2386	3.6	2.7333	7.56963203e-02	2.36383031e-02	1.22838515e-02
StatlogHeart	0.6879	0.7891	0.4895	2.2	1.9333	0.0670	0.0492	0.0084
AcuteInflamations	0.8424	0.9860	0.8045	2.0	1.6166	0.0614	0.0366	0.0139
CreditApproval	0.7711	0.8066	0.5814	2.6666	2.3166	8.07791650e-02	4.88995649e-02	2.57404996e-02
Hypothyroid	0.6213	0.8018	0.0408	5.9666	4.6166	1.81452612e-01	5.96296392e-02	1.02121212e-02
KidneyDisease	0.8513	0.9474	0.7525	2.2666	2.0333	0.0672	0.0376	0.02050
saheart	0.5561	0.6736	0.2591	2.1	2.0666	6.82366649e-02	2.80511196e-02	5.65278320e-03
australianCrx	0.8017	0.8121	0.6090	2.2	1.95	7.44184778e-02	5.13957111e-02	2.66894235e-02

Tabela 20 – Resultados Experimentos GARSD Dados Numéricos

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
Vehicle	0.6639	0.5226	0.3289	5.5666	2.2250	5.227968e-02	4.973299e-02	6.335447e-03
Ionosphere	0.3837	0.8803	0.2833	6.7333	3.6333	0.0957	0.0519	0.0155
Wine	0.8243	0.8097	0.6270	3.4333	1.5	0.0567	0.0580	0.0157
Iris	0.9382	0.9531	0.8124	3.0	1.0	0.0376	0.0182	0.0071
Glass	0.7214	0.6178	0.2616	8.0666	2.3055	4.249496e-02	6.034973e-02	3.568814e-03
BreastCancer	0.7061	0.8439	0.5409	2.8333	1.75	1.166983e-01	7.152832e-02	2.865274e-02
Ecoli	0.8433	0.5610	0.2616	8.1	1.6458	3.835334e-02	8.320135e-02	3.080479e-03
BreastTissue	0.4522	0.6590	0.2204	12.2666	2.0833	0.0473	0.0584	0.0054
Diabetes	0.5852	0.6980	0.3025	2.86666	1.4666	8.326919e-02	4.483682e-02	1.045753e-02
Appendicitis	0.4885	0.8786	0.2682	4.96666	2.55	0.0826	0.0392	0.0090

Tabela 21 – Resultados Médios Experimentos EASD Alta Dimensionalidade

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
gravier	0.5682	0.6965	0.2801	2.83	1.04	0.0794	0.0675	0.0735
nakayama	0.9065	0.7692	0.2616	10.03	2.11	0.0935	0.0666	0.0522
tian	0.6156	0.6878	0.2561	2.5	1.51	0.0473	0.0448	0.0367
yeoh	0.8688	0.6282	0.3973	6.13	1.9	0.0245	0.0363	0.0217
gordon	0.8402	0.8171	0.4154	2.03	1.47	0.0253	0.0126	0.0253
chiaretti	0.9080	0.8496	0.2516	6.1	2.52	0.0293	0.0389	0.0122
christensen	0.9841	0.9899	0.7375	3.03	1.98	0.0392	0.0617	0.0108
chin	0.7873	0.8709	0.6186	2.07	2.17	0.0842	0.0534	0.0198
alon	0.6595	0.8411	0.4780	2.3	1.36	0.0484	0.0641	0.0260
burczynski	0.7497	0.6895	0.4883	3.03	1.65	0.0912	0.0545	0.0326

Tabela 22 – Resultados Experimentos EASD Dados Numéricos

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
Vehicle	0.7589	0.5253	0.3843	4.1	1.99	0.0420	0.0324	0.0230
Ionosphere	0.6716	0.7985	0.4784	2.0	1.03	0.0659	0.0267	0.0174
Diabetes	0.7159	0.6995	0.3899	2.0	1.0	0.0534	0.0341	0.0435
Wine	0.9296	0.8857	0.7601	3.03	1.4	0.0140	0.0250	0.0163
Iris	0.9797	0.9313	0.8367	3.0	1.18	0.0344	0.0432	0.0129
Appendicitis	0.7186	0.8067	0.3875	2.03	1.02	0.0491	0.0319	0.0357
Glass	0.8301	0.5682	0.3096	6.07	1.73	0.0158	0.0237	0.0143
BreastCancer	0.8850	0.9193	0.7611	2.03	1.22	0.0578	0.0381	0.0234
Ecoli	0.9450	0.5572	0.2860	8.03	1.73	0.0583	0.0403	0.0319
BreastTissue	0.9744	0.6822	0.4785	6.0	1.88	0.0502	0.0287	0.0188

Tabela 23 – Resultados Experimentos EASD Dados Mistos

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
hepatitis	0.7660	0.9281	0.4143	2.23	2.11	0.0636	2.7510	0.0446
German	0.6269	0.6837	0.3184	2.0	2.0	1.110223e-16	2.842170e-14	2.775557e-16
StatlogHeart	0.7491	0.77	0.5243	2.0	1.5	3.330669e-16	0.000000e+00	1.110223e-16
AcuteInflamations	0.8938	0.9970	0.8667	2.0	1.62	0.0157	1.9672	0.0177
CreditApproval	0.8701	0.8675	0.7326	2.0	1.0	1.110223e-16	2.842170e-13	2.220446e-16
Hypothyroid	0.9585	0.7922	0.2084	2.0	1.33	1.518369e-03	1.295084e+00	1.099748e-03
KidneyDisease	0.9448	0.9877	0.9008	2.03	2.04	0.0089	5.5787	0.0343
saheart	0.6227	0.6647	0.2849	2.67	1.01	0.0179	2.7547	0.0097
australianCrx	0.8620	0.8584	0.7153	2.0	1.0	3.330669e-16	1.705302e-13	5.551115e-16

Tabela 24 – Resultados Experimentos EASD Dados Categoricos

Datasets	**Recall**	**Conf**	**Wracc**	**N-Regras**	**Size-Regras**	**Recall-std**	**Conf-std**	**WRACC-std**
solarflare	0.8595	0.5779	0.3744	6.0	3.01	0.0042	0.0032	0.0004
bridgesVersion2	0.9837	0.7551	0.3625	7.0	2.99	4.440892e-16	2.991619e-02	6.880866e-03
balancescale	0.5003	0.5652	0.3174	5.43	2.0	3.330669e-16	3.330669e-16	0
housevotes84	0.9439	0.9564	0.8645	2.0	1.0	4.440892e-16	4.440892e-16	2.220446e-16
lymphography	0.8519	0.9568	0.3567	4.0	3.42	7.670397e-04	4.988876e-05	7.608036e-04
Spect	0.7198	0.6446	0.1676	2.0	3.18	1.528288e-02	1.985896e-02	5.133116e-04
TicTacToe	0.5815	0.6819	0.2941	2.0	1.0	1.110223e-16	0.000000e+00	5.551115e-17
Monk2	0.7972	0.9583	0.7685	2.0	1.5	5.551115e-16	4.440892e-16	3.330669e-16
hayes$_r oth$	0.5647	0.7062	0.2998	4.33	1.47	3.499111e-16	5.551115e-16	5.551115e-17
primarytumor	0.9152	0.4015	0.1078	21.0	5.65	2.304830e-03	2.045107e-02	8.576262e-05