UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Programa de Pós-Graduação em Ciência da Computação

Vitor Hugo Monteiro Privatto

Classificação Associativa em Contextos Desbalanceados: Aspectos de Extração e Ranqueamento de Regras

Rio Claro – SP

2025

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Vitor Hugo Monteiro Privatto

Classificação Associativa em Contextos Desbalanceados: Aspectos de Extração e Ranqueamento de Regras

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Orientador: Profª. Drª. Veronica Oliveira de Carvalho

Rio Claro – SP
2025

P961c

Privatto, Vitor Hugo Monteiro
Classificação associativa em contextos desbalanceados: aspectos de extração e ranqueamento de regras / Vitor Hugo Monteiro Privatto. -- Rio Claro, 2025
76 p.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientadora: Veronica Oliveira de Carvalho

1. Classificação associativa. 2. Conjunto de dados desbalanceados. 3. Extração de regras de associação. 4. Ranquamentos de regras de associação. 5. Agrupamentos de medidas objetivas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Impacto potencial desta pesquisa

Esta dissertação de mestrado propõe um método de seleção dinâmica de medidas objetivas, denominado DyOMS, que possa ser incorporado a fluxos de indução de classificadores associativos em contextos desbalanceados. O aprimoramento de tais algoritmos é essencial em função de suas aplicações nas mais diversas áreas, como saúde e defeito de software. O método proposto torna os modelos mais precisos e interpretáveis, promovendo inovação e transparência em inteligência artificial.

Potential impact of this research

This master's thesis proposes a method for dynamic selection of objective measures, named DyOMS, which can be incorporated into induction flows of associative classifiers in imbalanced contexts. The improvement of such algorithms is essential due to their applications in the most diverse areas, such as health and software defects. The proposed method makes models more accurate and interpretable, promoting innovation and transparency in artificial intelligence.

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Vitor Hugo Monteiro Privatto

Classificação Associativa em Contextos Desbalanceados: Aspectos de Extração e Ranqueamento de Regras

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Comissão Examinadora

- Profª. Drª. Veronica Oliveira de Carvalho (Orientadora)
  Departamento de Estatística, Matemática Aplicada e Computação (DEMAC)
  Universidade Estadual Paulista "Júlio De Mesquita Filho" (UNESP) – Câmpus de Rio Claro
- Profª. Drª. Solange Oliveira Rezende
  Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC)
  Universidade de São Paulo (USP) – Câmpus de São Carlos
- Prof. Dr. Frank José Affonso
  Departamento de Estatística, Matemática Aplicada e Computação (DEMAC)
  Universidade Estadual Paulista "Júlio De Mesquita Filho" (UNESP) – Câmpus de Rio Claro

Conceito: Aprovado

Rio Claro (SP), 24 de fevereiro de 2025

AGRADECIMENTOS

Agradeço imensamente todas as pessoas que, direta ou indiretamente, contribuíram para a realização desta dissertação. Sem vocês nada disso seria possível, muito obrigado!

RESUMO

Em diversos problemas é interessante o uso de algoritmos inerentemente interpretáveis, uma vez que facilita o entendimento do conhecimento obtido e das predições realizadas. Dentre estes algoritmos encontram-se os classificadores associativos. Estes mesclam as tarefas de associação e classificação e são, portanto, induzidos em etapas, a saber: [a] extração de um conjunto de regras, [b] ranqueamento das regras via medidas objetivas e [c] poda das regras. Embora a classificação associativa, assim como outras técnicas, apresente bons resultados, quando aplicada a problemas desbalanceados o desempenho não se mantém o mesmo. O desbalanceamento ocorre quando o número de instâncias de uma dada classe, chamada de majoritária, supera em muito o número de instâncias da outra classe, chamada de minoritária. Deste modo, soluções vem sendo desenvolvidas de modo a diferenciar corretamente as instâncias de ambas as classes. Assim, este trabalho explora o uso de classificadores associativos quando aplicados em dados desbalanceados via abordagens internas, i.e., em nível de algoritmo. Para tanto, três objetivos são propostos. O primeiro se refere a execução de uma revisão sistemática da literatura a fim de identificar as abordagens internas que vêm sendo adotadas e/ou propostas a fim de fundamentar este trabalho, assim como identificar lacunas e oportunidades na área. Tendo como base as lacunas identificadas, o segundo objetivo explora o impacto das diferentes estratégias levantadas na revisão para se realizar a extração de regras (etapa [a]) visando identificar a mais adequada a ser utilizada no contexto aqui apresentado. Como resultado recomenda-se o uso da estratégia Apriori-C, já adotada pelo CBA2, algoritmo *baseline* neste contexto. Por fim, o terceiro objetivo é voltado a proposta de um método de seleção dinâmica de medidas objetivas, denominado DyOMS, a fim de ranquear as regras da melhor maneira possível (etapa [b]). A motivação ocorre em função dos trabalhos identificados na revisão utilizarem poucas das medidas objetivas existentes na literatura de maneira estática, i.e., pré-estabelecida, mesmo sabendo-se que não existe uma medida que seja adequada a todas as explorações, já que o seu desempenho depende das próprias características das regras extraídas e, portanto, do conjunto de dados utilizado. Como resultado foi possível notar que o DyOMS se apresenta como uma solução viável ao contexto aqui apresentado em relação a alguns *baselines*.

Palavras-chave: classificadores associativos; dados desbalanceados; estratégias de extração de regras; ranqueamento de regras; seleção dinâmica de medidas objetivas.

ABSTRACT

In several problems the use of inherently interpretable algorithms is interesting, as it facilitates the understanding of the obtained knowledge and the predictions made. Among these algorithms are the associative classifiers. They merge association and classification tasks and are, therefore, induced in steps, namely: [a] extraction of a set of rules, [b] ranking of the rules via objective measures and [c] pruning of the rules. Although associative classification, like other techniques, presents good results, when applied to imbalanced problems the performance does not remain the same. Imbalance occurs when the number of instances of a given class, named the majority class, greatly exceeds the number of instances of the other class, named the minority class. Therefore, solutions are being developed to correctly differentiate instances of both classes. Thus, this work explores the use of associative classifiers when applied to imbalanced data via internal approaches, i.e., at the algorithm level. To this end, three objectives are proposed. The first refers to the execution of a systematic review of the literature in order to identify the internal approaches that have been adopted and/or proposed in order to support this work, as well as to identify gaps and opportunities in the area. Based on the gaps identified, the second objective explores the impact of the different strategies raised in the review to carry out rule extraction (step [a]) with the aim of identifying the most appropriate one to be used in the context presented here. As a result, it is recommended to use the Apriori-C strategy, already adopted by CBA2, an algorithm baseline in this context. Finally, the third objective is aimed at proposing a dynamic selection method of objective measures, named DyOMS, in order to rank the rules in the best possible way (step [b]). The motivation occurs because the works identified in the review use few of the objective measures existing in the literature in a static way, i.e., pre-established, even knowing that there is no measure that is suitable for all explorations, since its performance depends on the characteristics of the rules extracted and, therefore, on the dataset used. As a result, it was possible to note that DyOMS presents itself as a viable solution to the context presented here in relation to some baselines.

Keywords: associative classifiers; imbalanced data; rule extraction strategies; rules ranking; dynamic selection of objective measures.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CA	*Classificação Associativa e/ou Classificador Associativo*
CAs	*Classificadores Associativos*
MO	*Medida Objetiva*
MOs	*Medidas Objetivas*
RA	*Regra de Associação*
RAs	*Regras de Associação*
RAC	*Regra de Associação Classificativa*
RACs	*Regras de Associação Classificativas*
RSL	*Revisão Sistemática da Literatura*

SUMÁRIO

1 Introdução

O avanço da tecnologia nos permitiu estar conectados de uma forma inédita, praticamente durante 24 horas por dia e, com isso, gerar uma grande quantidade de informações. Para coletar, pré-processar e explorar esse volume de dados, técnicas de mineração de dados têm sido utilizadas (ABDELLATIF; HASSINE; YAHIA, 2019). Por meio de um processo computacional o conhecimento útil, escondido nesse grande volume de informações, é descoberto. Uma variedade de tarefas de mineração de dados, como associação, classificação e agrupamento, foram desenvolvidas e aplicadas em vários domínios.

Uma das tarefas mais comuns é a classificação, a qual tem como objetivo a atribuição de uma categoria, chamada de classe, à instâncias não rotuladas. Dentre as técnicas que se destacam nesta tarefa encontra-se a classificação associativa (CA). A CA faz parte da família de algoritmos baseados em regras, os quais fazem uso das mesmas para representar o conhecimento extraído. Os algoritmos de CA utilizam-se de um tipo especial de regra de associação (RA) (TAN *et al.*, 2019), conhecida como regra de associação classificativa (RAC), para realizar a indução do modelo. Uma RAC é uma regra do tipo $A \Rightarrow c$, na qual o antecedente (A) contém um conjunto de pares <atributo-valor> e o consequente (c) uma das classes de um dado problema (LIU; HSU; MA, 1998). As mesmas são, em geral, extraídas via adaptação do algoritmo Apriori (AGRAWAL; SRIKANT, 1994), amplamente utilizado para obtenção de RAs. A utilização destes algoritmos é vantajosa, uma vez que as regras contidas nos modelos são interpretáveis por especialistas que podem avaliá-las e tomar decisões com auxílio computacional. Segundo Padillo, Luna e Ventura (2020) estudos recentes têm demonstrado que a classificação associativa apresenta vantagens sobre as abordagens de classificação interpretáveis tradicionais. Devido a esta característica, dentre outras, diversas áreas vêm aplicando classificadores associativos (CAs) nos últimos anos (PIRAN *et al.*, 2024; SEN *et al.*, 2022; AL-HAWARI; NAJADAT; SHATNAWI, 2021; BASHA, 2021; HAAS; MAIER; ROTHGANG, 2021; MOHAMMAD, 2020; SHAO *et al.*, 2017). Além disso, tal como outras famílias de algoritmos, esforços vêm sendo realizados no sentido de popularizar ainda mais a CA por meio da implementação de pacotes, como visto, por exemplo, em Hahsler *et al.* (2019), Padillo, Luna e Ventura (2020) e Azmi e Berrado (2020), os quais contribuem ao disponibilizar algoritmos para uso geral. Segundo Filip e Kliegr (2018), o CBA (LIU; HSU; MA, 1998) é o algoritmo mais utilizado da família e, portanto, o que se encontra disponível nos mais diversos pacotes.

Em geral, a construção de um CA ocorre em etapas, a saber: [a] extração de um conjunto de RACs; [b] ranqueamento das regras geradas; [c] poda das regras. No CBA, por exemplo, as RACs são extraídas (passo [a]) e então ranqueadas de acordo com a relação $\succ$. Dada duas regras,

r_i e r_j, $r_i \succ r_j$, i.e., r_i tem maior precedência em relação a r_j se: (i) a confiança de r_i é maior do que a de r_j; (ii) se as confianças são iguais, mas o suporte de r_i é maior do que o suporte de r_j; a confiança e o suporte são iguais, mas r_i foi gerada antes de r_j. O suporte (sup) e a confiança (conf) são medidas objetivas (MOs) que computam, respectivamente, a frequência (sup=$P(Ac)$) e a força da implicação (conf=$P(c|A)$) de uma dada regra $A \Rightarrow c$. Ambas constituem as medidas básicas quando se trabalha com RAs (TAN *et al.*, 2019) e, consequentemente, com RACs. Considerando esse ranqueamento[1], a poda ocorre. Para cada regra r, verifica-se as transações que ela cobre e se cobre corretamente pelo menos uma transação. Nesse caso, a regra é selecionada para ser incluída no modelo e todas as transações cobertas por ela são removidas. A etapa de poda finaliza a indução do modelo. Em relação a predição, dada uma instância não vista, o rótulo associado à primeira regra que casa com a instância é o que será atribuído a ela.

Embora a CA, assim como outras técnicas, apresente bons resultados, quando aplicada a problemas desbalanceados o desempenho não se mantém o mesmo. O desbalanceamento ocorre quando o número de instâncias de uma dada classe, chamada de majoritária, supera em muito o número de instâncias da outra classe, chamada de minoritária, a qual é, em geral, a classe de maior interesse e importância (FERNÁNDEZ *et al.*, 2018). Negligenciar este fato pode prejudicar todo o processo. Os classificadores padrão são, em geral, direcionados à classe majoritária em favor, por exemplo, da medida de acurácia (FERNÁNDEZ *et al.*, 2018). Assim, regras específicas voltadas para a classe minoritária acabam por serem ignoradas; portanto, instâncias da classe minoritária acabam por serem classificadas incorretamente com mais frequência do que as instâncias da classe majoritária. Diversos problemas do mundo real são desbalanceados: Shao *et al.* (2018) e Shao *et al.* (2020) estão interessados em descobrir defeitos em softwares; Hassine, Abdellatif e Yahia (2022) estão interessados em detectar ideação suicida a partir de textos coletados do Twitter; etc.

Soluções vêm sendo desenvolvidas de modo a diferenciar corretamente as instâncias da classe majoritária da minoritária. De acordo com Fernández *et al.* (2018), as técnicas voltadas para tratamento de dados desbalanceados podem ser divididas, em geral, em: abordagens em nível de dados (ou externas) e abordagens em nível de algoritmo (ou internas), foco deste trabalho. Maiores detalhes vide referência. O CBA2, proposto por Liu, Ma e Wong (2001), é uma adaptação do CBA para dados desbalanceados, também utilizado como *baseline* neste contexto, sendo, portanto, uma solução interna.

Diante do exposto, este trabalho tem por objetivo explorar o uso de CAs quando aplicados em dados desbalanceados via abordagens internas. Este estudo é importante devido ao uso dos CAs nos mais diversos domínios em função de sua interpretabilidade inerente. Assim, este trabalho concentra-se em três objetivos, a saber:

[1] Ranqueamento e ordenação são utilizados como sinônimos neste trabalho.

(Obj.1): realizar uma revisão sistemática da literatura (RSL) a fim de identificar as abordagens internas que vêm sendo adotadas e/ou propostas na literatura. Este objetivo contribui tanto para a fundamentação deste trabalho, quanto para a identificação de lacunas e oportunidades na área. Este mapeamento foi publicado no artigo "Associative Classifiers Algorithms for Imbalanced Data: A Systematic Literature Review" (PRIVATTO; CARVALHO, 2024). As contribuições do artigo e, consequentemente, da RSL, é (i) fornecer um melhor entendimento sobre as estratégias utilizadas em cada uma das etapas da indução dos modelos a fim de tratar o desbalanceamento, (ii) identificar lacunas e oportunidades na área, (iii) apoiar o desenvolvimento e/ou incremento de pacotes voltados a algoritmos de CA para dados desbalanceados.

(Obj.2): realizar uma análise sobre o impacto das diferentes estratégias utilizadas para se realizar a etapa de extração de regras visando identificar a mais adequada a ser utilizada no contexto aqui abordado. A motivação originou-se na identificação, por meio da RSL, de diferentes estratégias adotadas pelos algoritmos propostos na literatura, as quais baseiam-se, em geral, em múltiplos suportes visando obter regras pertencentes à classe minoritária e reduzir o número total de regras.

(Obj.3): propor um método de seleção dinâmica de MOs, denominado DyOMS, que possa ser incorporado a fluxos de indução de CAs. A ideia é que o método detecte a MO mais adequada, em tempo de execução, de modo que as regras sejam ordenadas da melhor maneira possível. A motivação originou-se na identificação, por meio da RSL, da utilização de um conjunto restrito de MOs (Lift, Confiança, Suporte e variações dessas) para realizar a etapa de ranqueamento das regras. Contudo, inúmeras delas (mais de 60) são encontradas na literatura, como as descritas em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022). Uma vez que o ranqueamento é uma etapa importante no processo de indução dos CAs, já que, em geral, a poda se baseia no ranqueamento gerado, é importante que exista uma maneira de se selecionar a MO mais adequada a um dado conjunto de regras. Dado que não existe uma MO que seja adequada a todas as explorações (SHARMA *et al.*, 2020), trabalhos foram realizados visando agrupá-las em função de sua similaridade de desempenho, como os trabalhos de Yang e Cui (2015) e Dall'Agnol e Carvalho (2023). Nestes trabalhos os autores sugerem um grupo de MOs adequadas ao contexto aqui abordado, não especificando, porém, como realizar a escolha da MO mais adequada dentro do grupo. Porém, é possível notar que, de fato, algumas MOs apresentam melhores desempenhos do que outras, e que a mais adequada depende das próprias características das regras extraídas e, portanto, do conjunto de dados utilizado.

A fim de cobrir os três objetivos apresentados, esta dissertação encontra-se estruturada da seguinte maneira: no Capítulo 2 os fundamentos necessários ao entendimento deste trabalho

são apresentados. No Capítulo 3 a revisão sistemática da literatura, referente ao **[Obj.1]**, é apresentada. No Capítulo 4 a metodologia de análise referente ao **[Obj.2]** é apresentada, assim como a configuração experimental, resultados e discussões. No Capítulo 5 o método DyOMS, referente ao **[Obj.3]**, é apresentado, assim como os experimentos, os resultados e as discussões. Por fim, no Capitulo 6 são apresentadas as conclusões e trabalhos futuros.

2 Fundamentos

Este capítulo apresenta os conceitos fundamentais que embasam este trabalho, a saber: regras de associação (Seção 2.1), classificação associativa (Seção 2.2), medidas objetivas (Seção 2.3) e critérios de avaliação (Seção 2.4).

2.1 Regras de Associação

As regras de associação (RAs) compõem uma das maneiras mais tradicionais em aprendizado de máquina para se extrair padrões dos dados (SHARMA *et al.*, 2020). Os algoritmos desta família são utilizados para extração de um conjunto de regras que buscam representar as relações intrínsecas dos atributos e seus valores em conjuntos de dados. Elas são classificadas dentro do aprendizado de máquina como não-supervisionada, tendo um caráter exploratório, com objetivo de realizar descobertas de padrões interessantes.

As RAs são expressas via implicação lógica, tendo um formato do tipo $A \Rightarrow B$, onde o antecedente A implica no consequente B. Esta representação torna a técnica transparente e de fácil entendimento, uma vez que as relações podem ser lidas, entendidas e validadas por um especialista. Uma regra de associação é definida formalmente da maneira descrita a seguir (AGRAWAL; SRIKANT, 1994): seja D um conjunto de dados composto por um conjunto de itens $I = \{i_1, ..., i_m\}$, ordenados lexicograficamente, e por um conjunto de transações $T = \{t_1, ..., t_n\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens (chamado *itemset*) tal que $t_i \subseteq I$. A regra de associação é uma implicação na forma $A \Rightarrow B$, em que $A \subset I$, $B \subset I$ e $A \cap B = \varnothing$. A regra $A \Rightarrow B$ ocorre no conjunto de transações T com **confiança** $conf$ e **suporte** sup, definidas pelas Equações 2.1 e 2.2, em que $n(AB)$ representa a frequência com que $(A \cup B)$ ocorrem conjuntamente no conjunto de transações, $|D|$ o número de transações e $|D(A)|$ o número de transações em que A ocorre. Estas equações também são chamadas de medidas objetivas ou medidas de interesse, detalhadas na Seção 2.3. Assim, os algoritmos de associação visam encontrar todas as possíveis regras (correlações) que apresentem valores de suporte e confiança superiores a valores previamente estabelecidos, denominados de suporte mínimo (**sup-min**) e confiança mínima (**conf-min**).

$$\text{Suporte} = \frac{n(AB)}{|D|} = P(AB) \tag{2.1}$$

$$\text{Confiança} = \frac{n(AB)}{|D(A)|} = P(B|A) \tag{2.2}$$

O problema de obtenção de regras de associação é decomposto em dois passos (AGRAWAL; IMIELIŃSKI; SWAMI, 1993):

1. Encontrar todos os *k-itemsets* (conjunto de k itens) que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário (**sup-min**). Os *itemsets* com suporte igual ou superior a **sup-min** são definidos como *itemsets* frequentes, os demais conjuntos são denominados de *itemsets* não-frequentes. O suporte de um *itemset* é dado por $P(Itemset) = \frac{n(Itemset)}{|D|}$;
2. Utilizar todos os *k-itemsets* frequentes, com $k \geq 2$, para gerar as regras de associação. Somente regras que apresentem confiança maior ou igual a confiança mínima especificada pelo usuário (**conf-min**) são extraídas.

A parte mais dispendiosa dos dois passos é a geração dos *itemsets* frequentes. Assim, a diferença entre os algoritmos se dá neste passo. Isso se deve devido ao número de *itemsets* que podem ser gerados, sendo proporcional a $2^{|I|}$, em que $|I|$ representa o número de itens distintos contidos em I. Assim, à medida que a quantidade de itens aumenta, o espaço de busca também aumenta. Neste sentido, os algoritmos buscam estratégias para encontrar os *itemsets* frequentes sem a necessidade de percorrer todo o espaço de busca. Diversos algoritmos são encontrados na literatura para se obter os *itemsets* frequentes, por conseguinte, gerar as regras de associação, sendo o Apriori (AGRAWAL; SRIKANT, 1994) um dos mais utilizados[1].

No algoritmo Apriori apenas um **sup-min** é especificado, considerando, portanto, que todos os itens contidos nos dados apresentam frequências semelhantes. Contudo, em diversos domínios, alguns itens aparecem com muita frequência (leite, por exemplo), enquanto outros raramente aparecem (caviar, por exemplo). Se as frequências dos itens variam muito, dois problemas podem ocorrer: (i) se o **sup-min** for definido muito alto, regras que envolvam itens pouco frequentes ou raros não serão extraídas; (ii) se o **sup-min** for definido muito baixo, a fim de encontrar regras que envolvam itens frequentes e raros, uma explosão combinatória poderá acontecer, já que os itens frequentes serão associados uns aos outros de todas as maneiras possíveis. Visando solucionar tal problema, Liu, Hsu e Ma (1999) propuseram o MS-Apriori, uma variação do algoritmo Apriori que permite que o usuário especifique múltiplos suportes mínimos, i.e., um suporte mínimo diferente para cada item (**MIS**). Assim, diferentes *itemsets* precisam satisfazer diferentes suportes mínimos dependendo dos itens que os compõem. Deste modo, é possível encontrar *itemsets* que envolvam itens raros sem fazer com que itens frequentes gerem muitos *itemsets* não interessantes. Devido a este contexto, o **sup-min** associado a uma regra (idem para os *itemsets*) depende dos **MIS** definidos para cada item que a compõe. Assim, tem-se que o suporte mínimo de uma regra R é o menor valor **MIS** entre os itens que a compõe,

[1] Diversas outras implementações encontram-se disponíveis em <http://www.philippe-fournier-viger.com/spmf/>.

i.e., para $R : i_1, i_2, ..., i_j \Rightarrow i_{j+1}, ...i_k$, o **sup-min** é dado por min(MIS(i_1), MIS(i_2), ..., MIS(i_k)). Deste modo, uma regra será extraída apenas se o seu suporte for maior ou igual ao **sup-min** definido como exposto. Note que cada regra (idem *itemset*) apresenta um **sup-min** distinto a ser atingido. O MS-Apriori permite, portanto, a obtenção de **sup-min** mais altos para regras que envolvem apenas itens frequentes e **sup-min** mais baixos para regras que envolvem itens menos frequentes. Por fim, vale mencionar que Liu, Hsu e Ma (1999) sugerem definir o valor de **MIS** de cada item em função de sua frequência no conjunto de dados da seguinte maneira: se a frequência do item i é $f(i)$, seu **MIS** é computado por $\beta * f(i)$, em que β é um valor no intervalo [0,1]. β pondera, portanto, a porcentagem do total da frequência associada a cada item. O valor sugerido pelos autores é de β=25% (0.25).

2.2 Classificação Associativa

A classificação associativa (CA) faz parte da família de algoritmos baseados em regras, os quais fazem uso das mesmas para representar o conhecimento extraído. No caso da CA, os algoritmos utilizam-se de um tipo especial de regra de associação (RA), conhecida como regra de associação classificativa (RAC), para realizar a indução do modelo. Uma RAC é uma regra do tipo $A \Rightarrow c$, na qual o antecedente (A) contém um conjunto de pares <atributo-valor> e o consequente (c) uma das classes de um dado problema (LIU; HSU; MA, 1998). As mesmas são, em geral, extraídas via adaptação do algoritmo Apriori. A utilização destes algoritmos é vantajosa, uma vez que as regras contidas nos modelos são interpretáveis por especialistas que podem avaliá-las e tomar decisões com auxílio computacional. Segundo Padillo, Luna e Ventura (2020) estudos recentes têm demonstrado que a classificação associativa apresenta vantagens sobre as abordagens de classificação interpretáveis tradicionais. Devido a esta característica, dentre outras, diversas áreas vêm aplicando classificadores associativos (CAs) nos últimos anos (PIRAN *et al.*, 2024; SEN *et al.*, 2022; AL-HAWARI; NAJADAT; SHATNAWI, 2021; BASHA, 2021; HAAS; MAIER; ROTHGANG, 2021; MOHAMMAD, 2020; SHAO *et al.*, 2017). Além disso, tal como outras famílias de algoritmos, esforços vêm sendo realizados no sentido de popularizar ainda mais a CA por meio da implementação de pacotes, como visto, por exemplo, em Hahsler *et al.* (2019), Padillo, Luna e Ventura (2020), e Azmi e Berrado (2020), os quais contribuem ao disponibilizar algoritmos para uso geral. Segundo Filip e Kliegr (2018), o CBA (LIU; HSU; MA, 1998) é o algoritmo mais utilizado da família e, portanto, o que se encontra disponível nos mais diversos pacotes.

A construção de um classificador associativo, via CBA, assim como na maioria dos demais algoritmos, ocorre em etapas, a saber: [a] extração de um conjunto de RACs; [b] ranqueamento das regras geradas; [c] poda das regras. No que se refere a etapa [a], em relação ao CBA, Liu, Hsu e Ma (1998) apresentam uma adaptação do algoritmo Apriori para obtenção das RACs. No

que se refere as demais etapas tem-se que:

Ranqueamento. A partir das RACs extraídas, as mesmas são ordenadas de acordo com a relação $\succ$. Dado duas regras, r_i e r_j, $r_i \succ r_j$, ou seja, r_i tem maior precedência que r_j se:

1. a confiança de r_i for maior que a de r_j, ou
2. se as confianças forem iguais, mas o suporte de r_i for maior que o suporte de r_j, ou
3. tanto a confiança como o suporte são iguais, mas r_i foi gerada primeiro que r_j.

Poda. Para cada regra r, verifica-se as transações que ela cobre e se cobre corretamente pelo menos uma transação. Nesse caso, a regra é selecionada para ser incluída no modelo e todas as transações cobertas por ela são removidas.

A partir do modelo obtido, dada uma instância não vista, a classe associada à primeira regra que casa com a instância é a que será atribuído a ela. A Figura 1 ilustra o fluxo descrito, incluindo a etapa de predição. Vale mencionar em relação a etapa de ranqueamento que diversos trabalhos, como alguns descritos e/ou mencionados ao longo deste trabalho (vide Seção 2.3, por exemplo), modificam a referida etapa da seguinte maneira: dada uma medida objetiva m (vide Seção 2.3), uma regra r_i precede uma regra r_j, em uma lista ordenada, se o valor de m em r_i é maior que r_j; em caso de empate, se o suporte de r_i é maior que r_j; em caso de empate, r_i foi gerada antes de r_j. Deste modo, é possível avaliar o quanto cada medida objetiva influencia o desempenho final do classificador quando utilizada na etapa de ranqueamento dos CAs. Essa mesma ideia é a utilizada nos Capítulos 4 e 5.

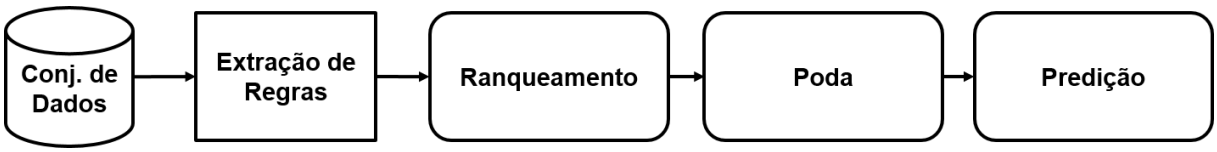


Figura 1 – Etapas envolvidas na indução de um CA.

Embora o CBA seja o algoritmo que mais se destaca, quando se trata de dados desbalanceados ele apresenta problemas em relação a extração de regras, já que um único suporte é especificado (LIU; MA; WONG, 2001): se esse único valor for muito alto, pode-se não encontrar regras suficientes das classes minoritárias; se o valor for muito baixo, muitas regras inúteis e de overfitting serão encontradas para as classes frequentes. Assim, Liu, Ma e Wong (2001) propuseram o CBA2, o qual faz uso de vários suportes mínimos (**sup-min**), de maneira que cada classe receba um **sup-min** diferente. O CBA2 é, portanto, uma adaptação do CBA para dados desbalanceados, também utilizado como *baseline* neste contexto. A fim de determinar o **sup-min** de cada classe, a Equação 2.3 é utilizada, a qual baseia-se em um **sup-min** total especificado pelo usuário, em que sup-min$_{c_i}$ representa o **sup-min** da classe c_i e $freq(c_i)$ a frequência de c_i no

conjunto de dados. A equação gera valores mais altos para as classes majoritárias e mais baixos para as classes minoritárias. Dessa maneira, garante-se que sejam geradas regras suficientes para as classes minoritárias e não muitas regras para as classes majoritárias. Nota-se, neste caso, que o **sup-min** da Equação 2.3 equivale ao β do MS-Apriori (vide Seção 2.1).

$$\text{sup-min}_{c_i} = \text{sup-min} \times freq(c_i) = \beta \times freq(c_i) \tag{2.3}$$

2.3 Medidas Objetivas

Como mencionado na Seção 2.1, as RAs compõem uma das maneiras mais tradicionais para se extrair padrões dos dados (SHARMA *et al.*, 2020). Além disso, como visto, as RAs são acompanhadas por medidas de interesse, como o Suporte (Equação 2.1) e a Confiança (Equação 2.2). As medidas de interesse visam apoiar a descoberta de padrões significativos, as quais podem ser classificadas em medidas objetivas (MOs)[2], medidas subjetivas e medidas semânticas (SHARMA *et al.*, 2020). O foco deste trabalho são as medidas objetivas, como o Suporte e a Confiança, uma vez que elas independem de conhecimento externo, i.e., são computadas apenas com base no conjunto de dados. Em geral, quanto maior o valor da medida em uma dada regra, melhor ranqueada ela estará. Assim, uma das aplicações mais comuns ao se utilizar uma medida é ranquear as regras de modo que as primeiras classificadas sejam as mais interessantes ao usuário/aplicação. Contudo, existem mais de 60 MOs disponíveis na literatura, como visto em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022).

Para se computar uma dada medida é necessário conhecer a matriz de contingência da regra, apresentada na Tabela 1. As quatro frequências que compõem a tabela, i.e., f_{11} ($P(AB)$), f_{10} ($P(A\overline{B})$), f_{01} ($P(\overline{A}B)$) e f_{00} ($P(\overline{AB})$), a partir das quais as medidas são derivadas, podem ser obtidas a partir de f_{11}, f_{1+} (P(A)), f_{+1} (P(B)) e N.

Tabela 1 – Matriz de contingência de uma dada regra $A \Rightarrow B$, em que N representa o número de transações/instâncias, f_{11} o suporte da regra (P(AB)), f_{1+} (P(A)) o suporte de A e f_{+1} o suporte de B (P(B)). Fonte: Somyanonthanakul e Theeramunkong (2022).

	B	$\overline{B}$	B $\cup$ $\overline{B}$
A	f_{11}	$f_{10} = f_{1+} - f_{11}$	f_{1+}
$\overline{A}$	$f_{01} = f_{+1} - f_{11}$	$f_{00} = N - f_{1+} - f_{+1} + f_{11}$	$f_{0+} = N - f_{1+}$
A $\cup$ $\overline{A}$	f_{+1}	$f_{+0} = N - f_{+1}$	N

Dado que não existe uma MO que seja adequada a todas as explorações (SHARMA *et al.*, 2020), realizar a escolha de uma ou mais medidas para explorar um conjunto de regras torna-se importante dado os benefícios que se obtêm ao optar por uma ou outra MO. No entanto,

[2] Neste trabalho, medidas se referem a medidas objetivas, podendo o termo ser usado indistintamente.

fazer a escolha de qual utilizar é um problema difícil (SHARMA *et al.*, 2020), visto que existem diferenças significativas em relação aos objetivos (semânticas), entre outros aspectos, que as distinguem. Para tanto, diferentes maneiras foram propostas para realizar esta escolha, sendo uma delas via agrupamento.

O objetivo das abordagens baseadas em agrupamento é formar grupos de medidas de modo que o usuário possa selecionar uma medida representativa de cada grupo. Deste modo, visa-se reduzir o espaço de busca. Uma das maneiras existentes para se realizar o agrupamento é computando a similaridade das MOs em relação ao ranqueamento por elas geradas em um dado conjunto de regras de associação. Neste contexto, os trabalhos de Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022) apresentam um estudo comportamental de um conjunto de 61 MOs quando usadas para ranquear RAs. Uma vez que os resultados de Somyanonthanakul e Theeramunkong (2022) são similares aos apresentados em Tew *et al.* (2014), ele não será aqui descrito.

De maneira geral, o método de agrupamento proposto por Tew *et al.* (2014) se dá extraindo um conjunto de regras de associação para cada conjunto de dados (110 no total) e computando-se os valores das 61 MOs para cada conjunto de regras. Os valores das MOs são então substituídos por ranks, de modo que regras que apresentam maiores valores recebam ranks menores, já que quanto maior o valor de uma dada medida em uma dada regra, mais interessante a regra é. A proposta de se trabalhar com ranks se dá pelo fato de os valores das MOs não serem comparáveis. Ao final, cada MO gera uma lista ordenada de regras, em cada conjunto de dados. Computa-se então a correlação de Spearman entre as MOs em cada conjunto de dados considerando os ranks gerados por elas. Em seguida, computa-se a matriz de distâncias entre as medidas considerando a correlação média entre as medidas em todos os conjuntos de dados. A partir da matriz de distâncias o algoritmo Complete Linkage é aplicado e os grupos obtidos. Aplicando o método por eles proposto, 21 grupos de MOs são obtidos. Das 61 MOs analisadas pelos autores, 50 foram utilizadas para realizar o agrupamento. As 11 medidas que ficaram de fora são matematicamente equivalentes a outras medidas e, portanto, foram removidas.

Seguindo a mesma linha de raciocínio dos trabalhos que agrupam MOs no contexto de RAs, trabalhos como o de Yang e Cui (2015) e Dall'Agnol e Carvalho (2023) foram desenvolvidos no sentido de agrupar as MOs no contexto de CAs. O objetivo é agrupá-las em função do desempenho obtido no modelo final quando as MOs são utilizadas na etapa de ranqueamento. Deste modo, uma vez que o método descrito no Capítulo 5 é inspirado nestes trabalhos, apresenta-se a seguir uma breve descrição dos mesmos.

Estudo de Yang e Cui (2015). Visando analisar o comportamento das MOs quando aplicadas aos CAs, Yang e Cui (2015) dividem o trabalho em duas partes, sendo a segunda parte voltada aos CAs. O processo proposto pelos autores ocorre em etapas e, de maneira geral, ocorre

da maneira descrita a seguir. Para cada conjunto de dados d (9 no total), k amostras são obtidas ($k = 10$), das quais s amostras são geradas ($s = 10$), cada uma contendo uma distribuição de classe $s \times 0.1$, i.e., para cada distribuição entre [0.1;1.0][passo=0.1], 10 amostras são geradas. Na sequência, para cada amostra d_{sk} de cada conjunto de dados (100 por conjunto), o algoritmo CBA é executado m vezes, em que m é o número de medidas objetivas avaliadas pelos autores (55). Os autores não mencionam como o CBA é modificado para realizar tais execuções, mas supõe-se que o ranqueamento das regras é trocado por cada uma das m medidas. Ao final de cada execução, armazena-se o desempenho P obtido pelo classificador via AUC (área abaixo da curva ROC). Em seguida, para cada distribuição s, o desempenho médio P_m^s de cada medida m em todas as amostras com distribuição s é computado. Por fim, para cada distribuição s, as medidas são ordenadas por desempenho (do maior para o menor) e as top 10, i.e., as 10 primeiras melhor ranqueadas de cada distribuição, são então selecionadas. Tendo como base essas s listas, uma lista final L é obtida realizando-se a intersecção entre todas as listas anteriormente obtidas. Das 55 medidas dadas como entrada, apenas 26 são selecionadas após o processo aqui descrito. Tendo como base essas 26 medidas, os autores aplicam um algoritmo de agrupamento hierárquico e as dividem em dois grupos: as medidas mais adequadas a serem utilizadas para conjuntos de dados com distribuição menor do que 0.4 e aquelas com distribuição maior do que 0.4. As medidas apresentadas em cada grupo foram as seguintes:

- $G_<$: Correlation Coefficient, Collective Strength, Kappa, Piatetsky-Shapiro, Putative Causal Dependency, Zhang, Intensity of Implication, Confirm Causal, Goodman–Kruskal, Entropic Implication Intensity 1, Implication Index, Leverage, Added Value
- $G_>$: Odd Multiplier, Complement Class Support, Conviction, Yule's Q, Sebag–Schoenauer, Yule's Y, Odds Ratio, Confidence Causal, Confirmed Confidence Causal, Example and Counterexample Rate, Ganascia, J-measure, Confidence

Segundo as análises, o grupo $G_<$ desempenha bem em dados extremamente desbalanceados e a maioria tem um desempenho ruim em dados balanceados; por outro lado, o grupo $G_>$ desemprenha bem em dados levemente desbalanceados e ficam melhores quando usados em dados mais balanceados.

Estudo de Dall'Agnol e Carvalho (2023). Assim com em Yang e Cui (2015), Dall'Agnol e Carvalho (2023) também visam agrupar as MOs de acordo com seu desempenho quando aplicadas em CAs. Contudo, o estudo dos autores é voltado apenas para conjuntos de dados balanceados ou levemente desbalanceados. Para tanto, os autores modificam o algoritmo CBA, denominado por eles de CBA', em que todas as etapas de extração do modelo são iguais ao CBA, com exceção da ordenação. Neste caso, dada uma medida m, uma regra r_i precede uma regra r_j, em uma lista ordenada, se o valor de m em r_i é maior que r_j; em caso de empate, r_i deve

ter sido gerada antes de r_j. Os autores afirmam que dessa maneira é possível acessar o quanto cada medida influencia o desemprenho final do classificador. O processo proposto pelos autores ocorre em etapas e, de maneira geral, ocorre da maneira descrita a seguir. Para cada conjunto de dados d (43 no total), uma validação cruzada estratificada de 10-folds é realizada (SCV_{10}). Em cada fold f, o CBA' é aplicado em cada medida m considerada. Na sequência, computa-se o desempenho médio P_m^d de cada medida m em cada conjunto de dados d ao longo de todos os folds. A medida de desempenho utilizada foi a F1-Macro. Em seguida, o coeficiente de correlação de Pearson é computado entre todos os pares de medidas, ao longo de todos os conjuntos, a fim de se obter uma matriz de distâncias. Tendo como base a matriz gerada, o agrupamento é obtido utilizando-se o algoritmo Complete Linkage, como nos demais trabalhos. Das 61 MOs apresentadas em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022), Dall'Agnol e Carvalho (2023) selecionaram 44. Após aplicação do método, as 44 MOs foram agrupadas em 15 grupos, os quais foram ranqueados de acordo com seu desempenho médio. Os grupos podem ser vistos na Tabela 2. Os resultados foram comparados com os obtidos por Yang e Cui (2015).

Tabela 2 – Grupos de medidas propostos por Dall'Agnol e Carvalho (2023).

Grupo	**Medidas**
1	Odd Multiplier, Complement Class Support, Confidence Causal, Loevinger, Added Value, One Way Support, Confirmed Confidence Causal, Lift, Confidence, Putative Causal Dependency, Leverage, Confirm Causal, TIC, DIR, Normalized Mutual Information
2	Klosgen, Implication Index, Accuracy, Correlation Coefficient, Kappa, Collective Strength
3	J-Measure, Chi-Square, Gini Index, Theil Uncertainty Coefficient, Goodman-Kruskal, Mutual Information
4	Conditional Entropy
5	Odds Ratio
6	Least Contradiction, Confirm Descriptive
7	2-Way Support, Piatetsky-Shapiro
8	Cosine, F-Measure
9	K-Measure
10	Kulczynski 2
11	Relative Risk
12	Specificity, Logical Necessity
13	Prevalence
14	Recall, Support
15	Coverage

Visão Geral dos Estudos. Uma visão geral de como os trabalhos de Yang e Cui (2015) e Dall'Agnol e Carvalho (2023) realizam o processo de agrupamento das MOs é apresentado na

Figura 2. Como visto, o processo de indução do CBA é alterado de modo que as RACs sejam ordenadas por diferentes MOs. Cada MO leva a obtenção de um modelo distinto associado a uma medida de avaliação, F1-Macro, por exemplo (na figura, VA significa "Valor de Avaliação"). Após executar o fluxo indicado na figura para uma coleção de conjuntos de dados, uma matriz MxN é obtida, a qual relaciona, para cada conjunto de dados m, o desempenho obtido pelo modelo gerado via uma dada MO n. A partir desta matriz, dada uma medida de dissimilaridade, é possível se gerar uma matriz NxN, i.e., medidas por medidas, de modo que o agrupamento das mesmas seja realizado.

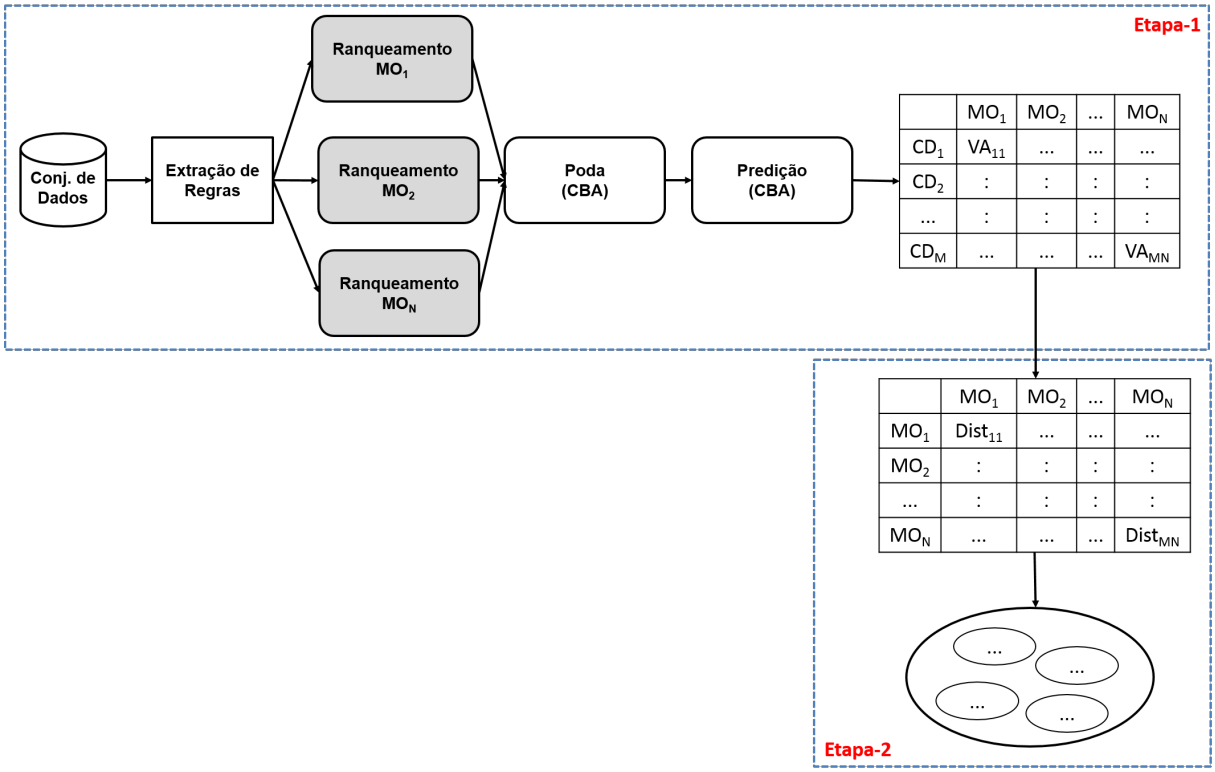


Figura 2 – Fluxo geral para obtenção de agrupamentos de MOs em CAs usado na literatura.

Embora interessantes, já que fornecem um subconjunto das MOs a serem exploradas no contexto de CAs, os autores não especificam como realizar a escolha de uma dada MO em um dado subconjunto. Contudo, é possível notar que, de fato, algumas MOs apresentam melhores desempenhos do que outras, e que a mais adequada depende das próprias características das regras extraídas. Assim, este trabalho propõe um método de seleção dinâmica de MOs que possa ser incorporado a fluxos de indução de CAs. O método é inspirado nos trabalhos descritos, porém levando em consideração as características de distribuição de ordenação das regras em ambas as classes (majoritária e minoritária), assim como as características do conjunto de dados.

2.4 Critérios de Avaliação

A tarefa de classificação tem por objetivo atribuir uma categoria, chamada de classe, para instâncias não rotuladas. Assim, uma vez que os modelos são obtidos, é necessário que avalie-os via uma dada medida de desempenho em dados ainda não vistos. Para tanto, medidas de avaliação são utilizadas, como Precisão (*Precision*), Revocação (*Recall*) e medida-F (versões micro e/ou macro e/ou macro ponderada). Além disso, a fim de garantir uma boa estimativa da medida de avaliação, assim como viabilizar a configuração dos hiperparâmetros, diferentes estratégias de validação podem ser utilizadas, como *holdout*, validação cruzada (*cross-validation*) e validação cruzada estratificada (*stratified cross-validation*). Assim, em relação ao desempenho, este trabalho avalia os resultados dos experimentos via F1 e G-Mean, nas versões macro, ambas estimadas via 2-fold cross-validation estratificado[3]. Ambas as medidas são amplamente utilizadas em contextos desbalanceados, como visto em Fernández *et al.* (2018) e nos demais trabalhos da literatura revisados no Capítulo 3. A medida F1, apresentada na Equação 2.4, visa analisar o equilíbrio (*trade-off*), por meio da média harmônica, entre a corretude (*Precision*) e a revocação (*Recall*) na classificação de instâncias positivas. Já a medida G-Mean, apresentada na Equação 2.5, visa o equilíbrio entre os desempenhos de classificação nas classes majoritárias e minoritárias por meio da média geométrica da taxa de verdadeiros positivos (TVP) e negativos (TVN). Nas equações, VP=Verdadeiros Positivos, FP=Falsos Positivos, FN=Falsos Negativos, VN=Verdadeiros Negativos, TVP=Taxa dos Verdadeiros Positivos e TVN=Taxa dos Verdadeiros Negativos. Por fim, vale mencionar que entende-se que estes conceitos apresentam-se como definições básicas na literatura de aprendizado de máquina, e, portanto, não serão aqui detalhados. Estes conceitos podem ser consultados em Tan *et al.* (2019) e Fernández *et al.* (2018).

$$\begin{aligned} &\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall},\ em\ que \\ &Precision = \frac{VP}{VP+FP},\ Recall = \frac{VP}{VP+FN} \end{aligned} \tag{2.4}$$

$$\begin{aligned} &\text{G-Mean} = \sqrt{TVP \times TVN},\ em\ que \\ &TVP = Recall = Recall_{(Pos)},\ TVN = Recall_{(Neg)} = \frac{VN}{VN+FP} \end{aligned} \tag{2.5}$$

Um outro critério importante a ser utilizado para se avaliar modelos induzidos via algoritmos baseados em regras (*rule-based*) é a interpretabilidade. De acordo com Margot e Luta (2021), a interpretabilidade vem se tornando cada vez mais importante no contexto de tarefas preditivas, embora ainda não haja um consenso sobre esta noção. Em geral, como nota-se em Rudin (2019), Margot e Luta (2021) e Molnar (2022), é possível se obter modelos de predição interpretáveis (i) via algoritmos não interpretáveis, como redes neurais profundas, sobre os

[3] Em contextos desbalanceados é comum setar o número de folds em 2 ou 5, já que o número de instâncias da classe minoritária é geralmente pequeno. Neste trabalho, a fim de garantir um número mínimo de instâncias de ambas as classes nos folds gerados utilizou-se k=2.

quais aplicam-se métodos de XAI, ou (ii) via algoritmos inerentemente interpretáveis, como os baseados em regras e árvores. Ainda de acordo com Margot e Luta (2021), embora algoritmos inerentemente interpretáveis pareçam fáceis de se entender, não existe uma definição matemática exata para o conceito de interpretabilidade. Assim, cada trabalho opta por avaliar este critério por meio de uma medida distinta.

Diante do exposto, neste trabalho optou-se por medir a interpretabilidade por meio do número de regras contidas no modelo, como em outros trabalhos recentes no contexto de CA (DALL'AGNOL; CARVALHO, 2024; MATTIEV; DAVITYAN; KAVSEK, 2023; BUI-THI; MEYSMAN; LAUKENS, 2022; MATTIEV; MEZA; KAVSEK, 2022; SOOD; ZAIANE, 2020; RAJAB, 2019; LAKKARAJU; BACH; LESKOVEC, 2016). Em todos estes trabalhos, quanto menor a quantidade de regras melhor o modelo induzido, i.e., mais interpretável ele é. Por fim, vale mencionar que a mesma é também estimada via 2-fold cross-validation estratificado, assim como adotado nas medidas de desempenho. Este critério foi denominado aqui de tamanho do modelo e representado por $\mathcal{L}$.

3 Revisão Sistemática da Literatura

Como visto no capítulo anterior, a CA tem sido utilizada em diversas áreas devido a sua característica inerentemente interpretável. Contudo, embora a CA, assim como outras técnicas, apresente bons resultados, quando aplicada a problemas desbalanceados o desempenho não se mantém o mesmo. O desbalanceamento ocorre quando o número de instâncias de uma dada classe, chamada de majoritária, supera em muito o número de instâncias da outra classe, chamada de minoritária, a qual é, em geral, a classe de maior interesse e importância (FERNÁNDEZ *et al.*, 2018). Negligenciar este fato pode prejudicar todo o processo. Os classificadores padrão são, em geral, direcionados à classe majoritária em favor, por exemplo, da medida de acurácia (FERNÁNDEZ *et al.*, 2018). Assim, regras específicas voltadas para a classe minoritária acabam por serem ignoradas; portanto, instâncias da classe minoritária acabam por serem classificadas incorretamente com mais frequência do que as instâncias da classe majoritária.

Diante do exposto, soluções vêm sendo desenvolvidas de modo a diferenciar corretamente as instâncias da classe majoritária da minoritária. Assim, uma revisão sistemática da literatura (RSL) sobre soluções em nível de algoritmo voltadas a CAs para dados desbalanceados foi realizada. A mesma foi publicada no artigo "Associative Classifiers Algorithms for Imbalanced Data: A Systematic Literature Review" (PRIVATTO; CARVALHO, 2024). O objetivo do artigo é (i) fornecer um melhor entendimento sobre as estratégias utilizadas em cada uma das etapas da indução dos modelos a fim de tratar o desbalanceamento, (ii) identificar lacunas e oportunidades na área, (iii) apoiar o desenvolvimento e/ou incremento de pacotes voltados a algoritmos de CA para dados desbalanceados. Este capítulo apresenta, portanto, grande parte do artigo apresentado na referência Privatto e Carvalho (2024), uma vez que detalhes encontram-se disponíveis na respectiva referência. Para tanto, a Seção 3.1 descreve o protocolo utilizado na RSL, a Seção 3.2 os resultados, assim como as análises e discussões para cada questão de pesquisa e, por fim, a Seção 3.3 as considerações finais.

3.1 Protocolo

Uma Revisão Sistemática da Literatura (SLR) é um processo no qual um conjunto de estudos disponíveis na literatura é analisado com base em uma pergunta de pesquisa (KITCHENHAM; CHARTERS, 2007; DERMEVAL; COELHO; BITTENCOURT, 2020). O objetivo é fornecer uma visão geral do estado da arte por meio da apresentação e discussão dos resultados, considerando as análises realizadas em estudos identificados como relevantes. Para isso, um protocolo é elaborado, o qual contém os seguintes passos (KITCHENHAM;

CHARTERS, 2007; DERMEVAL; COELHO; BITTENCOURT, 2020): (a) formulação de uma ou mais questões de pesquisa (Seção 3.1.1); (b) identificação dos estudos primários a serem considerados (Seção 3.1.2); (c) extração e síntese de dados (Seção 3.1.3); (d) resumo e discussão dos resultados (Seção 3.2). Para auxiliar na revisão, as ferramentas Parsifal[1] e Excel foram utilizadas.

3.1.1 Questões de Pesquisa

O objetivo desta RSL foi recuperar e analisar estudos primários que apresentam soluções internas[2], i.e., um fluxo algorítmico completo, em relação a classificadores associativos quando utilizados em conjuntos de dados desbalanceados. Em outras palavras, a RSL focou na identificação e compreensão de como os algoritmos propostos funcionam. Para isso, as perguntas abaixo foram formuladas.

QP-1. Quais passos típicos de um classificador associativo os algoritmos propostos cobrem? Esta pergunta visa identificar se os algoritmos propostos cobrem ou não os passos básicos de um algoritmo de classificação associativa, i.e., extração, ranqueamento e poda.

QP-2. Quais estratégias os algoritmos propostos usam para realizar a extração de regras? Esta pergunta busca identificar as estratégias usadas para extrair regras em contextos desbalanceados, já que se o suporte mínimo for definido muito alto, regras da classe minoritária podem não ser extraídas; por outro lado, se definido muito baixo, muitas regras irrelevantes são geradas, não apenas da classe majoritária, resultando em um grande volume de regras obtidas.

QP-3. Quais algoritmos de extração de regras os algoritmos propostos utilizam? Esta pergunta visa identificar os algoritmos de extração de regras utilizados, se são soluções proprietárias ou baseadas em algoritmos tradicionais, como o Apriori.

QP-4. Quais medidas objetivas são utilizadas quando a etapa de ranqueamento está presente? Existem outros critérios envolvidos? Esta pergunta busca identificar as MOs comumente usadas para ranquear as regras e se existem ou não outros critérios envolvidos.

QP-5. Se a etapa de poda existir, como é realizada? Esta pergunta visa identificar como a poda é realizada, as estratégias utilizadas, etc.

QP-6. Em relação à predição, os algoritmos propostos usam a estratégia "Lista de Regras" ("Rule List") ou "Conjunto de Regras" ("Rule Set")? Esta pergunta busca identificar se, no momento da predição, os algoritmos utilizam uma única regra para determinar a classe de uma determinada instância ("Rule List") ou um conjunto de regras ("Rule Set").

[1] <https://parsif.al/>.

[2] De acordo com Fernández *et al.* (2018), as técnicas voltadas para o tratamento de dados desbalanceados podem ser divididas, em geral, em: abordagens em nível de dados (ou externas) e abordagens em nível de algoritmo (ou internas), foco deste trabalho. Para mais detalhes, veja a referência.

QP-7. Os algoritmos propostos funcionam com classificação binária e/ou multiclasse? Esta pergunta visa identificar se os algoritmos funcionam com problemas binários e/ou multiclasse.

3.1.2 Identificação dos Estudos Primários

Para identificar os estudos primários relevantes para a extração de dados, é necessário definir a *string* de busca, as bases de dados para recuperação dos artigos, os critérios de inclusão e exclusão para selecionar ou não um artigo como relevante e os passos para realizar a seleção.

***String* de Busca.** A *string* considerada foi a seguinte: "*({association classification} OR {associative classification} OR {associative rule mining} OR {associative classifier} OR {associative classifiers} OR {predictive association rules} OR {predictive association rule} OR {class association rules} OR {class association rule} OR {classification association rules} OR {classification association rule}) AND ({imbalanced} OR {imbalance} OR {unbalanced} OR {unbalance} OR {skewed} OR {rare item} OR {rare items})*". A primeira parte da *string* inclui os classificadores associativos e a segunda parte os dados desbalanceados. Para formular essa *string*, avaliou-se as palavras frequentemente usadas em vários trabalhos, bem como seus sinônimos, e depois verificou-se os trabalhos recuperados com essa *string* nas bases de dados utilizadas a fim de calibrá-la.

Seleção das Fontes. A *string* de busca foi aplicada apenas em bibliotecas digitais, fazendo-se os ajustes necessários para a sintaxe de cada uma delas. As bibliotecas utilizadas foram: Scopus[3], Compendex[4], ISI Web of Science[5]. A *string* foi aplicada aos títulos, resumos e palavras-chave. O período considerado na busca foi de 01/01/2012 a 31/12/2023 (11 anos)[6].

Critérios de Inclusão e Exclusão. O propósito de se definir esses critérios é identificar os estudos primários que fornecem evidências diretas em relação às perguntas de pesquisa. Assim, os estudos a serem selecionados para a extração de dados são aqueles que não se encaixam em nenhum critério de exclusão. Os seguintes critérios de exclusão foram considerados: (i) o artigo está fora do escopo, ou seja, não propõe uma solução interna para lidar com o problema de desbalanceamento no contexto de classificadores associativos; (ii) o artigo atende ao escopo, mas trata não apenas do desbalanceamento, mas também de outros aspectos, como escalabilidade por meio de soluções paralelas e/ou distribuídas, soluções incrementais e/ou *lazy*, etc.; (iii) o artigo atende ao escopo, mas não induz realmente um classificador, i.e., não propõe um fluxo algorítmico completo; (iv) o artigo atende ao escopo, mas propõe uma solução interna específica

[3] <https://www.scopus.com>.
[4] <https://www.engineeringvillage.com>.
[5] <https://www.webofscience.com>.
[6] O mapeamento foi realizado no final de 2023 considerando o período de 01/01/2012 a 31/12/2022, pois o ano de 2023 não havia terminado. No entanto, no início de 2024 a *string* de busca foi executada novamente nas respectivas bases de dados e os artigos de 2023 que não haviam sido identificados até então foram adicionados e analisados. Por esse motivo, o mapeamento abrangeu 11 anos.

para um determinado problema que não se aplica a outros contextos; (v) o artigo é uma cópia ou outra versão de um artigo já considerado; (vi) o artigo não está em inglês.

Passos de Seleção. Os passos utilizados para selecionar os artigos relevantes aqui apresentados foram os seguintes:

- Passo 1: Identificação e organização dos artigos retornados das bibliotecas digitais utilizando a ferramenta Parsifal. O seguinte número de artigos foi obtido de cada base de dados, totalizando 124 artigos: Scopus=49; Compendex=38; Web of Science=37.
- Passo 2: Remoção automática e/ou manual de artigos duplicados usando a ferramenta Parsifal. Ao final deste passo, 53 artigos permaneceram.
- Passo 3: Revisão de títulos e resumos para aplicar os critérios de exclusão. Artigos que atendiam aos critérios de exclusão foram removidos e os restantes foram mantidos. Em caso de dúvida, o artigo foi mantido. Ao final deste passo, 29 artigos permaneceram.
- Passo 4: Revisão dos artigos completos. Artigos que atendiam aos critérios de exclusão foram removidos e os restantes foram mantidos. Ao final deste passo, 11 artigos permaneceram.

3.1.3 Extração de Dados

A extração de dados diz respeito à coleta de informações dos artigos selecionados para responder às perguntas de pesquisa. A extração de dados foi realizada por meio da leitura dos artigos selecionados. Os formulários de extração foram elaborados no Excel. Os resultados da extração são apresentados na Seção 3.2.

3.1.4 Ameaças à Validade

Uma das vantagens de se realizar uma RSL é apresentar uma visão geral do estado da arte por meio de um processo metodológico e não arbitrário. No entanto, mesmo neste caso, é possível que artigos relevantes não sejam incluídos. Em relação a este trabalho, podem ser mencionadas as seguintes questões: (i) embora as buscas tenham sido realizadas nas maiores bibliotecas digitais, algumas outras não foram utilizadas, como a Springer[7], o que pode ter resultado na perda de estudos; (ii) como foi decidido selecionar soluções focadas apenas em dados desbalanceados, soluções que envolvem conjuntamente outros aspectos não foram consideradas, o que pode ter resultado na perda de estudos.

[7] <https://link.springer.com/>.

3.2 Resultados, Análise e Discussão

Os 11 estudos (algoritmos) selecionados nesta RSL foram os seguintes: ARCID (ABDELLATIF *et al.*, 2018a), IARCID (ABDELLATIF *et al.*, 2018b), CBA Adaptado via ModifiedLift (HASSINE; ABDELLATIF; YAHIA, 2022), ACAR (SHAO *et al.*, 2018), CWCAR (SHAO *et al.*, 2020), CBA Adaptado via PoI (LIEWLOM, 2021), ACRIPPER (ABU-ARQOUB; HADI; ISHTAIWI, 2021), ACRE (CHEN; HSU, 2016), PCAR (CHEN; HSU; HSU, 2012), MMSCBA (HU *et al.*, 2016) e SSCR (WAIYAMAI; SUWANNARATTAPHOOM, 2014). As análises a seguir são baseadas na leitura desses estudos.

QP-1. Quais etapas típicas de um classificador associativo os algoritmos propostos cobrem? A maioria dos algoritmos (9/81,82%) cobrem as etapas básicas de um algoritmo de classificador associativo, como pode ser visto na Tabela 3. As exceções referem-se ao ACRIPPER

Tabela 3 – Etapas cobertas por cada algoritmo proposto.

	Extração	**Ranqueamento**	**Poda**
ARCID	X	X	X
IARCID	X	X	X
CBA Adaptado via ModifiedLift	X	X	X
ACAR	X	X	X
CWCAR	X	X	X
CBA Adaptado via PoI	X	X	X
ACRIPPER	X		
ACRE	X		X
PCAR	X	X	X
MMSCBA	X		
SSCR	X	X	X

e ao MMSCBA. Ambos cobrem apenas a etapa de extração. O ACRIPPER é, na verdade, mais uma adaptação do algoritmo RIPPER (COHEN, 1995) do que uma adaptação de um algoritmo de CA. Em relação às etapas de ranqueamento e poda, o ACRE cobre apenas a poda. Dos oito algoritmos que realizam tanto o ranqueamento quanto a poda, cinco deles primeiro ordenam e depois podam (IARCID, CBA Adaptado via ModifiedLift, ACAR, CWCAR e PCAR), um poda e depois ranqueia (SSCR), um mescla ranqueamento, poda e ranqueamento (ARCID) e um mescla poda, ranqueamento e poda (CBA Adaptado via PoI). Dadas as variações, é interessante explorar o impacto da ordem das etapas em relação ao desempenho do classificador, uma vez que cada uma desempenha um papel diferente no processo: o ranqueamento, em geral, visa apoiar a escolha das melhores regras a serem incluídas no modelo final, enquanto a poda reduz o espaço de busca.

QP-2. Quais estratégias os algoritmos propostos utilizam para realizar a extração

de regras? Os algoritmos de extração de regras são impactados em contextos desbalanceados, uma vez que, se o suporte mínimo é definido como muito alto, as regras da classe minoritária não são extraídas; por outro lado, se definido como muito baixo, muitas regras irrelevantes são geradas, resultando em um grande volume de regras. Os algoritmos selecionados basicamente utilizam duas estratégias, como pode ser visto na Tabela 4: (i) extrair as regras do conjunto de

Tabela 4 – Estratégias utilizadas para se extrair regras com respectivos algoritmos. Algoritmos marcados com "*" são variações do Apriori.

	Conjunto Total			**Conjunto Particionado**	
	Múltiplos sup-min (Classe)	**Único sup-min**	**Outros**	**Múltiplos sup-min (Classe)**	**Múltiplos sup-min (Item)**
ARCID				X (IGB)	
IARCID				X (IGB)	
CBA Adaptado via ModifiedLift		X (Apriori)			
ACAR				X (Apriori)	
CWCAR				X (Apriori*)	
CBA Adaptado via PoI		X (Apriori)			
ACRIPPER			X (RIPPER)		
ACRE	X (Apriori)				
PCAR	X (Apriori)				
MMSCBA					X (MS-Apriori*)
SSCR		X (Apriori)			

dados como um todo; (ii) particionar o conjunto de dados por classe e extrair as regras em cada partição, i.e., em cada classe. Além disso, em relação ao item (i), não considerando o algoritmo ACRIPPER, que apresenta uma estratégia de extração muito diferente dos outros algoritmos, nota-se o uso de (i.i) múltiplos suportes mínimos (**sup-min**) em dois algoritmos (18,18%), i.e., um para cada classe, ou (i.ii) um único **sup-min** em três algoritmos (27,27%), i.e., o mesmo para todas as classes. Em relação ao item (ii) nota-se o uso de (ii.i) múltiplos **sup-min** em quatro algoritmos (36,36%), i.e., um para cada classe, ou (ii.ii) múltiplos **sup-min** por item em um algoritmo (9,09%), i.e., para cada classe, cada um dos itens contém seu respectivo **sup-min**. Um aspecto relacionado a essas estratégias ((ii.i), (ii.ii)) é que as MOs computadas para as regras de cada partição levam em consideração apenas as instâncias da própria classe, fazendo com que os valores das medidas sejam "locais" e não "globais". Por exemplo, suponha que a regra $r : A \Rightarrow c_1$ foi gerada para a classe c_1. A confiança ($\frac{A \cup c_1}{A}$) dessa regra será 1 (100%), já que toda transação em que A ocorre, c_1 também ocorre. Em outras palavras, não se tem o valor de suporte

"global" de A, i.e., o suporte de A em todas as classes. Assim, em alguns casos, é necessária uma passada adicional pelo conjunto de dados, como visto no algoritmo MMSCBA. Vale mencionar que a diferença entre a estratégia (i.i) e (ii.i) é que, na primeira, apresentada em ACRE e PCAR, um único suporte é informado, que é usado como base para computar o suporte de cada classe; na segunda, apresentada em ARCID, IARCID, ACAR e CWCAR, cada classe pode ser configurada com um **sup-min** distinto. Portanto, dadas as variações, há necessidade de se explorar o impacto das estratégias mais utilizadas em relação ao desempenho do classificador.

QP-3. Quais algoritmos de extração de regras os algoritmos propostos utilizam? Observou-se que a maioria dos trabalhos (8/72,73%) utiliza algoritmos baseados no Apriori, conforme mostrado na Tabela 4. ARCID e IARCID utilizam IGB, mas não há detalhes sobre como ele é utilizado para extrair as regras, já que adaptações seriam necessárias (não extrai diretamente as regras).

QP-4. Quais medidas objetivas são usadas quando a etapa de ranqueamento está presente? Existem outros critérios envolvidos? Considerando os oito algoritmos apresentados na Tabela 3 que abrangem a etapa de ranqueamento, observa-se que a maioria deles (7/87,50%) utiliza MOs para ranquear as regras, como pode ser visto na Tabela 5, sendo "Confiança",

Tabela 5 – MOs usadas na etapa de ranqueamento. "+MO:ND" significa outra MO, embora **N**ão **D**efinida.

	Medidas	**Outros**	
ARCID	X		Lift, Laplace, +OM:ND
IARCID	X		Lift, Laplace, +OM:ND
CBA Adaptado via ModifiedLift	X		ModifiedLift, DM2, DM3, DM4
ACAR	X	X	Confiança, Suporte
CWCAR	X	X	Suporte Ponderado
CBA Adaptado via PoI	X		Confiança, Suporte
PCAR	X		Confiança
SSCR		X	

"Suporte" e "Lift" as mais comuns, assim como suas variações (as MOs utilizadas em cada trabalho estão apresentadas na tabela). Apenas o SSCR utiliza outra estratégia, baseada em técnicas sensíveis a custo. Vale mencionar que ACAR e CWCAR combinam MOs com critérios de cardinalidade e ordem de geração, assim como o SSCR, que também utiliza cardinalidade. Esses dois critérios são, em geral, usados como métodos de desempate. No entanto, várias MOs existem na literatura, como aquelas descritas em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022). Assim, é necessário explorar o impacto de outras MOs na etapa de ranqueamento de regras. O único trabalho encontrado nessa linha é o apresentado em Yang e Cui (2015), que de acordo com Dall'Agnol e Carvalho (2023) apresenta inconsistências (vide Seção 2.3).

QP-5. Se a etapa de poda existe, como ela é realizada? Não foi possível observar um padrão entre os nove algoritmos apresentados na Tabela 3 que abrangem a etapa de poda. No entanto, pode-se mencionar que:

- no caso de ARCID e IARCID, propostos pelos mesmos autores, a poda ocorre pela interseção de listas obtidas de diferentes ranqueamentos, a fim de manter regras relevantes para ambas as classes, ou seja, minoritária e majoritária. Os autores não apresentam nenhum tipo de análise em relação ao resultado obtido após a poda, uma vez que a interseção pode ser vazia ou gerar um conjunto de regras muito reduzido, o que impacta o desempenho final do classificador;
- no caso de ACAR e CWCAR, propostos pelos mesmos autores, a poda é realizada através de regras específicas (regras conflitantes, regras redundantes, etc.);
- no caso de CBA Adaptado via ModifiedLift, CBA Adaptado via PoI e PCAR, a poda ocorre de maneira semelhante a do CBA;
- no caso de ACRE a poda ocorre por meio de um processo que mistura cobertura (como no CBA) com taxa de erro;
- no SSCR a poda é realizada através de testes estatísticos.

Assim, observou-se, ao contrário da etapa de ranqueamento, que a poda é mais dependente da solução algorítmica proposta.

QP-6. Em relação à predição, os algoritmos propostos utilizam a estratégia "Rule List" ou "Rule Set"? Foi observado um empate entre as estratégias utilizadas pelos algoritmos, como pode ser visto na Tabela 6. MMSCBA não foi contado aqui já que utiliza ambas as estratégias. Na estratégia "Rule List" uma única regra é usada para determinar a classe, geralmente a de melhor precedência que casa com a instância, enquanto na estratégia "Rule Set" um conjunto de regras é utilizado para se definir a classe. No entanto, de acordo com Kliegr (2019), a estratégia "Rule List" garante melhor interpretabilidade, uma vez que apenas uma regra é ativada e a explicação do porquê a predição ocorreu é direta. Assim, maiores esforços poderiam ser feitos em relação às soluções "Rule List".

QP-7. Os algoritmos propostos funcionam com classificação binária e/ou multiclasse? Nota-se que a maioria dos algoritmos (8/72,73%) é viável apenas para problemas binários, com CBA Adaptado via ModifiedLift, ACRIPPER e MMSCBA sendo as exceções (multiclasse), como pode ser visto na Tabela 7. Assim, há uma falta de soluções multiclasse. Além disso, a maneira como a saída do classificador é apresentada também foi identificada, i.e., de maneira discreta ou probabilística (pontuação). A maioria deles (9/81,82%) trabalha indicando exatamente a classe à

Tabela 6 – Estratégias usadas na predição.

	Rule List	**Rule Set**
ARCID	X	
IARCID	X	
CBA Adaptado via ModifiedLift	X	
ACAR		X
CWCAR		X
CBA Adaptado via PoI	X	
ACRIPPER		X
ACRE		X
PCAR	X	
MMSCBA	X	X
SSCR		X

Tabela 7 – Características dos algoritmos.

	Classificador Binário	**Saída Discreta**
ARCID	X	X
IARCID	X	X
CBA Adaptado via ModifiedLift		X
ACAR	X	X
CWCAR	X	X
CBA Adaptado via PoI	X	X
ACRIPPER		X
ACRE	X	
PCAR	X	
MMSCBA		X
SSCR	X	X

qual a instância pertence (saída discreta). Por outro lado, ACRE e PCAR são "ranqueadores", uma vez que apenas relatam uma pontuação, indicando a probabilidade de a instância pertencer à classe minoritária. Neste caso, cabe ao usuário definir um limite que determine se a instância pertence ou não à classe minoritária. No entanto, os autores dos trabalhos citados não mencionam como isso é feito.

3.3 Considerações Finais

Este capítulo apresentou uma RSL com o objetivo de identificar e analisar estudos disponíveis na literatura focados em soluções de classificação associativa algorítmicas para lidar com dados desbalanceados. Considerando os resultados, é possível entender como os algoritmos propostos identificados na RSL funcionam. É possível mencionar que, na maioria dos casos, os algoritmos propostos:

- cobrem as etapas básicas de um algoritmo de classificador associativo, i.e., extração, ranqueamento e poda;
- utilizam diferentes estratégias de extração de regras, baseando-se, em geral, em múltiplos suportes para obter regras pertencentes à classe minoritária e reduzir o número total de regras;
- utilizam o algoritmo Apriori como base para extrair regras;
- utilizam MOs para ranquear as regras, focando em poucas delas (Lift, Confiança, Suporte e variações dessas);
- utilizam diferentes estratégias de poda, sem um padrão definido;
- utilizam estratégias de "Rule List" e "Rule Set" para realizar a predição;
- são viáveis para problemas binários, utilizando uma predição discreta como saída.

Considerando o exposto, há algumas lacunas que podem ser exploradas. Uma delas refere-se ao impacto de diferentes estratégias levantadas para se realizar a extração de regras em relação ao impacto das mesmas no desempenho dos classificadores. Outra questão é a exploração do impacto de outras MOs na etapa de ranqueamento das regras, já que vários delas são encontradas na literatura, como as descritas em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022). Em relação à interpretabilidade dos modelos, metade dos algoritmos utiliza a estratégia "Rule Set", fazendo com que um conjunto de regras seja utilizado durante a predição. No entanto, dessa maneira, uma das principais características dos classificadores associativos é reduzida, já que se perde a oportunidade de usar uma única regra para fornecer a justificativa da predição. Assim, é interessante avaliar este impacto e pensar mais sobre a estratégia "Rule List". Por fim, há poucos trabalhos focados em problemas multiclasse, bem como em saídas probabilísticas. Dependendo do problema, ambos os aspectos tornam-se relevantes, exigindo um esforço adicional nesta área.

Diante do exposto, considera-se que esta RSL contribui para a área, apresentando detalhes sobre o tema e suas lacunas. Vale mencionar que uma dificuldade observada nesta área é a falta de

pacotes (e/ou implementações) que incorporem algoritmos desse tipo. Portanto, seria interessante desenvolver e/ou aumentar pacotes voltados para algoritmos de classificação associativa para dados desbalanceados.

4 Análise das Estratégias de Extração de Regras

Como levantado no capítulo anterior, os trabalhos da literatura, relacionados aos algoritmos de classificação associativa em contextos desbalanceados, utilizam diferentes estratégias de extração de regras a fim de garantir a obtenção de regras pertencentes à classe minoritária, assim como reduzir o número total de regras geradas (evitar explosão combinatória). Ademais, utilizam em geral o algoritmo Apriori (ou variações) para realizar a extração das regras em conjunto com a estratégia adotada. Assim, este capítulo realiza uma análise das estratégias de extração adotadas a fim de verificar se existe, de fato, diferença entre elas no que se refere ao desempenho e a interpretabilidade do modelo final obtido. A análise visa apoiar pesquisas futuras no sentido de direcionar os trabalhos para a estratégia mais adequada. Para tanto, a Seção 4.1 apresenta a metodologia adotada, a Seção 4.2 a configuração experimental e a Seção 4.3 os resultados e discussões.

4.1 Metodologia

A metodologia de análise proposta foi elaborada tendo como base o algoritmo CBA que, como já mencionado, é o utilizado como *baseline* para comparação com novas propostas/soluções. Ademais, ele contempla todas as etapas utilizadas, em geral, pelos demais algoritmos da literatura (vide Capítulo 3). A Figura 3 apresenta a metodologia adotada: dado um conjunto de dados, uma estratégia de extração de regras é executada e as regras obtidas. Na sequência, as etapas de ranqueamento, poda e predição são executadas como no CBA. Em outras palavras, o CBA foi modificado de modo a realizar a extração de regras de quatro maneiras diferentes, a fim de avaliar o impacto das mesmas no desempenho e na interpretabilidade dos modelos gerados. Contudo, como um método automático de seleção de MOs é proposto no próximo capítulo, a etapa de ranqueamento acabou sendo alterada também, a fim de considerar não apenas a ordenação por meio da medida confiança (CBA), mas também de outras 47 medidas analisadas posteriormente (Capítulo 5) (o símbolo "||" na figura indica o conectivo OU). O objetivo foi verificar se os resultados obtidos pelas estratégias de extração se mantinham invariáveis em relação a MO utilizada na etapa de ranqueamento, já que uma seleção automática de medidas é realizada no próximo capítulo.

As quatro estratégias exploradas foram as levantadas no capitulo anterior, a saber:

- Apriori-T: o algoritmo Apriori é executado em todo o conjunto de dados e a frequência de

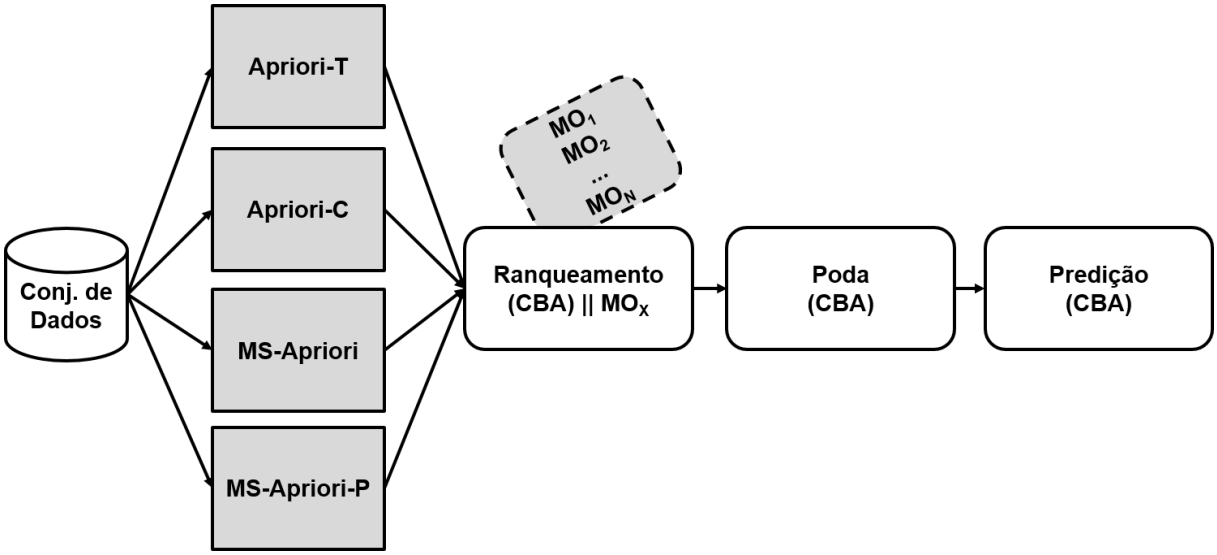


Figura 3 – Metodologia de análise proposta.

ocorrência mínima dos *itemsets* é constante independente da classe. Neste caso, a frequência mínima é calculada utilizando-se o suporte mínimo especificado (denominado aqui de β) aplicado sobre o número de instâncias pertencentes a classe minoritária. Por exemplo, suponha um conjunto de dados com 80 instâncias pertencentes a classe majoritária e 20 a classe minoritária. Se $\beta = 10\%$, a frequência mínima de um dado *itemset* referente a qualquer uma das classes será de 2 ocorrências. Essa é a estratégia adotada nos algoritmos CBA Adaptado via ModifiedLift, CBA Adaptado via PoI e SSCR.

- Apriori-C: o algoritmo Apriori é executado em todo o conjunto de dados, porém a frequência de ocorrência mínima dos *itemsets* é dependente da classe. Neste caso, a frequência mínima é calculada utilizando-se o suporte mínimo especificado (β) aplicado sobre o número de instâncias pertencentes a cada classe. Por exemplo, suponha um conjunto de dados com 80 instâncias pertencentes a classe majoritária e 20 a classe minoritária. Se $\beta = 10\%$, a frequência mínima dos *itemsets* refentes a classe minoritária será de 2 ocorrências e a dos *itemsets* refentes a classe majoritária será de 8 ocorrências. Essa é a estratégia adotada nos algoritmos ACRE e PCAR. De fato, esta estratégia é a utilizada pelo CBA2, i.e., a execução aqui se refere ao CBA2, baseline do CBA para contextos desbalanceados.

- MS-Apriori: o algoritmo MS-Apriori é executado em todo o conjunto de dados, porém a frequência de ocorrência mínima dos *itemsets* é dependente da frequência mínima dos itens individuais (vide MS-Apriori na Seção 2.1). Neste caso, a frequência mínima de cada item é calculada utilizando-se o suporte mínimo especificado (β) aplicado sobre a frequência total do item, i.e., sua ocorrência em todo o conjunto de dados. Embora essa estratégia não tenha sido adotada em nenhum dos algoritmos levantados, optou-se por utilizá-la por ser uma variação da estratégia abaixo descrita utilizada pelo MMSCBA.

- MS-Apriori-P: a diferença desta estratégia em relação a anterior, adotada pelo algoritmo MMSCBA, é que em vez de se aplicar o MS-Apriori no todo, as instâncias são particionadas em grupos, referente a cada uma das classes, e o processo ocorre dentro da partição. A diferença então é que a frequência mínima dos itens individuais é computada na partição e não no todo. Por exemplo, caso um item i apareça tanto na classe majoritária quanto na minoritária, ele tem uma frequência mínima diferente em cada partição e afeta, portanto, a frequência mínima dos *itemsets* gerados em cada classe. Na estratégia anterior a frequência mínima é calculada independente da classe, i.e., em todo o conjunto de dados.

Por fim, vale mencionar que a estratégia adotada pelos algoritmos ARCID, IARCID, ACAR e CWCAR não foi considerada. Nestes algoritmos as instâncias são primeiramente particionadas em grupos, referente a cada uma das classes, para depois aplicar-se o CBA2 (estratégia Apriori-C). Contudo, realizar ou não o particionamento não afeta o resultado, já que a frequência mínima de ocorrência é afetada pela classe; assim, a frequência é a mesma tanto no todo quanto na partição.

4.2 Configuração Experimental

A fim de executar a metodologia apresentada na Figura 3, é necessário que alguns detalhes, relacionados a execução dos experimentos, sejam aqui descritos.

Conjuntos de Dados. 87 conjuntos de dados, disponíveis no repositório da ferramenta KEEL[1], foram utilizados. O repositório disponibiliza 145 conjuntos ao todo (seção "Imbalanced data sets for classification"). Contudo, apenas aqueles referentes a problemas binários foram selecionados (classificação binária)[2]. O motivo se deve ao fato da maior parte dos algoritmos levantados no capítulo anterior trabalharem com classificação binária e do método proposto no próximo capítulo ter sido proposto para contextos binários. A Tabela 8 apresenta os conjuntos selecionados. Os conjuntos apresentam características diversas, sendo as apresentadas as seguintes: número de instâncias (#Transações), número de atributos (#Atributos), valores únicos para cada atributo (#Itens-Distintos), IR (Imbalanced Ratio) referente a taxa de desbalanceamento dos conjuntos ($IR = \frac{\#Maj}{\#Min}$).

[1] <https://sci2s.ugr.es/keel/datasets.php>.

[2] Dos 145 conjuntos, 100 são referentes a problemas binários. Contudo, 13 deles não foram considerados por não terem instâncias suficientes de ambas as classes para se executar o Apriori nos βs especificados. Os 13 conjuntos são dermatology-6, kddcup-buffer_overflow_vs_back, kddcup-guess_passwd_vs_satan, kddcup-land_vs_portsweep, kddcup-land_vs_satan, kddcup-rootkit-imap_vs_back, lymphography-normal-fibrosis, page-blocks0, segment0, vehicle0, vehicle1, vehicle2 e vehicle3.

Tabela 8 – Características dos conjuntos de dados utilizados nos experimentos.

Conjunto de Dado	#Transações	#Atributos	#Itens-Distintos	IR
abalone-17_vs_7-8-9-10	2338,0	8,0	110,0	39,31
abalone-19_vs_10-11-12-13	1622,0	8,0	99,0	49,69
abalone-20_vs_8-9-10	1916,0	8,0	83,0	72,69
abalone-21_vs_8	581,0	8,0	45,0	40,5
abalone-3_vs_11	502,0	8,0	5,0	32,47
abalone19	4174,0	8,0	149,0	129,44
abalone9-18	731,0	8,0	88,0	16,4
car-good	1728,0	6,0	1728,0	24,04
car-vgood	1728,0	6,0	1728,0	25,58
cleveland-0_vs_4	177,0	10,0	103,0	12,62
ecoli-0-1-3-7_vs_2-6	281,0	6,0	38,0	39,14
ecoli-0-1-4-6_vs_5	280,0	5,0	31,0	13,0
ecoli-0-1-4-7_vs_2-3-5-6	336,0	6,0	25,0	10,59
ecoli-0-1-4-7_vs_5-6	332,0	6,0	37,0	12,28
ecoli-0-1_vs_2-3-5	244,0	7,0	35,0	9,17
ecoli-0-1_vs_5	240,0	5,0	30,0	11,0
ecoli-0-2-3-4_vs_5	202,0	5,0	30,0	9,1
ecoli-0-2-6-7_vs_3-5	224,0	5,0	25,0	9,18
ecoli-0-3-4-6_vs_5	205,0	5,0	15,0	9,25
ecoli-0-3-4-7_vs_5-6	257,0	5,0	27,0	9,28
ecoli-0-3-4_vs_5	200,0	5,0	28,0	9,0
ecoli-0-4-6_vs_5	203,0	5,0	26,0	9,15
ecoli-0-6-7_vs_3-5	222,0	5,0	21,0	9,09
ecoli-0-6-7_vs_5	220,0	5,0	38,0	10,0
ecoli-0_vs_1	220,0	5,0	17,0	1,86
ecoli1	336,0	5,0	30,0	3,36
ecoli2	336,0	6,0	67,0	5,46
ecoli3	336,0	5,0	35,0	8,6
ecoli4	336,0	6,0	44,0	15,8
flare-F	1066,0	11,0	287,0	23,79
glass-0-1-2-3_vs_4-5-6	214,0	9,0	62,0	3,2
glass-0-1-4-6_vs_2	205,0	9,0	92,0	11,6
glass-0-1-5_vs_2	172,0	9,0	56,0	9,12
glass-0-1-6_vs_2	192,0	9,0	78,0	10,29

Continuação da Tabela 8 .

Conjunto de Dado	#Transações	#Atributos	#Itens-Distintos	IR
glass-0-1-6_vs_5	184,0	9,0	55,0	19,44
glass-0-4_vs_5	92,0	9,0	31,0	9,22
glass-0-6_vs_5	108,0	9,0	30,0	11,0
glass0	214,0	9,0	62,0	2,06
glass1	214,0	9,0	85,0	1,82
glass2	214,0	9,0	84,0	11,59
glass4	214,0	9,0	58,0	15,47
glass5	214,0	9,0	63,0	22,78
glass6	214,0	9,0	63,0	6,38
haberman	306,0	3,0	8,0	2,78
iris0	150,0	4,0	6,0	2,0
kr-vs-k-one_vs_fifteen	2244,0	6,0	2244,0	27,77
kr-vs-k-three_vs_eleven	2935,0	6,0	2935,0	35,23
kr-vs-k-zero-one_vs_draw	2901,0	6,0	2901,0	26,63
kr-vs-k-zero_vs_eight	1460,0	6,0	1460,0	53,07
kr-vs-k-zero_vs_fifteen	2193,0	6,0	2193,0	80,22
led7digit-0-2-4-5-6-7-8-9_vs_1	443,0	3,0	23,0	10,97
new-thyroid1	215,0	5,0	25,0	5,14
new-thyroid2	215,0	5,0	22,0	5,14
page-blocks-1-3_vs_4	472,0	10,0	106,0	15,86
pima	768,0	8,0	209,0	1,87
poker-8-9_vs_5	2075,0	10,0	118,0	82,0
poker-8-9_vs_6	1485,0	9,0	1309,0	58,4
poker-8_vs_6	1477,0	10,0	1271,0	85,88
poker-9_vs_7	244,0	10,0	225,0	29,5
shuttle-2_vs_5	3316,0	9,0	78,0	66,67
shuttle-6_vs_2-3	230,0	9,0	43,0	22,0
shuttle-c0-vs-c4	1829,0	9,0	57,0	13,67
shuttle-c2-vs-c4	129,0	9,0	24,0	20,5
vowel0	988,0	13,0	257,0	9,98
winequality-red-3_vs_5	691,0	11,0	160,0	68,1
winequality-red-4	1599,0	11,0	290,0	29,17
winequality-red-8_vs_6	656,0	11,0	167,0	35,44
winequality-red-8_vs_6-7	855,0	11,0	241,0	46,5
winequality-white-3-9_vs_5	1482,0	11,0	139,0	58,28

Continuação da Tabela 8 .

Conjunto de Dado	#Transações	#Atributos	#Itens-Distintos	IR
winequality-white-3_vs_7	900,0	11,0	73,0	44,0
winequality-white-9_vs_4	168,0	11,0	116,0	32,6
wisconsin	683,0	9,0	284,0	1,86
yeast-0-2-5-6_vs_3-7-8-9	1004,0	8,0	63,0	9,14
yeast-0-2-5-7-9_vs_3-6-8	1004,0	8,0	72,0	9,14
yeast-0-3-5-9_vs_7-8	506,0	8,0	57,0	9,12
yeast-0-5-6-7-9_vs_4	528,0	8,0	46,0	9,35
yeast-1-2-8-9_vs_7	947,0	8,0	44,0	30,57
yeast-1-4-5-8_vs_7	693,0	8,0	61,0	22,1
yeast-1_vs_7	459,0	7,0	39,0	14,3
yeast-2_vs_4	514,0	6,0	61,0	9,08
yeast-2_vs_8	482,0	8,0	45,0	23,1
yeast1	1484,0	8,0	120,0	2,46
yeast3	1484,0	8,0	67,0	8,1
yeast4	1484,0	8,0	77,0	28,1
yeast5	1484,0	8,0	74,0	32,73
yeast6	1484,0	8,0	49,0	41,4
zoo-3	101,0	16,0	59,0	19,2

Pré-processamento. Todos os conjuntos de dados utilizados foram pré-processados. Atributos numéricos foram discretizados, tanto os reais quanto os inteiros contendo mais de dez valores únicos. Vale mencionar que a fim de evitar *data leakage*, a discretização foi realizada somente nos *folds* utilizados como treinamento em uma dada execução da validação cruzada. Em outras palavras, as transformações foram realizadas somente no conjunto de treinamento e aplicadas, posteriormente, no conjunto de teste. O algoritmo de discretização utilizado foi o proposto por Fayyad e Irani (1993). Após a discretização, atributos que apresentassem valores únicos eram excluídos.

Extração de Regras. Conforme visto na Figura 3, quatro estratégias distintas foram selecionadas, tendo como base os algoritmos levantados no capítulo anterior, para realizar a extração das regras, a saber: Apriori-T, Apriori-C, MS-Apriori e MS-Apriori-P. Em todas elas, para que as regras sejam extraídas, é necessário que se especifique, como visto na Seção 2.1, o suporte mínimo (β) e a confiança mínima. A fim de evitar interferências nos resultados apresentados optou-se por zerar o valor da confiança mínima, extraindo-se, portanto, todas as possíveis regras dentro do β especificado. Já em relação ao suporte mínimo, a fim de padronizar os experimentos e analisar o seu impacto nas estratégias de extração, variou-se o mesmo da

seguinte maneira: $\beta = 5\%$, $\beta = 10\%$, $\beta = 15\%$, $\beta = 20\%$ e $\beta = 25\%$. Conforme descrito na metodologia (Seção 4.1), esse valor relativo (porcentagem) é usado para se computar a frequência de ocorrência mínima dos *itemsets* em cada um dos casos explorados. Por fim, definiu-se que o tamanho máximo do *itemset* seria igual a 5 itens, o que implica em regras com no máximo 4 itens no antecedente.

Medidas Objetivas. Das 61 MOs apresentadas em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022), este trabalho utilizou 44, uma vez que existem equivalências entre elas (11) (sublinhadas a seguir) e que algumas apresentam alto custo computacional (6)[3], já que apresentam aspectos combinatoriais em suas definições, e, portanto, não foram consideradas. As 44 MOs consideradas foram (1) Odds Ratio (=Yule's Y, Yule's Q); (2) F-Measure (=Kulczynski 1, Jaccard); (3) Lift (=Information Gain); (4) Loevinger (=Conviction); (5) Odd Multiplier (=Zhang); (6) Confidence (=Ganascia, Example and Counterexample Rate, Sebag–Schoenauer, Laplace Correction); (7) Support; (8) Prevalence; (9) K-Measure; (10) Least Contradiction; (11) Confirm Descriptive; (12) Complement Class Support; (13) Leverage; (14) Confidence Causal; (15) Confirmed Confidence Causal; (16) Directed Information Ratio; (17) Confirm Causal; (18) Putative Causal Dependency; (19) Klosgen; (20) Added Value; (21) 1-Way Support; (22) Kulczynski 2; (23) Goodman–Kruskal; (24) Accuracy; (25) Cosine; (26) Piatetsky-Shapiro; (27) 2-Way Support; (28) Collective Strength; (29) Kappa; (30) Correlation Coefficient; (31) Theil Uncertainty Coefficient; (32) Mutual Information; (33) Chi-Square; (34) J-Measure; (35) Gini Index; (36) Normalized Mutual Information; (37) Recall; (38) Specificity; (39) Relative Risk; (40) Logical Necessity; (41) Conditional Entropy; (42) Coverage; (43) Implication Index; (44) TIC. As equações referentes às suas definições não são apresentadas aqui, mas podem ser encontradas em Tew *et al.* (2014). Além dessas, mais 4 MOs foram consideradas, a saber: ModifiedLift, DM_2, DM_3 e DM_4, apresentadas em Hassine, Abdellatif e Yahia (2022) referente ao algoritmo CBA Adaptado via ModifiedLift identificado nos trabalhos relacionados apresentados no capítulo anterior. A ModifiedLift é uma adaptação da medida Lift e as demais adaptações da própria medida por eles proposta. Uma vez que este algoritmo é utilizado como um dos *baselines* no próximo capítulo, as mesmas foram também incluídas. Deste modo, ao todo, 48 MOs foram utilizadas.

Critério de Avaliação. Como já mencionado na Seção 2.4, as medidas consideradas foram a F1 e a G-Mean, no quesito desempenho, e o tamanho do modelo $\mathcal{L}$, no quesito interpretabilidade, estimadas via 2-fold cross-validation estratificado. A fim de comparar os resultados obtidos, testes estatísticos foram realizados, utilizando-se para tanto o teste de Friedman com $\alpha = 0{,}05$ e o teste post-hoc de Nemenyi, juntamente com os diagramas de diferença crítica (CD). A escolha dos mesmos foi baseada em Demsar (2006).

[3] (1) Indice Probabiliste d'Ecart d'Equilibre; (2) EII1; (3) EII2; (4) Dilated Chi-Square; (5) Intensity of Implication; (6) Interestingness Weighting Dependency.

4.3 Resultados e Discussão

Considerando os fluxos possíveis derivados da Figura 3, obtém-se ao todo 960 possibilidades (4 estratégias de extração * 5 βs distintos * 48 MOs). Cada uma delas induz um modelo distinto, o qual é avaliado em termos de desempenho (F1, G-Mean) e interpretabilidade (tamanho do modelo). Os 960 fluxos foram explorados em 87 conjuntos de dados, totalizando ao todo 167.040 experimentos (960*87*2), já que as medidas foram estimadas, em cada conjunto, via 2-fold cross-validation estratificado. Assim, devido a grande quantidade de experimentos realizados, a análise foi divida em etapas, as quais são apresentadas a seguir.

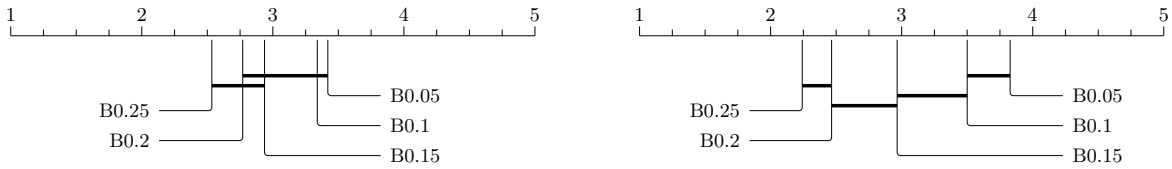
Inicialmente avaliou-se o impacto do valor de β em cada uma das estratégias de extração de regras. De fato, optou-se por variar β a fim de selecionar um valor adequado a todos os experimentos visando as demais análises, i.e., buscou-se garantir que este valor não interferisse nas conclusões das demais análises. Para tanto, dividiu-se esta análise em duas partes: uma considerando apenas o fluxo do CBA, i.e., em que o ranqueamento é realizado por meio da medida Confiança (Confidence), e a outra considerando diferentes ranqueamentos, um para cada MO considerada.

No que se refere a análise dos βs em relação ao fluxo do CBA, para cada estratégia de extração, e medida de avaliação, uma tabela como a apresentada na Tabela 9 foi gerada. Os resultados se referem a média de desempenho (neste caso, F1) dos modelos gerados em ambos os folds. Para cada tabela (12 no total (4 estratégias * 3 medidas de avaliação)), o teste de Friedman foi aplicado a fim de verificar a existência ou não de diferença significativa entre os βs. Quando detectada diferença, o teste post-hoc de Nemenyi foi aplicado. Para as medidas de desempenho, i.e., F1 e G-Mean, o teste de Friedman não detectou diferença entre os β's em nenhuma das estratégias de extração. Já para a medida tamanho do modelo ($\mathcal{L}$), o teste de Friedman detectou diferenças entre os β's em todas as estratégias de extração e, portanto, o teste post-hoc de Nemenyi foi aplicado e os gráficos de diferença crítica (CD) gerados, os quais encontram-se apresentados na Figura 4. Nos gráficos de CD as linhas que saem do eixo enumerado indicam o rank médio da respectiva configuração. As linhas que encontram-se conectadas por uma barra correspondem as configurações que não apresentam diferença crítica entre si. A análise realizada, por exemplo, na Figura 4a, referente a estratégia de extração Apriori-T, indica que o β=25% apresenta aproximadamente 2,5 de rank médio, não apresentando diferença em relação aos β's iguais a 20% e 15%; contudo, apresenta diferença entre os β's 10% e 5%. Assim, analisando os gráficos de CD, nota-se que nenhum dos β's se destacou, i.e., apareceu em um grupo isolado e separado dos demais. Contudo, é possível observar, em todas as estratégias, a seguinte ordem (via rank médio) em termos de desempenho: β=25%, β=20%, β=15%, β=10% e β=5%. Este resultado para a medida tamanho do modelo pode ser explicado pelo seguinte: quanto maior a frequência de ocorrência mínima, menos *itemsets* são gerados e, consequentemente menos

regras. Ademais, como as regras passam por um processo de ranqueamento e poda para geração do modelo final, pode ser que regras mais específicas geradas pelos βs menores acabem sendo ranqueadas primeiro e mais regras acabem tendo de ser selecionadas após a poda para obtenção do modelo. Assim, uma vez que não existe diferença entre as estratégias em relação ao F1 e G-Mean, pode-se dizer que em relação ao tamanho do modelo um valor de β=25% se mostra adequado a todas as estratégias avaliadas no fluxo do CBA.

Tabela 9 – Parte dos resultados obtidos via estratégia de extração Apriori-C, no fluxo do CBA (Confiança), na medida F1.

Conjunto de Dado	**β=5%**	**β=10%**	**β=15%**	**β=20%**	**β=25%**
abalone-17_vs_7-8-9-10	0.5086493718	0.5086493718	0.5086493718	0.5086493718	0.4937202252
abalone-19_vs_10-11-12-13	0.49501868	0.49501868	0.49501868	0.49501868	0.49501868
abalone-20_vs_8-9-10	0.4960541502	0.4960541502	0.4960541502	0.4965843405	0.4965843405
abalone-21_vs_8	0.6822426298	0.6318458495	0.6609699567	0.6609699567	0.6609699567
abalone-3_vs_11	0.9828199863	0.9828199863	0.9828199863	0.9828199863	0.9828199863
...	...	...	...	...	...
zoo-3	0.5818513746	0.5818513746	0.5818513746	0.5818513746	0.5818513746



(a) Apriori-T

(b) Apriori-C

(c) MS-Apriori

(d) MS-Apriori-P

Figura 4 – Gráficos de diferença crítica, para comparação dos βs, em relação ao tamanho dos modelos gerados.

Visando verificar se o mesmo comportamento se mantém, independente da MO utilizada para se realizar o ranqueamento, uma análise mais abrangente foi realizada. Neste caso, 576 tabelas como a Tabela 9 foram geradas: 4 estratégias * 3 medidas de avaliação * 48 MOs distintas para se realizar o ranqueamento. Considerando as tabelas geradas, o mesmo procedimento anteriormente descrito foi aplicado, i.e., teste de Friedman seguido, quando possível, do teste post-hoc de Nemenyi. Como fica inviável discutir tantos resultados como anteriormente, os mesmos foram sumarizados, os quais encontram-se apresentados na Tabela 10. Essa tabela foi construída da seguinte maneira: para uma dada estratégia de extração e medida de avaliação, considerando uma certa MO, avalia-se seu gráfico de CD correspondente. Por exemplo, a

Figura 4b se refere a estratégia Apriori-C avaliada pela medida $\mathcal{L}$, considerando a MO Confiança (fluxo CBA). Para cada gráfico uma correspondência entre ele e as sub-matrizes da Tabela 10 é encontrada. As células destacadas em cinza na tabela correspondem a configuração da Figura 4b. Vale mencionar que todos os gráficos de CD referente a combinação Apriori-C, em relação ao tamanho do modelo, independente da MO utilizada, estão relacionados a esta sub-matriz em destaque (são 48 gráficos de CD por sub-matriz, uma para cada MO explorada). Para cada gráfico, a seguinte análise é realizada: caso um dado β seja estatisticamente diferente de outros βs e seu rank médio seja menor que os mesmos, i.e., apresente um resultado melhor em relação aos demais, soma-se 1 na célula correspondente a fim de computar o seu ganho em relação aos demais; caso não haja diferença entre um dado β e os demais nenhum computo é realizado. Por exemplo, tomando como base o exemplo da Figura 4b, destacada na Tabela 10, tem-se que o β=25% apresenta diferença em relação aos βs 5%, 10% e 15%, e, portanto, as posições [B25%, B5%], [B25%, B10%] e [B25%, B15%] da tabela referente a estratégia Apriori-C receberiam um ponto. As matrizes, portanto, começam zeradas. Desta maneira, é possível verificar, em cada linha, para cada estratégia e medida de avaliação, os βs que "ganharam" dos demais, de modo a verificar o mais adequado. Por exemplo, na Tabela 10, estratégia Apriori-C, nota-se que o β=25% é melhor que o β=5% 21 vezes, do β=10% 19 vezes e do β=15% 13 vezes, quando a avaliação é feita pela medida tamanho do modelo. Olhando-se um dado β por coluna, tem-se o oposto, i.e., quantas vezes ele perdeu dos demais βs. Por exemplo, na Tabela 10, estratégia Apriori-C, nota-se que o β=5% perde do β=15% 17 vezes, do β=20% 20 vezes e do β=25% 21 vezes, quando a avaliação é feita pela medida tamanho do modelo. O valor máximo de cada célula é 48, número máximo de MOs exploradas. As células em branco indicam que não houve diferença estatística, i.e., que ninguém ganhou nem perdeu (as células estariam zeradas, mas para evitar uma "confusão" visual optou-se por deixá-las em branco). Deste modo, é possível notar que embora quase não haja diferença entre os βs em relação as medidas F1 e G-Mean, no que se refere ao tamanho do modelo o β=25% tende a apresentar melhores resultados que os demais, sendo, portanto, considerado o mais adequado a todas as estratégias avaliadas independente da MO utilizada. Assim, todas a demais análises, daqui em diante, são feitas apenas em cima deste valor de β.

Na sequência, como uma segunda análise, avaliou-se o desempenho e o $\mathcal{L}$ entre as estratégias de extração de regras a fim de verificar qual delas, de fato, gera o melhor impacto nos modelos finais. Assim como na análise anterior, dividiu-se esta análise em duas partes: uma considerando apenas o fluxo do CBA, i.e., em que o ranqueamento é realizado por meio da medida de Confiança, e a outra considerando diferentes ranqueamentos, uma para cada MO considerada.

No que se refere a análise em relação ao fluxo do CBA, para cada medida de avaliação, uma tabela como a apresentada na Tabela 11 foi gerada. Os resultados se referem a média de

Tabela 10 – Análise "Ganha" x "Perde" em relação aos βs ao longo das 48 MOs.

		F1					G-Mean					Tamanho ($\mathcal{L}$)				
		β5%	β10%	β15%	β20%	β25%	β5%	β10%	β15%	β20%	β25%	β5%	β10%	β15%	β20%	β25%
Apriori-C	β5%		1	1	1	1		1	1	1	2					
	β10%										1					
	β15%											17	3			
	β20%											20	16			
	β25%											21	19	13		
Apriori-T	β5%					1					1					
	β10%											1				
	β15%											10	1			
	β20%											18	11			
	β25%											19	17	2		
MS-Apriori	β5%					1					1					
	β10%					1					1					
	β15%											8	1			
	β20%											17	8			
	β25%											20	15	9		
MS-Apriori-P	β5%															
	β10%															
	β15%											1				
	β20%											1	1			
	β25%											8	1			

desempenho (neste caso, F1) dos modelos gerados em ambos os folds. Para cada tabela (3 no total, uma para cada medida de avaliação), o teste de Friedman foi aplicado a fim de verificar a existência ou não de diferença significativa entre as estratégias. Quando detectada diferença, o teste post-hoc de Nemenyi foi aplicado. Para as medidas de desempenho, i.e., F1 e G-Mean, o teste de Friedman não detectou diferença entre as estratégias de extração. Já para a medida $\mathcal{L}$, o teste de Friedman detectou diferenças e, portanto, o teste post-hoc de Nemenyi foi aplicado e o gráfico CD gerado, o qual encontra-se apresentado na Figura 5. Nota-se claramente que a estratégia Apriori-C se destaca significativamente das demais, sendo, portanto, a mais adequada a ser utilizada em contextos desbalanceados segundo a análise realizada.

Tabela 11 – Parte dos resultados obtidos via β=25%, fluxo CBA (Confiança), na medida F1.

Conjunto de dado	**Apriori-C**	**Apriori-T**	**MS-Apriori**	**MS-Apriori-P**
abalone-17_vs_7-8-9-10	0.5086493718	0.5086493718	0.4937202252	0.4937202252
abalone-19_vs_10-11-12-13	0.49501868	0.49501868	0.49501868	0.49501868
abalone-20_vs_8-9-10	0.4953882632	0.4960541502	0.4953882632	0.4965843405
abalone-21_vs_8	0.6818002488	0.6818002488	0.6818002488	0.6609699567
abalone-3_vs_11	0.9828199863	0.9828199863	0.9828199863	0.9828199863
...	...	...	...	...
zoo-3	0.5818513746	0.5818513746	0.5818513746	0.5818513746

Assim como anteriormente (Análise-1), visando verificar se o mesmo comportamento se mantém, independente da MO utilizada para se realizar o ranqueamento, uma análise mais abrangente foi realizada. Neste caso, 144 tabelas como a Tabela 11 foram geradas: 3 medidas de avaliação * 48 MOs distintas para se realizar o ranqueamento. Considerando as tabelas geradas, o mesmo procedimento descrito anteriormente (Análise-1) para geração da tabela "Ganha"

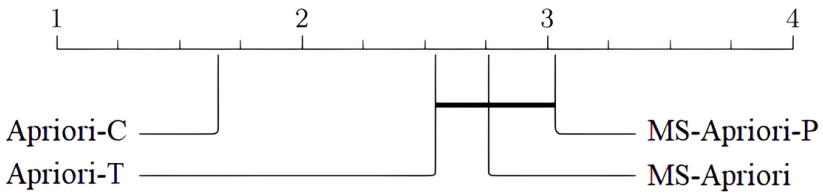


Figura 5 – Gráfico de diferença crítica, para comparação das estratégias, em relação ao tamanho dos modelos gerados.

x "Perde" (Tabela 10) foi aplicado, gerando a tabela "Ganha" x "Perde" entre as estratégias analisadas (Tabela 12). Observa-se pelos resultados obtidos que:

- a estratégia MS-Apriori-P, no que se refere as medidas de desempenho, i.e., F1 e G-Mean, apresenta um pequeno destaque em relação as demais, já que os valores apresentados são pequenos (2, 1 e 2 para F1 e 3, 1 e 1 para G-Mean);
- a estratégia Apriori-C, no que se refere a medida tamanho do modelo, se destaca em relação as demais, como no experimento anterior.

Tabela 12 – Análise "Ganha" x "Perde" em relação as estratégias de extração de regras.

	F1				G-Mean				$\mathcal{L}$			
	Apriori-C	Apriori-T	MS-Apriori	MS-Apriori-P	Apriori-C	Apriori-T	MS-Apriori	MS-Apriori-P	Apriori-C	Apriori-T	MS-Apriori	MS-Apriori-P
Apriori-C										12	14	22
Apriori-T	2				2						1	13
MS-Apriori	1				1							12
MS-Apriori-P	2	1	2		3	1	1					

Diante do exposto, nota-se, segundo as análise realizadas, que a estratégia Apriori-C foi a que se mostrou mais adequada ao contexto apresentado em relação as medidas F1, G-Mean e tamanho do modelo. Como mecionado anteriormente, esta estratégia é a adotada pelo CBA2, algoritmo *baseline* no contexto aqui abordado. Assim, embora os algoritmos propostos na literatura tenham variado esta etapa, sugere-se utilizar a estratégia do CBA2 nesta etapa e trabalhar em estratégias diferentes nas demais etapas.

5 Seleção Dinâmica de Medidas Objetivas

Este capítulo apresenta o método de seleção dinâmica de MOs proposto, denominado de DyOMS (Dynamic Objective Measures Selection). A motivação para proposição do mesmo fundamenta-se nas discussões anteriormente apresentadas nos Capítulos 2 e 3, a partir das quais é possível notar que:

- A etapa de ranqueamento se apresenta como uma etapa de grande importância na indução dos CAs, uma vez que influencia as demais etapas e, consequentemente, o desempenho e a interpretabilidade do modelo gerado;
- Ao longo dos anos diversas MOs foram propostas, como visto em Tew *et al.* (2014) e Somyanonthanakul e Theeramunkong (2022) (Seção 2.3);
- Em contextos desbalanceados, as MOs mais utilizadas pelos algoritmos levantados na literatura são Lift, Confiança, Suporte e/ou variações destas, mantendo-se, em geral, o padrão referente a regras de associação (Seção 3.2);
- Dado que não existe uma MO que seja adequada a todas as explorações (SHARMA *et al.*, 2020), trabalhos foram realizados visando agrupá-las em função de sua similaridade de desempenho, como o trabalho de Yang e Cui (2015), voltado também para CAs em contextos desbalanceados (Seção 2.3);
- No trabalho de Yang e Cui (2015), assim como em outros trabalhos (vide Seção 2.3), os autores sugerem um grupo de MOs adequadas ao contexto aqui abordado, não especificando como realizar a escolha da MO mais adequada dentro do grupo. Contudo, é possível notar que, de fato, algumas MOs apresentam melhores desempenhos do que outras, e que a mais adequada depende das próprias características das regras extraídas e, portanto, do conjunto de dados utilizado.

Diante do exposto, este capítulo apresenta um método de seleção dinâmica de MOs que possa ser incorporado a fluxos de indução de CAs. O método é inspirado nos trabalhos que realizam o agrupamento das MOs neste contexto, como Yang e Cui (2015), porém levando em consideração as características de distribuição de ranqueamento das regras em ambas as classes (majoritária e minoritária), assim como as características do conjunto de dados.

5.1 DyOMS: Dynamic Objective Measures Selection

Tendo como base os trabalhos de agrupamento de MOs em CAs, descritos na Seção 2.3, a Figura 6 apresenta parte do método DyOMS, referente a um agrupamento "estático " das medidas, uma vez que o mesmo é baseado em um conjunto de regras pré-definidas, e não por meio de um algoritmo de agrupamento. Contudo, como pode ser visto, a saída do fluxo é também uma matriz, como nos trabalhos da literatura (vide Figura 2, Seção 2.3). Neste caso, uma matriz Mx9 é obtida, a qual relaciona, para cada conjunto de dados *m*, o desempenho obtido pelos modelos gerados em cada um dos 9 grupos possíveis considerados (vide abaixo), os quais são compostos por um conjunto de MOs, juntamente com seus respectivos valores de avaliação (VA). Cada grupo aglomera um conjunto de MOs que apresenta um comportamento de ranqueamento semelhante no que se refere a importância (precedência) que as mesmas atribuem as regras das classes majoritária e minoritária. Tendo como base essa matriz, diferentemente dos trabalhos da literatura, que já disponibilizam grupos "estáticos" de MOs, identifica-se, em tempo de execução, para um dado conjunto de dados *m*, a MO mais adequada ao ranqueamento. Essa etapa é apresentada na Figura 7. Assim, pode-se dividir o método em duas etapas: geração da matriz (Figura 6) e utilização da matriz para identificação da MO mais adequada (Figura 7).

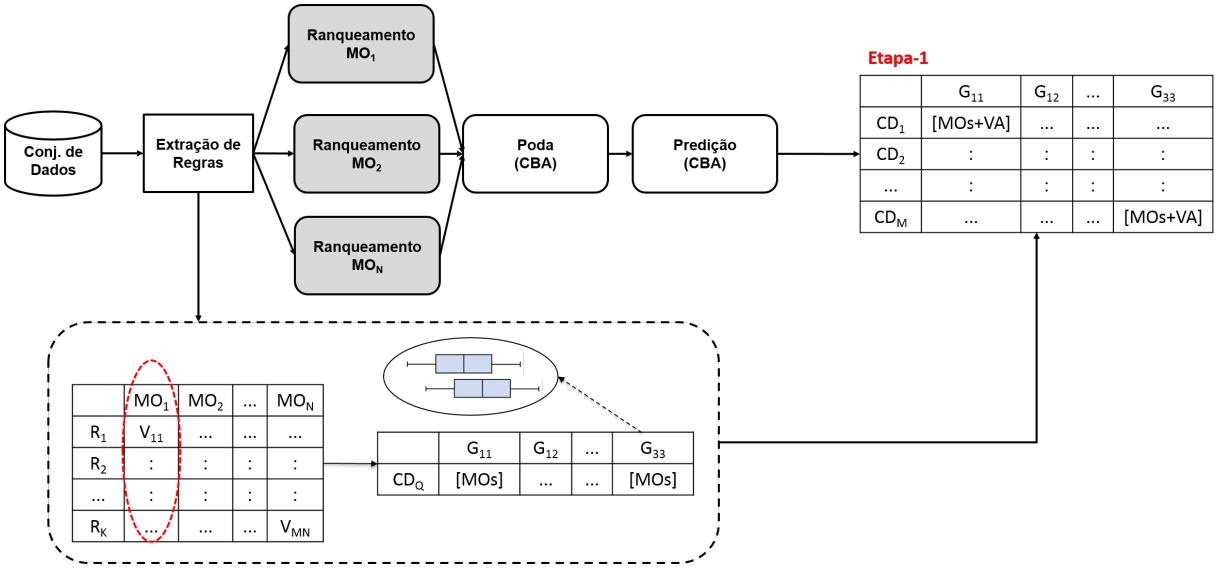


Figura 6 – Etapa de geração dos grupos de distribuição das MOs via método DyOMS.

No que se refere a primeira etapa, a construção da matriz ocorre da seguinte maneira:

- dado um conjunto de dados *m*, as RACs são extraídas, um conjunto de *n* MOs computadas e uma matriz KxN gerada. Para cada MO *n* tem-se, então, seu valor em cada uma das regras *k* (visão por coluna (circunferência em vermelho)). Tendo como base essa matriz, as MOs são agrupadas "estaticamente" em um dos 9 grupos possíveis considerados, os quais são descritos a seguir. Para que uma MO *n* seja alocada em um dado grupo, é necessário que se obtenha a distribuição das regras nas classes majoritária e minoritária na referida medida.

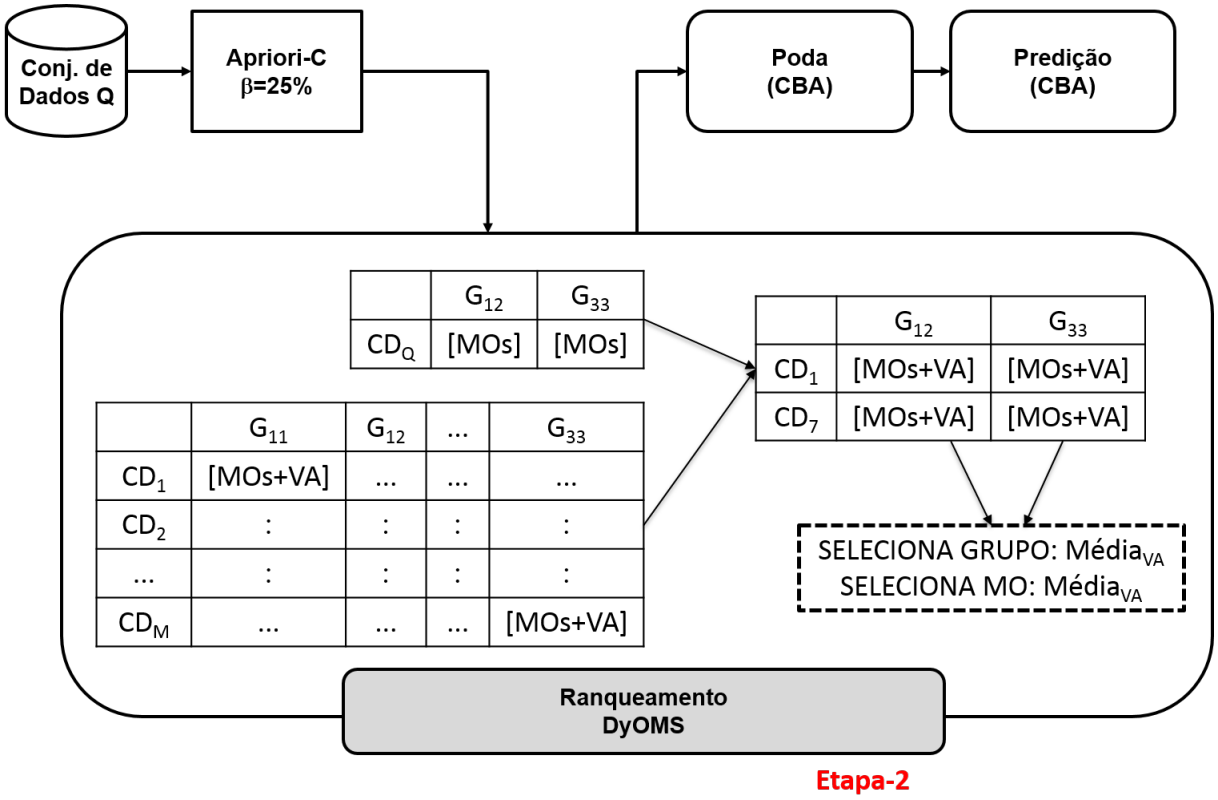


Figura 7 – Utilização dos grupos de distribuição das MOs via método DyOMS para indução dos modelos.

Assim, dada uma MO k, divide-se as regras em ambas as classes e gera-se um gráfico de distribuição (boxplot) que represente o ranqueamento das regras em cada uma das classes. Os gráficos são gerados em função dos valores da medida em cada classe. Vale mencionar que o boxplot não é de fato gerado, uma vez que os critérios de análise são baseados em seus valores (limite superior, limite inferior) e não na sua representação gráfica em si. Contudo, é mais didático explicar os critérios usando uma visualização gráfica. A fim de entender a análise realizada sobre os gráficos de distribuição, considere a Figura 9a referente ao Grupo-2.1 (G_{21}). Cada gráfico de distribuição representa o ranqueamento das regras em uma dada classe. Como quanto maior o valor de uma MO melhor ranqueada ela é, se uma comparação for realizada entre os gráficos de ambas as classes, é possível observar se a MO prioriza ou não mais as regras de uma dada classe do que de outra. Por este motivo, os boxplots foram plotados inversamente, i.e., do limite superior para o inferior. No caso da referida figura, nota-se que um subconjunto das regras da classe minoritária serão ranqueadas primeiro analisando-se os limites superiores de ambas as distribuições (quanto maior, melhor), embora não haja distinção (prioridade) entre o subconjunto final das regras de ambas as classes. Deste modo, tendo como base os gráficos de distribuição, o critério para se alocar uma MO em um dado grupo é dado a seguir:

Grupo-1.1 (G_{11}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja igual ao limite superior (L_{sup}) da distribuição da classe majoritária, assim como o

limite inferior (L_{inf}) da distribuição da classe minoritária seja igual ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{11}). Este grupo representa as MOs que não priorizam, durante o ranqueamento, regras de nenhuma das classes, i.e., uma regra terá maior precedência do que outra em função unicamente de seu valor (vide Figura 8a).

Grupo-1.2 (G_{12}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja igual ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja maior ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{12}). Este grupo representa as MOs que, a princípio, não priorizam, durante o ranqueamento, regras de nenhuma das classes, porém que apresentam uma leve tendência para regras da classe minoritária, já que ao final do ranqueamento um subconjunto das mesmas terão precedência sobre as da classe majoritária (vide Figura 8b).

Grupo-1.3 (G_{13}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja igual ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja menor ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{13}). Este grupo representa as MOs que, a princípio, não priorizam, durante o ranqueamento, regras de nenhuma das classes, porém que apresentam uma leve tendência para regras da classe majoritária, já que ao final do ranqueamento um subconjunto das mesmas terão precedência sobre as da classe minoritária (vide Figura 8c).

Grupo-2.1 (G_{21}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja maior ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja igual ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{21}). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe minoritária, embora, ao final, não haja distinção de precedência entre as classes (vide Figura 9a).

Grupo-2.2 (G_{22}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja maior ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja maior ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{22}). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe minoritária (vide Figura 9b).

Grupo-2.3 (G_{23}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja maior ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja menor ao limite inferior (L_{inf}) da

distribuição da classe majoritária, a MO é alocada a este grupo (G_{23}). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe minoritária, embora, ao final, regras da classe majoritária apresentem uma maior precedência em relação as da classe minoritária (vide Figura 9c).

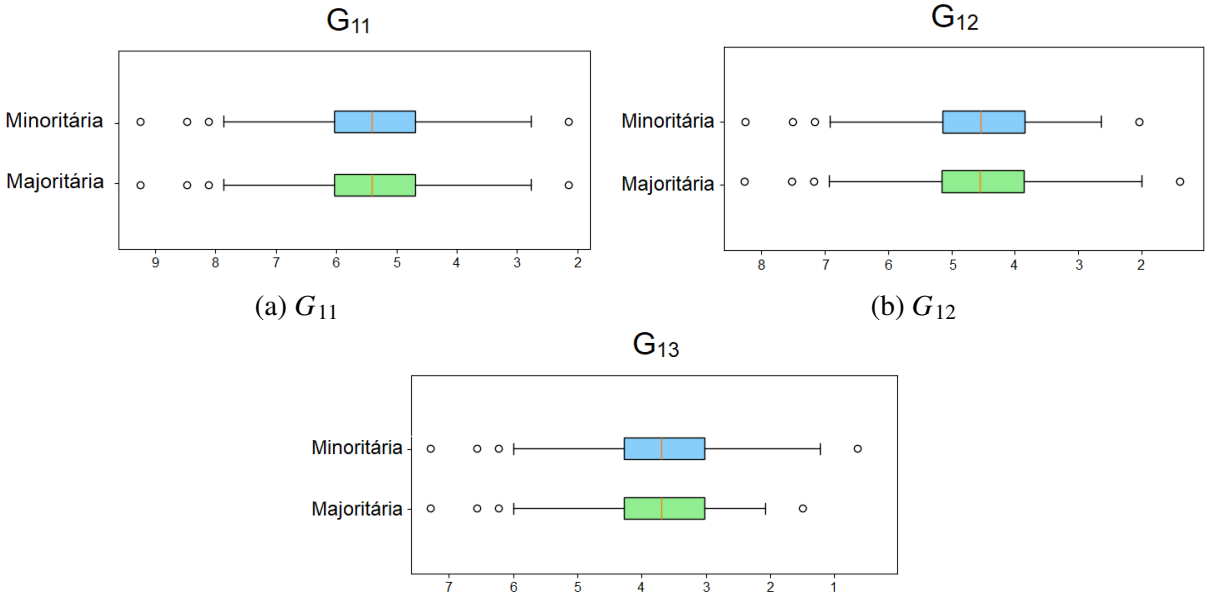
Grupo-3.1 (G_{31}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja menor ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja igual ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{31}). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe majoritária, embora, ao final, não haja distinção de precedência entre as classes (vide Figura 10a).

Grupo-3.2 (G_{32}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja menor ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja maior ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{32}). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe majoritária, embora, ao final, regras da classe minoritária apresentem uma maior precedência em relação as da classe majoritária (vide Figura 10b).

Grupo-3.3 (G_{33}). Caso o limite superior (L_{sup}) da distribuição da classe minoritária seja menor ao limite superior (L_{sup}) da distribuição da classe majoritária, e o limite inferior (L_{inf}) da distribuição da classe minoritária seja menor ao limite inferior (L_{inf}) da distribuição da classe majoritária, a MO é alocada a este grupo (G_{33}). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe majoritária (vide Figura 10c).

A Tabela 13 sumariza as regras de cada um dos grupos tendo como base os limites superiores e inferiores das distribuições.

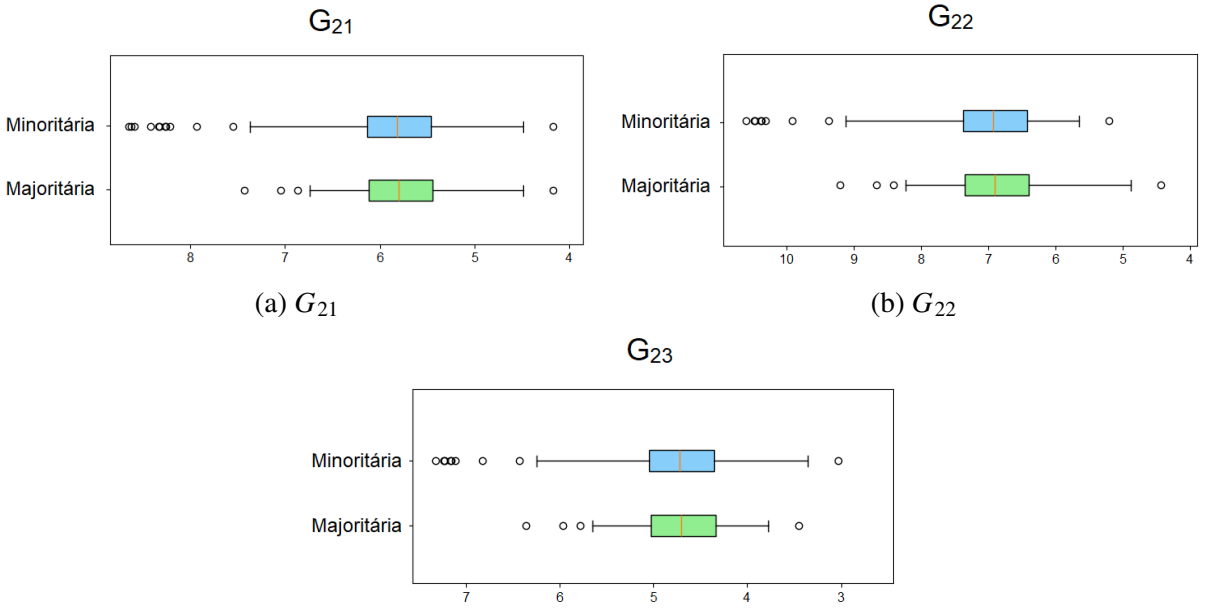
- Após extração das RACs, assim como geração dos grupos de MOs por meio de suas distribuições, a indução do modelo prossegue e é realizada considerando as diferentes MOs, como nos trabalhos da literatura. Cada MO leva a obtenção de um modelo distinto associado a uma medida de avaliação (na figura, VA significa "Valor de Avaliação"). Assim, após executar o fluxo indicado na figura para uma coleção de conjuntos de dados, uma matriz Mx9 é obtida, a qual relaciona, para cada conjunto de dados m, o desempenho obtido pelos modelos gerados em cada um dos 9 grupos considerados, os quais são compostos por um conjunto de MOs, juntamente com seus respectivos valores de avaliação (VA). De fato, atrelado a cada grupo encontra-se um conjunto de MOs com seus respectivos VA em cada um de seus respectivos modelos. Vale mencionar que cada conjunto de dados



(a) G_{11} (b) G_{12}

(c) G_{13}

Figura 8 – Gráficos dos grupos G_{11}, G_{12} e G_{13}.

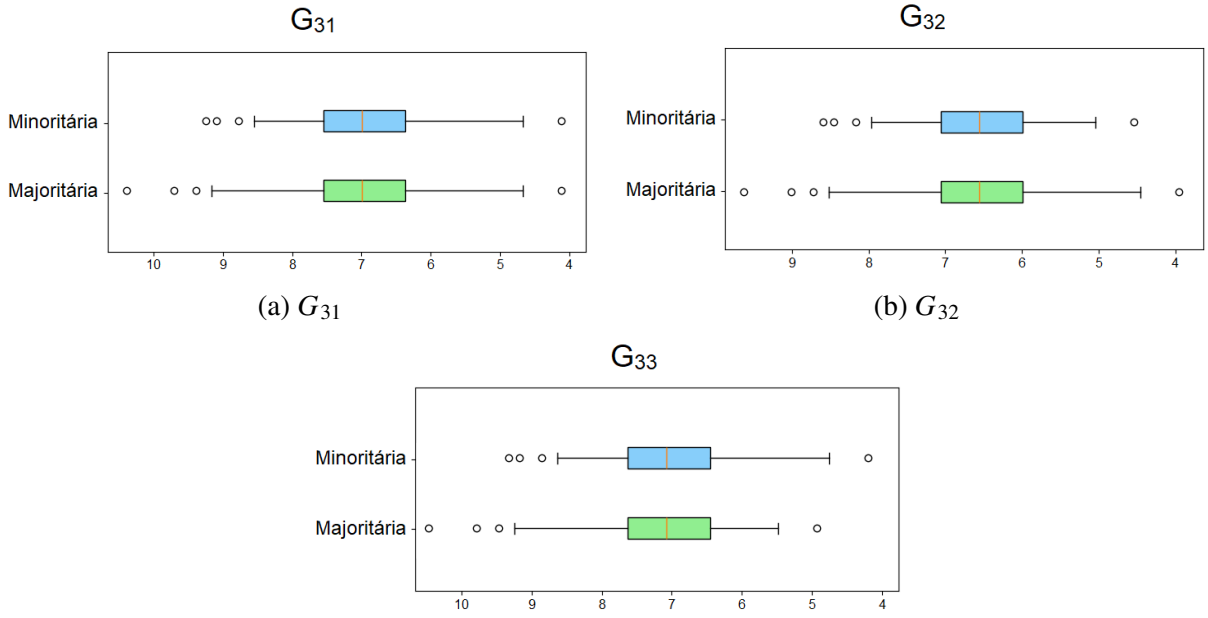


(a) G_{21} (b) G_{22}

(c) G_{23}

Figura 9 – Gráficos dos grupos G_{21}, G_{22} e G_{23}.

leva a um conjunto de regras distinto e, portanto, cada MO se comporta de uma maneira em relação as distribuições de ranqueamento entre as classes a depender do conjunto de regras extraídas. Assim, é possível que em um dado conjunto de dados uma MO pertença a um dado grupo e em outro conjunto de dados a outro grupo. Além disso, é possível que um dado conjunto de dados não contenha todos os grupos, pois é possível que algumas distribuições não aconteçam e, portanto, nestes casos, o vetor de MOs associado ao grupo é vazio. Deste modo, nota-se que as MOs pertencentes aos grupos levam em consideração



(a) G_{31} (b) G_{32}

(c) G_{33}

Figura 10 – Gráficos dos grupos G_{31}, G_{32} e G_{33}.

Tabela 13 – Regras dos grupos referente a distribuição das MOs.

Grupo	L_{Sup} **Min x** L_{Sup} **Maj**	L_{Inf} **Min x** L_{Inf} **Maj**
G_{11}	=	=
G_{12}	=	>
G_{13}	=	<
G_{21}	>	=
G_{22}	>	>
G_{23}	>	<
G_{31}	<	=
G_{32}	<	>
G_{33}	<	<

não apenas as distribuições em relação aos ranqueamentos, mas também as características implícitas dos conjuntos de dados.

Tendo como base a matriz Mx9 obtida na etapa anterior, diferentemente dos trabalhos da literatura, que já disponibilizam grupos "estáticos" de MOs, identifica-se então, em tempo de execução, para um dado conjunto de dados m, a MO mais adequada ao ranqueamento. Para tanto, a matriz é utilizada da maneira descrita a seguir (Figura 7):

- dado um conjunto de dados Q[1], em que um modelo deve ser induzido, extrai-se as RACs a fim de realizar o ranqueamento das mesmas. Contudo, em vez de se utilizar uma MO fixa

[1] Q de *Query*.

específica, como no CBA2, decidi-se por uma dada MO. Para tanto, os grupos referentes ao conjunto de dados Q são encontrados, como anteriormente descrito. Na sequência, a matriz Mx9 é consultada visando identificar conjuntos de dados que apresentem, no mínimo, os mesmos grupos de distribuição do conjunto Q. A ideia é utilizar, neste caso, o princípio dos vizinhos mais próximos, i.e., conjuntos de dados que apresentem grupos de distribuição semelhantes ao conjunto corrente Q, de modo que a escolha seja impactada apenas por conjuntos de dados com características similares ao conjunto corrente Q. Assim, diferentemente dos trabalhos da literatura, apenas conjuntos de dados similares ao conjunto corrente Q são utilizados no processo de escolha de uma dada medida.

- dado um subconjunto de conjuntos de dados similares S e uma dada medida de avaliação (VA), verifica-se o valor médio da mesma em cada um dos grupos de distribuição contidos em Q, tendo como base o conjunto S. A fim de facilitar o entendimento, tome como base a Figura 7. Nela, os conjuntos de distribuição de Q são G_{12} e G_{33}. Os conjuntos de dados similares a Q, neste exemplo, são CD_1 e CD_7. Computa-se então a média do VA nestes referidos grupos, i.e., G_{12} e G_{33}, ao longo dos conjuntos de dados S. O grupo que apresentar o melhor valor médio é o escolhido. Este é o grupo que contém as MOs que, em tese, melhor ranqueiam as regras referente ao conjunto de dados Q. Na sequência, das MOs pertencentes a este grupo de distribuição em Q, escolhe-se a MO que apresenta o maior VA médio ao longo dos conjuntos de dados S. Por fim, vale mencionar que caso o conjunto S seja vazio[2], os grupos G_{23}, G_{22} ou G_{21} são escolhidos, nesta ordem, já que todos eles priorizam regras da classe minoritária, com sutis diferenças em relação a classe majoritária. A medida a ser selecionada é então escolhida considerando o maior VA médio ao longo de todos os conjuntos de dados que apresentem o grupo selecionado.

Por fim, vale mencionar algumas considerações finais sobre o método aqui apresentado:

- a etapa referente a geração da matriz Mx9 é realizada apenas uma única vez, assim como nos trabalhos da literatura. Uma vez gerada, a mesma é apenas utilizada para identificação da MO mais adequada. A ideia é considerar na escolha apenas conjuntos de dados com distribuições de ranqueamento similares a um conjunto de dados corrente Q, e não uma coleção de conjuntos de dados como na literatura, em que um conjunto de MOs fixas é sempre sugerido.

- nos experimentos, a matriz Mx9 foi gerada a partir de 87 conjuntos de dados, número de conjuntos de dados comparável aos utilizados na literatura, a saber: Tew *et al.* (2014) 110 conjuntos, Yang e Cui (2015) 9 conjuntos e Dall'Agnol e Carvalho (2023) 43 conjuntos. Assim, embora a matriz Mx9 seja utilizada como uma "base de conhecimento", a mesma foi

[2] Esta condição aconteceu apenas duas vezes (2,30%) ao longo dos 87 conjuntos de dados.

gerada da mesma maneira que os trabalhos da literatura, com a diferença de não utilizá-la diretamente em um processo de agrupamento.

- nos experimentos, no que se refere a escolha de uma dada medida de avaliação (VA), as seguintes medidas foram utilizadas: F1, G-Mean e tamanho do modelo.
- o método foi elaborado para problemas de classificação binária, já que a análise dos grupos é realizada entre as distribuições das classes majoritária e minoritária, i.e., supõe-se a existência de apenas duas classes.

5.2 Configuração Experimental

A fim de avaliar o método proposto, experimentos foram realizados. Assim, é necessário que alguns detalhes, relacionados a execução dos experimentos, sejam aqui descritos.

Conjuntos de Dados. Os mesmos 87 conjuntos de dados apresentados na Seção 4.2 do Capítulo 4 foram utilizados.

Pré-processamento. O mesmo pré-processamento descrito na Seção 4.2 do Capítulo 4 foi realizado.

Extração de Regras. Em função dos resultados apresentados no Capítulo 4, em relação as estratégias de extração de regras, apenas o algoritmo Apriori-C, setado com β=25%, foi utilizado. Assim como no capítulo anterior, a fim de evitar interferências nos resultados apresentados optou-se por zerar o valor da confiança mínima, extraindo-se, portanto, todas as possíveis regras dentro do β especificado. Ademais, assim como anteriormente, definiu-se que o tamanho máximo do *itemset* seria igual a 5 itens, o que implica em regras com no máximo 4 itens no antecedente.

Medidas Objetivas. As mesmas 48 MOs apresentadas na Seção 4.2 do Capítulo 4 foram utilizadas.

Critério de Avaliação. Os mesmos critérios de avaliação apresentados na Seção 4.2 do Capítulo 4 foram utilizados, a saber: F1, G-Mean e tamanho do modelo ($\mathcal{L}$), todos estimados via 2-fold cross-validation estratificado. Ademais, assim como anteriormente, testes estatísticos foram realizados, utilizando-se para tanto o teste de Friedman com $\alpha = 0{,}05$ e o teste post-hoc de Nemenyi, juntamente com os diagramas de diferença crítica (CD).

Configuração do DyOMS. Em relação ao método proposto, o único parâmetro a ser especificado é o referente a última etapa, no que se refere a escolha de uma dada medida de avaliação (VA). Como mencionado anteriormente, as seguintes medidas foram utilizadas: F1, G-Mean e tamanho do modelo ($\mathcal{L}$).

Baselines. A fim de verificar se o método proposto produz, de fato, bons resultados

em relação aos trabalhos da literatura, dois baselines foram utilizados, a saber: CBA2 e CBA Adaptado via ModifiedLift. Como mencionado na Seção 2.2, o CBA2 é o algoritmo *baseline* de CAs em contextos desbalanceados. O mesmo utiliza como medida de ranqueamento a Confiança. Já o CBA Adaptado via ModifiedLift foi identificado na revisão sistemática apresentada no Capítulo 3. O mesmo segue o mesmo fluxo do CBA, com a diferença de utilizar como medida de ranqueamento a ModifiedLift, uma medida por eles proposta baseada na medida Lift. Os autores afiram que a mesma é adequada para CAs em contextos desbalanceados. Assim, nos experimentos realizados, o método DyOMS é executado no fluxo do CBA, porém alterando-se a medida de ranqueamento pela selecionada pelo referido método.

5.3 Resultados e Discussão

Ao todo, 870 experimentos foram realizados: 5 fluxos distintos (CBA2, CBA-ModifiedLift, CBA-DyOMS (F1, G-Mean, $\mathcal{L}$)) * 87 conjuntos de dados * 2-fold cross-validation estratificado. Diferentemente do capítulo anterior, em que a média de F1, G-Mean e $\mathcal{L}$ foi computada ao longo dos folds, a análise aqui apresentada foi feita fold a fold. O motivo se deve ao fato do método DyOMS selecionar, em tempo de execução, uma dada MO, a qual pode variar, para um mesmo conjunto de dados, em folds diferentes. Ademais, em função de uma análise complementar de outros aspectos, apresentada na Seção 5.4, decidiu-se por uma análise fold a fold.

Tabela 14 – Parte dos resultados obtidos em relação ao DyOMS setado com a medida de avaliação (VA) F1, referente ao fold-1, em comparação aos *baselines*.

Conjunto de Dados	CBA-DyOMS	CBA2	CBA-ModifiedLift
abalone-17_vs_7-8-9-10	0.3833826654523377	0.49372022520571673	0.49372022520571673
abalone-19_vs_10-11-12-13	0.4950186799501868	0.4950186799501868	0.4950186799501868
abalone-20_vs_8-9-10	0.49658434051497635	0.49658434051497635	0.49658434051497635
abalone-21_vs_8	0.7804325955734406	0.4930313588850174	0.4930313588850174
abalone-3_vs_11	1.0	1.0	1.0
...	...	...	...
zoo-3	0.7397959183673469	0.6845360824742268	0.6845360824742268

Os resultados dos experimentos foram agrupados então por fold e medida de avaliação (VA) e, portanto, tabelas como a apresentada na Tabela 14 foram geradas para que os testes estatísticos pudessem ser realizados. As Figuras 11 a 16 apresentam os gráficos de CD resultantes. Pode-se notar que:

- em relação ao fold-1 tem-se que (Figuras 11 a 13):
  - quando o método DyOMS é setado com medida de avaliação (VA) F1 para escolha do grupo/medida (Figura 11), não há diferença de desempenho nos modelos em relação a F1 e G-Mean, embora na média (rank médio) o método DyOMS se destaque em

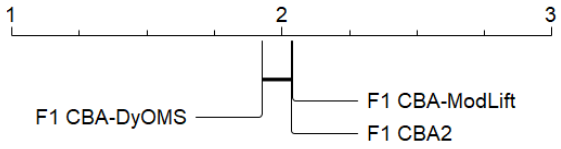
relação aos *baselines*, sempre ficando em primeiro. Já em relação ao tamanho do modelo, o método DyOMS se destaca, i.e., há diferença estatística em relação aos *baselines*.

  - quando o método DyOMS é setado com medida de avaliação (VA) G-Mean para escolha do grupo/medida (Figura 12), não há diferença de desempenho nos modelos em relação a F1 e G-Mean, embora na média (rank médio) o método DyOMS se destaque em relação aos *baselines*, sempre ficando em primeiro. Já em relação ao tamanho do modelo, o método DyOMS se destaca, i.e., há diferença estatística em relação aos *baselines*.
  - quando o método DyOMS é setado com medida de avaliação (VA) tamanho do modelo para escolha do grupo/medida (Figura 13), não há diferença de desempenho nos modelos em relação a F1 e G-Mean, embora na média (rank médio) o método DyOMS sempre fique em último. Já em relação ao tamanho do modelo, o método DyOMS se destaca, i.e., há diferença estatística em relação aos *baselines*.

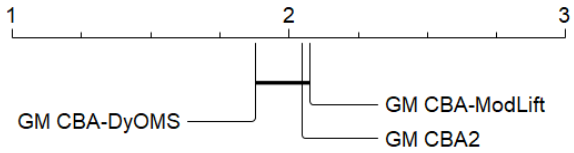
- em relação ao fold-2 tem-se que (Figuras 14 a 16):

  - quando o método DyOMS é setado com medida de avaliação (VA) F1 para escolha do grupo/medida (Figura 14), não há diferença de desempenho nos modelos em relação a F1 e G-Mean, embora na média (rank médio) o método DyOMS se destaque em relação aos *baselines*, sempre ficando em primeiro. Já em relação ao tamanho do modelo, o método DyOMS se destaca, i.e., há diferença estatística, porém apenas em relação ao CBA2. Contudo, na média (rank médio) o método DyOMS se destaca em relação aos *baselines*, sempre ficando em primeiro.
  - quando o método DyOMS é setado com medida de avaliação (VA) G-Mean para escolha do grupo/medida (Figura 15), não há diferença de desempenho nos modelos em relação a F1 e G-Mean, embora na média (rank médio) o método DyOMS se destaque em relação aos *baselines*, sempre ficando em primeiro. Já em relação ao tamanho do modelo, o método DyOMS se destaca, i.e., há diferença estatística, porém apenas em relação ao CBA2. Contudo, na média (rank médio) o método DyOMS se destaque em relação aos *baselines*, sempre ficando em primeiro.
  - quando o método DyOMS é setado com medida de avaliação (VA) tamanho do modelo para escolha do grupo/medida (Figura 16), não há diferença de desempenho nos modelos em relação a F1 e G-Mean; contudo, em relação ao G-Mean, na média (rank médio) o método DyOMS se destaca em relação aos *baselines*. Já em relação ao tamanho do modelo, o método DyOMS se destaca, i.e., há diferença estatística em relação aos *baselines*.

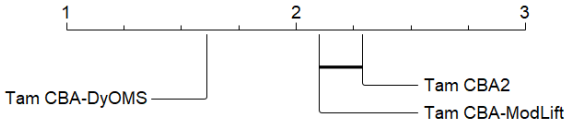
Diante do exposto, pode-se notar que em relação ao fold-1, utilizar o método DyOMS setado com as medidas de avaliação (VA) F1 e G-Mean é uma boa opção, já que o mesmo mantém o desempenho dos modelos (F1 e G-Mean), sempre se destacando na média (rank médio) em relação aos *baselines*, assim como melhora a interpretabilidade dos modelos com diferença estatística em relação aos *baselines*. O mesmo padrão se mantém em relação ao fold-2, embora em relação a interpretabilidade diferenças estatísticas não tenham sido detectadas; contudo, o mesmo sempre se destaca na média (rank médio) em relação aos *baselines*. Considerando o apresentado, nota-se que o método DyOMS se apresenta como uma solução viável ao contexto aqui apresentado, i.e., ao ranqueamento de RACs em fluxos de CAs em contextos desbalanceados.



(a) F1

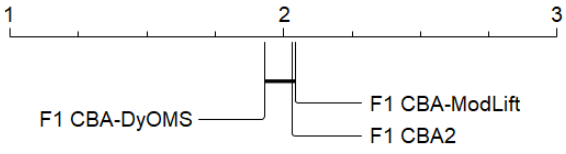


(b) G-Mean

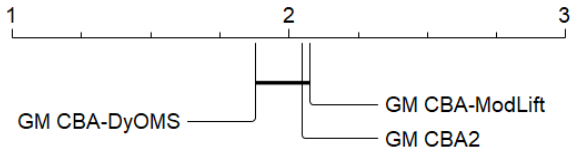


(c) Tamanho do modelo

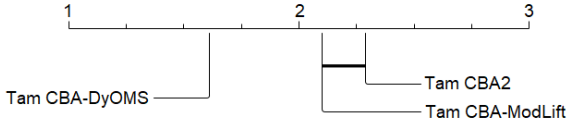
Figura 11 – Gráficos de diferença crítica referentes ao fold-1 em relação ao F1, G-Mean e tamanho do modelo computados com DyOMS setado com medida de avaliação (VA) F1.



(a) F1

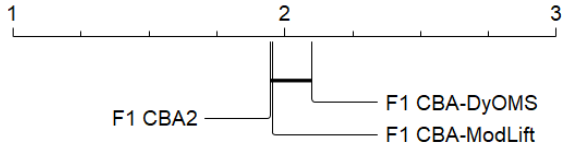


(b) G-Mean

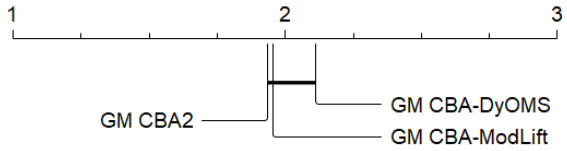


(c) Tamanho do modelo

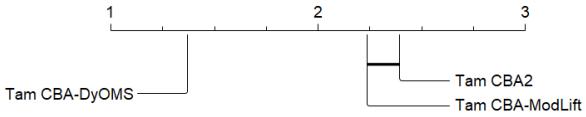
Figura 12 – Gráficos de diferença crítica referentes ao fold-1 em relação ao F1, G-Mean e tamanho do modelo computados com DyOMS setado com medida de avaliação (VA) G-Mean.



(a) F1

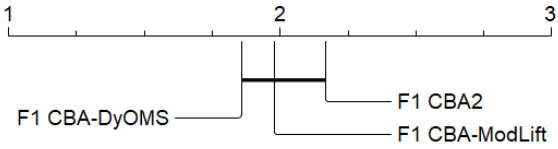


(b) G-Mean

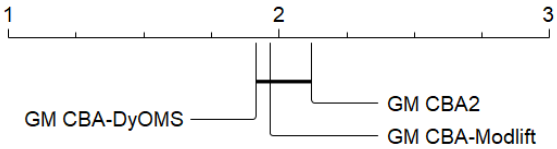


(c) Tamanho do modelo

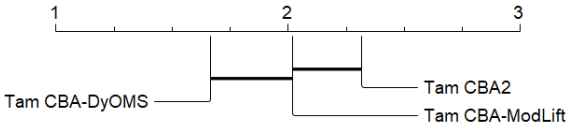
Figura 13 – Gráficos de diferença crítica referentes ao fold-1 em relação ao F1, G-Mean e tamanho do modelo computados com DyOMS setado com medida de avaliação (VA) tamanho do modelo.



(a) F1

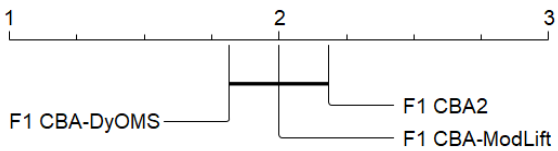


(b) G-Mean

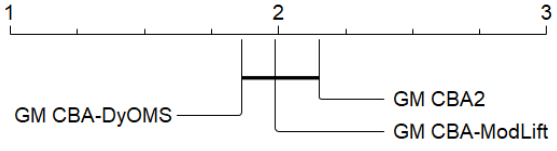


(c) Tamanho do modelo.

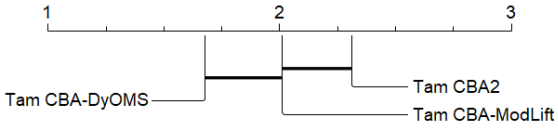
Figura 14 – Gráficos de diferença crítica referentes ao fold-2 em relação ao F1, G-Mean e tamanho do modelo computados com DyOMS setado com medida de avaliação (VA) F1.



(a) F1

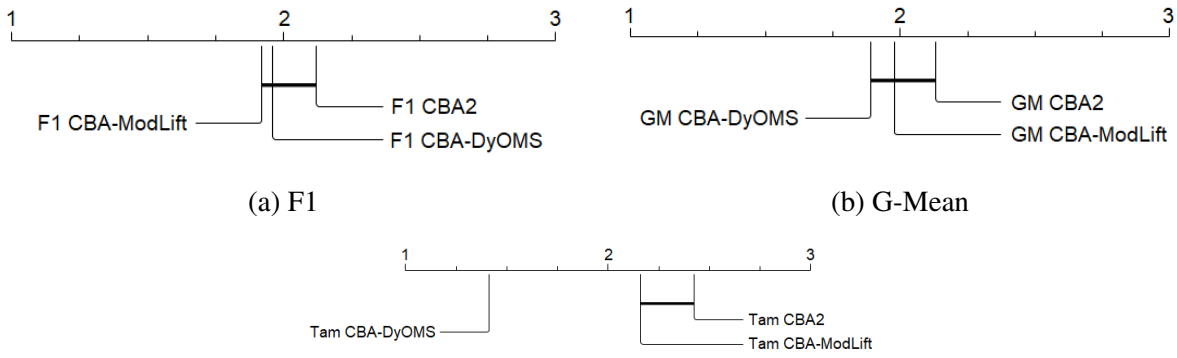


(b) G-Mean



(c) Tamanho do modelo

Figura 15 – Gráficos de diferença crítica referentes ao fold-2 em relação ao F1, G-Mean e tamanho do modelo computados com DyOMS setado com medida de avaliação (VA) G-Mean.



(a) F1

(b) G-Mean

(c) Tamanho do modelo

Figura 16 – Gráficos de diferença crítica referentes ao fold-2 em relação ao F1, G-Mean e tamanho do modelo computados com DyOMS setado com medida de avaliação (VA) tamanho do modelo.

5.4 Análise Complementar

A fim de complementar os resultados apresentados, esta seção realiza uma análise adicional a partir do experimentos realizados. Para tanto, alguns gráficos foram gerados, os quais são apresentados e discutidos a seguir.

Visando identificar os grupos mais selecionados ao longo dos conjuntos de dados, as Figuras 17 a 22 foram elaboradas. As Figuras 17 a 19 se referem ao fold-1 e as Figuras 20 a 22 se referem ao fold-2. Pode-se notar que:

- em relação ao uso do DyOMS setado com as medidas de avaliação (VA) F1 e G-Mean, tanto no fold-1 quanto no fold-2, tem-se que o grupo G_{23} se destaca em relação aos demais, cobrindo majoritariamente a maioria dos casos (81/87 (93,10%) na Figura 17, 79/87 (90,80%) na Figura 18, 74/87 (85,06%) na Figura 20, 73/87 (83,91%) na Figura 21). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe minoritária, embora, ao final, regras da classe majoritária apresentem uma maior precedência em relação as da classe minoritária. Assim, nota-se a importância do uso de MOs que tendem a priorizar regras da classe minoritária.

- em relação ao uso do DyOMS setado com a medida de avaliação (VA) tamanho do modelo, tanto no fold-1 quanto no fold-2, tem-se que o grupo G_{22} se destaca em relação aos demais, cobrindo majoritariamente a maioria dos casos (77/87 (88,51%) na Figura 19, 76/87 (87,36%) na Figura 22). Este grupo representa as MOs que priorizam, durante o ranqueamento, mais fortemente as regras da classe minoritária. Assim, nota-se aqui também, a importância do uso de MOs que tendem a priorizar regras da classe minoritária.

Por outro lado, visando identificar as MOs mais selecionadas ao longo dos conjuntos de dados, as Figuras 23 a 28 foram elaboradas. As Figuras 23 a 25 se referem ao fold-1 e as

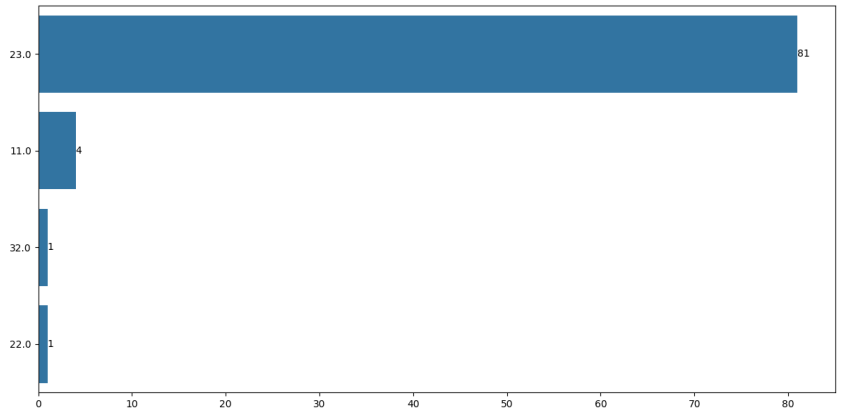


Figura 17 – Frequência dos grupos escolhidos no fold-1 com DyOMS setado com medida de avaliação (VA) F1.

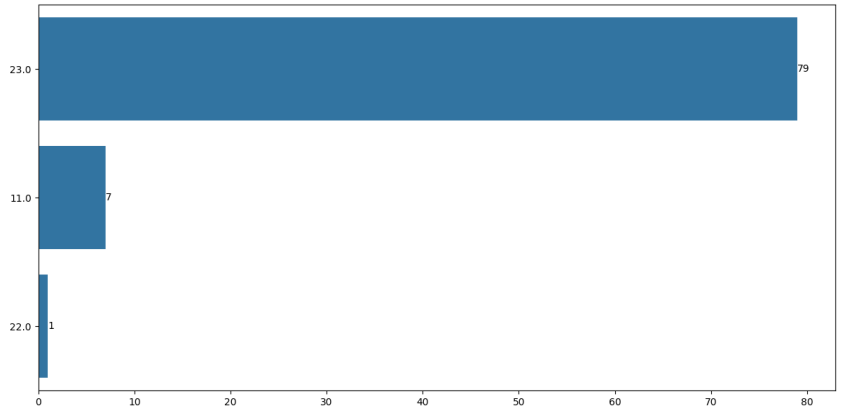


Figura 18 – Frequência dos grupos escolhidos no fold-1 com DyOMS setado com medida de avaliação (VA) G-Mean.

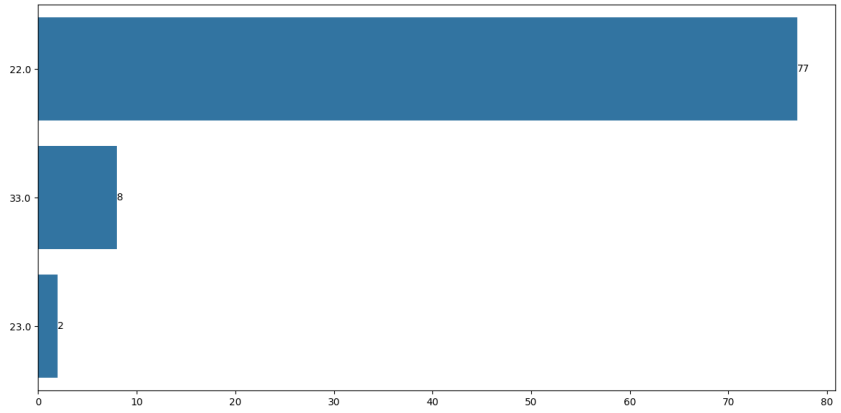


Figura 19 – Frequência dos grupos escolhidos no fold-1 com DyOMS setado com medida de avaliação (VA) tamanho do modelo.

Figuras 26 a 28 se referem ao fold-2. Pode-se notar que:

- diferentemente dos grupos, não há uma medida que se destaca majoritariamente na maioria dos casos, ressaltando a importância de se avaliar cada conjunto de regras individualmente. Contudo, algumas MOs se destacam, a saber: Implication Index (26/87

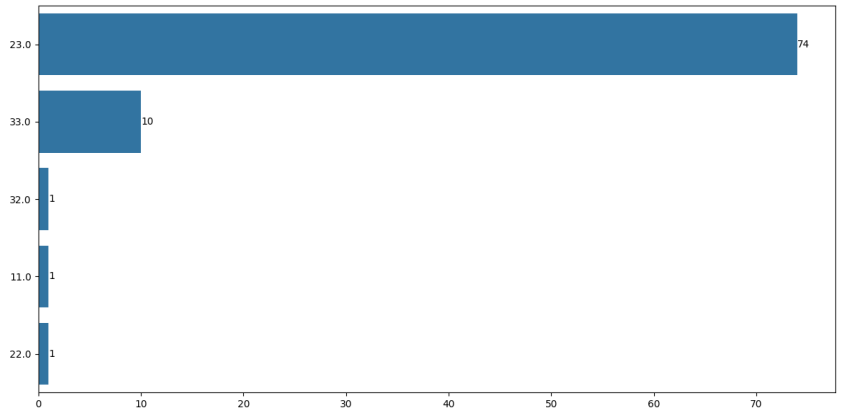


Figura 20 – Frequência dos grupos escolhidos no fold-2 com DyOMS setado com medida de avaliação (VA) F1.

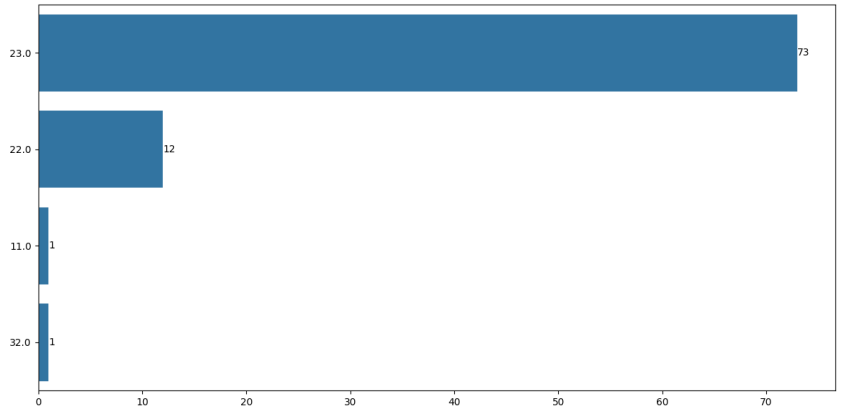


Figura 21 – Frequência dos grupos escolhidos no fold-2 com DyOMS setado com medida de avaliação (VA) G-Mean.

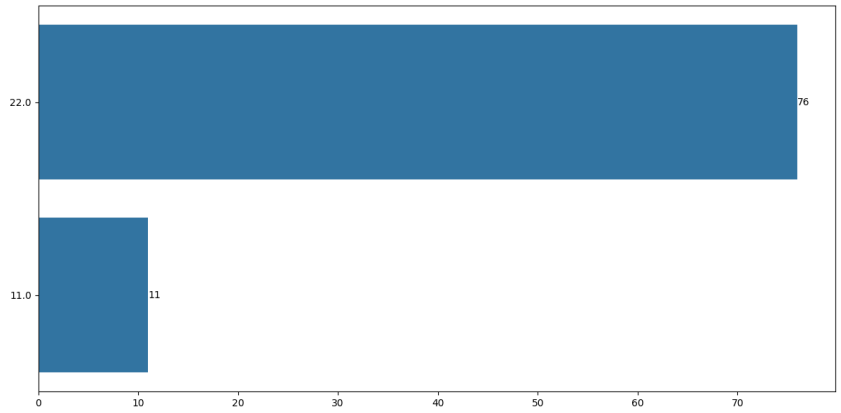


Figura 22 – Frequência dos grupos escolhidos no fold-2 com DyOMS setado com medida de avaliação (VA) tamanho do modelo.

(29,89%)), Collective Strength (14/87 (16,09%)), Loevinger (10/87 (11,49%)) e Confirmed Confidence Causal (8/87 (9,20%)) com DyOMS setado com as medidas de avaliação (VA) F1 e G-Mean no fold-1 (Figuras 23 e 24) e Putative Causal Dependency (23/87 (26,44%)) e Collective Strength (14/87 (16,01%)) com DyOMS setado com as medidas de avaliação

(VA) F1 e G-Mean no fold-2 (Figuras 26 e 27). Já em relação a medida de avaliação (VA) tamanho do modelo, as medidas Accuracy (22/87 (25,29%); 14/87 (16,09%)) e Confirm Causal (12/87 (13,79%); 11/87 (12,64%)) tanto no fold-1 (Figura 25) quanto no fold-2 (Figura 28).

- as MOs mais utilizadas na literatura, a saber, Lift, Confiança, Suporte e/ou variações destas, aparecem com uma baixa frequência ou nem aparecem, ressaltando a importância de se avaliar cada conjunto de regras individualmente.

- em relação as MOs sugeridas por Yang e Cui (2015) (vide Seção 2.3), algumas delas pertencentes ao grupo $G_<$ aparecem em destaque como Implication Index e Collective Strength. Contudo, medidas como Loevinger (Conviction) e Confirmed Confidence Causal, pertencentes ao grupo $G_>$ também aparecem. Este fato demonstra, de certa maneira, que a escolha da MO é, portanto, dependente do conjunto de regras extraído e, portanto, do conjunto de dados utilizado. Ademais, outras MOs, como a Accuracy (acima em destaque), não aparecem nos grupos propostos por Yang e Cui (2015).

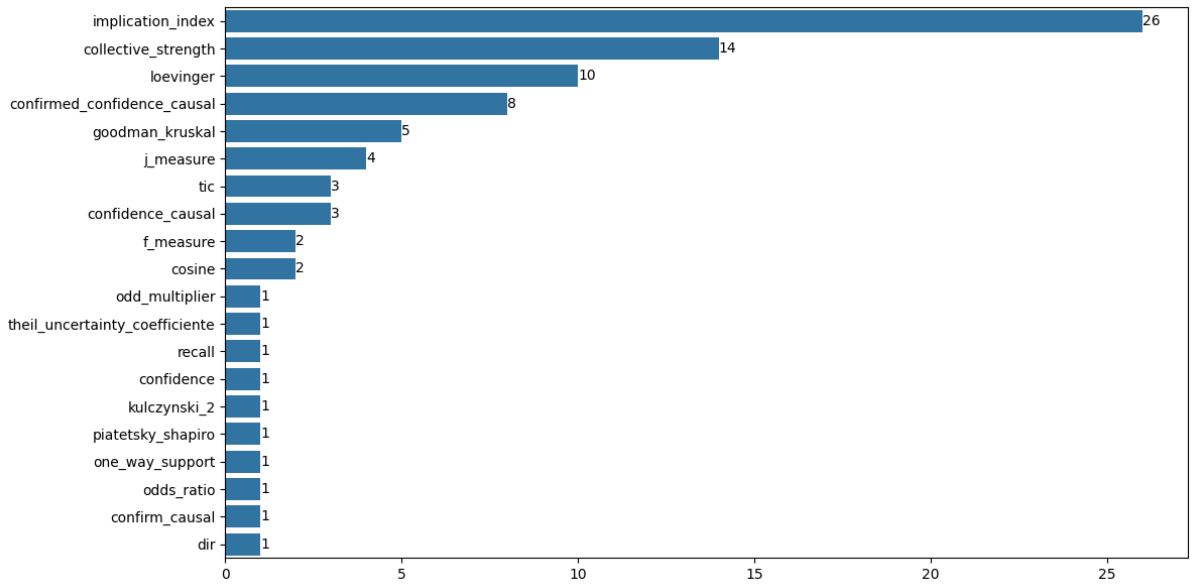


Figura 23 – Frequência das MOs escolhidas no fold-1 com DyOMS setado com medida de avaliação (VA) F1.

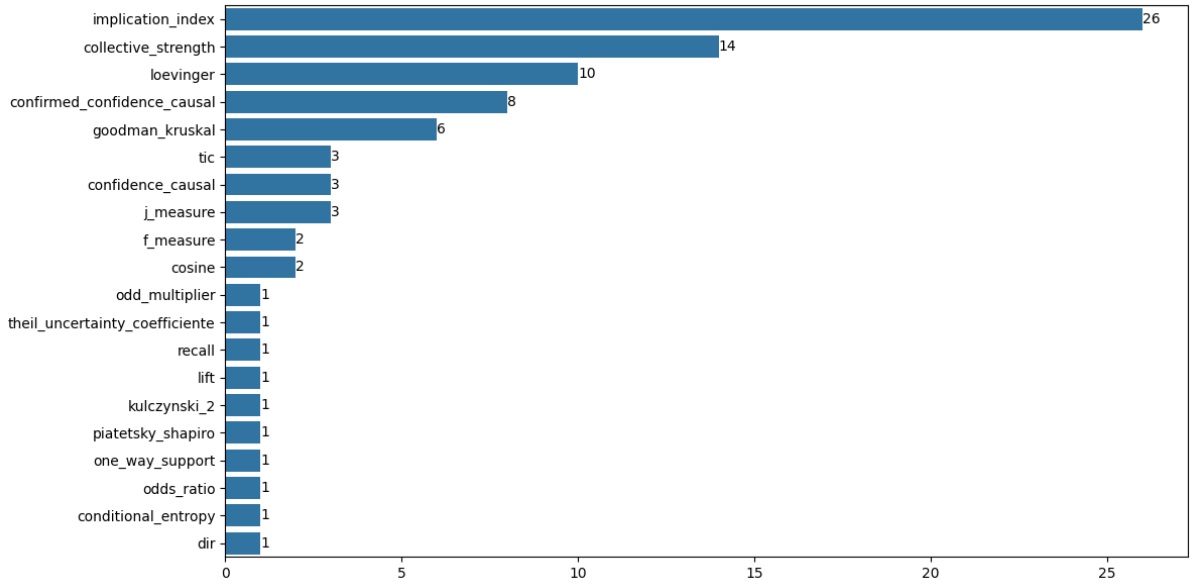


Figura 24 – Frequência das MOs escolhidas no fold-1 com DyOMS setado com medida de avaliação (VA) G-Mean.

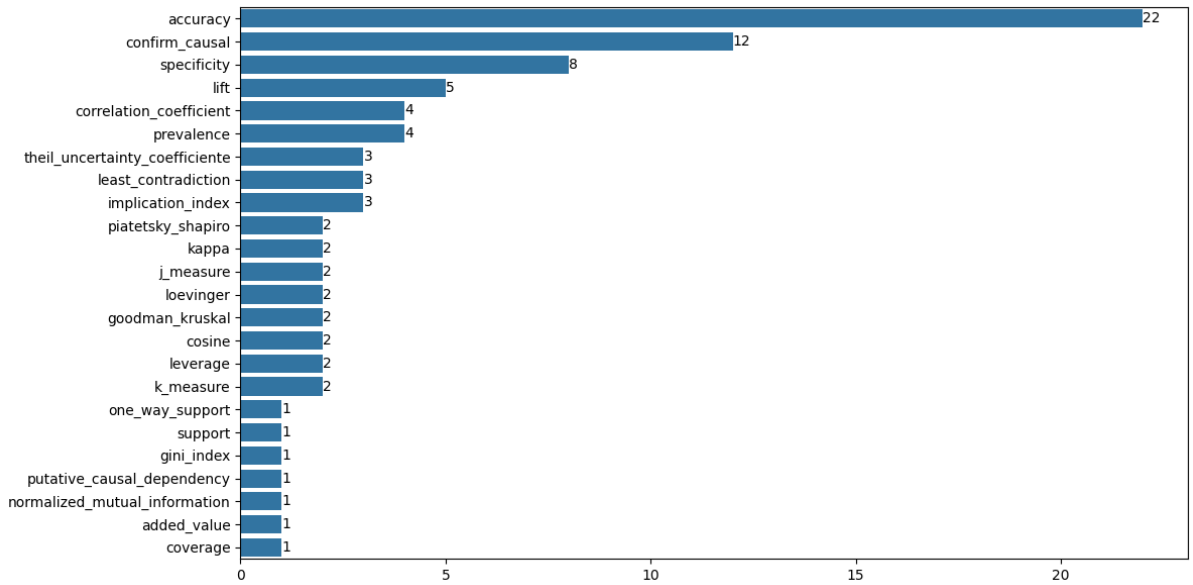


Figura 25 – Frequência das MOs escolhidas no fold-1 com DyOMS setado com medida de avaliação (VA) tamanho do modelo.

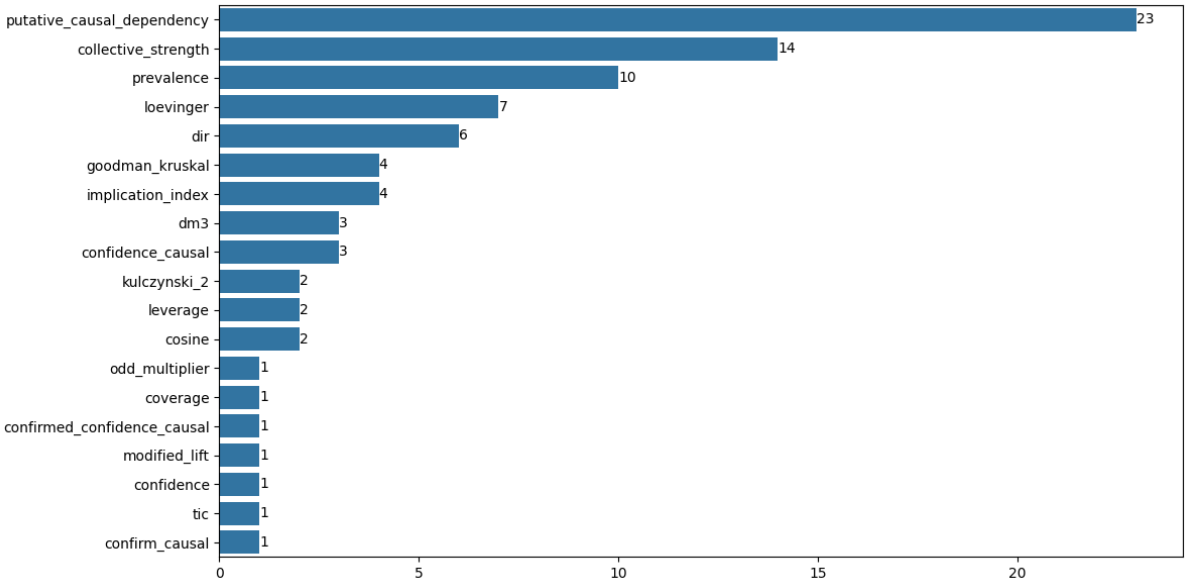


Figura 26 – Frequência das MOs escolhidas no fold-2 com DyOMS setado com medida de avaliação (VA) F1.

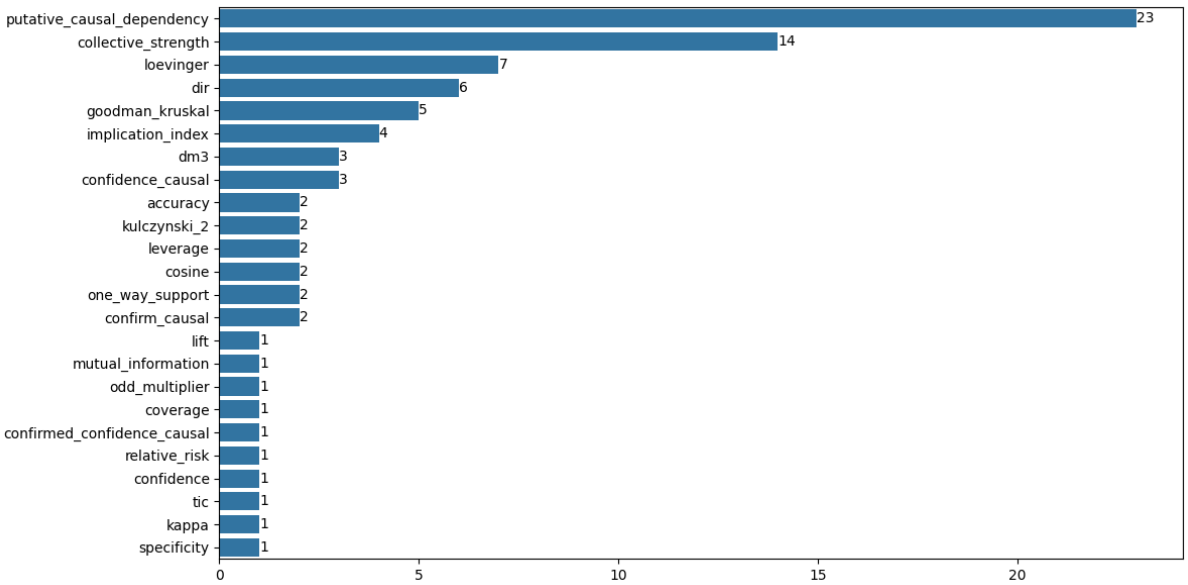


Figura 27 – Frequência das MOs escolhidas no fold-2 com DyOMS setado com medida de avaliação (VA) G-Mean.

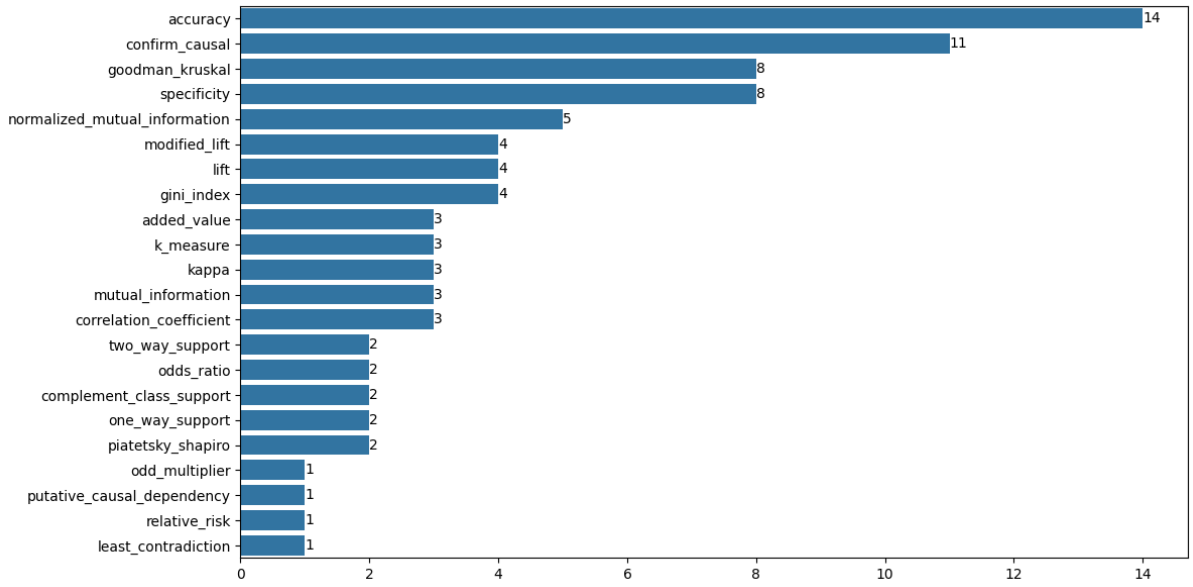


Figura 28 – Frequência das MOs escolhidas no fold-2 com DyOMS setado com medida de avaliação (VA) tamanho do modelo.

6 Conclusão

A tarefa de classificação é uma das mais conhecidas na área de aprendizado de máquina. Assim, diversas famílias de algoritmos foram propostas de modo a possibilitar a indução de modelos voltados a predição. Contudo, em diversos domínios busca-se pela utilização de algoritmos interpretáveis, i.e., por modelos em que seja possível compreender por quais razões uma determinada predição ocorreu. Mesmo com diversos métodos de XAI (Explainable Artificial Intelligence) disponíveis hoje em dia, como LIME e SHAP, há trabalhos como o de Rudin (2019) que propõem o uso de classificadores mais interpretáveis em detrimento da tentativa de criar modelos de explicação para algoritmos caixa-preta via tais métodos.

A classificação associativa (CA) representa uma família de algoritmos que faz uso de um conjunto de regras para representar o conhecimento extraído. Para tanto, tais algoritmos mesclam as tarefas de associação e classificação e são, portanto, induzidos em etapas, a saber: [a] extração de um conjunto de regras de associação classificativas, [b] ranqueamento das regras via medidas objetivas (MOs) e [c] poda das regras. A utilização destes algoritmos é, portanto, vantajosa, uma vez que as regras contidas nos modelos são interpretáveis por especialistas que podem avaliá-las e tomar decisões com auxílio computacional. Contudo, embora a CA, assim como outras famílias, apresente bons resultados, quando aplicada a problemas desbalanceados o desempenho não se mantém o mesmo. Os classificadores padrão são, em geral, direcionados à classe majoritária em favor, por exemplo, da medida de acurácia (FERNÁNDEZ *et al.*, 2018). Assim, regras específicas voltadas para a classe minoritária acabam por serem ignoradas; portanto, instâncias da classe minoritária acabam por serem classificadas incorretamente com mais frequência do que as instâncias da classe majoritária.

Diante do exposto, este trabalho explorou o uso de CAs quando aplicados em dados desbalanceados via abordagens internas, i.e., em nível de algoritmo. Para tanto, três objetivos foram propostos, a saber:

- a realização de uma revisão sistemática da literatura (RSL) a fim de identificar as abordagens internas que vêm sendo adotadas e/ou propostas na literatura (**[Obj.1]**) (Capítulo 3);
- a realização de uma análise sobre o impacto das diferentes estratégias utilizadas para se realizar a etapa de extração de regras visando identificar a mais adequada a ser utilizada no contexto aqui abordado (**[Obj.2]**) (Capítulo 4);
- a proposta de um método de seleção dinâmica de MOs, denominado DyOMS, que possa ser incorporado a fluxos de indução de CAs (**[Obj.3]**) (Capítulo 5).

A motivação para realização do **[Obj.1]** foi a de apoiar a fundamentação deste trabalho, assim como identificar lacunas e oportunidades na área. A RSL foi publicada no artigo "Associative Classifiers Algorithms for Imbalanced Data: A Systematic Literature Review" (PRIVATTO; CARVALHO, 2024). Deste modo, este trabalho contribui com a comunidade por meio da apresentação não apenas do que tem sido realizado neste contexto, mas também com a identificação de problemas a serem ainda explorados para um maior avanço nesta área.

A motivação para realização do **[Obj.2]** se deu em função dos diferentes algoritmos propostos na literatura utilizarem diferentes estratégias a fim de realizar a etapa de extração de regras. Para tanto, uma metodologia de análise foi proposta a fim de explorar quatro estratégias de extração, a saber: Apriori-T, Apriori-C, MS-Apriori e MS-Apriori-P. As mesmas foram avaliadas no fluxo base do CBA considerando diferentes valores de suporte mínimo (β) e MOs. A estratégia Apriori-C, já adotada pelo algoritmo CBA2, é a que se mostrou mais adequada segundo os experimentos realizados. Deste modo, este trabalho contribui com a comunidade direcionado-a para a estratégia de extração mais apropriada a ser considerada no desenvolvimento de novas soluções no contexto apresentado.

A motivação para realização do **[Obj.3]** se deu em função dos diferentes algoritmos propostos na literatura utilizarem um conjunto restrito de MOs (Lift, Confiança, Suporte e variações dessas) para realizar a etapa de ranqueamento das regras, embora inúmeras delas existam. Contudo, a MO mais adequada depende das próprias características das regras extraídas e, portanto, do conjunto de dados utilizado. Deste modo, o método proposto detecta a MO mais adequada, em tempo de execução, de modo que as regras sejam ordenadas da melhor maneira possível. O método apresenta um hiperparâmetro, medida de avaliação (VA), o qual trata-se de uma medida de avaliação de modelos. Foi possível notar que o DyOMS setado com as medidas de avaliação (VA) F1 e G-Mean se apresenta como uma opção viável, já que o mesmo mantém o desempenho dos modelos (F1 e G-Mean), sempre se destacando na média (rank médio) em relação aos *baselines* considerados (CBA2, CBA Adaptado via ModifiedLift), assim como melhora a interpretabilidade dos modelos com diferença estatística em relação aos *baselines*. Diante do exposto, este trabalho contribui com a comunidade por meio da proposição do DyOMS, método a ser considerado no desenvolvimento de novas soluções no que se refere ao ranqueamento de regras de associação classificativas em fluxos de CAs em contextos desbalanceados. Vale mencionar que análises complementares foram realizadas tendo sido possível observar que MOs que priorizam regras da classe minoritária (G_{22} e G_{23}) são mais adequadas ao contexto. Em outras palavras, é importante que novas soluções considerem, de alguma maneira, MOs que priorizem regras da classe minoritária. Além disso, foi possível notar que diversas MOs podem ser adequadas ao ranqueamento, não sendo conveniente o uso de uma medida específica, como nos *baselines* utilizados. Assim, novos métodos dinâmicos, como o aqui apresentado, são interessantes de serem explorados. Por fim, ressalta-se que o DyOMS leva em consideração tanto

as características de distribuição de ranqueamento das regras em ambas as classes (majoritária e minoritária), assim como as características do conjunto de dados.

No que se refere a trabalhos futuros, pode-se mencionar:

- a exploração de modificações no método DyOMS aqui proposto, como, por exemplo, o uso simultâneo de várias MOs dentro de um dado grupo de distribuição. O uso agregado de MOs em CAs é explorado, por exemplo, em Dall'Agnol e Carvalho (2024);
- trabalhos que busquem contribuir com as demais lacunas identificadas na RSL, uma vez que apenas um subconjunto das mesmas foi aqui explorado.

REFERÊNCIAS

ABDELLATIF, S.; HASSINE, M. A. B.; YAHIA, S. B. Novel interestingness measures for mining significant association rules from imbalanced data. In: SPRINGER. *Workshops of the International Conference on Advanced Information Networking and Applications*. [S.l.], 2019. p. 172–182. Citado na página 11.

ABDELLATIF, S. *et al.* ARCID: a new approach to deal with imbalanced datasets classification. In: SPRINGER. *International Conference on Current Trends in Theory and Practice of Informatics*. [S.l.], 2018. p. 569–580. Citado na página 30.

ABDELLATIF, S. *et al.* Fuzzy aggregation for rule selection in imbalanced datasets classification using Choquet integral. In: IEEE. *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.], 2018. p. 1–7. Citado na página 30.

ABU-ARQOUB, M.; HADI, W.; ISHTAIWI, A. ACRIPPER: A New Associative Classification Based on RIPPER Algorithm. *Journal of Information & Knowledge Management*, v. 20, n. 01, p. 2150013, 2021. Citado na página 30.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. [S.l.]: World Scientific, 1993. p. 207–216. Citado na página 16.

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. p. 487–499. Citado 3 vezes na(s) página(s) 11, 15 e 16.

AL-HAWARI, A.; NAJADAT, H.; SHATNAWI, R. Classification of application reviews into software maintenance tasks using data mining techniques. *Software Quality Journal*, v. 29, n. 3, p. 667–703, 2021. Citado 2 vezes na(s) página(s) 11 e 17.

AZMI, M.; BERRADO, A. RCAR Framework: Building a Regularized Class Association Rules Model in a Categorical Data Space. In: *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*. New York, NY, USA: Association for Computing Machinery, 2020. Citado 2 vezes na(s) página(s) 11 e 17.

BASHA, M. S. Early Prediction of Cardio Vascular Disease by Performing Associative Classification on Medical Datasets and Using Genetic Algorithm. In: *Intelligent Computing and Innovation on Data Science*. Singapore: Springer Singapore, 2021. p. 393–402. Citado 2 vezes na(s) página(s) 11 e 17.

BUI-THI, D.; MEYSMAN, P.; LAUKENS, K. MoMAC: Multi-objective optimization to combine multiple association rules into an interpretable classification. *Applied Intelligence*, Kluwer Academic Publishers, Usa, v. 52, n. 3, p. 3090–3102, feb 2022. ISSN 0924-669x. Citado na página 25.

CHEN, W.-C.; HSU, C.-C. An associative classification approach for enhancing prediction of imbalance data. In: *The Fifth International Conference on Informatics and Applications (ICIA2016)*. [S.l.: s.n.], 2016. p. 105. Citado na página 30.

CHEN, W.-C.; HSU, C.-C.; HSU, J.-N. Adjusting and generalizing CBA algorithm to handling class imbalance. *Expert Systems with Applications*, v. 39, n. 5, p. 5907–5919, 2012. Citado na página 30.

COHEN, W. W. Fast Effective Rule Induction. In: *The 20th International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 1995. p. 115. Citado na página 30.

DALL'AGNOL, M.; CARVALHO, V. O. Clustering the behavior of objective measures in associative classifiers. In: *18th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.]: IEEE, 2023. p. 6p. Citado 7 vezes na(s) página(s) x, 13, 20, 21, 22, 32 e 56.

DALL'AGNOL, M.; CARVALHO, V. O. d. AC.RankA: Rule ranking method via aggregation of objective measures for associative classifiers. *IEEE Access*, v. 12, p. 88862–88882, 2024. Citado 2 vezes na(s) página(s) 25 e 71.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Citado na página 43.

DERMEVAL, D.; COELHO, J. A. P. d. M.; BITTENCOURT, I. I. Mapeamento Sistemático e Revisão Sistemática da Literatura em Informática na Educação. In: *Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa*. [S.l.: s.n.], 2020. v. 2. Citado 2 vezes na(s) página(s) 26 e 27.

FAYYAD, U. M.; IRANI, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. In: *International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 1993. p. 1022–1029. Citado na página 42.

FERNÁNDEZ, A. *et al. Learning from imbalanced data sets*. [S.l.]: Springer, 2018. v. 10. Citado 5 vezes na(s) página(s) 12, 24, 26, 27 e 69.

FILIP, J.; KLIEGR, T. *Classification based on Associations (CBA) - a performance analysis*. 2018. Citado 2 vezes na(s) página(s) 11 e 17.

HAAS, O.; MAIER, A.; ROTHGANG, E. Rule-Based Models for Risk Estimation and Analysis of In-hospital Mortality in Emergency and Critical Care. *Front Med (Lausanne)*, v. 8, p. 785711, 2021. Citado 2 vezes na(s) página(s) 11 e 17.

HAHSLER, M. *et al.* Associative classification in R: Arc, arulesCBA, and rCBA. *R Journal*, v. 11, n. 2, p. 254–267, 2019. Citado 2 vezes na(s) página(s) 11 e 17.

HASSINE, M. A. B.; ABDELLATIF, S.; YAHIA, S. B. A novel imbalanced data classification approach for suicidal ideation detection on social media. *Computing*, v. 104, n. 4, p. 741–765, 2022. Citado 3 vezes na(s) página(s) 12, 30 e 43.

HU, L. *et al.* Building an associative classifier with multiple minimum supports. *SpringerPlus*, v. 5, n. 528, 2016. Citado na página 30.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. [S.l.], 2007. Citado 2 vezes na(s) página(s) 26 e 27.

KLIEGR, T. QCBA: Postoptimization of quantitative attributes in classifiers based on association rules. *arXiv:1711.10166*, 2019. Citado na página 33.

LAKKARAJU, H.; BACH, S. H.; LESKOVEC, J. Interpretable decision sets: A joint framework for description and prediction. In: KRISHNAPURAM, B. *et al.* (Ed.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. [S.l.]: Acm, 2016. p. 1675–1684. Citado na página 25.

LIEWLOM, P. Class-association-rules pruning by the profitability-of-interestingness measure: Case study of an imbalanced class ratio in a breast cancer dataset. v. 12, n. 3, p. 246–252, 2021. Citado na página 30.

LIU, B.; HSU, W.; MA, Y. Integrating classification and association rule mining. In: *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98),*. [S.l.: s.n.], 1998. p. 27–31. Citado 2 vezes na(s) página(s) 11 e 17.

LIU, B.; HSU, W.; MA, Y. Mining association rules with multiple minimum supports. In: *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 1999. p. 337—341. Citado 2 vezes na(s) página(s) 16 e 17.

LIU, B.; MA, Y.; WONG, C.-K. Classification using association rules: weaknesses and enhancements. In: *Data mining for scientific and engineering applications*. [S.l.: s.n.], 2001. p. 591–605. Citado 2 vezes na(s) página(s) 12 e 18.

MARGOT, V.; LUTA, G. A new method to compare the interpretability of rule-based algorithms. *Ai*, v. 2, n. 4, p. 621–635, 2021. Citado 2 vezes na(s) página(s) 24 e 25.

MATTIEV, J.; DAVITYAN, M.; KAVSEK, B. ACMKC: A compact associative classification model using K-Modes clustering with rule representations by Coverage. *Mathematics*, v. 11, n. 18, 2023. ISSN 2227-7390. Citado na página 25.

MATTIEV, J.; MEZA, C.; KAVSEK, B. The effect of "Directness" of the distance metric to produce compact and accurate associative classification models. *Applied Sciences*, v. 12, n. 18, 2022. Citado na página 25.

MOHAMMAD, R. M. A. An improved multi-class classification algorithm based on association classification approach and its application to spam emails. *IAENG International Journal of Computer Science*, v. 47, n. 2, p. 187–198, 2020. Citado 2 vezes na(s) página(s) 11 e 17.

MOLNAR, C. *Interpretable Machine Learning*: A guide for making black box models explainable. 2. ed. [S.l.: s.n.], 2022. Citado na página 24.

PADILLO, F.; LUNA, J. M.; VENTURA, S. LAC: Library for associative classification. *Knowledge-Based Systems*, v. 193, 2020. Citado 2 vezes na(s) página(s) 11 e 17.

PIRAN, N. *et al.* Diabetic foot ulcers risk prediction in patients with type 2 diabetes using classifier based on associations rule mining. *Sci Rep*, v. 14, n. 635, 2024. Citado 2 vezes na(s) página(s) 11 e 17.

PRIVATTO, V. H. M.; CARVALHO, V. O. Associative classifiers algorithms for imbalanced data: A systematic literature review. In: *19th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.: s.n.], 2024. p. No prelo. Citado 3 vezes na(s) página(s) 13, 26 e 70.

RAJAB, K. D. New associative classification method based on rule pruning for classification of datasets. *IEEE Access*, v. 7, p. 157783–157795, 2019. Citado na página 25.

RUDIN, C. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. [S.l.]: Springer US, 2019. 206–215 p. Citado 2 vezes na(s) página(s) 24 e 69.

SEN, D. *et al.* Associative classifier coupled with unsupervised feature reduction for dengue fever classification using gene expression data. *IEEE Access*, v. 10, p. 88340–88353, 2022. Citado 2 vezes na(s) página(s) 11 e 17.

SHAO, Y. *et al.* Software defect prediction based on class-association rules. In: IEEE. *2017 Second International Conference on Reliability Systems Engineering (ICRSE)*. [S.l.], 2017. p. 1–5. Citado 2 vezes na(s) página(s) 11 e 17.

SHAO, Y. *et al.* A novel software defect prediction based on atomic class-association rule mining. *Expert Systems with Applications*, v. 114, p. 237–254, 2018. Citado 2 vezes na(s) página(s) 12 e 30.

SHAO, Y. *et al.* Software defect prediction based on correlation weighted class association rule mining. *Knowledge-Based Systems*, v. 196, p. 105742, 2020. Citado 2 vezes na(s) página(s) 12 e 30.

SHARMA, R. *et al.* Expected vs. unexpected: Selecting right measures of interestingness. In: SONG, M. *et al.* (Ed.). *Big Data Analytics and Knowledge Discovery - 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14-17, 2020, Proceedings*. [S.l.]: Springer, 2020. (Lecture Notes in Computer Science, v. 12393), p. 38–47. Citado 5 vezes na(s) página(s) 13, 15, 19, 20 e 49.

SOMYANONTHANAKUL, R.; THEERAMUNKONG, T. Scenario-based analysis for discovering relations among interestingness measures. *Information Sciences*, v. 590, p. 346–385, 2022. Citado 9 vezes na(s) página(s) x, 13, 19, 20, 22, 32, 35, 43 e 49.

SOOD, N.; ZAIANE, O. *Building a Competitive Associative Classifier*. [S.l.]: arXiv, 2020. Citado na página 25.

TAN, P.-N. *et al. Introduction to Data Mining*. 2. ed. [S.l.: s.n.], 2019. Citado 3 vezes na(s) página(s) 11, 12 e 24.

TEW, C. *et al.* Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, v. 28, n. 4, p. 1004–1045, 2014. Citado 9 vezes na(s) página(s) 13, 19, 20, 22, 32, 35, 43, 49 e 56.

WAIYAMAI, K.; SUWANNARATTAPHOOM, P. A Cost-Sensitive Based Approach for Improving Associative Classification on Imbalanced Datasets. In: SPRINGER. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. [S.l.], 2014. p. 31–42. Citado na página 30.

YANG, G.; CUI, X. A study of interestingness measures for associative classification on imbalanced data. In: *Trends and Applications in Knowledge Discovery and Data Mining*. [S.l.]: Springer, 2015. p. 141–151. Citado 8 vezes na(s) página(s) 13, 20, 21, 22, 32, 49, 56 e 65.

DADOS CURRICULARES

Identificação

Nome completo: Vitor Hugo Monteiro Privatto
Data de nascimento: 27/01/1987
Nacionalidade: Brasileira
Nome em citações bibliográficas: Privatto, V. H. M.

Formação Acadêmica

2016 - 2019: Sistemas de Informação (Bacharel) – FHO - Fundação Hermínio Ometto

Produção Bibliográfica

PRIVATTO, V. H. M.; CARVALHO, V. O. Associative classifiers algorithms for imbalanced data: A systematic literature review. In: 19th Iberian Conference on Information Systems and Technologies (CISTI). [S.l.: s.n.], 2024. p. No prelo.