UNIVERSIDADE ESTADUAL DE PONTA GROSSA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM
COMPUTAÇÃO APLICADA

LUÍS GUILHERME RIBEIRO

ESQUEMA DE SELEÇÃO DINÂMICA PARA A IDENTIFICAÇÃO DE BACTÉRIAS A PARTIR DE DADOS DE *M/Z* VIRTUAIS DE PROTEÍNAS RIBOSSOMAIS

PONTA GROSSA

2020

LUÍS GUILHERME RIBEIRO

ESQUEMA DE SELEÇÃO DINÂMICA PARA A IDENTIFICAÇÃO DE BACTÉRIAS A PARTIR DE DADOS DE *M/Z* VIRTUAIS DE PROTEÍNAS RIBOSSOMAIS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, como parte dos requisitos para obtenção do título de mestre em Computação Aplicada na área de Computação, Automação e Gestão de Dados em Agricultura.

Orientador: Prof. Dr. José Carlos Ferreira da Rocha
Coorientador: Prof Dr. Rafael Mazer Etto

PONTA GROSSA

2020

Ribeiro, Luís Gulherme
R484 Esquemas de seleção dinâmica para a identificação de bactérias a partir de dados de m/z virtuais de proteínas ribossomais / Luís Gulherme Ribeiro. Ponta Grossa, 2020.
74 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientador: Prof. Dr. José Carlos Ferreira da Rocha.
Coorientador: Prof. Dr. Rafael Mazer Etto.

1. Meta-aprendizagem. 2. Aprendizado de máquina. 3. Metacaracterística. 4. Maldi-tof. I. Rocha, José Carlos Ferreira da. II. Etto, Rafael Mazer. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 004

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - https://uepg.br

TERMO

TERMO DE APROVAÇÃO

Luís Guilherme Ribeiro

ESQUEMAS DE SELEÇÃO DINÂMICA PARA A IDENTIFICAÇÃO DE BACTÉRIAS A PARTIR DE DADOS DE M/Z VIRTUAIS DE PROTEÍNAS RIBOSSOMAIS

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Prof. Dr. José Carlos Ferreira da Rocha - UEPG

Prof(a). Dr(a). Carolina Weigert Galvão - UEPG

Prof(a). Dr(a). Mauren Louise Sguario - UTFPR

Ponta Grossa, 10 de agosto de 2020.

Documento assinado eletronicamente por **MAUREN LOUISE SGUARIO COELHO DE ANDRADE**, **Usuário Externo**, em 10/08/2020, às 12:06, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Jose Carlos Ferreira da Rocha**, **Coordenador(a) do Programa de Pós-Graduação em Computação Aplicada - Mestrado**, em 10/08/2020, às 12:09, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Carolina Weigert Galvao**, **Professor(a)**, em 10/08/2020, às 15:24, conforme art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando

o código verificador **0251714** e o código CRC **DB9FC2B1**.

Dedico este trabalho a minha mãe Izabel Rodrigues
que por amor fez de tudo por mim.

AGRADECIMENTOS

À Deus, que esteve presente em todos os momentos, me guiou com sua luz divina, ouviu minhas preces e me fortaleceu. A Ti, meu Deus, toda honra e toda glória eternamente.

A minha mãe Izabel Rodrigues, que durante minha vida sempre me deu amor, apoio e acreditou em meus sonhos.

Aos meus irmãos Caroline e Fábio que sempre estiveram presente, me deram apoio e momentos de lazer.

Ao meu pai Gilmar Ribeiro, que mesmo distante sempre me incentivou com sábias palavras.

A minha madrinha Shirlene que me ajudou minha vida toda, principalmente através de orações.

A minha prima Luciane que sempre me ouve, me dá conselhos e me ajuda em todas as situações.

A minha namorada Thaís Kruger que me ouve e me dá conselhos todos os dias, além de ser o amor de minha vida.

Ao meu orientador Prof. Dr. José Carlos, que desde a graduação tem me acompanhado e acreditado em meu potencial. Além disso, por sempre estar presente, compartilhando seu tempo e conhecimento na área acadêmica.

Ao meu amigo Everaldo, que desde a infância esteve presente em todos os momentos da minha vida.

Aos meus amigos Murilo, Bruno, Renann, Rodrigo, Marlom, Fábio, Alisson, Henike, entre outros. Que estiveram presentem em minha vida, me proporcionando conselhos, risos e momentos de lazer.

“Não existe triunfo sem perda, não há vitória sem sofrimento e nem liberdade sem sacrifício. A vitória pertence aquele que acredita nela por mais tempo. Cada minuto que passa é uma nova chance para mudar tudo para sempre. Lembre- se que o mundo está nas mãos daqueles que tem coragem de viver seus sonhos. Cada vez que vencemos um obstáculo, descobrimos que valeu a pena. Não há nada impossível, porque os sonhos de ontem são as esperanças de hoje e podem converter- se em realidade amanhã. Alcançar o sucesso na vida é a capacidade de enfrentar o fracasso sem perder o entusiasmo. Existe tempo para tudo, só basta acreditar que podemos ser capazes.” (Autor Desconhecido)

RESUMO

Um grama de solo pode conter até 8,3 milhões de diferentes espécies bacterianas, entre elas existem as que favorecem a produtividade agrícola, promovendo o crescimento vegetal e protegendo contra pragas e doenças. Uma das abordagens que tem sido utilizada na identificação destes microrganismos é por meio da impressão digital de proteínas ribossomais de bactérias em espectros de massa extraídos através de uma técnica de química analítica chamada de *MALDI-TOF*. Esta técnica extrai informações de massa carga *(m/z)* das moléculas de uma amostra, que podem ser submetidas à classificadores digitais que rotulam os dados da amostra de acordo com seu nível de taxonomia. Porém, na maioria dos casos, estes conjuntos de dados possuem múltiplas classes e alto índice de desbalanceamento, o que tem dificultado o desenvolvimento de classificadores digitais. Assim, os sistemas de múltiplos classificadores proveem maneiras de tratar estes problemas. Neste contexto, os esquemas de seleção dinâmica que usam meta-aprendizagem exploram um conjunto de metacaracteristicas, extraídas do conjunto de treinamento, para estimar o nível de competência dos classificadores base e então selecionar o conjunto de classificadores mais aptos para predizer uma amostra. Considerando o exposto, este trabalho avalia o desempenho de diferentes esquemas de seleção dinâmica para identificação de gêneros bacterianos a partir de sua impressão digital em termos de *m/z* de proteínas ribossomais. O trabalho também apresenta o esquema denominado de *METADES-i* que estende o esquema *METADES*, por meio do uso de metacaracterísticas sensíveis a dados desbalanceados. O desempenho dos esquemas de seleção dinâmica foi mensurado em termos de acurácia média, acurácia balanceada, média geométrica e *overfitting* sobre um conjunto de dados sintéticos, chamado de *PUKYU*. Nos experimentos foram utilizados diferentes cenários, definidos como subconjuntos do conjunto de dados *PUKYU*. Além disso, foi analisado a influência da composição dos classificadores base: homogênea ou heterogênea. Os resultados dos experimentos demonstram que o esquema proposto foi significativamente superior aos demais em termos de acurácia balanceada e média geométrica. Em termos de acurácia média, o esquema *METADES-i* foi superior somente quando utilizou a composição homogênea. No que tange ao *overfitting*, os esquemas com melhor desempenho foram o *METADES-i*, *KNOP* e *KNORA-U*. A análise multiobjetivo entre a acurácia balanceada e o *overfitting* indicou que o esquema *METADES-i* participou da fronteira de dominância em todos os cenários. O resultado relacionado ao procedimento de seleção de metacaracterísticas, indicou que ao aplicar o método *Relief* ao *METADES-i* obteve uma melhora nas métricas de assertividade, porém houve queda no desempenho em termos de *overfitting*. Com relação a composição dos classificadores, a composição heterogênea mostrou-se superior na maioria dos casos. Finalmente, os resultados sugerem que a adequação do subconjunto de metacaracterísticas de esquemas de seleção dinâmica baseado em meta-aprendizagem, pode incrementar o desempenho de sistemas com múltiplos classificadores em termos de assertividade em conjunto de dados desbalanceados.

Palavras-chave: Meta-aprendizagem. Aprendizado de Máquina. Metacaracterística. MALDI-TOF

ABSTRACT

A gram of soil can contain up to 8.3 million different bacterial species, among which are those that favor agricultural productivity, promoting plant growth and protecting against pests and diseases. One approach that has been used in identifying these microorganisms is through the fingerprint ribosomal proteins in bacterial mass spectra extracted by a chemical analytical technique called *MALDI-TOF*. This technique extracts charge mass information *(m/z)* from the molecules of a sample, which can be subjected to digital classifiers that label the sample data according to their taxonomy level. However, in most cases, these data sets have multiple classes and high imbalance ratio, making it difficult to develop digital classifiers. Thus, the multiple classifier systems provide ways to treat these problems. In this context, dynamic selection schemes that use meta-learning explore a set of meta-features, extracted from the training set, to estimate the level of competence of the base classifiers and then select the most suitable set of classifiers to predict a sample. Considering the above, this work evaluates the performance of different dynamic selection schemes for the identification of bacterial genera from their fingerprint in terms of *m/z* of ribosomal proteins. The work also presents the scheme called *METADES-i* that extends the *METADES* scheme, through the use of meta-features sensitive to imbalanced data. The performance of the tested dynamic selection schemes was measured in terms of average accuracy, balanced accuracy, geometric mean and overfitting in a PUKYU synthetic data set. In the experiments different scenarios were used, defined as subsets of the data set *PUKYU*. In addition, the influence of the composition of the base classifiers: homogeneous or heterogeneous was analyzed. The results of the experiments show that the proposed scheme was significantly superior to the others in terms of balanced accuracy and geometric mean. In terms of average accuracy, the *METADES-i* scheme was superior only when using the homogeneous composition. Regarding overfitting, the schemes with the best performance were *METADES-i*, *KNOP* and *KNORA-U*. The multiobjective analysis between balanced accuracy and overfitting indicated that the *METADES-i* scheme participated in the dominance frontier in all scenarios. The result related to the meta-featus selection procedure, indicated that when applying the *Relief* method to *METADES-i*, there was an improvement in assertiveness metrics, but there was a decrease in performance in terms of overfitting. Regarding the composition of the classifiers, the heterogeneous composition proved to be superior in most cases. Finally, the results suggest that the adequacy of the meta-features subset of dynamic selection schemes based on meta-learning, can increase the performance of multiple classifier systems in terms of assertiveness in imbalanced data sets.

Keywords: Meta-Learning. Machine Learning. Meta-features. MALD-TOF

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

MALDI-TOF	*Matrix Assisted Laser Dessortion Ionization - Time of Flight*
m/z	Massa/Carga
SMC	Sistema de Múltiplos Classificadores
ESD	Esquema de Seleção Dinâmica
PMF	*Peptide Mass Fingerprint*
SVM	*Support Vector Machine*
MLP	*Multi-Layer Perceptron*
CDSD	Conjunto de Dados de Seleção Dinâmica
KNORA	*K-Nearest Oracles*
KNOP	*K-Nearest Output Profiles*
META-DES	*Meta-learning Dynamic Ensemble Selection*
LCA	*Local Class Accuracy*
DESMI	*Dynamic Ensemble Selection for Multi-class Imbalanced*
RUS	*Random Under-Sampling*
ROS	*Random Over-Sampling*
SMOTE	*Synthetic Minority Oversampling Technique*
VP	Verdadeiro Positivo
FP	Falso Positivo
VN	Verdadeiro Negativo
FN	Falso Negativo
ID	Índice de Desbalanceamento
AM	Acurácia Média
AB	Acurácia Balanceada
OVAM	*Overfitting* - Acurácia Média
OVAB	*Overfitting* - Acurácia Balanceada
NB	*Naïve Bayes*

RL	Regressão Logística
CART	*Classification and Regression Tree*

SUMÁRIO

1 INTRODUÇÃO

Segundo Gans, Wolinsky e Dunbar (2005), um grama de solo pode conter 8,3 milhões de diferentes espécies bacterianas, dentre estas espécies, existem aquelas que contribuem para o crescimento vegetal, fornecendo nutrientes, reduzindo o uso de água e insumos, e protegendo contra pragas e doenças (VRIEZE, 2015). Neste sentido, o emprego de sistemas de identificação automática de bactérias em amostras de solos possibilita a detecção de microrganismos que influenciam no desenvolvimento vegetal (FOLEY *et al.*, 2011). Este tipo de informação é relevante porque provê meios para a tomada de decisão, ou seja, a identificação de bactérias fornece dados para a administração de nutrientes de forma apropriada que pode aumentar a fertilidade do solo bem como a reduzir à aplicação de agroquímicos (LABUSCHAGNE, 2003).

Uma maneira de abordar a identificação automática de bactérias em amostras de solo é utilizar algoritmos que sejam capazes de analisar dados extraídos das amostras para determinar a categoria taxonômica dos microrganismos nela contido (LEMAîTRE; NOGUEIRA; ARIDAS, 2016). Neste contexto, o termo categorização taxonômica é definido como um problema de classificação com múltiplas classes (politômica), pelo qual o objetivo é determinar a espécie, gênero, família, ordem, classe ou filo de uma bactéria a partir das características da amostra.

Neste aspecto, existem determinadas técnicas que extraem informações relacionada as bactérias. Uma das técnicas que tem se destacado pelo baixo custo e precisão na identificação, é a espectometria de massa do tipo MALDI-TOF (*Matrix Assisted Laser Dessortion Ionization - Time Of Flight*) (PASTERNAK, 2012). Pela qual extrai informações da massa carga (m/z), que são conhecidas como impressão digital de proteínas. Estas informações são comparadas a um banco de dados contendo sequências de proteínas ribossomais, e assim é composto um conjunto de dados bacterianos por espectometria de massa virtual de proteínas ribossomais (TOMACHEWSKI *et al.*, 2018a).

Em termos de modelos de aprendizagem de máquina, uma das dificuldades ao utilizar dados biológicos é que em geral estes dados são volumosos e desbalanceados (WEI *et al.*, 2017). O problema de desbalanceamento dificulta à aprendizagem do modelo de classificação porque, segundo Barella (2015), classificadores treinados em conjuntos de dados desbalanceados tendem a gerar modelos bastante acurados na categorização de instâncias das classes majoritárias e baixo desempenho nas classes minoritárias. Adicionalmente, como observa Galar *et al.* (2013), a maioria dos modelos de classificação supõe que a distribuição de amostra entre várias classes é balanceada. Assim, a escolha de uma função de classificação para determinado problema é um paradigma (WICHARD, 2006), pois cada problema é específico e exige especialidades de um modelo de classificação.

Segundo Galar *et al.* (2012), uma maneira de abordar simultaneamente o problema do desbalanceamento de classes e facilitar a escolha de uma função de classificação é o uso de sistemas de múltiplos classificadores (SMC), que combinam classificadores a fim de aumentar a assertividade na classificação. Seu funcionamento pode ser abstraído em três fases: geração, seleção e integração. Na fase de geração, um conjunto de classificadores é gerado. Na fase de seleção, um único classificador ou um subconjunto com os melhores classificadores do conjunto é selecionado. E na fase de integração, os classificadores selecionados são combinados para obter a decisão (KITTLER; HATER; DUIN, 1996).

Uma das estratégias de seleção empregadas por um SMC é a seleção dinâmica. Nela a escolha de um classificador ou de um subconjunto de classificadores é realizada de acordo com a nova amostra a ser classificada, ao contrário da seleção estática que seleciona os classificadores base durante o estágio de treinamento, e no estágio de teste utiliza os mesmos classificadores para predizer todas as amostras não vistas. Além disso, as técnicas de seleção dinâmica visam selecionar os classificadores mais competentes para a região local onde a amostra de teste está localizada (CRUZ *et al.*, 2015). Portanto, um dos propósitos de um esquema de seleção dinâmica (ESD) é definir um critério para medir o nível de competência de um classificador base.

Para selecionar os classificadores mais competentes, os esquemas de seleção dinâmica utilizam meta-aprendizagem, onde é extraído conhecimento por meio de metadados. Neste caso, a tarefa de seleção dos classificadores é tratado como um metaproblema que possuí metacaracterísticas, em que utilizam diferentes critérios para mensurar o nível de competência dos classificadores, considerando a região local onde se encontra a amostra. As metacaracterísticas são usadas como entrada para um metaclassificador que decide se um classificador base é competente ou não para realizar uma predição. Dentre as metacaracterísticas mais usadas na literatura, temos as baseadas em acurácia (geral e local) (WOODS; KEGELMEYER; BOWYER, 1997), probabilidade a posteriori (GIACINTO; ROLI, 1999), perfil de saída dos classificadores (CAVALIN; SABOURIN; SUEN, 2013), oráculo (KO; SABOURIN; JR, 2008), etc.

Considerando o exposto, este trabalho tem como objetivo avaliar a eficiência de esquemas de seleção dinâmica para categorização automática de gêneros bacterianos. A categoria taxonômica de gêneros foi escolhido por ser um fator relevante para identificar bactérias que contribuem para o crescimento vegetal, como é o caso do gênero *RHIZOBIUM*. Neste sentido, este trabalho considera a tarefa de identificar o gênero das bactérias a partir de dados extraídos do espectro de massa do tipo MALDI-TOF. Para tanto, o atingimento deste do objetivo depende da análise de diversos fatores, entre eles:

1. A arquitetura do SMC;

2. A composição dos classificadores base;
3. As metacaracterísticas empregadas;
4. O índice de desbalanceamento do conjunto de dados;

Assim, a fim de atingir tais objetivos, foi proposto um esquema nomeado de *METADES-i* (*METADES-imbalanced*), que é uma extensão do *METADES* incorporando metacaracterísticas baseadas em métricas de desempenho sensíveis ao desbalanceamento de classes. Assim, foram realizados sete experimentos na comparação dos esquemas de seleção dinâmica mais conhecidos na literatura, utilizando as composições de classificadores: homogêneas e heterogêneas.

Durante os experimentos, foram criados cenários variando os índices de desbalanceamento de subconjuntos do conjunto de dados *PUKYU*. Nos três primeiros experimentos foram avaliados os esquemas em termos de acurácia média, acurácia balanceada e média geométrica. No quarto e quinto experimento, os esquemas foram avaliados em termos de *overfitting*, utilizando as métricas de acurácia média e balanceada. No sexto experimento, foi aplicado um algoritmo para a seleção de metacaracterísticas na fase de treinamento do esquema *METADES-i*. E no último experimento foi comparado o desempenho do esquema *METADES-i* com sua versão após aplicar o procedimento de seleção de metacaracterísticas.

Este trabalho está organizado da seguinte maneira. O Capítulo 3 contém a revisão bibliográfica sobre o problema da classificação automática de bactérias, sistemas de classificadores múltiplos e esquemas de seleção dinâmica. O Capítulo 4 (metodologia) será abordado a arquitetura do esquema *METADES-i*, e os protocolos experimentais realizados. O Capítulo 5 aborda os resultados e discussões dos experimentos. E por fim, o Capítulo 6 contém a conclusão e trabalhos futuros.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Avaliar o desempenho de esquemas de seleção dinâmica baseados em meta-aprendizagem para identificação de bactérias em nível de gênero a partir de dados de *m/z* virtuais de proteínas ribossomais.

2.2 OBJETIVOS ESPECÍFICOS

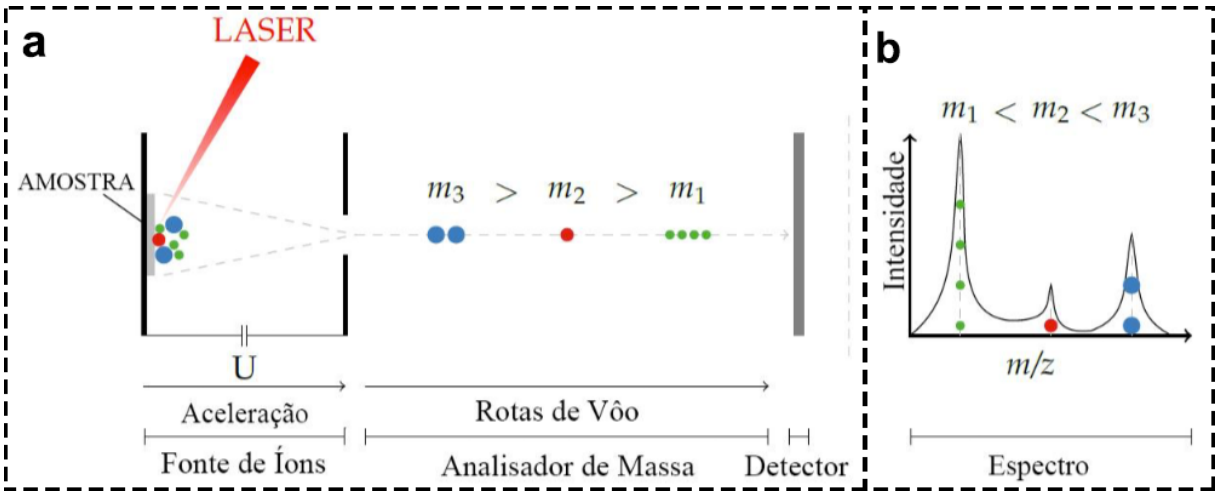
- Comparar o desempenho de esquemas de seleção dinâmica sobre diferentes cenários, considerando o índice de desbalanceamento de subconjuntos do conjunto de dados *PUKYU*.
- Determinar se existem diferenças nos desempenhos dos ESDs em relação ao emprego de composições do conjunto de classificadores base homogêneos ou heterogêneos;
- Estender um esquema de seleção dinâmica utilizando metacaracterísticas sensíveis a dados desbalanceados.
- Estimar a eficácia do procedimento de seleção de metacaracterísticas no desempenho do esquema proposto.

3 REVISÃO DA LITERATURA

3.1 IDENTIFICAÇÃO BACTERIANA A PARTIR DA ESPECTROMETRIA DE MASSA

O valores de massa/carga (*m/z*) das proteínas celulares, obtidos por espectrometria de massa do tipo *MALDI-TOF* fornecem dados que pode ser usados na identificação bacteriana . De acordo com Wieser *et al.* (2012) esta técnica de química analítica consiste na deposição de uma determinada amostra sobre uma matriz de reagentes químicos capaz de fornecer prótons para o processo de ionização das moléculas da amostra. Quando esta matriz absorve a energia emitida por um laser, ocorre a transferência de prótons da matriz para as moléculas e ao mesmo tempo desencadeia-se um processo de dessorção, possibilitando a passagem da amostra do estado sólido para o gasoso. Estas moléculas são submetidas a um campo elétrico que as acelera em direção a um tubo metálico capilar de vácuo até atingirem um detector responsável para captar os sinais para a geração do espectro, conforme é mostrado na Figura 1.

Figura 1: Esquema da técnica de espectro de massa MALDI-TOF.



Fonte: Gibb (2015).

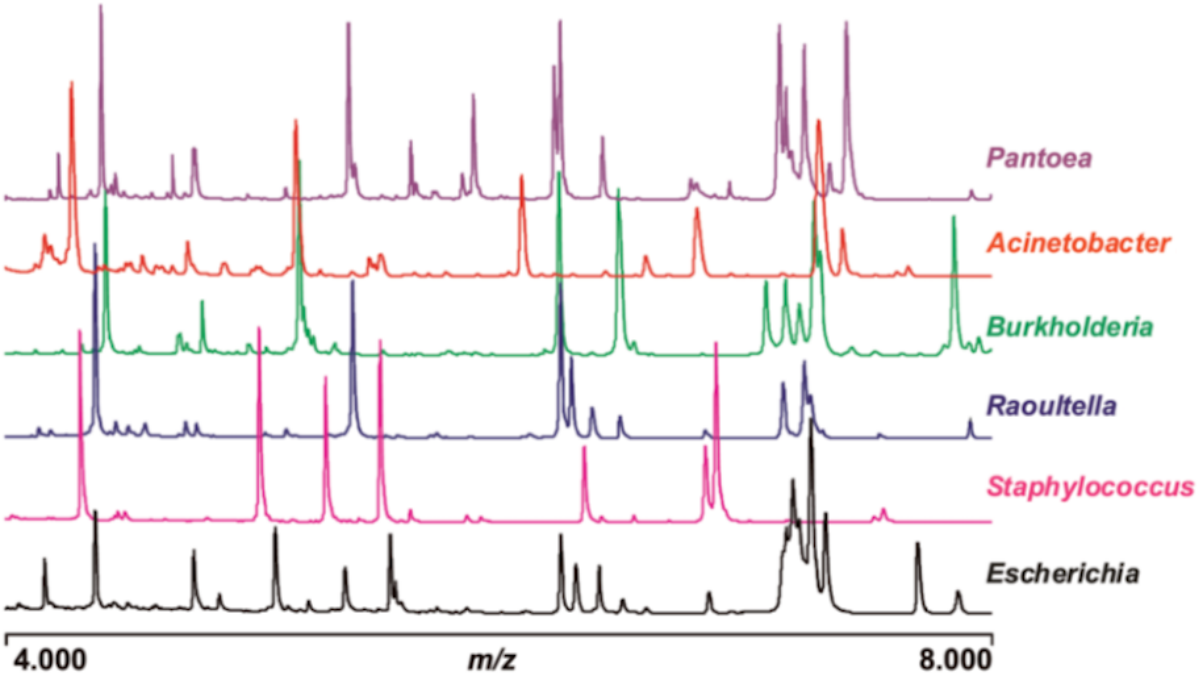
Em resumo, o espectro de massa obtido por *MALDI-TOF* é um histograma de duas dimensões em que o eixo das abcissas indica a relação *m/z* e o eixo das ordenadas indica a intensidade do pico (JURGEN, 2018), conforme ilustrado na Figura 1b. Como o valor de *m/z* é proporcional ao tempo de voo, um valor maior que zero em uma abcissa indica a presença de uma determinada partícula na amostra dando origem ao espectro (TODD, 1995).

Segundo Tamura, Hotta e Sato (2013), uma maneira de identificar espectros de massa gerados pelo MALDI-TOF é por meio de um procedimento chamado de "impressão digital do mapa peptídico", do inlgês *Peptide Mass Fingerprint* (PMF), que realiza uma análise espectrométrica dos peptídeos, fazendo com que seja possível identificar proteínas em bancos de dados sequenciais sem a necessidade de sequenciar a proteína em estudo (MELO, 2014). Assim, cada

pico é associado a uma proteína ribossomal.

De acordo com Teramoto *et al.* (2007), a vantagem de utilizar proteínas ribossomais é devido sua alta conservação e sua expressão constitutiva em diferentes condições fisiológicas, independente de meios de crescimento ou ciclo celular. Assim, as proteínas ribossomais podem ser consideradas confiáveis para identificação bacteriana. Na Figura 2 é ilustrado algumas bactérias a nível taxonômico do gênero, pelo qual sua identificação é caracterizado por meio dos picos referentes à proteínas ribossomais.

Figura 2: Exemplo de diferentes gêneros bacterianos, que são identificadas por meio dos picos referentes à proteínas ribossomais.



Fonte: García *et al.* (2012).

3.2 PROBLEMA DE CLASSIFICAÇÃO BACTERIANA

Um classificador é uma função $\boldsymbol{f} : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ em que $\boldsymbol{X} = \{\boldsymbol{X}_1 \ldots \boldsymbol{X}_p\}$ é um conjunto de variáveis que representam os atributos ou características dos objetos de um domínio de interesse e $\boldsymbol{Y}$ é uma variável cujos valores $\boldsymbol{y}_1 \ldots \boldsymbol{y}_n$ simbolizam os rótulos que identificam a classe ou categoria desses objetos. É assumido que $\boldsymbol{Y}$ provê uma enumeração exaustiva e mutuamente exclusiva dos rótulos de classe, pelo qual se define como um problema multiclasse (HAIXIANG *et al.*, 2016). Assim, dada uma instância $\boldsymbol{x} = \{\boldsymbol{x}_1 \ldots \boldsymbol{x}_p\} \in \boldsymbol{X}$ um classificador $\boldsymbol{f}$ retorna um rótulo de classe $\boldsymbol{y}_*$, que é chamada de classe de predição. A aprendizagem automática de classificadores tem como objetivo induzir funções de classificação a partir da inspeção de um conjunto de dados cujas instâncias registram os valores de atributos descritivos e o rótulo

de classe de uma coleção objetos observados (CARVALHO *et al.*, 2011). Tais conjuntos de dados são ditos supervisionados.

Neste trabalho, a aprendizagem de classificadores para identificação de bactérias objetiva em determinar uma função $\boldsymbol{f}(\boldsymbol{U})$ tal que $\boldsymbol{U} = \{\boldsymbol{u_1} \ldots \boldsymbol{u_p}\}$ enumera as intensidades de picos dos espectros de massa MALDI-TOF relativos à proteínas ribossomais (HOTTA *et al.*, 2010). O valor retornado por $\boldsymbol{f}$ é um rótulo $\boldsymbol{y}$ que informa o gênero da amostra analisada, dado o modelo de predição implementado pela função.

Segundo (BARELLA, 2015), o problema de múltiplas classes junto ao alto índice de desbalanceamento, afetam negativamente o desempenho dos modelos de classificação em termos de assertividade, pois quanto maior o número de classes tende a aumentar o índice de desbalanceamento, e dificultar que um classificador $\boldsymbol{f}(\boldsymbol{U})$ tenha um alto desempenho em termos de assertividade. No entanto, Pushpa e Karpagavalli (2017) afirmam que ao combinar a função de vários classificadores, que é o caso dos sistemas de múltiplos classificadores, o tratamento de problemas de classificação com múltiplas classes se tornam mais eficazes. Adicionalmente, segundo Krawczyk *et al.* (2017), a combinação de classificadores e o uso de métodos de reamostragem podem transformar a distribuição de amostras dos dados originais em classes mais balanceadas.

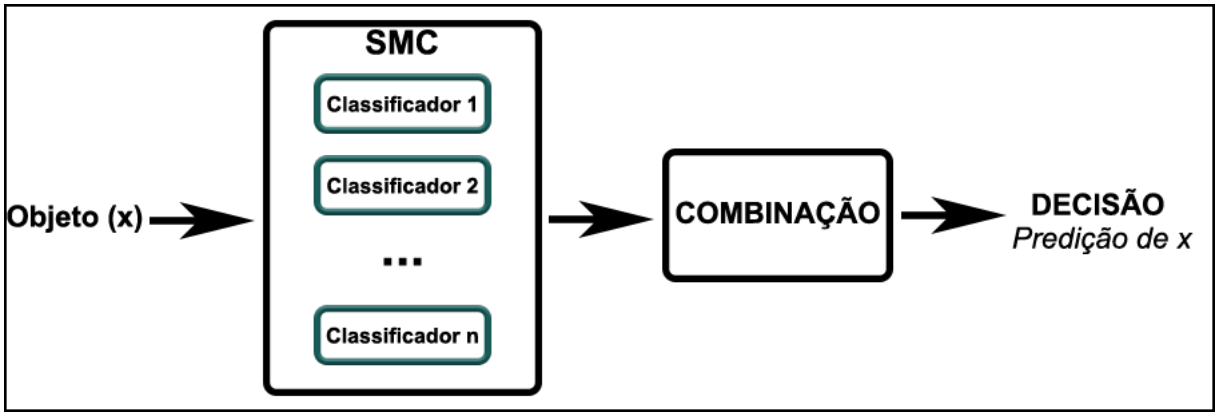
Recentemente na literatura diversos trabalhos têm abordado os classificadores digitais para identificar micro-organismos presentes no solo. No trabalho de Fan *et al.* (2018) é mostrado os principais algoritmos de classificação utilizados na literatura para classificação baseada no espectro de massa, os autores levantam: *Support Vector Machine* (SVM), Regressão Logística, Árvores de Decisão, Rede Neural *Multi-Layer Perceptron* (MLP), *k-Nearest Neighbors* (kNN) e *Random Forest* (*RF*). Adicionalmente, os autores afirmam que os SMCs têm demonstrado alta assertividade na classificação de espectro de massa, não apenas os métodos tradicionais como *AdaBoost* e *Gradient Boosting*, mas também métodos híbridos que usam diferentes heurísticas para combinar classificadores bases.

Em termos de dados extraídos pela técnica de espectro de massa do tipo MALDI-TOF, os autores Rossel e Arbizu (2018) propuseram um modelo para a identificação de espécies de crustáceos, utilizando o algoritmo do *Random Forest*. Já no trabalho de Tomachewski (2017), foi utilizado o algoritmo de classificação *Naïve Bayes* (*NB*), para predizer bactérias a nível de espécie e gênero, por meio de picos referentes a proteínas ribossomais, em seguida no trabalho Tomachewski *et al.* (2018b) foi implementado uma ferramenta de apoio para a identificação bacteriana através do picos de *m/z*.

3.3 SISTEMA DE MÚLTIPLOS CLASSIFICADORES

A intenção de um sistema de múltiplos classificadores é combinar classificadores, aumentando a capacidade preditiva de forma que o esquema composto tenha um desempenho superior a qualquer um dos classificadores que o constitui (ROKACH, 2010). A Figura 3 ilustra a estrutura básica de um SMC. Como pode ser observado, ele é composto por diversos classificadores tal que quando uma entrada $\boldsymbol{x}$ é dada, ela é processada por todos ou por alguns dos classificadores que fazem parte do SMC. O rótulo final atribuído a $\boldsymbol{x}$ é o obtido com a combinação dos resultados individuais. A escolha dos classificadores a serem usados no processamento de uma instância depende da especificação de um critério de seleção. Da mesma forma, o resultado final depende do procedimento aplicado na combinação dos resultados individuais.

Figura 3: Exemplo de um sistema de múltiplos classificadores.



Fonte: O autor.

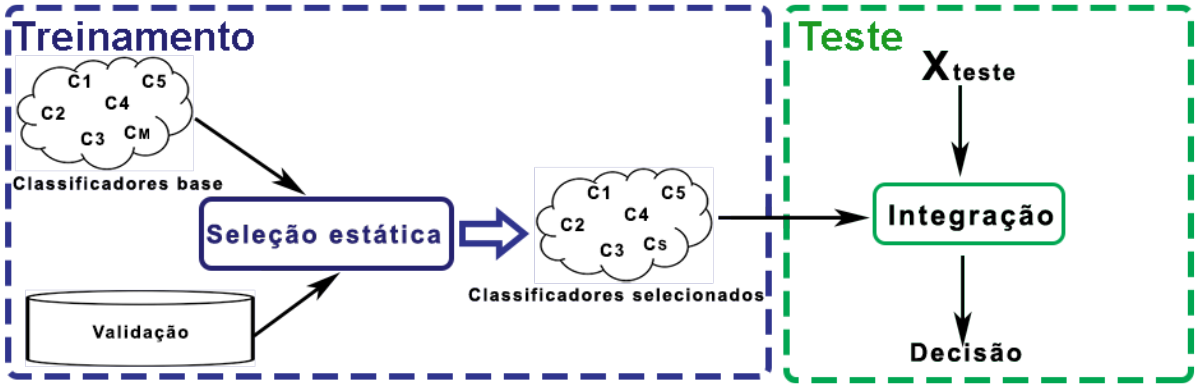
Em geral, o ciclo de vida de um SMC é composto por três etapas (JR; SABOURIN; OLIVEIRA, 2014): geração, seleção e integração. Na fase de geração, um conjunto de classificadores (*pool*) é gerado, estes classificadores são denominados de classificadores base e o sua composição pode ser homogênea ou heterogênea. Quando a composição é homogênea, todos os classificadores implementam a mesma função de classificação, por sua vez, na heterogênea cada classificador base tem sua função específica (SABZEVARI; MARTÍNEZ-MUÑOZ; SUÁREZ, 2018).

Na fase de seleção, um único classificador ou um subconjunto com os melhores classificadores do conjunto é selecionado. Existem duas abordagens básicas de seleção (JR; SABOURIN; OLIVEIRA, 2014): a seleção estática e a seleção dinâmica.

Na seleção estática, a escolha dos classificadores base é realizada durante o estágio de treinamento, conforme é ilustrado na Figura 4. Neste esquema de seleção é realizado o treinamento com os classificadores base e um conjunto de validação, então aplica-se uma certa heurística a fim de combinar os melhores classificadores base, para obter um esquema (classificadores selecionados). Assim no estágio de teste, sempre é usado o mesmo esquema para

predizer uma amostra não vista.

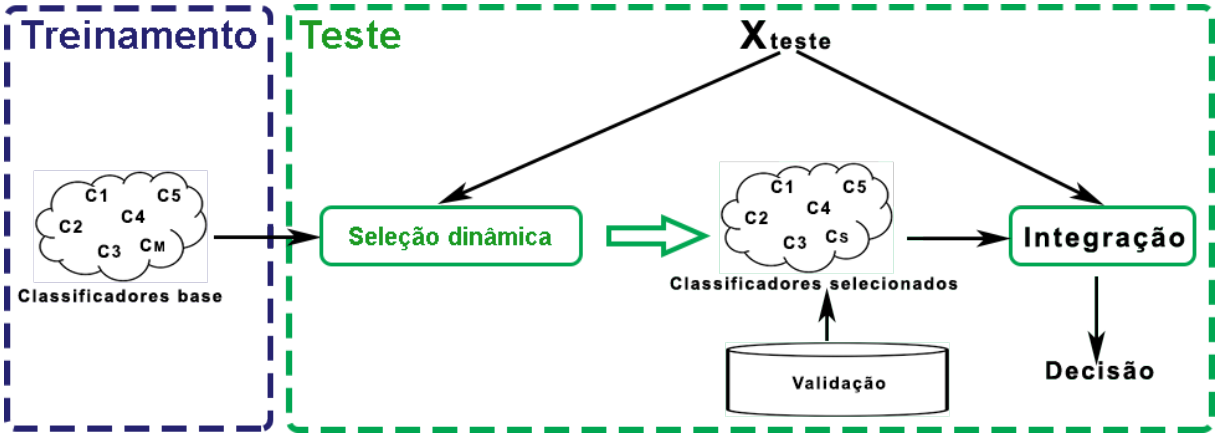
Figura 4: Exemplo de seleção estática.



Fonte: O autor.

Na abordagem dinâmica, a seleção dos classificadores ocorre no estágio de teste, para todas as instâncias não vistas (CRUZ; SABOURIN; CAVALCANTI, 2018). Como pode ser observado na Figura 5, a fase de treinamento desta abordagem é usada apenas para consultar os modelos de classificação gerados pelos classificadores base. No estágio de teste, para cada nova instância (X_{teste}) é construído um esquema de classificação por meio de uma heurística dinâmica que usa o conjunto de validação e os classificadores base. Em geral, esta abordagem visa selecionar os classificadores mais competentes na região local onde se encontra a instância a ser predita.

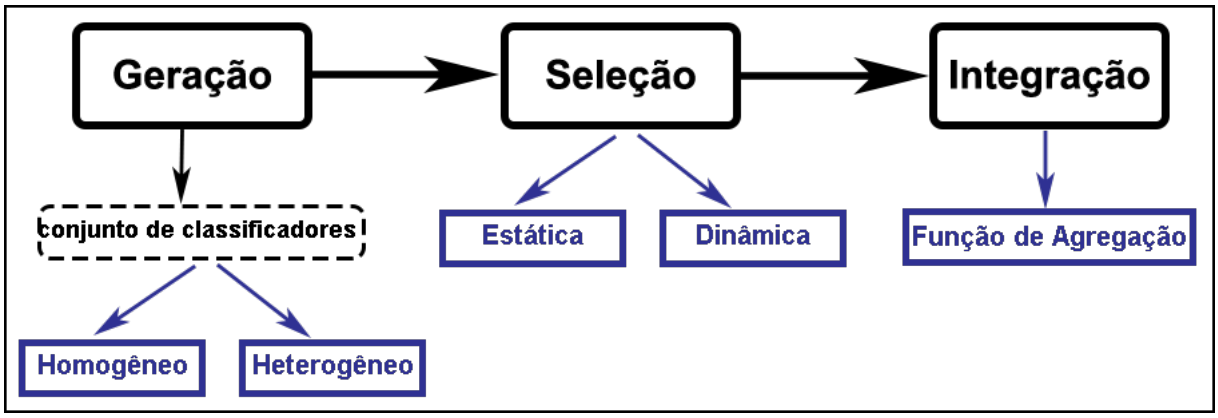
Figura 5: Exemplo de seleção dinâmica.



Fonte: O autor.

E por fim, a fase de integração, as predições dos classificadores selecionados para processar a instância a ser categorizada são combinadas para obter a decisão final, isto é feito por meio de uma função de agregação. Em suma, a Figura 6 ilustra o esquema completo de um SMC, especificando as particularidades de cada fase.

Figura 6: Estrutura de um SMC.



Fonte: O autor.

3.4 ESQUEMA DE SELEÇÃO DINÂMICA

Sabendo que um SMC é composto por um conjunto de classificadores base, ao qual nos referimos como $\boldsymbol{C}$. O objetivo de um esquema de seleção dinâmica é encontrar o melhor subconjunto de classificadores $\boldsymbol{C_i}$, em que $\boldsymbol{C_i} \subset \boldsymbol{C}$ para classificar a amostra de teste $\boldsymbol{x_i}$, tal subconjunto é definido como $\boldsymbol{C'}$, e na maioria dos casos sua escolha é fundamentada dentro de regiões específicas (CAVALIN; SABOURIN; SUEN, 2013). A literatura agrupa os métodos de ESD em duas categorias, a seleção dinâmica de um classificador e seleção dinâmica de conjuntos de classificadores. Na seleção dinâmica de um classificador, apenas um classificador é designado para cada amostra a ser classificada (ZHU; WU; YANG, 2004). Já na seleção dinâmica de conjuntos de classificadores, um conjunto de classificadores é escolhido para cada amostra (KO; SABOURIN; JR, 2008). Entretanto independente da categoria, o principal objetivo destes esquemas é encontrar o melhor subconjunto de classificadores $\boldsymbol{C_i}$ para classificar $\boldsymbol{x_i}$, tal melhor conjunto é associado ao nível mais alto de competência, pelo qual é estimado em regiões locais.

A definição das regiões locais é uma etapa relevante para a construção de um ESD, uma vez que a seleção dos classificadores é fundamentada no seu desempenho nestas regiões, a caracterização destas regiões afeta o desempenho de um ESD (CRUZ; SABOURIN; CAVALCANTI, 2017a). Geralmente, as regiões locais são definidas usando o algoritmo *kNN* (WOODS; KEGELMEYER; BOWYER, 1997), algum método de agrupamento (KUNCHEVA, 2000), as decisões dos classificadores de base (CAVALIN; SABOURIN; SUEN, 2012) ou um mapa de competência baseado em regressão (WOLOSZYNSKI; KURZYNSKI, 2011). A execução destes procedimentos depende da disponibilidade de um conjunto de amostras rotuladas, que pode ser o conjunto de treinamento ou validação. Os autores Cruz *et al.* (2015) denominam este conjunto como o conjunto de dados de seleção dinâmica (CDSD), pelo qual neste trabalho foi obtido por meio do algoritmo *kNN*, assim o valor de $\boldsymbol{K}$ é o tamanho de vizinhos contidos em um espaço de recurso

(região de competência).

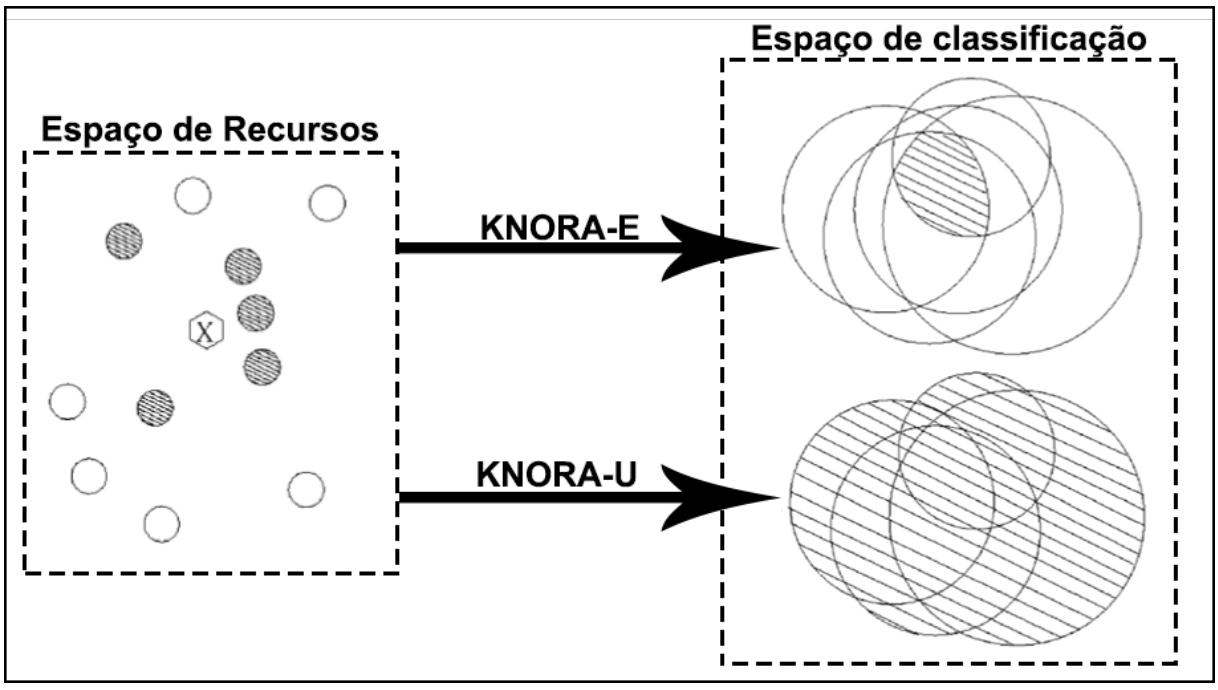
3.4.1 Esquemas de Seleção Dinâmica

A literatura descreve variados esquemas de seleção dinâmica. Todos têm o objetivo de selecionar os classificadores mais relevantes para realizar a predição, contudo, os esquemas utilizam diferentes critérios de seleção como (CRUZ; SABOURIN; CAVALCANTI, 2018): acurácia, probabilidade, oráculo, comportamento dos perfis de saídas, meta-aprendizagem, entre outros. A seguir é mostrado alguns esquemas de seleção dinâmica e explicado seus critérios de seleção.

3.4.1.1 K-Oráculos Mais Próximos (KNORA)

O esquema conhecido como K-Oráculos mais próximos, ou em inglês *K-Nearest Oracles* (KNORA) explora o critério de seleção do oráculo, pelo qual tem como objetivo verificar se uma amostra é classificada corretamente ou não. Assim, o esquema KNORA KNORA realiza a seleção dos classificadores mais aptos por meio de sua capacidade de acerto. De modo geral, este esquema simplesmente encontra seus $\boldsymbol{K}$ vizinhos mais próximos em seu espaço de recurso, descobre quais classificadores base classificam corretamente esses vizinhos e os utiliza para classificar por meio do voto majoritário (KO; SABOURIN; JR, 2008). Na literatura são definidas duas abordagens da técnica KNORA, nomeadas como *KNORA-U* (*Union*) e *KNORA-E* (*Eliminate*).

Figura 7: Processo de seleção do espaço de classificação pelo esquema *KNORA*.



Fonte: Ko, Sabourin e Jr (2008).

A Figura 7 é ilustrado o processo de seleção das abordagens. Conforme visto naquela figura, o esquema do *KNORA-E* seleciona os classificadores que reconhecem corretamente todas as amostras pertencentes à região de competência θ_j. Isto é, apenas os classificadores que atingiram 100% de acerto nessa região são selecionados para compor o conjunto C'. Porém, na abordagem do *KNORA-U*, o processo de seleção escolhe os classificadores que são capazes de reconhecer corretamente pelo menos uma amostra na região de competência θ_j. Nesta abordagem é levado em conta que um classificador base pode participar mais de uma vez no esquema de votação quando classifica corretamente mais de uma instância na região de competência.

3.4.1.2 K-Perfis de Saída Mais Próximos (KNOP)

O esquema K-Perfis de saída mais próximos, ou em inglês *K-Nearest Output Profiles* (KNOPKNOP) utiliza o critério do comportamento dos perfis de saída. Este esquema é semelhante ao *KNORA*, porém o *KNOP* trabalha no espaço de decisão ao invés do espaço de recursos (CRUZ; SABOURIN; CAVALCANTI, 2018). O espaço de decisão pode ser definido como as decisões tomadas (saídas) pelos classificadores base sobre o conjuntos de teste e validação. Isto é, as saídas dos classificadores são agrupadas como um conjunto de informações, que são denominadas como perfis de saída. Deste modo, a ideia desta abordagem é usar a decisão dos classificadores como o resultado de uma transformação que tem como entrada as informações das amostras de teste e validação.

Primeiramente, é aplicado uma transformação sobre a entrada x_j fornecendo seu perfil de saída, que é denotado por $\tilde{x_j} = [\tilde{x_{j,1}}, \tilde{x_{j,2}}, \tilde{x_{j,3}}, ..., \tilde{x_{j,Kp}}]$, em que cada $x_{j,i}$ é a decisão gerada pelo classificador base c_i para a amostra x_j. Em seguida, a semelhança entre $\tilde{x_j}$ e os perfis de saída do conjunto de dados de seleção dinâmica é calculada e armazenada no conjunto ϕ_j. Assim, cada vez que um classificador base executa uma predição correta, para uma amostra pertencente a ϕ_j obtém um voto, que são agregados para obter a decisão final. Por fim, os classificadores que obtiverem mais acertos, são os escolhidos para classificar a instância em questão.

3.5 SELEÇÃO DINÂMICA BASEADO EM META-APRENDIZAGEM

A meta-aprendizagem tem como objetivo usar metadados para aprender, selecionar, alterar ou combinar diferentes algoritmos de aprendizado para resolver um problema (LEMKE; BUDKA; GABRYS, 2015). Os metadados são dados que fornecem informações sobre um ou mais aspectos de um conjunto de dados e de seus atributos (GARTNER, 2016). Assim, metadados podem ser informações como métricas de desempenho, predições, combinações, partições ou

procedimentos definidos sobre um conjunto de dados previamente fornecido.

Em geral, os algoritmos de classificação utilizam metadados com o objetivo de se adaptar em diferentes cenários (SOUZA, 2010), no que se refere ao desbalanceamento de dados, múltiplas classes e problemas de *overfitting*. Os autores Clavera *et al.* (2018) definiram três categorias de aplicações de meta-aprendizagem em algoritmos de classificação:

- Na etapa de treinamento: através do particionamento do conjunto de dados, combinação de classificadores e comparação métricas.
- Na etapa teste: por meio de buscas, comparação do desempenho e comportamento de classificadores.

A literatura mostra que vários autores têm desenvolvido SMC baseado em meta-aprendizagem, como é o caso do trabalho de Zhang, Jiang e Li (2017) que utiliza metadados de assertividade (oráculo) das instâncias na etapa de treinamento para combinar os classificadores individuais, além disso, os autores utilizam o algoritmo *kNN* para buscar a instância mais próxima na etapa de teste.

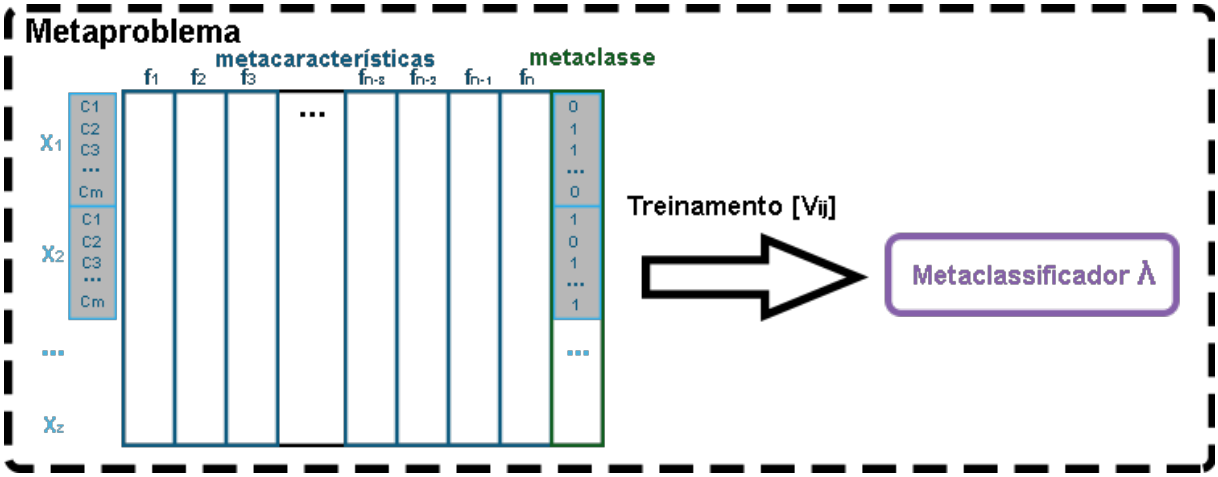
Em termos de seleção dinâmica, alguns trabalhos também exploram o emprego de métodos de meta-aprendizagem (CRUZ *et al.*, 2015), (CRUZ; SABOURIN; CAVALCANTI, 2015), (CRUZ; SABOURIN; CAVALCANTI, 2017b). No que se refere a seleção dinâmica, a meta-aprendizagem apresenta uma perspectiva diferente de como a competência dos classificadores base pode ser aprendida por meio de diferentes fontes de informação. Neste sentido, esta abordagem visa criar um metaproblema para extrair informações a fim de medir o nível de competência dos classificadores base. Na seção seguinte é mostrado um esquema específico de seleção dinâmica baseado em meta-aprendizagem.

3.5.1 Esquema do *META-DES*

Os autores Cruz *et al.* (2015) desenvolveram um esquema para seleção dinâmica de conjuntos de classificadores chamado de *META-DESMETA-DES* (*Meta-learning Dynamic Ensemble Selection*), pelo qual utiliza meta-aprendizagem como critério de seleção. Resumidamente, o *META-DES* gera um conjunto de metadados dando origem a uma coleção de informações que é definido como um metaproblema de aprendizado. Tal metaproblema é usado para selecionar os classificadores c_i mais aptos de acordo com seu nível de competência. A Figura 8 ilustra a estrutura do metaproblema gerado pelo *META-DES* por meio dos metadados, cujos elementos principais são:

- O atributo alvo (metaclasse) é binário, ou seja, um classificador base pode ser competente (1) ou incompetente (0) para predizer uma amostra $\boldsymbol{x_j}$.
- As metacaracterísticas $\boldsymbol{f_i}$ correspondem a deferentes critérios que procuram medir o nível de competência de um classificador base $\boldsymbol{c_i}$;
- O vetor de metacaracterísticas $\boldsymbol{v_{i,j}}$, em que o índice $\boldsymbol{i}$ indica as metacaracterísticas, e $\boldsymbol{j}$ as instâncias de treinamento.
- Um metaclassificador $\boldsymbol{\lambda}$ que é treinado com base nas metacaracterísticas $\boldsymbol{v_{i,j}}$ para prever se um classificador base $\boldsymbol{c_i}$ é competente o suficiente para estimar a instância $\boldsymbol{x_j}$.

Figura 8: Estrutura do metaproblema, criado na fase de treinamento do esquema *META-DES*.



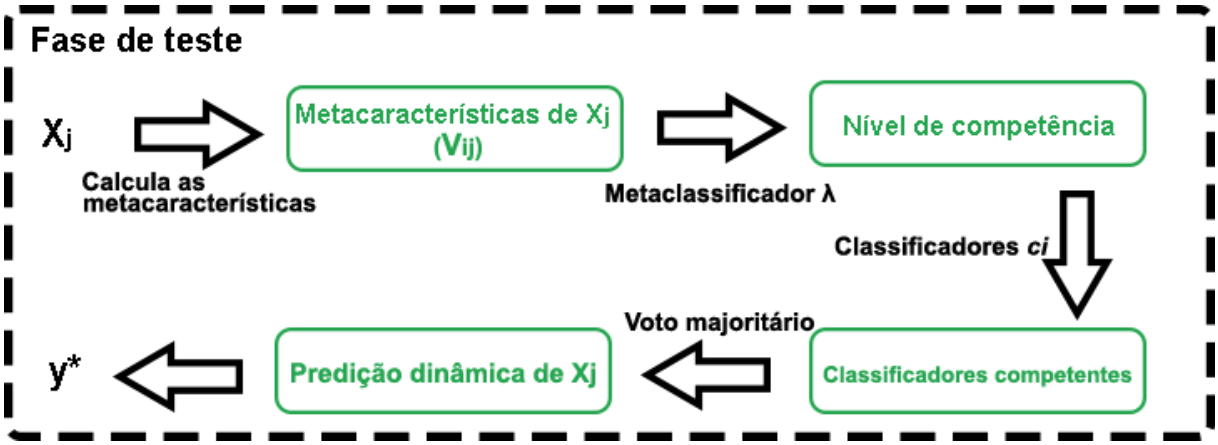
Fonte: O autor.

A vantagem de usar meta-aprendizagem é que vários critérios podem ser codificados como conjuntos diferentes de metacaracterísticas, a fim de estimar o nível de competência dos classificadores de base. Assim, um ESD é criado através do metaclassificador $\boldsymbol{\lambda}$ e das metacaracterísticas extraídas na etapa de treinamento (CRUZ; SABOURIN; CAVALCANTI, 2018).

A Figura 9 mostra o processamento de uma amostra de teste pelo *METAD-DES*. Como pode ser visto, quando uma amostra desconhecida $\boldsymbol{x_j}$ é apresentada ao esquema, as metacaracterísticas são calculadas de acordo com a instância $\boldsymbol{x_j}$ e apresentadas ao metaclassificador $\boldsymbol{\lambda}$. O classificador então estima o nível de competência $\boldsymbol{\delta_{i,j}}$ para cada classificador base $\boldsymbol{c_i}$. Ao fim, os classificadores mais competentes são escolhidos para predizer $\boldsymbol{x_j}$ por meio do voto majoritário.

Uma vez que o metaclassificador $\boldsymbol{\lambda}$ é essencial neste esquema, os autores Cruz *et al.* (2015) realizaram uma análise da escolha do classificador, onde quatro modelos de classificadores foram avaliados: *MLP*, *SVM*, *NB* e *RF*. Resultados experimentais indicaram que o desempenho do *MLP*, *SVM* e *RF*, foram estatisticamente equivalentes. Porém, o classificador *NB* obteve o maior número de vitórias na análise experimental.

Figura 9: Procedimentos realizados durante a seleção dinâmica no estágio de teste do *META-DES*.



Fonte: O autor.

Assim sendo, a essência do esquema *META-DES* é avaliar o nível de competência dos classificadores base no estágio de treinamento, para que ao realizar a predição no estágio de teste, o esquema consegue escolher os classificadores mais aptos que posteriormente participam do voto majoritário, para efetuar a predição da classe. Considerando isso, é trivial notar que as metacaracterísticas utilizadas são primordiais na composição destes esquemas. Na seção seguinte é descrito a respeito das metacaracterísticas abordadas pelo esquema *META-DES*.

3.5.1.1 Metacaracterísticas do *META-DES*

As metacaracterísticas de um ESDs, são geradas por metadados que usualmente utilizam métricas de desempenho dos classificadores base. Na literatura é abordado diferentes técnicas para obter tais metadados, entre elas as baseadas em acurácia (geral e local) (WOODS; KEGELMEYER; BOWYER, 1997), probabilidade a posteriori (GIACINTO; ROLI, 1999), perfil de saída dos classificadores (CAVALIN; SABOURIN; SUEN, 2013), oráculo (KO; SABOURIN; JR, 2008), entre outras. Mas em especial no esquema *META-DES* são utilizadas as seguintes metacaracterísticas.

1. ***LCA***: A acurácia da classe local, ou do inglês *Local Class Accuracy* (LCA) é uma metacaracterística que se baseia na acurácia local estimada em relação às classes das **K** instâncias mais próxima, denominada por ω_l. Então, para cada amostra da região de competência é calculado o percentual de acerto dos classificadores base em relação ao acerto dos vizinhos na região de competência. Assim, o classificador base c_i que apresentar o melhor nível de competência $\delta_{i,j}$, será escolhido para predizer amostras da região de competência θ_j.

$$\delta_{i,j} = \frac{\sum_{x_k \in \omega_l} P(\omega_l | x_k, c_i)}{\sum_{k=1}^{K} P(\omega_l | x_k, c_i)} \tag{1}$$

Conforme definido na Equação 1, esta metacaracterística gera um nível de competência para cada amostra da região local $\boldsymbol{\theta}$, isto é, para cada amostra é gerado $\boldsymbol{K}$ atributos. Assim quanto maior o tamanho de uma região local, mais atributos terá para estimar o nível de competência dos classificadores base.

2. ***A Posteriori***: Esta metacaracterística é semelhante a *LCA*, onde é criado um vetor com $\boldsymbol{K}$ elementos, fazendo com que para cada instância $\boldsymbol{x_k}$ pertencente à região de competência $\boldsymbol{\theta_j}$, a probabilidade posterior de $\boldsymbol{c_i}$, $\boldsymbol{P(\omega_l|x_k)}$ seja calculada e inserida na k-ésima posição do vetor. Porém, esta metacaracterística utiliza um coeficiente de ponderação $\boldsymbol{W_k}$, que é inversamente proporcional à distância euclidiana da instância avaliada com relação aos seus $\boldsymbol{K}$ vizinhos (Equação 2).

$$\delta_{i,j} = \frac{\sum_{x_k \in \omega_l} P(\omega_l|x_k, c_i) W_k}{\sum_{k=1}^{K} P(\omega_l|x_k, c_i) W_k} \tag{2}$$

3. ***OLA***: Nesta metacaracterística objetiva estimar à acurácia de cada classificador individual em regiões locais do espaço de recurso (WOODS; KEGELMEYER; BOWYER, 1997). Sua determinação começa com o cálculo do nível de competência $\boldsymbol{\delta_{i,j}}$ de um classificador de base $\boldsymbol{c_i}$ na região $\boldsymbol{\theta_j}$, como é descrito na Equação 3. Assim, para cada classificador base $\boldsymbol{c_i}$ é estimado um nível de competência para cada amostra $\boldsymbol{x_j}$, que faz parte de uma região de competência $\boldsymbol{\theta_j}$, onde possui um espaço de recurso com $\boldsymbol{K}$ vizinhos. Então, como $\boldsymbol{\omega_l}$ é a classe predita do classificador $\boldsymbol{c_i}$, é calculado a probabilidade do classificador $\boldsymbol{c_i}$ acertar a instância $\boldsymbol{x_k}$ dentro de sua região de competência, obtendo assim uma média de acerto.

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} P(\omega_l|x_k \in \omega_l, c_i) \tag{3}$$

4. $\boldsymbol{K_p}$ **perfis de saída**: Nesta metacaracterísticas é gerado um vetor com $\boldsymbol{K_p}$ elementos. Então, para cada membro $\boldsymbol{\tilde{x_k}}$ pertencente ao conjunto de perfis de saída $\boldsymbol{\theta_j}$, se o rótulo produzido por $\boldsymbol{c_i}$ para $\boldsymbol{\tilde{x_k}}$ for igual ao rótulo $\boldsymbol{\omega_{l,k}}$ de $\boldsymbol{\tilde{x_k}}$, a k-ésima posição do vetor é ajustada para 1, caso contrário é 0. Um total de $\boldsymbol{K_p}$ atributos são extraídos usando perfis de saída.

5. **Confiança do classificador**: esta metacaracterística compreende na distância perpendicular entre o campo de entrada de dados e o limite de decisão do classificador base $\boldsymbol{c_i}$, a qual é calculada e codificada como um vetor. Tal vetor é normalizado para um intervalo de 0 à 1 usando a normalização de mínimo e máximo.

3.5.2 Esquema do *DESMI*

A seleção dinâmica de conjuntos para multiclasse desbalanceada, ou em inglês *Dynamic Ensemble Selection for Multi-class Imbalanced* (DESMI) (GARCÍA *et al.*, 2018) é um esquema inspirado no *META-DES*, e tem como objetivo gerar conjuntos de dados de treinamento balanceado e selecionar os classificadores apropriados. Para isso, os autores García *et al.* (2018) desenvolveram um procedimento de pré-processamento para balancear o conjunto de dados por meio de dados sintéticos. Para selecionar os classificadores no cenário de múltiplas classes, desenvolveram um mecanismo de ponderação para destacar a competência dos classificadores mais aptos na classificação de amostras em regiões de classes minoritárias. Baseado nisso, os seguintes procedimentos são realizados neste esquema:

1. No conjunto de treinamento é utilizado técnicas híbridas para obter o balanceamento aleatório. As técnicas utilizadas são: *Random Under-Sampling* (RUS), *Random Over-Sampling* (ROS) e *Synthetic Minority Oversampling Technique* (SMOTE).

2. A avaliação de competência dos classificadores é feita pela ponderação da amostra consultada. Assim, amostras de classes minoritárias fornecem um maior nível de competência aos classificadores que obtém desempenho superiores na região local.

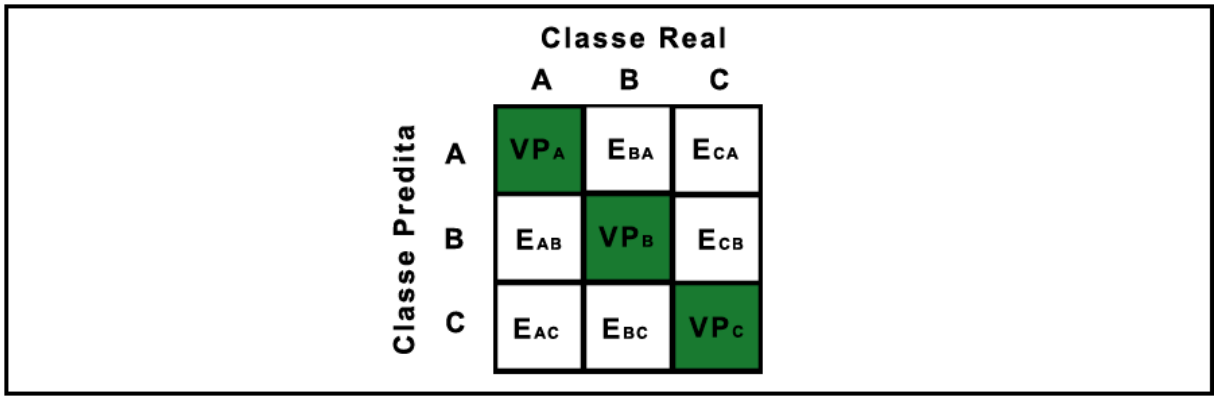
3. Por fim, os classificadores selecionados são submetidos ao voto majoritário para realizar a predição da amostra avaliada.

3.6 MÉTRICAS DE DESEMPENHO PARA MODELOS DE CLASSIFICAÇÃO

Existem diferentes métricas para avaliar o desempenho de classificadores, neste trabalho as métricas são calculadas a partir da matriz de confusão, pela qual mostra quatro resultados possíveis de predição (ALBERTO; ALMEIDA, 2012): verdadeiro positivo (**VP**), falso positivo (**FP**), verdadeiro negativo (**VN**) e falso negativo (**FN**). A Figura 10 ilustra a matriz de confusão para um problema de classificação multiclasse com três classes (**A, B e C**). Conforme observado, $\boldsymbol{VP_A}$ é o número de amostras positivas verdadeiras na classe **A**, ou seja, o número de amostras corretamente classificadas na classe $\boldsymbol{A}$ e $\boldsymbol{E_{AB}}$ são as amostras da classe $\boldsymbol{A}$ que foram incorretamente classificadas na classe $\boldsymbol{B}$. Assim, o falso negativo na classe $\boldsymbol{A}$ ($\boldsymbol{FN_A}$) é a soma de $\boldsymbol{E_{AB}}$ e $\boldsymbol{E_{AC}}$ ($\boldsymbol{FN_A = E_{AB} + E_{AC}}$) que indica a soma de todas as amostras da classe $\boldsymbol{A}$ que foram classificadas incorretamente como classe $\boldsymbol{B}$ ou $\boldsymbol{C}$, representando os erros de uma coluna. Enquanto o falso positivo para qualquer classe prevista localizada em uma linha representa a soma de todos os erros nessa linha. Por exemplo, o falso positivo na classe **A** ($\boldsymbol{FP_A}$)

é calculado da seguinte forma: $\boldsymbol{FP_A = E_{BA} + E_{CA}}$. Baseado nisso, é descrito os cálculos de cada métrica de avaliação, levando em consideração o problema de classificação multiclasse.

Figura 10: Matriz de confusão para o problema de classificação multiclasse.



Fonte: ALBERTO e ALMEIDA (2012).

- **Acurácia Média**: é a medida de desempenho global, onde mostra a proporção de classificações corretas, tanto de casos positivos quanto negativos (Equação 4).

$$\textit{Acurácia-Média} = \frac{VP + VN}{(VP + FN) + (VN + FP)} \tag{4}$$

- **Acurácia Balanceada**: quando se trata de classes binárias, à acurácia balanceada é a média entre a sensibilidade e a especificidade. Mas ao se tratar de um problema multiclasse é média da sensibilidade de todas as classes, definida como:

$$\textit{Acurácia-Balanceada} = \frac{1}{M}\sum_{i=1}^{M}\frac{VP_i}{VP_i + FN_i} \tag{5}$$

- **Média Geométrica**: mais conhecida como *G-mean*, esta é uma métrica exclusivamente de conjuntos desbalanceados, pois mensura a raiz quadrada do produto da precisão e sensibilidade, para classificação binária, mas em problemas de multiclasse é a raiz n-ésima do produto da sensibilidade para cada classe (ESPÍNDOLA; EBECKEN, 2005) (Equação 6).

$$\textit{Média-Geométrica (multiclasse)} = \left(\prod_{i=1}^{N}\frac{VP_i}{VP_i + FN_i}\right)^{1/N} \tag{6}$$

- ***Overfitting***: é a produção de um modelo muito próximo ou exatamente a um conjunto de dados específico (ANDERSON; BURNHAM, 2004). Não existem métricas específicas para avaliar o *overfitting*, mas alguns autores criaram certas definições, como a diferença entre

o desempenho no conjunto de treinamento e o conjunto de teste (REUNANEN, 2003), outros mensuram como a razão entre o desempenho no conjunto de teste pelo treinamento (Equação 7).

$$Overfitting = \frac{Escore_{teste}}{Escore_{treinamento}} \tag{7}$$

Por fim, com o intuito de realizar a seleção de atributos relevantes para o problema de classificação com múltiplas classes, o algoritmo ***Relief*** é apropriado segundo Kira e Rendell (1992). Este algoritmo realiza um ranqueamento levando em conta a relevância dos atributos de um conjunto de dados, e tem a capacidade de lidar com ruídos contendo múltiplas classes. Sua premissa é lidar com a amostragem aleatória de instâncias e a distância do vizinho mais próximo da mesma classe e da classe oposta, então os valores dos atributos dos vizinhos mais próximos são comparados aos da classe amostrada para atualizar os pesos de relevância ($\boldsymbol{W_i}$) de cada atributo em relação a classe. A ideia do *Relief* é que atributos importantes devem ser capazes de diferenciar amostras de classes diferentes e possuir valores similares aos da mesma classe. A Equação 8 descreve o procedimento para atualizar a relevância de cada atributo, onde o vetor $\boldsymbol{x_i}$ é uma instância aleatória, ***nearHit*** é a instância mais próxima da mesma classe, e ***missHit*** a mais próxima da classe oposta.

$$W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \tag{8}$$

Este procedimento é repetido $\boldsymbol{m}$ vezes, e a pontuação final de cada atributo é a média da divisão cada elemento do vetor por $\boldsymbol{m}$. Então os atributos são selecionados se sua relevância for maior que um limite $\boldsymbol{\tau}$, que conforme os autores Urbanowicz *et al.* (2018), este limiar pode ser estimado pela mediana do ranqueamento final.

4 METODOLOGIA

A fim de avaliar o desempenho de diferentes esquemas de seleção dinâmica para a identificação de bactérias a partir de dados de *m/z* virtuais de proteínas ribossomais, este capítulo apresenta um conjunto de experimentos que mensuram o desempenho de diferentes composições de classificadores base, sobre variados cenários do conjunto de dados *PUKYU*. Também é apresentado um ESD, nomeado como *METADES-i* (*METADES-imbalanced*), que estende o esquema *META-DES* ao incluir metacaracterísticas preditivas sensíveis ao desbalanceamento de dados.

Este capítulo está dividido da seguinte maneira. A Seção 4.1 descreve as metacaracterísticas utilizadas no desenvolvimento do *METADES-i*. A Seção 4.2 descreve o conjunto de dados utilizado nos experimentos. A Seção 4.3 descreve as rotinas de pré-processamento do conjunto de dados. E a Seção 4.4 descreve o procedimento experimental, análise de resultados e também os parâmetros utilizados nos experimentos.

4.1 METACARACTERÍSTICAS DO *METADES-i*

Como descrito anteriormente, as metacaracterísticas devem possuir informações que facilitem a determinação do nível de competência $\boldsymbol{\delta_{i,j}}$ de um classificador $\boldsymbol{c_i}$. O propósito é empregar metacaracterísticas que permitam a escolha de classificadores eficazes em termos de assertividade para o tratamento de uma nova instância. Neste caso, esquemas de seleção dinâmica com o uso de metacaracterísticas tendem a contribuir de forma positiva para o desempenho de um SMC.

Neste contexto, a metacaracterística *OLA*, que determina a média geral da métrica de acurácia média dos classificadores em seu espaço de recurso (região local). Este é um índice relevante, mas em conjuntos de dados desbalanceados a acurácia média pode superestimar o desempenho de classificadores (BRODERSEN *et al.*, 2010). Então, neste trabalho dez metacaracterísticas são propostas baseadas em *OLA*, mas ao invés de usar a métrica de acurácia média, são utilizadas métricas sensíveis a dados desbalanceados. Tais métricas foram selecionadas baseado em análises de métricas sensíveis ao desbalanceamento (BRANCO; TORGO; RIBEIRO, 2017), (WARDHANI *et al.*, 2019) e (LUQUE *et al.*, 2019). O intuito é melhorar a estimativa de competência dos classificadores base em conjunto de dados desbalanceados. Portanto o esquema *METADES-i* propõe metacaracterísticas baseados em *OLA*, com as seguintes métricas:

1. **Acurácia balanceada**: conforme definido na Equação 5, à acurácia balanceada realiza a ponderação da sensibilidade de acordo com a quantidade de instâncias por classe. Com

isso, é possível extrair informações relevantes quando se trabalha com dados desbalanceados. Ao se tratar de metacaracterística, a *OLA* foi usada para a extração de informações de acurácia balanceada em regiões específicas (Equação 9).

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} \textit{Acurácia-Balanceada}(\omega_l | x_k \in \omega_l, c_i) \tag{9}$$

2. **Precisão**: é definido como o percentual de padrões classificados como pertencentes a classe positiva (**VP**) e que realmente pertencem a classe positiva (**VP+FP**), conforme mostrado na Equação 10. Esta métrica visa extrair informações individuais de cada classe e ao final é obtido a média de precisão de todas as classes. Em termos de metacaracterística, é extraído a média geral dentro de uma região local (Equação 11).

$$\textit{Precisão} = \frac{VP}{VP + FP} \tag{10}$$

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} \textit{Precisão}(\omega_l | x_k \in \omega_l, c_i) \tag{11}$$

3. **Sensibilidade**: é a razão entre os padrões da classe positiva identificada corretamente (**VP**) e a soma dos padrões identificados corretamente para a classe positiva e negativa (**VP+FN**), conforme é definido na Equação 12. Por meio desta métrica é obtido informações individuais para cada classe em seguida a média de todas as classes, e como uma metacaracterística dentro de uma região local é definido na Equação 13.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{12}$$

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} Sensibilidade(\omega_l | x_k \in \omega_l, c_i) \tag{13}$$

4. **F-medida**: é a média harmônica entre precisão e sensibilidade (Equação 14), como metacaracterística estas informações são obtidas dentro de uma região local para que assim possa obter dados de uma região de competência, conforme mostrado na Equação 15.

$$\textit{F-medida} = 2 \times \frac{\textit{Precisão} \times Sensibilidade}{\textit{Precisão} + Sensibilidade} \tag{14}$$

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} \textit{F-medida}(\omega_l | x_k \in \omega_l, c_i) \tag{15}$$

5. **Índice de *Jaccard***: também conhecido como coeficiente de similaridade, é usado para comparar os rótulos previstos com os verdadeiros (**VP**), é definido como razão entre o tamanho da interseção e o tamanho da união dos conjuntos de rótulos (Equação 16). Em termos de metacaracterística, é extraído a média geral na região local, conforme a Equação 17.

$$Jacc = \frac{VP}{VP + FP + FN} \tag{16}$$

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} Jacc(\omega_l | x_k \in \omega_l, c_i) \tag{17}$$

6. **Medida F-Beta**: é similar ao f-medida, ou seja, informa a média harmônica entre a precisão e sensibilidade, mas pondera a precisão que é tratada com mais relevância através do coeficiente β (Equação 18), e para extrair o nível de competência desta metacaracterística é definido na Equação 19.

$$F_\beta = (1 + \beta^2) \frac{\textit{Precisão} \times Sensibilidade}{\beta^2 \times \textit{Precisão} + Sensibilidade} \tag{18}$$

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} F_\beta(\omega_l | x_k \in \omega_l, c_i) \tag{19}$$

7. **Média Geométrica**: conforme já observado na Equação 6, esta métrica possui alta relevância com dados desbalanceados, ao ponderar as classes por meio da sensibilidade. Quando se trata de metacaracterística é definida na Equação 20.

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} \textit{Média-Geométrica}(\omega_l | x_k \in \omega_l, c_i) \tag{20}$$

8. **Especificidade**: conforme definido na Equação 21, esta métrica corresponde a taxa de acerto na classe negativa (**VN**) pela soma da quantidade de classes identificadas erroneamente (**VN+FP**). Em termos de metacaracterística é extraído a média global da Equação 21, conforme descrito na Equação 22.

$$Especificidade = \frac{VN}{VN + FP} \tag{21}$$

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{K} Especificidade(\omega_l | x_k \in \omega_l, c_i) \tag{22}$$

9. **Perda *hamming***: é a fração das classes previstas incorretamente, ou seja, a fração das classes erradas $\boldsymbol{L}$ pelo número total de classes $\boldsymbol{N}$. Conforme é definido na Equação 23, onde $\boldsymbol{y_{i,j}}$ é o alvo e $\boldsymbol{z_{i,j}}$ é a predição. Em termos de metacaracterística é atribuído a equação na média geral global (Equação 24).

$$\textit{Perda-Hamming} = \frac{1}{N.L}\sum_{i=1}^{N}\sum_{j=1}^{L} xor(y_{i,j}, z_{i,j}) \tag{23}$$

$$\delta_{i,j} = \frac{1}{K}\sum_{k=1}^{K} \textit{Perda-Hamming}(\omega_l | x_k \in \omega_l, c_i) \tag{24}$$

10. **Índice de média geométrica balanceada (IMGB)**: Este índice pondera qualquer equação de acordo com o balanceamento do conjunto de dados, neste caso em específico é realizado a ponderação na média geométrica (Equação 6), em termos de metacaracterística é definido na Equação 25.

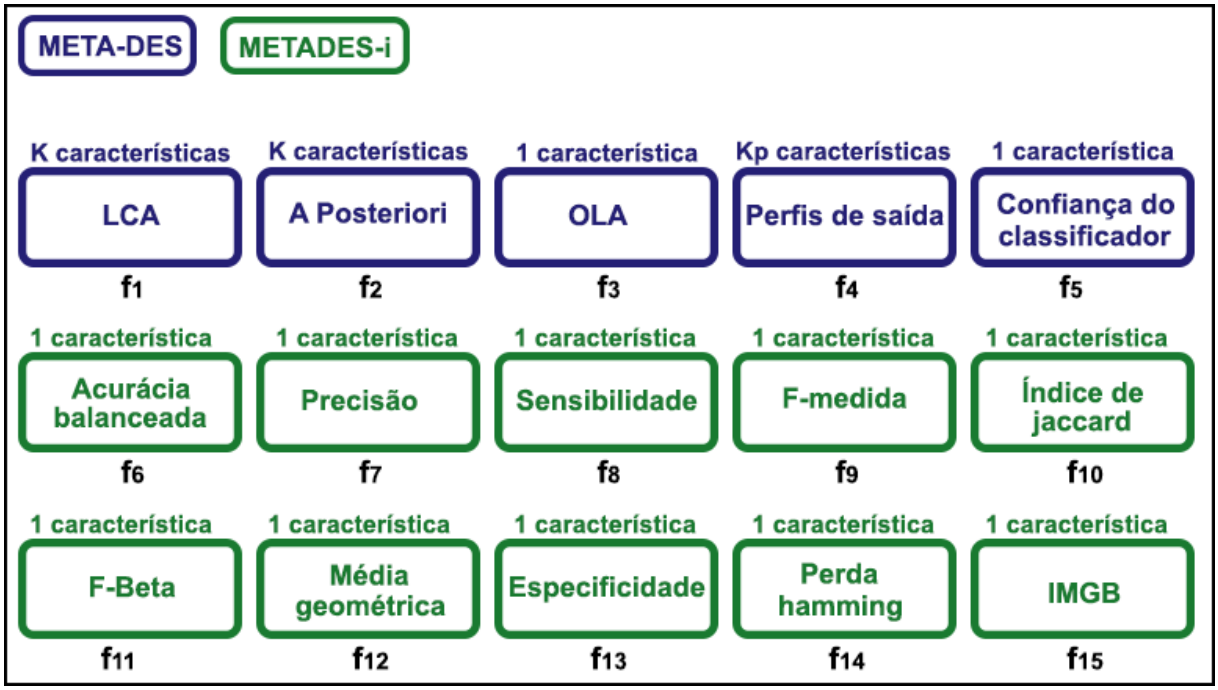
$$\delta_{i,j} = \frac{1}{K}\sum_{k=1}^{K} \textit{IMGB}(\omega_l | x_k \in \omega_l, c_i) \tag{25}$$

A Figura 11 descreve as metacaracterísticas utilizadas no *META-DES* e também aquelas adicionadas pelo *METADES-i*. Nota-se que a quantidade de metacaracterísticas propostas pelo esquema *METADES-i* é então $\boldsymbol{2K + Kp + 12}$, onde $\boldsymbol{K}$ é o número de vizinhos da região (espaço de recursos) e $\boldsymbol{Kp}$ é o número de vizinhos dos perfis de saída (espaço de decisão).

4.2 CONJUNTO DE DADOS

Os dados usados durante os experimentos foram obtidos a partir do conjunto de dados *PUKYU* de Tomachewski *et al.* (2018a). Este conjunto é composto por 32.026 registros categorizados em 2.422 gêneros distintos de bactérias, e 60 atributos. Cada registro se refere a uma bactéria a nível taxonômico de gênero e espécie. Cada bactéria é caracterizada por até sessenta atributos, em que são referenciados a uma determinada proteína ribossomal por meio de sua m/z. As proteínas ribossomais que compõem este conjunto de dados são: ***L1, L2, L3, L4 , L5, L6, L7, L7a, L7ae, L7.L12, L9, L10, L11, L12, L13, L14, L15, L16, L17, L18, L19, L20, L21, L22, L23, L24, L25, L27, L28, L29, L30, L31, L32, L33, L34, L35, L36, S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22 e S31e***.

Figura 11: Metacaracterísticas utilizadas no *META-DES* e no *METADES-i*, ao todo obteve-se $2K + KP + 12$ características.



Fonte: O autor.

4.3 PRÉ-PROCESSAMENTO

Como o conjunto de dados é desbalanceado, multiclasse e possui dados faltantes, seu uso na realização dos experimentos demandou rotinas de pré-processamento dos dados. Primeiramente foram removidos os registros que possuíam menos de quatro atributos com valores observados. Em seguida, foi executado um procedimento de imputação de dados, este procedimento foi executado por meio do algoritmo *kNN* com a função de agregação da mediana e k=5. Esta configuração está em acordo com aquela indicada por Santos *et al.* (2018) a partir de resultados obtidos em testes empíricos.

Um conjunto de dados é desbalanceado quando o número de registros para cada classe é desigual. Para medir o quanto um conjunto de dados é desbalanceado aplica-se a métrica chamada de índice de desbalanceamento (**ID**), que é definido como a proporção de registros a partir do número de classe majoritária em relação ao número de classe minoritária (NOORHALIM; ALI; SHAMSUDDIN, 2019), conforme exposto na Equação 26.

$$\text{ID} = \frac{\text{Classe Majoritária}}{\text{Classe Minoritária}} \tag{26}$$

Conforme visto, o treinamento de modelos de classificação é afetado negativamente quando utiliza-se um conjunto de dados desbalanceado (LEMAîTRE; NOGUEIRA; ARIDAS, 2016).

Assim, com o propósito de contornar esta dificuldade e avaliar o desempenho dos esquemas, foram criados diferentes cenários do conjunto de dados, variando o índice de desbalanceamento.

Para realizar tal procedimento, se fez necessário a remoção de alguns registros de acordo com um limiar pré-definido. Este limiar indica o número mínimo de registros por classe, por exemplo, ao aplicar um limiar de valor 20, serão excluídas todas as instâncias que possuírem menos de 20 registros por classe. Desta maneira, os limiares utilizados nestes experimentos foram: 20, 30, 50, 70 e 100.

A Tabela 1 mostra o nome de cada cenário (subconjunto), gerado a partir de seu limiar, observa-se que a aplicação de tal limiar afeta a distribuição dos gêneros, o número de instancias e o ID. Além disso, a Figura 12 ilustra que ao aumentar o limiar o número de gêneros e o ID tendem a diminuir.

Tabela 1: Diferentes cenários do conjunto *PUKYU*, construído por meio da variação dos limiares de desbalanceamento.

Nome	**Limiar**	**Nº Gêneros**	**Nº Instâncias**	**ID**
PUKYU	-	2.422	32.026	4.425
PK20	20	134	22.596	211
PK30	30	86	21.377	140
PK50	50	60	20.391	84
PK70	70	46	19.559	60
PK100	100	36	18.497	42

Fonte: O autor.

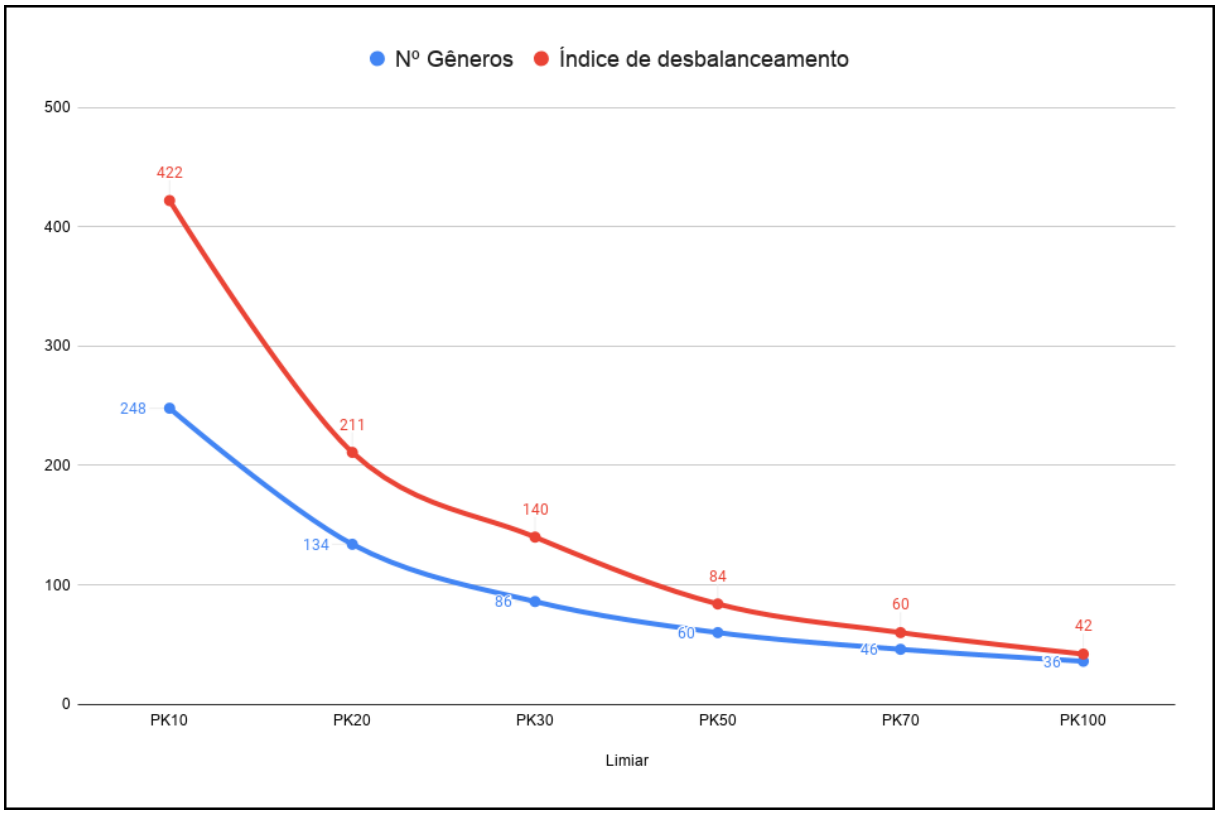
4.4 PROTOCOLOS EXPERIMENTAIS

Ao todo foram realizados sete experimentos cujos protocolos experimentais são descritos nas subseções seguintes. A Tabela 2 contém um resumo de todos os experimentos, com seus devidos protocolos e equações.

4.4.1 Protocolo Experimental 1

O protocolo experimental 1 visa avaliar o desempenho dos esquemas de seleção dinâmica, variando a composição dos classificadores base, cenários e as métricas. Conforme mostrado no Algoritmo 1, este protocolo experimental possui três parâmetros de entrada. O primeiro é um vetor de conjunto de dados $\boldsymbol{BD}$ contendo os diferentes cenários (*PK20*, *PK30*, *PK50*, *PK70* e *PK100*) com subconjuntos extraídos a partir do conjunto *PUKYU*. O segundo parâmetro refere-se ao conjunto dos ESDs a serem avaliados, denotados por $\boldsymbol{C}$. E o terceiro parâmetro são as métricas $\boldsymbol{M}$ para avaliarem os ESDs, são elas: acurácia média, acurácia ba-

Figura 12: Análise do número gêneros e índice de desbalanceamento de acordo com a variação do limiar.



Fonte: O autor.

lanceada e média geométrica. O algoritmo aplica um procedimento de validação cruzada para cada base de $\boldsymbol{BD}$, tal validação divide em dez subconjuntos, denotados por $\boldsymbol{BDP}$. Assim, para cada subconjunto é escolhido como base de validação e o desempenho da métrica $\boldsymbol{M}$ é calculado para o esquema $\boldsymbol{C}$. Ao final é calculado o desempenho médio de cada esquema sobre os subconjuntos.

Neste protocolo experimental foram executados três experimentos que mensuram o desempenho variando os esquemas e seus respectivos conjuntos de classificadores: homogêneos (*HM*) e heterogêneos (*HT*). No primeiro experimento, é avaliada a métrica de acurácia média (Equação 4). No segundo experimento é avaliado a métrica de acurácia balanceada (Equação 5). E para o terceiro experimento é mensurado o desempenho preditivo em termos de média geométrica (Equação 6).

4.4.2 Protocolo Experimental 2

O protocolo experimental 2 visa avaliar o desempenho dos ESDs em termos de *overfitting*, sore os diferentes cenários. Para extrair estas informações, os valores de *overfitting* foram calculados conforme a Equação 7 usando as métricas de desempenho obtidas no conjuntos de treinamento ($\boldsymbol{Escore_{treino}}$) e teste ($\boldsymbol{Escore_{teste}}$).

Algoritmo 1: PROTOCOLO EXPERIMENTAL 1

Entrada: BD: Diferentes bases do *PUKYU*; C: Esquemas a serem avaliados; M: Métricas para avaliar;

```
1  início
2      para cada base B ∈ BD faça
           // Particiona o conjunto em 10 partições
3          BDP = 10-Folds(B);
4          para cada partição P ∈ BDP faça
5              para cada esquema c ∈ C faça
                   // Treina em 9 partições e testa na partição
                      P
                   // Para a composição homogênea e heterogênea
6                  Treinar(c, BDP - P);
7                  Testar(c, P);
                   // Calcula métrica(s) para o classificado c
8                  M(c);
9              fim
10         fim
11     fim
12 fim
```

Conforme ilustrado no Algoritmo 2, os parâmetros referentes ao protocolo experimental 2 são semelhantes ao do Algoritmo 1, exceto o parâmetro $\boldsymbol{E}$ que recebe o escore a ser avaliado em termos de *overfitting*.

Neste protocolo foram executados dois experimentos, no primeiro é passado para o parâmetro $\boldsymbol{E}$ a métrica de acurácia média, a fim de se obter o desempenho de *overfitting* dos esquemas em termos de acurácia média (***OVAM***). Para o segundo experimento, a métrica utilizada no parâmetro $\boldsymbol{E}$ foi a acurácia média, assim obtendo o desempenho de *overfitting* em termos de acurácia média (***OVAB***).

Visto que o desempenho em termos de *overfitting* é um valor entre 0 e 1. Foi adotado uma alteração na escala dos resultados a fim de se igualar as demais métricas. Desta maneira, os valores dos resultados foram multiplicados por cem.

4.4.3 Protocolo Experimental 3

O protocolo experimental 3 tem como objetivo avaliar a relevância das metacaracterísticas do esquema *METADES-i*. Para isso, foi aplicado o algoritmo de seleção de atributos, chamado de *Relief*. Onde foi executado sobre o conjunto de dados relacionado ao metaproblema subjacente *METADES-i* na estimativa do nível de competência dos classificadores base.

Algoritmo 2: Protocolo Experimental 2

Entrada: BD: Diferentes bases do *PUKYU*; C: Esquemas a serem avaliados; E: escore a ser medido

início
  para *cada base* $B \in BD$ **faça**
    `// Particiona o conjunto em 10 partições`
    $BDP = 10 - Folds(B)$; **para** *cada partição* $P \in BDP$ **faça**
      para *cada esquema* $c \in C$ **faça**
        `// Treina em 9 partições e testa na partição` P
        Treinar(c, BDP - P);
        `// Calcula escore no conjunto de treinamento`
        $Escore(E)_{Treino} = Testar(c, BDP)$;
        `// Calcula escore no conjunto de teste`
        $Escore(E)_{Teste} = Testar(c, P)$;
        `// Calcula` *overfitting* `para` c
        Overffiting(c) = $\frac{Escore(E)_{Teste}}{Escore(E)_{Treino}} \times 100$;
      fim
    fim
  fim
fim

Conforme mostrado no Algoritmo 3, este protocolo possui apenas o parâmetro BD que é um vetor contendo os diferentes cenários do conjunto de dados *PUKYU*. Assim, para cada cenário é aplicado o procedimento do algoritmo *Relief*, a fim de obter um ranqueamento dos atributos mais relevantes.

Algoritmo 3: PROTOCOLO EXPERIMENTAL 3

Entrada: BD: Diferentes bases do *PUKYU*;

início
  para *cada base* $B \in BD$ **faça**
    `// Aplica ranqueamento por 10 vezes`
    Ranqueamento = Relief(B);
  fim
  `// Obtém a média do ranqueamento para cada base`
  Média = Ranqueamento/10;
fim

Para o protocolo experimental 3, o procedimento *Relief* foi executado dez vezes para obter a pontuação média de cada metacaracterística e assim, realizar o ranqueamento das mesmas, para ambas as composições de classificadores (homogênea e heterogênea), sobre o esquema *METADES-i*. Nos casos em que existem atributos compostos, como aqueles baseados em $\boldsymbol{K}$ ou $\boldsymbol{K_p}$ instâncias, foi considerado o valor da média dos vizinhos sobre a região local.

Por exemplo, a metacaracterística LCA, possui $\boldsymbol{K}$ índices de competência, assim será obtido a média dos mesmos a fim de se obter o índice de relevância desta metacaracterística.

Ao final, para cada composição foi obtido um esquema de classificação semelhante ao *METADES-i*, porém foram removidos as metacaracterísticas irrelevantes conforme o critério de seleção do algoritmo *Relief*. Deste modo, o último experimento avaliou o desempenho do esquema *METADES-i* com relação a sua versão com seleção das metacaracterísticas, nomeado como *METADES-ii*. Analisou-se em termos das métricas de acurácia média (AM), acurácia balanceada (AB), *overfitting* com relação a acurácia média (OVAM) e acurácia balanceada (OVAB). O objetivo deste experimento foi analisar o uso do algoritmo *Relief* na seleção de metacaracterísticas, avaliando as métricas de assertividade e *overfitting*.

4.5 ANÁLISE DOS RESULTADOS

Inicialmente, o teste de *Friedman* foi aplicado para verificar a existência de diferenças estatísticas entre os desempenhos dos esquemas. Em seguida, quando houve diferença, foi aplicado o teste não paramétrico de *wilcoxon*, a fim de averiguar os pareamentos dos esquemas que obtiveram diferenças estatística. Para ambos os testes foram considerados os níveis de confiança de 95% ($\boldsymbol{\alpha}$ **= 0,05**).

Estes procedimentos foram aplicados isoladamente para cada experimento. Conforme mostrado na Tabela 2, os experimentos 1, 2 e 3 avaliam as métricas de acurácia média, acurácia balanceada e média geométrica, respectivamente, seguindo o protocolo experimental 1. Nos experimentos 4 e 5, é avaliado o desempenho de *overfitting* dos esquemas com relação a acurácia média ($OVAM$) e acurácia balanceada ($OVAB$), seguindo o protocolo experimental 2. No experimento 6 é realizado o ranqueamento através do índice de relevância das metacaracterísticas, por meio do algoritmo *Relief*, seguindo o protocolo experimental 3. O critério de seleção utilizado neste experimento foi com relação aos escores que foram superiores ao limiar estimado pela mediana do ranqueamento. No experimento 7, foram avaliados as métricas de acurácia média, acurácia balanceada, $OVAM$ e $OVAM$, seguindo o protocolo experimental 1.

4.6 SELEÇÃO DOS CLASSIFICADORES BASE

O critério de seleção dos classificadores base foi realizado sobre o estado da arte dos sistemas de múltiplos classificadores, levando em consideração os seguintes trabalhos: (FAN *et al.*, 2018), (CRUZ; SABOURIN; CAVALCANTI, 2018), (FERNÁNDEZ-DELGADO *et al.*, 2014) e (SAGAWA *et al.*, 2018). Após este levantamento foram selecionados os seguintes classificadores:

Tabela 2: Resumo dos experimentos realizados

Experimento	Protocolo	Métrica	Equação
1	1	Acurácia Média	4
2	1	Acurácia Balanceada	5
3	1	Média Geométrica	6
4	2	*Overfitting* (AM)	7 e 4
5	2	*Overfitting* (AB)	7 e 5
6	3	Ranqueamento *Relief*	8
7	1	*AM, AB, OVAM e OVAB*	4,5 e 7

Fonte: O autor.

Naïve Bayes (NB), *Multi-Layer Perceptron* (MLP), *Support Vector Machine* (SVM), Regressão Logística (RL) e Árvore de Decisão (CART).

A partir disso, estes classificadores foram comparados em termos de acurácia média e acurácia balanceada, sobre os diferentes cenários ($PK20$, $PK30$, $PK50$, $PK70$ e $PK100$) do conjuntos de dados *PUKYU*. Conforme mostrado na Tabela 3, na média dos cenários, o classificador CART foi aquele que obteve o desempenho superior em todas as métricas avaliadas.

Baseado na análise de desempenho dos classificadores, foi elaborado o critério de escolha dos classificadores base para as composições (homogênea e heterogênea) dos esquemas. Na composição homogênea foi aplicado o método *bagging*, pelo qual particiona o conjunto de treinamento em *bags* e cria um modelo para cada um. Neste trabalho, o *bagging* foi particionado em dez *bags*, utilizando o classificador CART, a fim de se obter dez estimadores para compor o conjunto homogêneo. Foi escolhido o classificador CART pelo motivo do mesmo obter desempenho superior com relação aos demais classificadores.

Para a composição heterogênea, com o intuito de obter diferentes funções de classificação do SMC, foram considerados os classificadores pertencentes a diferentes categorias de classificação: redes bayesianas, redes neurais, máquinas de vetores de suporte e logísticos. Considerando isto e à análise de desempenho, foram selecionados classificadores: *RL*, *NB*, *SVM* e *MLP*. Na seção seguinte é apresentado os hiperparâmetros utilizados pelos classificadores base.

4.7 PARAMETRIZAÇÃO DE CLASSIFICADORES E MÉTODOS

Os métodos e algoritmos descritos neste trabalho foram implementados na linguagem *Python* com o uso da biblioteca *scikit-learn* (PEDREGOSA *et al.*, 2011) e *imbalanced-learn* (LEMAîTRE; NOGUEIRA; ARIDAS, 2016). Para os esquemas de seleção dinâmica foi utilizado a biblioteca *DESlib* de Cruz *et al.* (2018). Os demais parâmetros foram configurados como segue:

Tabela 3: Acurácia média e balanceada dos classificadores base.

Base	Classificadores	Acurácia Média	Acurácia Balanceada
PK20	**CART**	**71,15 ± 0,45**	**64,67 ± 1,50**
	RL	69,14 ± 0,45	27,78 ± 1,47
	NB	59,82 ± 0,45	63,36 ± 1,37
	SVM	53,64 ± 0,45	15,91 ± 1,76
	MLP	19,03 ± 0,45	12,75 ± 1,49
PK30	**CART**	**79,27 ± 0,33**	**67,04 ± 0,89**
	RL	72,85 ± 0,49	39,63 ± 1,09
	NB	63,76 ± 0,84	65,67 ± 1,61
	SVM	54,72 ± 0,98	20,17 ± 1,90
	MLP	57,19 ± 0,42	23,64 ± 1,55
PK50	**CART**	**85,66 ± 0,33**	**71,78 ± 0,89**
	RL	75,46 ± 0,49	50,76 ± 1,09
	NB	68,46 ± 0,84	68,31 ± 1,61
	MLP	59,61 ± 0,42	30,55 ± 1,55
	SVM	58,61 ± 0,98	26,43 ± 1,90
PK70	**CART**	**87,23 ± 0,33**	**77,18 ± 0,89**
	RL	76,79 ± 0,49	56,03 ± 1,09
	NB	69,59 ± 0,84	68,61 ± 1,61
	MLP	76,11 ± 0,42	60,22 ± 1,55
	SVM	60,47 ± 0,98	30,71 ± 1,90
PK100	**CART**	**92,57 ± 0,33**	**82,11 ± 0,89**
	RL	79,76 ± 0,49	64,99 ± 1,09
	MLP	78,44 ± 0,42	64,84 ± 1,55
	NB	70,05 ± 0,84	67,48 ± 1,61
	SVM	62,53 ± 0,98	39,99 ± 1,90

Fonte: O autor.

- O tamanho da região de competência $\boldsymbol{K}$ foi utilizado **10** vizinhos.
- O tamanho da região de competência para os perfis de saída $\boldsymbol{Kp}$ foi utilizado **7** vizinhos.
- O limiar para estimar se um classificador base é competente foi adota o valor de **0,5** pelo qual estava como padrão nos ESD.
- O coeficiente $\boldsymbol{\beta}$ utilizado na Equação 18, foi de **0,5**, valor recomendado na literatura.
- Para a escolha do metaclassificador $\boldsymbol{\lambda}$ foi usado a árvore de decisão $\boldsymbol{CART}$, conforme a análise empírica (Tabela 3).
- No classificador base $\boldsymbol{NB}$, foi utilizado sua versão **multinomial**;
- Para o classificador $\boldsymbol{SVM}$ foi usando o *kernel* de função de base **radial**, pelo qual leva em consideração dois parâmetros: $\boldsymbol{C = 1}$ e $\boldsymbol{gama = auto}$. O parâmetro $\boldsymbol{C}$ negocia erros de classificação de exemplos de treinamento contra a simplicidade da superfície de

decisão. O $gama$ define quanta influência um único exemplo de treinamento tem, neste caso como foi configurado $auto$ onde leva em conta 1/Nº de características.

- Para o classificador base MLP, foi considerado uma camada oculta com **100 neurônios**, a taxa de aprendizado de **0,001**, a função de ativação foi $f(x) = max(0, x)$ ($relu$), e o número máximo de iterações foi de **200**.
- Para o classificador RL foi considerado o critério de penalidade $L2$, a função de otimização foi a $bfgs$, e o número máximo de iterações foi de 100.
- Para o algoritmo $Relief$, o número de iterações m foi definido em **10000**.
- O limiar τ utilizado para escolher os atributos selecionados pelo algoritmo $Relief$ foi definido como **mediana** do ranqueamento.
- Para a árvore $CART$ usada como $bagging$, foi considerado o índice $gini$ como critério de seleção dos nós, e quanto a profundidade foi deixado como padrão, ou seja, os nós são expandidos até que todas as folhas contenham o mínimo de amostras.

5 RESULTADOS E DISCUSSÃO

Este capítulo reporta os resultados obtidos com a execução dos experimentos listados na Tabela 2. Inicialmente, as seções 5.1, 5.2 e 5.3 comparam os resultados dos diferentes ESDs em relação à acurácia média (Experimento 1), acurácia balanceada (Experimento 2) e média geométrica (Experimento 3). A Seção 5.4 relata os resultados dos experimentos 4 e 5, que avaliaram o desempenho dos esquemas em termos de *overfitting*. Na seção 5.5 é mostrado os resultados da análise de relevância das metacaracterísticas utilizadas durante a fase de treinamento do esquema *METADES-i*.

5.1 ANÁLISE DE DESEMPENHO EM RELAÇÃO À ACURÁCIA MÉDIA

A Tabela 4 contém os resultados do Experimento 1, o qual mensurou o desempenho dos ESDs em termos de acurácia média nos diferentes cenários. A primeira coluna da tabela se refere ao cenário avaliado, a segunda coluna ao esquema de seleção dinâmica empregado nos testes. E a terceira e quarta colunas trazem a média e o desvio padrão da acurácia média no emprego das composições homogênea (HM) e heterogênea (HT), respectivamente. O teste não paramétrico de *wilcoxon* foi aplicado sobre todos os experimentos, afim de avaliar se um esquema obteve diferença estatística com relação aos demais, quando isso ocorreu a média do desempenho deste esquema é marcado com um símbolo na forma de um círculo opaco ($\bullet$).

Conforme mostrado na Tabela 4, os testes estatísticos constataram que o esquema *METADES-i*, com a composição homogênea dos classificadores base, obteve uma acurácia média significativamente superior aos demais esquemaS em todos os cenários. Porém, ao analisar os ESDs com a composição heterogênea, nota-se que o *METADES-i* obteve seu desempenho superior em três dos cinco cenários (*PK30*, *PK50* e *PK70*). Os esquemas *KNOP* e *KNORA-E* foram aqueles que obtiveram os maiores valores de acurácia média nos testes referentes aos cenários *PK20* e *PK100*, respectivamente. Deve ser observado que, em cada cenário com a composição heterogênea, nenhum dos esquemas de seleção que foram superiores durante os testes, apresentaram diferença estatística em relação a todos os demais esquemas.

A Figura 13 mostra os diagramas de caixa referentes aos dados de desempenho dos ESDs e ilustram os resultados do teste estatístico de *wilcoxon*. Os diagramas à esquerda da figura ilustram os resultados dos testes dos ESDs com a composição homogênea e os que estão à direita com a composição heterogênea. Em cada diagrama, a caixa relativa ao ESD com o maior desempenho nos testes, denominado $s*$, é preenchida em cinza. As caixa associadas aos esquemas de seleção cujo desempenho não foi estatisticamente diferente daquele de $s*$ são

Tabela 4: Resultado de acurácia média (AM) com as composições homogênea (HM) e heterogênea (HT).

Cenário	Esquema	AM-HM	AM-HT
PK20	**knorae**	85,94±0,45	89,31±0,39
	knorau	87,99±0,66	88,65±0,37
	knop	88,32±0,63	**89,83±0,49**
	desmi	85,52±0,66	85,16±0,64
	metades	88,86±0,58	89,61±0,40
	metades-i	**89,65±0,48** •	89,69±0,21
PK30	**knorae**	88,57±0,14	91,27±0,23
	knorau	90,39±0,66	90,51±0,32
	knop	90,84±0,43	91,45±0,48
	desmi	88,52±0,39	87,12±0,55
	metades	91,14±0,42	91,23±0,45
	metades-i	**92,44±0,41** •	**91,46±0,26**
PK50	**knorae**	90,29±0,76	92,51±0,40
	knorau	91,91±0,64	92,06±0,54
	knop	92,00±0,57	92,76±0,52
	desmi	90,91±0,64	88,60±0,81
	metades	92,41±0,50	92,57±0,53
	metades-i	**93,48±0,41** •	**92,71±0,32**
PK70	**knorae**	92,21±0,47	94,04±0,35
	knorau	93,44±0,56	93,64±0,17
	knop	93,71±0,51	94,14±0,34
	desmi	92,26±0,34	90,22±0,42
	metades	94,15±0,35	93,98±0,30
	metades-i	**95,17±0,63** •	**94,29±0,34**
PK100	**knorae**	93,76±0,30	**95,13±0,28**
	knorau	94,88±0,20	94,45±0,42
	knop	95,13±0,20	94,99±0,35
	desmi	93,70±0,25	91,34±0,52
	metades	95,33±0,26	95,02±0,23
	metades-i	**96,49±0,25** •	95,08±0,32

Fonte: O autor.

preenchidas em branco, e aquelas que obtiveram um desempenho estatisticamente inferior a $s*$ são preenchidas em verde.

Em conformidade com o que foi observado na Tabela 4, a Figura 13 destaca o fato de que o esquema *METADES-i* teve um desempenho significativamente superior aos demais esquemas da composição homogênea. Também foi possível notar que, em ESDs com composição heterogênea, os esquemas *KNORA-E*, *KNOP* e *METADES* obtiveram maiores desempenhos médios. Além disso, na maioria dos cenários não há diferença estatística entre estes esquemas. Ao contrário dos modelos *DESMI* e *KNORA-U*, que foram estatisticamente inferiores em todos os cenários.

Em termos de dados desbalanceados a métrica de acurácia média pode superestimar o desempenho dos esquemas, pelo fato do alto índice de desbalanceamento. Assim, classes que possuem muitas instâncias elevam a média de desempenho desta métrica. Por isso, em casos com conjuntos de dados desbalanceados a métrica de acurácia balanceada pode estimar um resultado mais fundamentado.

5.2 ANÁLISE DE DESEMPENHO EM RELAÇÃO À ACURÁCIA BALANCEADA

A Tabela 5 enumera o resultado da comparação dos ESDs em termos de acurácia balanceada. Os dados desta tabela revelam que o esquema *METADES-i* obteve seu desempenho de acurácia balanceada superior entre todos os esquemas avaliados. Além disso, nos testes os ESDs com a composição homogênea, o *METADES-i* foi estatisticamente superior aos demais no que se refere àquele escore. Nos testes dos ESDs com composição heterogênea não foram detectadas diferenças significativas nos desempenhos atingidos pelos esquemas.

A Figura 14 ilustra os resultados do teste estatístico de *wilcoxon*, através de caixa referentes aos dados de desempenho dos ESDs em termos de acurácia média (Experimento 2). Constatou-se que ao se tratar de ESDs com a composição heterogênea, não houve diferença estatística com o esquemas *KNOP* e *METADES*, sobre todos os cenários. Uma possível explicação para isso são as metacaracteríststcas utilizadas pelos esquemas. Pois o *KNOP* utiliza os Kp perfis de saída, que demonstra estimar os classificadores com precisão em diferentes índices de desbalanceamento dos subconjuntos. Já no caso do esquema *METADES*, além da precisão na estimativa, este esquema supera-se pela quantidade de metacaracterísticas. Entretanto, o resultado do teste de *wilcoxon* reportou que o esquema *METADES-i* foi significativamente superior aos demais esquemas, isso indica que as metacaracterísticas propostas pelo *METADES-i* descrevem de forma mais precisa a competência dos classificadores base em conjuntos de dados desbalanceados, o que contribui para um incremento no desempenho de ESDs.

De modo geral, a média dos desempenhos de acurácia balanceada foi inferior ao da acurácia média. Conforme discutido anteriormente, a acurácia média superestima as classes majoritárias, ao contrário da acurácia balanceada que calcula a média de assertividade das classes majoritárias e minoritárias com devidas ponderações.

5.3 ANÁLISE DE DESEMPENHO EM RELAÇÃO À MÉDIA GEOMÉTRICA

A Tabela 6 relaciona os resultados obtidos no Experimento 3, o qual mensurou o desempenho dos ESDs em termos de média geométrica. Primeiramente, foi possível constatar que

Tabela 5: Resultado de acurácia balanceada (AB) com as composições homogênea (HM) e heterogênea (HT).

Cenário	Esquema	AB-HM	AB-HT
PK20	**knorae**	62,05±0,90	71,03±1,76
	knorau	64,48±1,20	67,65±1,49
	knop	64,92±1,07	72,08±1,47
	desmi	61,01±0,91	69,24±1,50
	metades	67,72±1,44	71,99±1,37
	metades-i	**70,08±0,93** •	**72,67±0,31**
PK30	**knorae**	68,79±0,76	76,04±1,17
	knorau	72,52±2,12	72,03±0,81
	knop	73,63±1,40	76,53±0,56
	desmi	69,33±1,06	73,26±1,29
	metades	75,62±0,62	76,62±0,31
	metades-i	**77,55±0,68** •	**76,62±0,31**
PK50	**knorae**	77,28±1,44	83,07±0,66
	knorau	80,74±1,17	81,47±0,94
	knop	80,74±1,26	80,74±1,26
	desmi	77,95±1,57	79,78±1,24
	metades	81,51±1,01	81,51±1,01
	metades-i	**83,04±0,41** •	**83,82±0,34**
PK70	**knorae**	82,51±1,28	87,28±0,89
	knorau	85,07±1,09	86,45±0,57
	knop	85,76±1,59	87,97±1,13
	desmi	83,07±1,09	83,65±0,73
	metades	86,57±1,16	87,87±0,89
	metades-i	**87,15±1,45** •	**88,66±0,31**
PK100	**knorae**	88,71±0,97	91,44±0,50
	knorau	90,40±0,65	89,65±0,96
	knop	90,76±0,39	91,12±0,62
	desmi	88,57±0,76	87,16±0,77
	metades	91,17±0,66	91,37±0,51
	metades-i	**92,16±0,62** •	**91,63±0,53**

Fonte: O autor.

a diferença entre o esquema como maior e menor desempenho foi inferior a 3,35% para os ESD com composição homogênea e 2,91 % para os ESD com composição heterogênea. Em seguida observou-se que, tanto nos ESDs com composição homogênea e heterogênea, o uso do esquema *METADES-i* propiciou um ganho em relação aos demais esquemas. Este resultado ocorreu pelo fato do esquema *METADES-i* extrair características relevantes de amostras tanto das classes majoritárias quanto das minoritárias, pois conforme verificado por Espíndola e Ebecken (2005), a média geométrica pondera o desempenho de acordo o número de classes da instância.

A Figura 15 ilustra os resultados do teste estatístico de *wilcoxon*, através de caixas referentes aos dados de desempenho dos ESDs em termos de média geométrica (Experimento

Tabela 6: Resultado de média geométrica (MG) com as composições: homogênea (HM) e heterogênea (HT).

Cenário	Esquema	MG-HM	MG-HT
PK20	**knorae**	92,52±0,20	94,39±0,21
	knorau	93,59±0,37	94,02±0,20
	knop	93,77±0,33	94,67±0,27
	desmi	92,43±0,34	92,19±0,35
	metades	94,13±0,30	94,55±0,21
	metades-i	**95,87±0,46** •	**95,01±0,12** •
PK30	**knorae**	93,99±0,18	95,43±0,12
	knorau	94,94±0,35	95,00±0,16
	knop	95,18±0,23	95,52±0,26
	desmi	93,97±0,20	93,24±0,29
	metades	95,37±0,23	95,40±0,25
	metades-i	**96,05±0,52** •	**96,04±0,17** •
PK50	**knorae**	94,90±0,40	96,08±0,21
	knorau	95,73±0,34	95,81±0,29
	knop	95,78±0,31	96,20±0,28
	desmi	94,95±0,48	94,02±0,43
	metades	96,03±0,26	96,10±0,29
	metades-i	**96,58±0,32** •	**96,93±0,08** •
PK70	**knorae**	95,91±0,24	96,88±0,18
	knorau	96,54±0,29	96,66±0,09
	knop	96,68±0,26	96,93±0,18
	desmi	95,93±0,17	94,88±0,23
	metades	96,93±0,18	96,85±0,16
	metades-i	**97,95±0,56** •	**97,54±0,27** •
PK100	**knorae**	96,71±0,17	97,43±0,14
	knorau	97,29±0,11	97,06±0,22
	knop	97,43±0,11	97,36±0,19
	desmi	96,69±0,13	95,46±0,27
	metades	97,56±0,14	97,38±0,12
	metades-i	**98,36±0,35** •	**98,18±0,23** •

Fonte: O autor.

3). Verificou-se que em todas as situações o esquema *METADES-i* demonstrou superioridade com relação aos demais esquemas de seleção.

Esta métrica descreve a média ponderada da sensibilidade dos esquemas, portanto possui alta relevância ao avaliar o desempenho sobre os subconjuntos da base de dados *PUKYU*. Assim, observou-se que existe mínima diferença entre o cenário mais desbalanceado para o menos desbalanceado, tanto para a composição homogênea quanto para heterogênea. Portanto esta métrica demonstra que mesmo se tratando de cenários com alto índices de desbalanceamentos, os esquemas de seleção dinâmica têm demonstrado ótimos desempenhos, principalmente o esquema *METADES-i* que explora o uso de metadados desbalanceados.

5.4 ANÁLISE DE DESEMPENHO EM TERMOS DE *OVERFITTING*

A seguir são descritos os resultados do Algoritmo 2 relativos a avaliação do *overfitting* dos ESDs no que diz respeito à acurácia média (OVAM) e à acurácia balanceada (OVAB). Visto que os modelos de classificação visam minimizar o *overfitting* e conforme descrito no algoritmo 2, quanto maior o escore menor o *overfitting*, então os resultados a seguir podem ser normalizados $(100 - escore)$;.

5.4.1 *Overfitting* com Relação à Acurácia Média

A Tabela 7 lista os resultados obtidos pelos esquemas de seleção dinâmica em termos de OVAM (Experimento 4). Observou-se que ESDs com composição homogênea, o esquema *KNOP* obteve o maior desempenho em todos os cenários, porém nos cenários $PK50$ e $PK100$ não houve diferença estatística com determinados esquemas. Nos testes dos ESDs com composição heterogênea, o *overfitting* da acurácia media do *METADES-i* foi a melhor registrada em três dos cinco cenários, *PK20*, *PK70* e *PK100*. Nos demais cenários, os esquemas com as maiores médias de *overfitting* foram o *KNORA-U* (*PK30*) e KNOP (*PK50*). Em nenhum dos cenários, o ESD com melhor desempenho de *overfitting* demonstrou diferença estatística sobre os demais esquemas.

A Figura 16 ilustra os resultados do teste estatístico de *wilcoxon*, através de caixas referentes aos dados de desempenho dos ESDs em termos de OVAM (Experimento 4). Observou-se que nos ESDs com composição homogênea o esquema *KNOP* foi significativamente superior na maioria dos cenários, porém o esquema *KNORA-U* demonstrou competitividade em certos cenários. Eventualmente, tratando-se d ESDs com composição heterogênea, não houve diferença estatística entre os esquemas: *KNOP*, *KNORA-U* e *METADES-i*.

Uma relevante análise levantada pelos autores Urbanowicz *et al.* (2018) é que um nível de confiança de até 5% em termos de *overfitting* é considerado aceitável. Deste modo, levando em consideração a minimização de *overfitting* por meio da normalização $(100 - escore)$, todos os esquemas obtiveram resultados aceitáveis em termos de OVAM.

5.4.2 *Overfitting* com Relação à Acurácia Balanceada

A Tabela 8 mostra o resultado de desempenho dos ESD em termos de OVAB (Experimento 5). Similarmente aos resultados de *overfitting* da acurácia média, o esquema *KNOP* obteve o melhor desempenho nos testes dos ESD com composição homogênea. Porém, houve diferença estatística em três dos cinco cenários ($PK30$, $PK50$ e $PK70$). Nos cenários dos

Tabela 7: Resultado de *Overfitting* em termos de Acurácia média (OVAM), com conjunto de classificadores homogêneo (HM) e heterogêneo (HT).

Cenário	Esquema	OVAM-HM	OVAM-HT
PK20	**knorae**	91,34±0,29	95,01±0,69
	knorau	98,77±0,52	99,54±0,68
	knop	**99,14±0,30** •	99,52±0,74
	desmi	98,09±0,26	98,99±0,67
	metades	95,46±0,46	97,89±0,71
	metades-i	96,33±0,28	**99,71±0,28**
PK30	**knorae**	92,48±0,20	95,90±0,31
	knorau	98,82±0,55	**99,31±0,58**
	knop	**99,31±0,32** •	99,19±0,57
	desmi	98,07±0,35	98,57±0,53
	metades	95,93±0,46	97,73±0,51
	metades-i	96,76±0,40	98,93±0,70
PK50	**knorae**	93,27±0,82	95,86±0,4
	knorau	98,70±0,74	**98,87±0,67**
	knop	**98,80±0,58**	98,81±0,80
	desmi	97,71±0,89	97,44±0,96
	metades	96,13±0,38	97,48±0,65
	metades-i	96,92±0,40	98,46±0,86
PK70	**knorae**	94,81±0,46	96,94±0,37
	knorau	99,33±0,53	99,59±0,31
	knop	**99,56±0,43** •	99,52±0,23
	desmi	98,59±0,39	98,57±0,32
	metades	97,54±0,31	98,20±0,21
	metades-i	97,95±0,48	**99,68±0,27**
PK100	**knorae**	95,47±0,43	97,38±0,43
	knorau	99,62±0,38	99,43±0,58
	knop	**99,71±0,27**	99,53±0,51
	desmi	98,75±0,35	98,52±0,48
	metades	97,81±0,34	98,33±0,35
	metades-i	98,93±0,18	**99,62±0,27**

Fonte: O autor.

ESD com composição heterogênea o melhor desempenho foi observado quando no emprego do ESD *METADES-i*, sobre todos cenários. No entanto, em nenhum dos cenários houve diferença estatística ao aplicar o teste não paramétrico.

A Figura 17 ilustra os resultados do Experimento 5, pelo qual aplica o teste estatístico de *wilcoxon* e o referencia por meio de caixas de diagramas. Observou-se que em todos os cenários o esquema *KNOP* foi superior ou não há diferença estatística quando o mesmo não foi superior. Com relação aos piores desempenhos em termos e $OVAB$, o esquemas *DESMI*, *KNORA-E* e *KNORA-U*, em todos os cenários são significativamente inferiores ao melhor esquema. Com relação ao esquema *KNORA-U*, na maioria dos cenários não há diferença estatís-

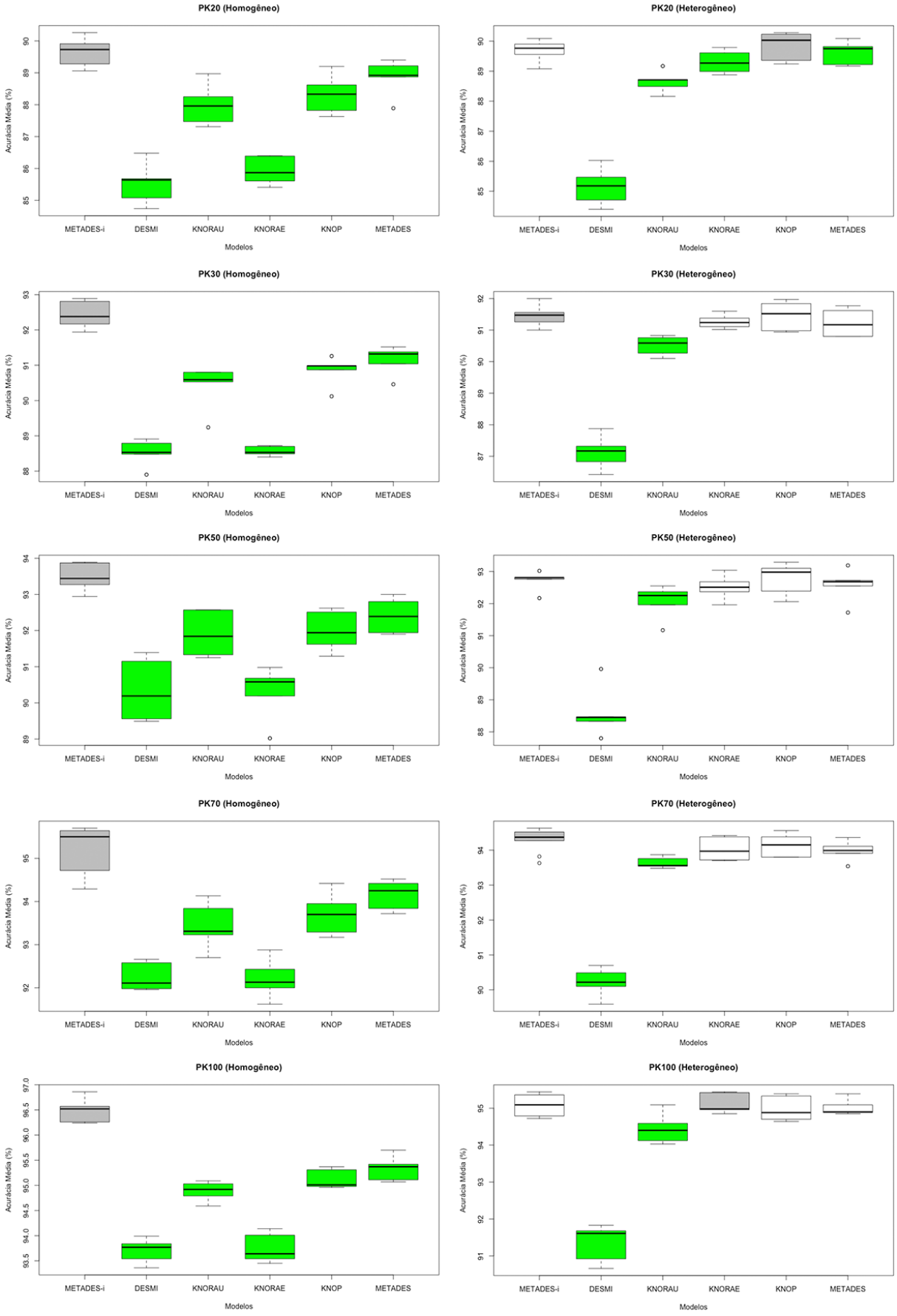
tica do desempenho deste esquema com relação ao esquema superior. Uma relevante análise com relação ao esquema *METADES-i*, ao utilizar a composição homogênea o modelo foi estatisticamente inferior, mas com a composição heterogênea, o modelo foi superior em todos cenários, mesmo não houve diferença estatística com determinados esquemas.

Tabela 8: Resultado de *overfitting* em termos de acurácia balanceada (OVAB), com conjunto de classificadores homogêneo (HM) e heterogêneo (HT).

Cenário	Esquema	OVAB-HM	OVAB-HT
PK20	**knorae**	74,43±1,03	86,48±1,92
	knorau	96,44±1,35	98,49±1,14
	knop	**96,68±1,38**	99,17±1,17
	desmi	91,19±1,05	97,16±2,04
	metades	84,27±0,81	94,21±1,78
	metades-i	88,15±1,13	**99,41±0,31**
PK30	**knorae**	78,35±1,24	87,85±1,39
	knorau	95,91±1,75	97,50±1,44
	knop	**97,41±2,06** •	97,59±0,85
	desmi	91,15±2,27	95,28±2,19
	metades	87,55±1,75	93,40±1,38
	metades-i	90,20±1,53	**98,05±0,71**
PK50	**knorae**	84,12±1,15	90,67±0,98
	knorau	96,12±1,50	97,07±1,32
	knop	**96,99±1,98** •	97,50±0,96
	desmi	93,86±1,66	94,19±1,64
	metades	90,70±1,52	94,59±1,20
	metades-i	92,12±2,18	**98,04±0,50**
PK70	**knorae**	88,79±0,41	93,35±0,97
	knorau	98,40±0,53	98,82±0,69
	knop	**98,91±0,85** •	98,84±0,79
	desmi	96,18±0,41	95,39±0,57
	metades	94,33±0,69	96,59±1,12
	metades-i	95,31±1,01	**98,88±1,05**
PK100	**knorae**	91,70±1,05	95,34±0,93
	knorau	99,04±0,71	98,70±1,04
	knop	**99,39±0,65**	99,06±0,86
	desmi	97,00±0,84	96,72±1,10
	metades	96,16±0,61	97,25±0,89
	metades-i	96,60±0,76	**99,17±0,70**

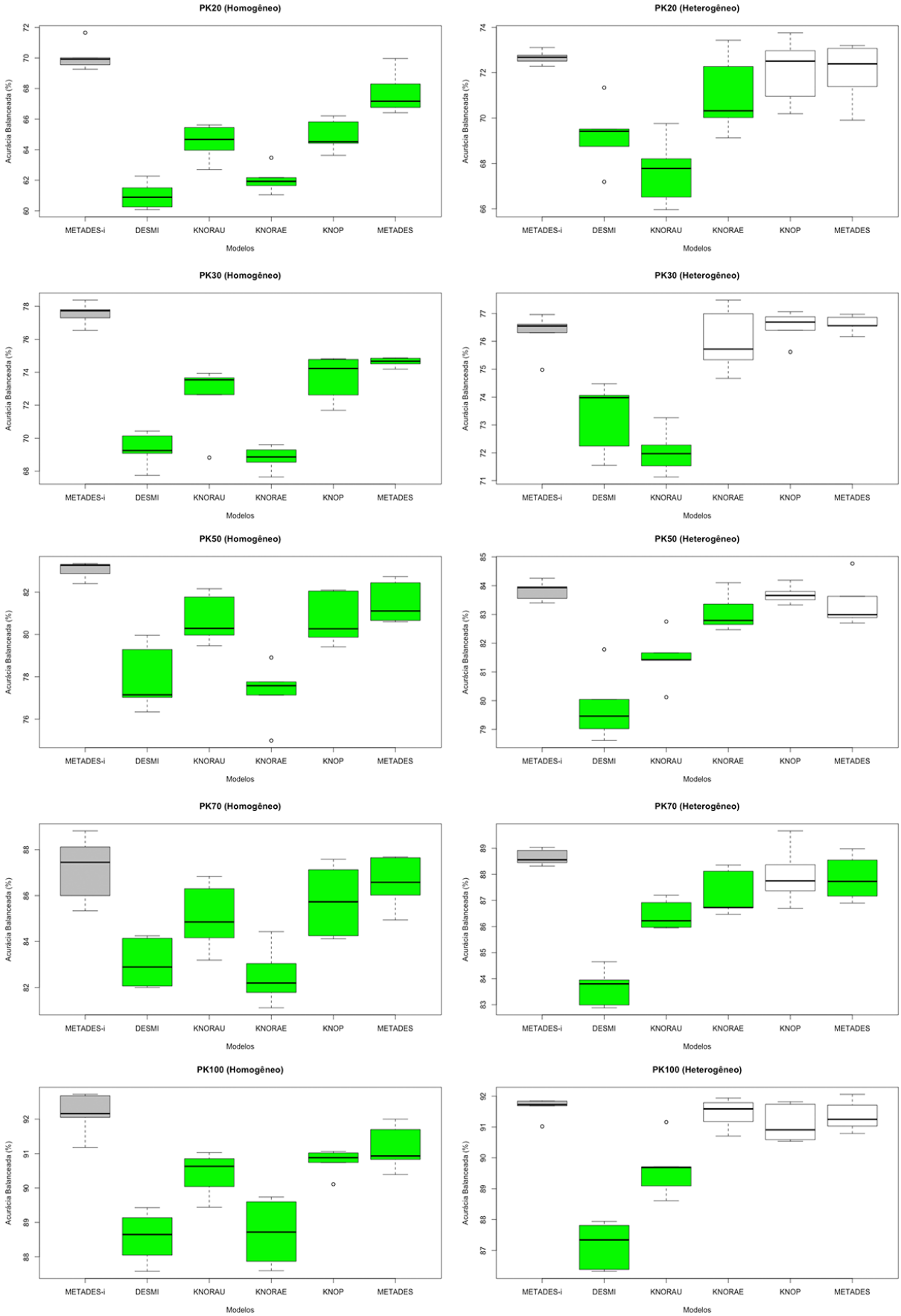
Fonte: O autor.

Figura 13: Análise do pareamento de *wilcoxon* em termos de acurácia média.



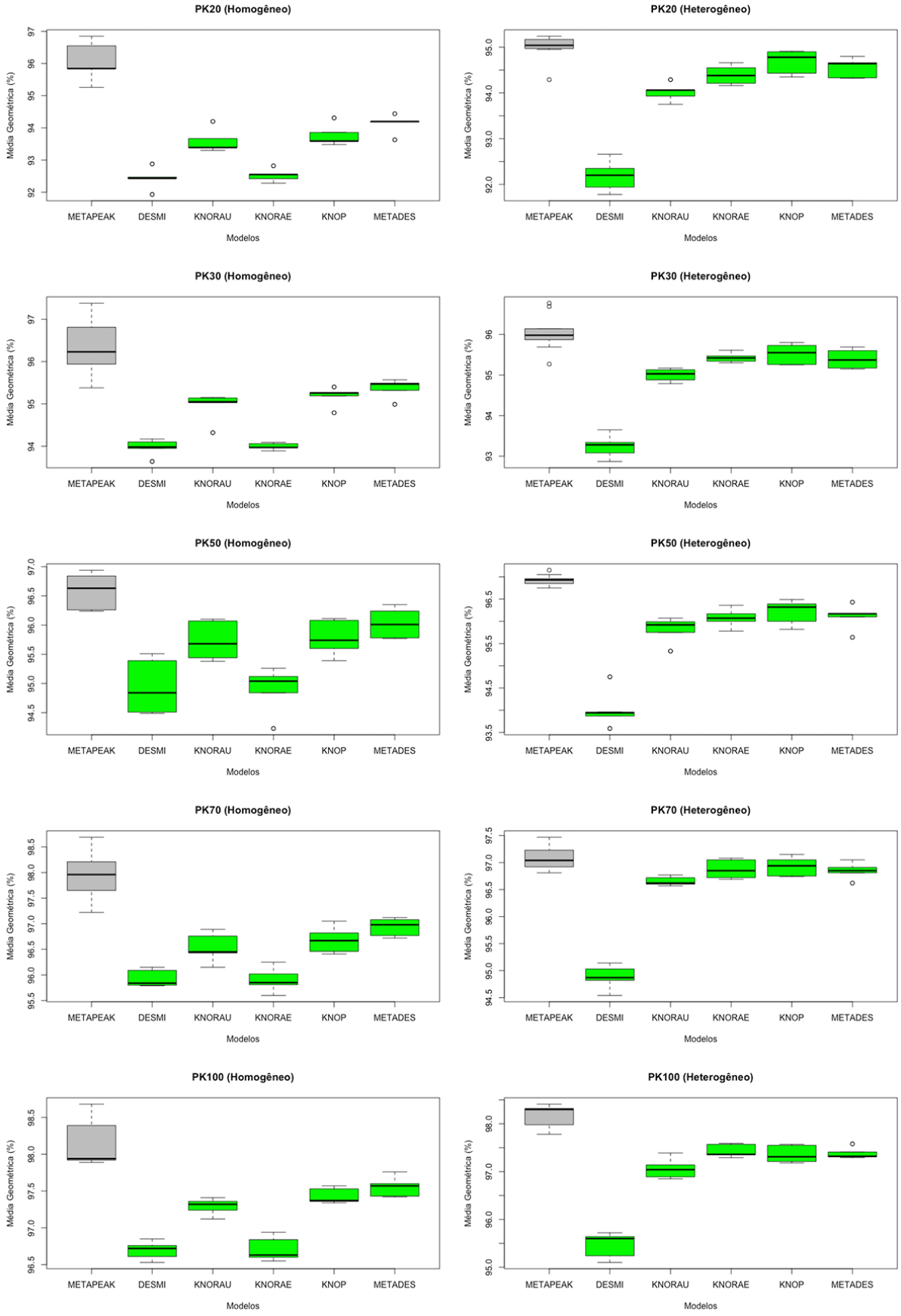
Fonte: O autor.

Figura 14: Análise do pareamento de *wilcoxon* em termos de acurácia balanceada.



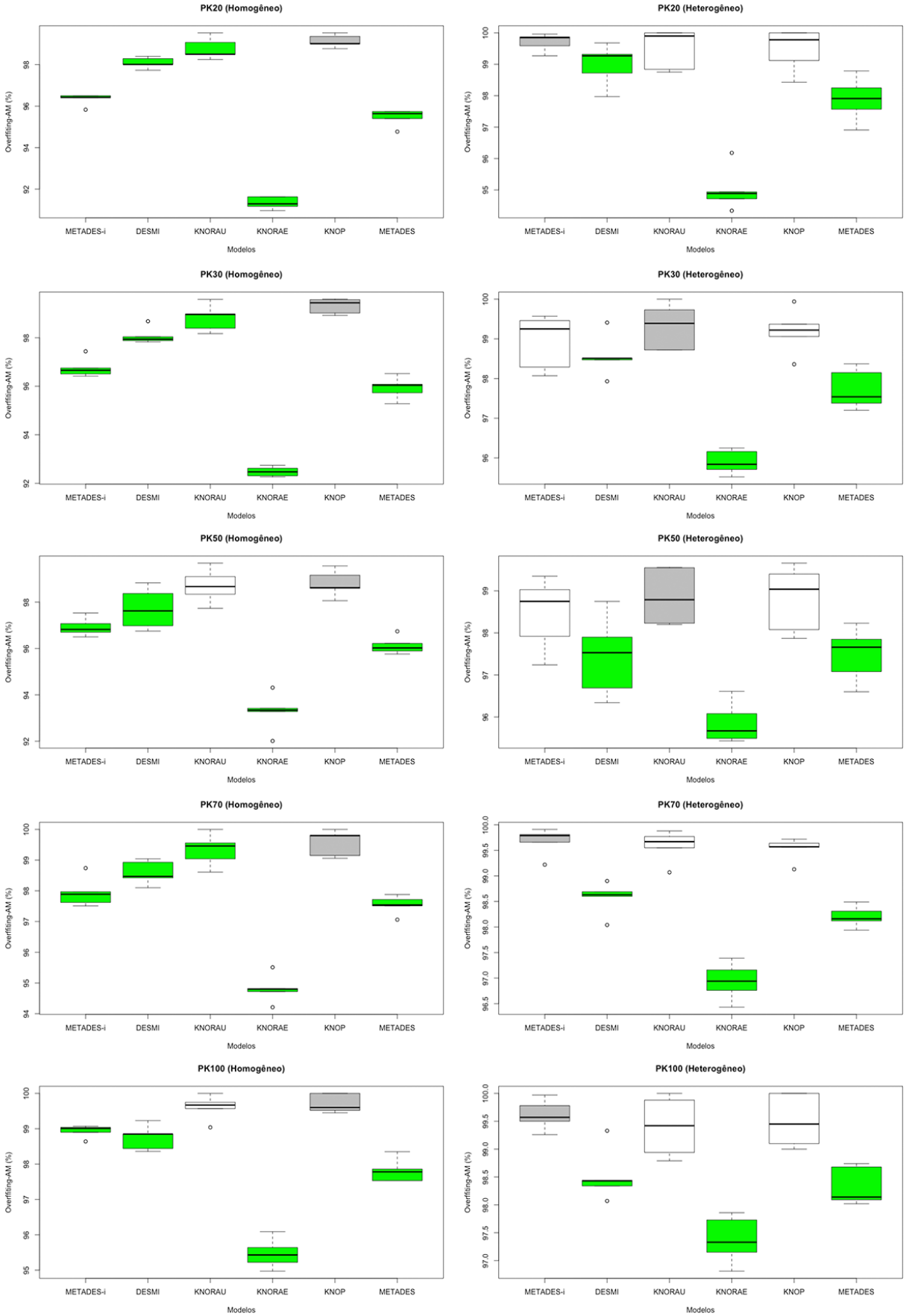
Fonte: O autor.

Figura 15: Análise do pareamento de *wilcoxon* em termos de média geométrica.



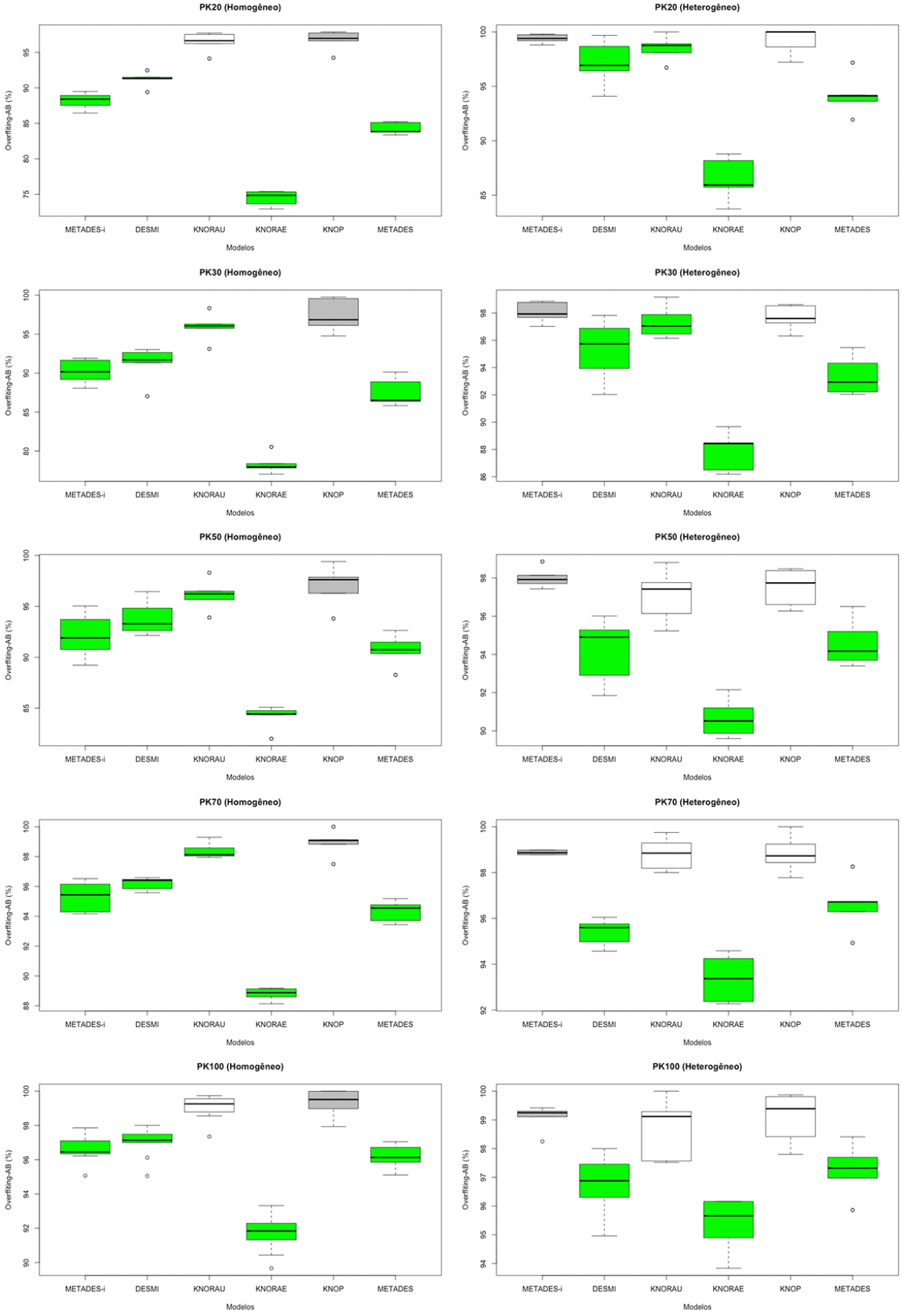
Fonte: O autor.

Figura 16: Gráfico de análise do teste de *wilcoxon* em termos de OVAM do conjunto homogêneo e heterogêneo



Fonte: O autor.

Figura 17: Gráfico de análise do teste de *wilcoxon* em termos de OVAB do conjunto homogêneo e heterogêneo



Fonte: O autor.

5.4.3 Análise Multiobjetivo em Termos de *Overfitting* e Acurácia Balanceada

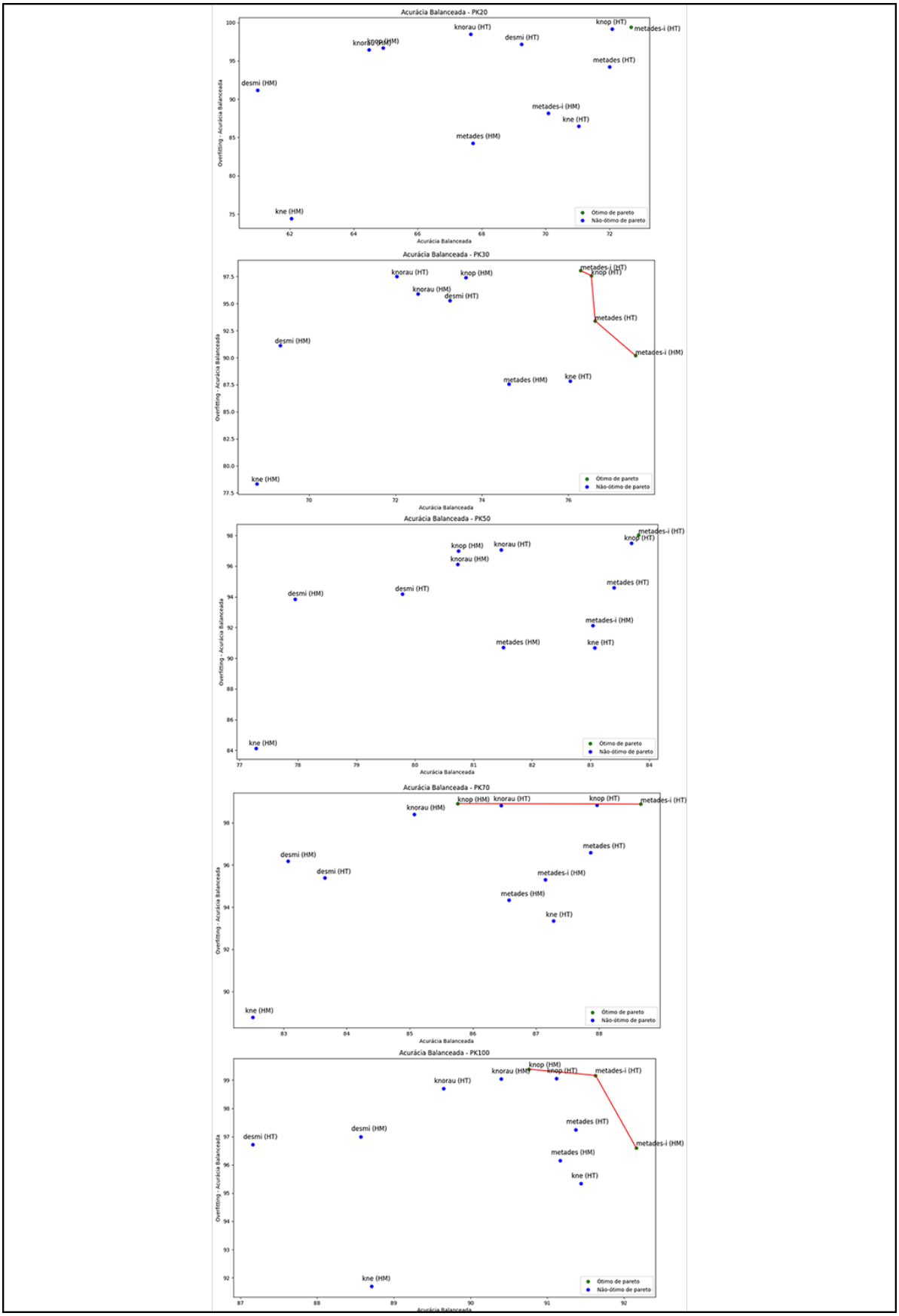
A Figura 18 mostra os gráficos da análise multiobjetivo de Pareto (PARETO, 1964), pelo qual leva em consideração o desempenho dos ESDs no que se refere a acurácia balanceada e OVAB. Quando um esquema é dominante em algum dos objetivos então é dito que ele é ótimo de Pareto e ilustrado pela cor verde, caso contrário é ilustrado pela cor azul, assim os esquemas ótimos de Pareto formam uma fronteira de dominância. Nesta análise foi levado em conta os ESDs com composição homogênea e heterogênea, conforme indicado na figura por HM e HT.

Conforme visto na Figura 18, o esquema *METADES-i* com composição heterogênea fez parte da fronteira de Pareto em todos cenários. Em particular, nos cenários *PK20* e *PK50* este esquema dominou todas as demais abordagens. Com relação a comparação do esquema *METADES-i* e *MESTADES*, em quatro dos cinco cenários o *METADES-i* obteve o desempenho superior ao *METADES* em ambos os escores, apenas no cenário *PK30* o *METADES* fez parte da fronteira de dominância.

O esquema *KNOP* com composição homogênea foi um dos ESD da fronteira de Pareto em dois cenários (*PK70* e *PK100*). Entretanto, sua versão com composição heterogênea fez parte da fronteia no cenário *PK30*. Importante analisar que quando este esquema faz parte da fronteira de Pareto, seu objetivo de dominância foi em termos de OVAB, ou seja, na maior parte dos cenários o esquema foi dominante em termos de *overfitting*.

De modo geral, foi possível destacar que os esquemas *METADES-i* e *KNOP* estão presentes na fronteira de Pareto em todos os cenários, independente de suas composições. Ainda que o *KNOP* possui dominância em termos de OVAB, o *METADES-i* possui dominância em termos de acurácia balanceada e em certos cenários possui dominância em termos de OVAB. Assim, levando em consideração que a métrica de acurácia balanceada é mais adequada para mensurar escores com dados desbalanceados, estes esquemas demonstraram-se eficiente em análise multiobjetivo sobre os cenários do conjunto de dados $PUKYU$.

Figura 18: Análise multiobjetivo de Pareto em termos de acurácia balanceada e OVAB.



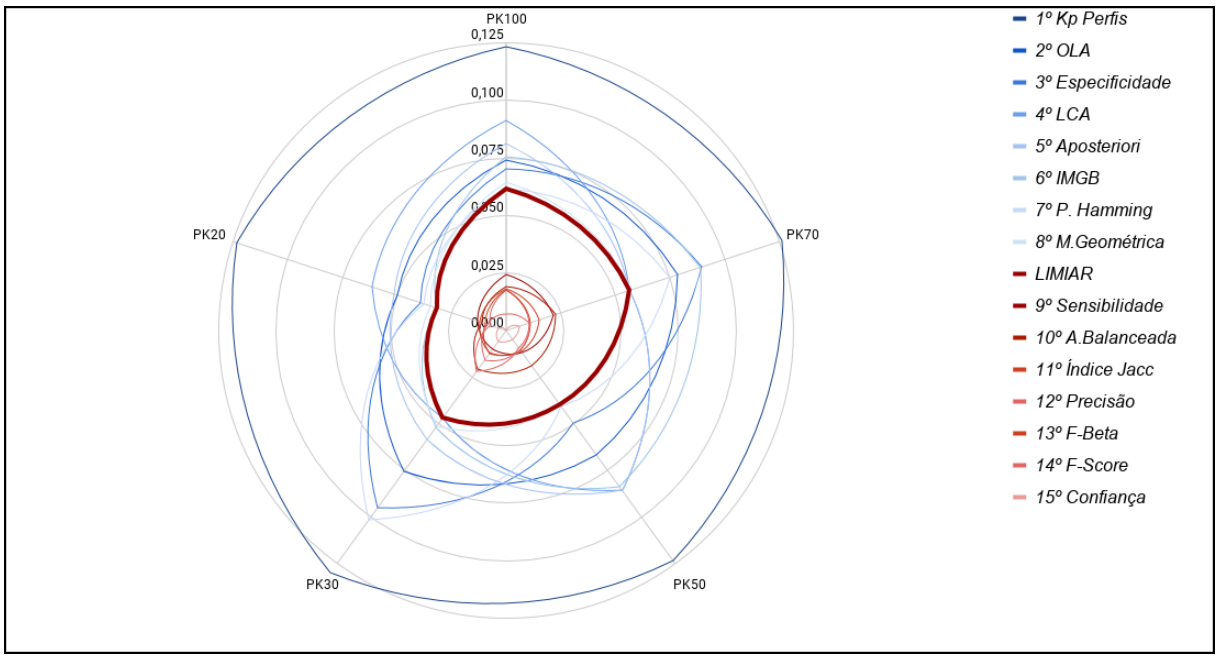
Fonte: O autor.

5.5 ANÁLISE DAS METACARACTERÍSTICAS DO *METADES-i*

Esta seção mostra os resultados da análise de seleção de metacaracterísticas obtidos por meio dos experimentos 6 e 7. O procedimento de seleção das metacaracterísticas, foi realizado por meio do algoritmo *Relief* (KIRA; RENDELL, 1992), que tem como objetivo ranquear atributos levando em consideração sua relevância para estimar o atributo alvo. Neste caso, os atributos utilizados para definir o ranqueamentos, foram as metacaracterísticas extraídas na fase de treinamento do esquema *METADES-i*.

A Figura 19 ilustra através de um gráfico de radar o ranqueamento do algoritmo *Relief* sobre as metacaracterísticas do esquema *METADES-i* com a composição homogênea. Neste gráfico além de plotar os escores que as metacaracterísticas obtiveram em cada cenário, ainda ilustra o valor do limiar da mediana sobre os cenários, conforme sugerido por Urbanowicz *et al.* (2018). Ao analisar este gráfico foi possível notar a diferença dos escores das metacaracterísticas que estão inferiores ao limiar da mediana. Assim, as metacaracterísticas do *METADES-i* com composição homogênea que foram superiores ao limiar são: *Kp perfis*, *LCA*, *Aposteriori*, *IMGB*, *OLA*, *especificidade* e *perda hamming*.

Figura 19: Ranqueamento do algoritmo *Relief* sobre as metacaracterísticas do esquema *METADES-i* com composição homogênea.

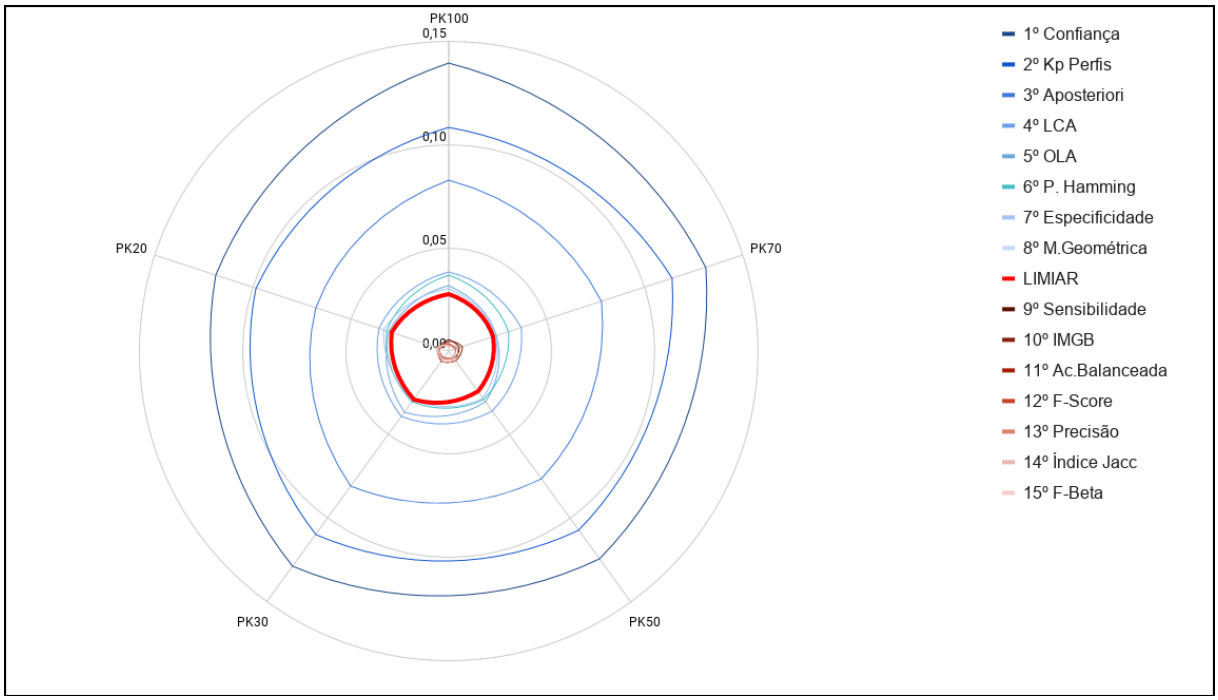


Fonte: O autor.

A Figura 20 ilustra o ranqueamento das metacacterísticas do esquema *METADES-i* com a composição heterogênea. Neste ranqueamento, as metacaracterísticas superiores ao limiar foram: *confiança*, *Kp perfis*, *Aposteriori*, *LCA*, *OLA*, *perda hamming*, *especificidade* e *média geométrica*. O gráfico da Figura 20 permite observar que nos ESDs em composições

heterogêneas, a ordenação do escores das metacaracterísicas, calculados conforme o algoritmo *Relief*, não foi alterada pela variação dos cenários. Assim, segundo aquele algoritmo, a *confiança* foi a metacaracterística com o maior escore em todas os cenários. As metacaracterísticas sensíveis a dados desbalanceados, a *perda hamming*, *especificidade* e *média geométrica*, obtiveram escores acima do limiar de seleção (mediana), assim sendo selecionadas como relevantes através do algoritmo *Relief*.

Figura 20: Ranqueamento do algoritmo *Relief* sobre as metacaracterísticas do esquema *METADES-i* com composição heterogênea.



Fonte: O autor.

O gráfico da 19 mostra que, segundo o algoritmo *Relief*, a metacaracterística *Kp perfis* obteve seu escore superior em todos os cenários nos ESDs com composição homogênea. No que diz respeito às metacarecterísticas sensíveis a dados desbalanceados, os resultados mostram que o procedimento *Relief* selecionou o mesmo subconjunto de elementos da composição heterogênea com a adição do *IMGB*.

As Tabelas 9 e 10 mostram os resultados do Experimento 7 que procedeu de uma análise comparativa do desempenho do esquema *METADES-i* com uma versão sua, denominada *METADES-ii*, que utiliza o algoritmo *Relief* para proceder uma seleção das metacaracterísticas a serem usadas durante a fase de seleção. Para tanto, o teste de *wilcoxon* foi aplicado sobre os cenários e os escores obtidos nas métricas de AM, AB, OVAM e OVAB. Adicionalmente, foi mostrado o valor do do *p-value*, identificando em quais pares houve diferença esta estatística.

Os resultados obtidos no Experimento 7 com relação aos esquema com composição homogênea são ilustrados na Tabela 9, onde foi possível constatar, pelo teste de *wilcoxon*, que

não houve diferença estatística na maioria dos casos. Isso mostra que o desempenho dos esquemas *METADES-i* e *METADES-ii* são semelhantes ao utilizar os esquemas com composição homogênea.

Tabela 9: Análise comparativa do desempenho dos esquema *METADES-i* e *METADES-ii*, com composição homogênea.

Cenário	Métrica	*METADES-i*	*METADES-ii*	p-value
PK20	**AM**	88,24±0,36	**88,52±0,48**	0,1674
	AB	73,73±0,35	**74,16±0,32**	0,7352
	OVAM	93,16±0,43	**93,33±0,34**	0,5062
	OVAB	86,39±1,06	**87,28±0,38**	0,0360 •
PK30	**AM**	90,78±0,37	**91,07±0,49**	0,1674
	AB	75,86±0,36	**76,29±0,33**	0,0735
	OVAM	95,84±0,44	**96,02±0,35**	0,5062
	OVAB	88,88±1,09	**89,80±0,39**	0,0360 •
PK50	**AM**	**92,44±0,52**	92,21±0,51	0,0049 •
	AB	**81,68±1,07**	81,14±1,01	0,0049 •
	OVAM	**97,30±0,37**	97,07±0,57	0,0047 •
	OVAB	**92,92±1,46**	92,65±1,65	0,3313
PK70	**AM**	**94,26±0,27**	94,13±0,29	0,0633
	AB	86,82±1,13	**86,83±1,23**	0,3847
	OVAM	97,99±0,30	**98,06±0,32**	0,0360 •
	OVAB	95,35±1,06	**95,71±1,16**	0,0049 •
PK100	**AM**	**95,11±0,27**	95,08±0,32	0,2371
	AB	**90,85±0,87**	90,81±1,04	0,5738
	OVAM	**98,17±0,23**	98,08±0,14	0,5062
	OVAB	**96,33±0,49**	96,20±0,72	0,8781

Fonte: O autor.

Entretanto, conforme ilustrado pela Tabela 10, ao utilizar os esquemas com composição heterogênea verificou-se, através do teste de *wilcoxon*, que em termos das métricas de AM e AB, o desempenho do esquema *METADES-ii* foi significativamente superior àquele do *MESTADES-i*. Mas ao comparar as métricas de OVAM e OVAB, constatou-se que o esquema *METADES-i* obteve o desempenho estatisticamente superior, isso ocorreu em todos os cenários de desbalanceamento.

Por fim, foi aplicado à análise multiobjetivo de *Pareto* sobre os desempenhos dos esquemas *METADES-i*, *METADES-ii* e *KNOP*, avaliando as métricas de acurácia balanceada e OVAB, para ambas as composições. Conforme é ilustrado na Figura 21, foi possível constatar a superioridade do esquema *KNOP* em termos de OVAB, principalmente sua versão com composição heterogênea que foi superior em todos os cenários. O esquema *METADES-ii* com composição heterogênea pertenceu a fronteira de *Pareto* em todos os cenários, devido a sua superioridade em termos de AB. Mas o esquema *METADES-i* não ficou na fronteira em nenhuma cenários, isso ocorreu pela dominância de objetivos específicos do demais esquemas *KNOP* e

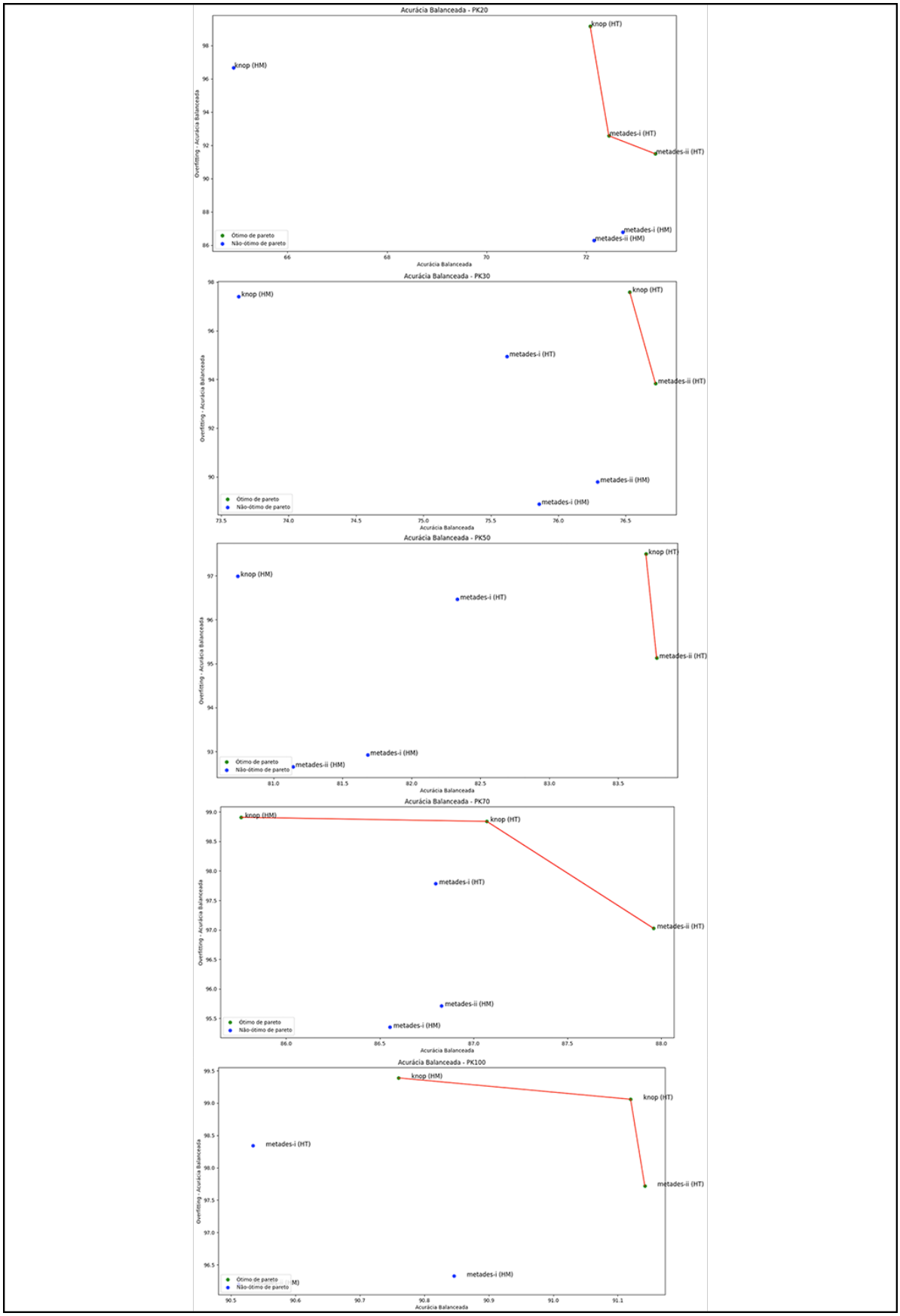
Tabela 10: Análise comparativa do desempenho dos esquema *METADES-i* e *METADES-ii*, com composição heterogênea.

Cenário	Métrica	*METADES-i*	*METADES-ii*	p-value
PK20	**AM**	89,59±0,49	**90,21±0,42**	0,0284 •
	AB	72,45±0,41	**73,39±0,80**	0,0382 •
	OVAM	**95,59±0,70**	94,51±0,78	0,0122 •
	OVAB	**92,58±0,45**	91,50±0,22	0,0049 •
PK30	**AM**	90,11±0,22	**90,86±0,28**	0,0049 •
	AB	75,62±0,58	**76,72±0,33**	0,0049 •
	OVAM	**97,54±0,71**	93,44±0,80	0,0122 •
	OVAB	**94,95±0,46**	93,84±0,22	0,0049 •
PK50	**AM**	92,16±0,57	**92,39±0,52**	0,0049 •
	AB	82,33±0,66	**82,98±0,72**	0,0049 •
	OVAM	**98,46±0,72**	97,84±0,73	0,0047 •
	OVAB	**96,47±1,12**	95,13±1,25	0,0049 •
PK70	**AM**	93,54±0,25	**93,84±0,30**	0,0047 •
	AB	86,80±0,82	**87,46±0,91**	0,0049 •
	OVAM	**99,08±0,27**	98,53±0,28	0,0049 •
	OVAB	**97,78±0,94**	97,03±1,20	0,0049 •
PK100	**AM**	94,67±0,23	**94,91±0,24**	0,0049 •
	AB	90,53±0,34	**91,14±0,43**	0,0049 •
	OVAM	**99,18±0,42**	98,65±0,32	0,0049 •
	OVAB	**98,34±1,04**	97,72±0,74	0,0122 •

Fonte: O autor.

METADES-ii. Assim, foi possível constatar que o procedimento de seleção de metacaracterísticas pelo algoritmo *Relief*, causa ganho em desempenho em termos de acurácia balanceada, fazendo com que esteja entre os esquemas dominantes.

Figura 21: Análise multiobjetivo de Pareto sobre os esquemas *METADES-i* e *METADES-ii*, em termos de acurácia balanceada e OVAB.



Fonte: O autor.

6 CONCLUSÕES

Este trabalho analisou o uso de esquemas de seleção dinâmica em sistemas de múltiplos classificadores para identificação de gêneros bacterianos a partir da relação *m/z* de proteínas ribossomais. Nesta análise foi avaliada a eficácia de diferentes esquemas que empregavam técnicas de meta-aprendizagem. Também foi proposto e avaliado um esquema de seleção dinâmica, nomeado de *METADES-i*, que estende o *META-DES* ao incorporar metacaracterísticas que são sensíveis ao desbalanceamento de dados.

A análise dos resultados experimentais constatou que o esquema de seleção dinâmica *METADES-i* obteve um desempenho médio superior aos esquemas descritos em termos de acurácia balanceada e média geométrica, para composições homogêneas e heterogêneas. No que diz respeito à acurácia média, o *METADES-i* foi o esquema de melhor desempenho em composições homogêneas, mas quando do seu emprego em composições heterogêneas seu desempenho foi superior em três dos cinco cenários. Os resultados obtidos sugerem que, para a aplicação proposta vale a afirmação de Cruz, Sabourin e Cavalcanti (2018) sobre o desempenho dos ESDs e o desbalanceamento dos subconjuntos. Segundo aqueles autores, a capacidade preditiva dos ESDs fornecem evidência de que a eficiência de um esquema baseado em meta-aprendizagem está relacionada à adequação das metacaracterísticas que ele explora em relação a fatores conjunturais do conjunto de dados, entre eles o desbalanceamento de classes.

Quanto as métricas relacionadas ao *overfitting*, os esquemas *METADES-i* e *KNOP* foram aqueles que obtiveram o desempenho médio superior em ambas as composições em todos os cenários. Este resultado, também é observado na análise de otimização multiobjetivo, em termos de acurácia balanceada e *overfitting* sobre acurácia balanceada.

Com relação a composição dos classificadores base na fase de geração, em termos de média geométrica e acurácia média, os esquemas que utilizaram a composição heterogênea obtiveram seus melhores desempenhos nos cenários em que os subconjuntos de dados apresentaram os menores índices de desbalanceamento ($PK20$, $PK30$ e $PK50$). Entretanto, em bases de dados mais desbalanceadas ($PK70$ e $PK100$) os esquemas que utilizaram as composições homogêneas foram superiores. Em termos de acurácia balanceada e *overffiting* ($OVAB$ e $OVAM$), os esquemas com composição heterogênea foram significativamente superiores.

O emprego do procedimento *Relief* para seleção de metacaracterísticas do esquema *METADES-i* aponta para uma dependência entre a relevância das metacarecterísticas e a composição dos classificadores base. Pois ao variar a composição dos classificadores alterna-se a relevância das metacracterísticas e consequentemente o desempenho preditivo do esquema.

Além disso, a análise de otimização multiobjetivo constatou que a utilização do procedimento de seleção de metacaracterísticas depende de um exame de relação de custo por benefício. Isto porque, se por um lado, a seleção de metacaracterísticas proporcionou um esquema que levou a um aumento do desempenho preditivo, ela também levou à perda em termos de *overfitting*.

Em trabalhos futuros, primeiramente pretende-se estender a análise de otimização multiobjetivo com a incorporação de diferentes métricas sensíveis a dados desbalanceados, tais como sensibilidade, f-medida e especificidade. Em segundo momento, pretende-se avaliar a aplicabilidade de outros procedimentos na seleção de metacaracterísticas.

REFERÊNCIAS

ALBERTO, B.; ALMEIDA, P. *Abordagens de pré-processamento de dados em problemas de classificaçao com classes desbalanceadas*. Tese (Doutorado) — Master's Thesis, Centro Federal de Educaçao Tecnológica de Minas Gerais . . . , 2012.

ANDERSON, D.; BURNHAM, K. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, v. 63, 2004.

BARELLA, V. H. Técnicas para o problema de dados desbalanceados em classificaćão herérquica. 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06012016145045/publico/VictorHugoBarella$_{d}issertacao_{r}evisada.pdf$>.

BRANCO, P.; TORGO, L.; RIBEIRO, R. P. Relevance-based evaluation metrics for multi-class imbalanced domains. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2017. p. 698–710.

BRODERSEN, K. H. *et al.* The balanced accuracy and its posterior distribution. In: IEEE. *2010 20th International Conference on Pattern Recognition*. [S.l.], 2010. p. 3121–3124.

CARVALHO, A. *et al.* Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2011.

CAVALIN, P. R.; SABOURIN, R.; SUEN, C. Y. Logid: An adaptive framework combining local and global incremental learning for dynamic selection of ensembles of hmms. *Pattern recognition*, Elsevier, v. 45, n. 9, p. 3544–3556, 2012.

CAVALIN, P. R.; SABOURIN, R.; SUEN, C. Y. Dynamic selection approaches for multiple classifier systems. *Neural Computing and Applications*, Springer, v. 22, n. 3-4, p. 673–688, 2013.

CLAVERA, I. *et al.* Model-based reinforcement learning via meta-policy optimization. *arXiv preprint arXiv:1809.05214*, 2018.

CRUZ, R. M. *et al.* Deslib: A dynamic ensemble selection library in python. *arXiv preprint arXiv:1802.04967*, 2018.

CRUZ, R. M.; SABOURIN, R.; CAVALCANTI, G. D. Meta-des. h: a dynamic ensemble selection technique using meta-learning and a dynamic weighting approach. In: IEEE. *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2015. p. 1–8.

CRUZ, R. M.; SABOURIN, R.; CAVALCANTI, G. D. Analyzing different prototype selection techniques for dynamic classifier and ensemble selection. In: IEEE. *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2017. p. 3959–3966.

CRUZ, R. M.; SABOURIN, R.; CAVALCANTI, G. D. Meta-des. oracle: Meta-learning and feature selection for dynamic ensemble selection. *Information fusion*, Elsevier, v. 38, p. 84–103, 2017.

CRUZ, R. M.; SABOURIN, R.; CAVALCANTI, G. D. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, Elsevier, v. 41, p. 195–216, 2018.

CRUZ, R. M. *et al.* Meta-des: A dynamic ensemble selection framework using meta-learning. *Pattern recognition*, Elsevier, v. 48, n. 5, p. 1925–1935, 2015.

ESPÍNDOLA, R.; EBECKEN, N. On extending f-measure and g-mean metrics to multi-class problems. *WIT Transactions on Information and Communication Technologies*, WIT Press, v. 35, 2005.

FAN, Z. *et al.* Intelligence algorithms for protein classification by mass spectrometry. *BioMed research international*, Hindawi, v. 2018, 2018.

FERNÁNDEZ-DELGADO, M. *et al.* Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 3133–3181, 2014.

FOLEY, J. A. *et al.* Solutions for a cultivated planet. *Nature*, Nature Publishing Group, v. 478, n. 7369, p. 337, 2011.

GALAR, M. *et al.* A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 42, n. 4, p. 463–484, 2012.

GALAR, M. *et al.* Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, Elsevier, v. 46, n. 12, p. 3460–3471, 2013.

GANS, J.; WOLINSKY, M.; DUNBAR, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, American Association for the Advancement of Science, v. 309, n. 5739, p. 1387–1390, 2005.

GARCÍA, P. *et al.* Identificación bacteriana basada en el espectro de masas de proteínas: Una nueva mirada a la microbiología del siglo xxi. *Revista chilena de infectología*, Sociedad Chilena de Infectología, v. 29, n. 3, p. 263–272, 2012.

GARCÍA, S. *et al.* Dynamic ensemble selection for multi-class imbalanced datasets. *Information Sciences*, Elsevier, v. 445, p. 22–37, 2018.

GARTNER, R. What metadata is and why it matters. In: *Metadata*. [S.l.]: Springer, 2016. p. 1–13.

GIACINTO, G.; ROLI, F. Methods for dynamic classifier selection. In: IEEE. *Proceedings 10th International Conference on Image Analysis and Processing*. [S.l.], 1999. p. 659–664.

GIBB, S. *Entwicklung einer flexiblen bioinformatischen Plattform zur Analyse von Massenspektrometriedaten*. Tese (Doutorado) — Verlag nicht ermittelbar, 2015.

HAIXIANG, G. *et al.* Bpso-adaboost-knn ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 49, p. 176–193, 2016.

HOTTA, Y. *et al.* Classification of genus pseudomonas by maldi-tof ms based on ribosomal protein coding in s10- spc- alpha operon at strain level. *Journal of proteome research*, ACS Publications, v. 9, n. 12, p. 6722–6728, 2010.

JR, A. S. B.; SABOURIN, R.; OLIVEIRA, L. E. Dynamic selection of classifiers—a comprehensive review. *Pattern Recognition*, Elsevier, v. 47, n. 11, p. 3665–3680, 2014.

JURGEN, H. G. *Mass spectrometry: a textbook*. [S.l.]: SPRINGER INTERNATIONAL PU, 2018.

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: *Machine Learning Proceedings 1992*. [S.l.]: Elsevier, 1992. p. 249–256.

KITTLER, J.; HATER, M.; DUIN, R. P. Combining classifiers. In: IEEE. *Proceedings of 13th international conference on pattern recognition*. [S.l.], 1996. v. 2, p. 897–901.

KO, A. H.; SABOURIN, R.; JR, A. S. B. From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition*, Elsevier, v. 41, n. 5, p. 1718–1731, 2008.

KRAWCZYK, B. *et al.* Ensemble learning for data stream analysis: A survey. *Information Fusion*, Elsevier, v. 37, p. 132–156, 2017.

KUNCHEVA, L. I. Clustering-and-selection model for classifier combination. In: IEEE. *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)*. [S.l.], 2000. v. 1, p. 185–188.

LABUSCHAGNE, N. Plant growth promoting rhizobacteria as biofertilizers. 2003.

LEMAîTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *CoRR*, abs/1609.06570, 2016. Disponível em: <http://arxiv.org/abs/1609.06570>.

LEMKE, C.; BUDKA, M.; GABRYS, B. Metalearning: a survey of trends and technologies. *Artificial intelligence review*, Springer, v. 44, n. 1, p. 117–130, 2015.

LUQUE, A. *et al.* The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, Elsevier, v. 91, p. 216–231, 2019.

MELO, L. F. de. A utilização da espectrometria de massa maldi-tof na identificação de microrganismos no controle de qualidade farmacêutico. Universidade Federal de Minas Gerais, 2014.

NOORHALIM, N.; ALI, A.; SHAMSUDDIN, S. M. Handling imbalanced ratio for class imbalance problem using smote. In: SPRINGER. *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. [S.l.], 2019. p. 19–30.

PARETO, V. *Cours d'économie politique*. [S.l.]: Librairie Droz, 1964.

PASTERNAK, J. Novas metodologias de identificação de micro-organismos: Maldi-tof. *Einstein*, v. 10, p. 118–119, 2012.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PUSHPA, M.; KARPAGAVALLI, S. Multi-label classification: Problem transformation methods in tamil phoneme classification. *Procedia computer science*, Elsevier, v. 115, p. 572–579, 2017.

REUNANEN, J. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, v. 3, n. Mar, p. 1371–1382, 2003.

ROKACH, L. Ensemble-based classifiers. *Artificial Intelligence Review*, Springer, v. 33, n. 1-2, p. 1–39, 2010.

ROSSEL, S.; ARBIZU, P. M. Automatic specimen identification of harpacticoids (crustacea: Copepoda) using random forest and maldi-tof mass spectra, including a post hoc test for false positive discovery. *Methods in Ecology and Evolution*, Wiley Online Library, v. 9, n. 6, p. 1421–1434, 2018.

SABZEVARI, M.; MARTÍNEZ-MUÑOZ, G.; SUÁREZ, A. Pooling homogeneous ensembles to build heterogeneous ensembles. *arXiv preprint arXiv:1802.07877*, 2018.

SAGAWA, T. *et al.* Logistic regression of ligands of chemotaxis receptors offers clues about their recognition by bacteria. *Frontiers in bioengineering and biotechnology*, Frontiers, v. 5, p. 88, 2018.

SANTOS, F. d. *et al.* Algoritmo knn na imputação de dados de espectros de massa do tipo maldi-tof: uma análise da influência da imputação com knn sobre o desempenho de classificadores logísticos para identificação de bactérias. Universidade Estadual de Ponta Grossa, 2018.

SOUZA, B. F. d. *Meta-aprendizagem aplicada à classificação de dados de expressão gênica*. Tese (Doutorado) — Universidade de São Paulo, 2010.

TAMURA, H.; HOTTA, Y.; SATO, H. Novel accurate bacterial discrimination by maldi-time-of-flight ms based on ribosomal proteins coding in s10-spc-alpha operon at strain level s10-germs. *Journal of the American Society for Mass Spectrometry*, Springer, v. 24, n. 8, p. 1185–1193, 2013.

TERAMOTO, K. *et al.* Phylogenetic classification of pseudomonas putida strains by maldi-ms using ribosomal subunit proteins as biomarkers. *Analytical chemistry*, ACS Publications, v. 79, n. 22, p. 8712–8719, 2007.

TODD, J. F. Recommendations for nomenclature and symbolism for mass spectroscopy (including an appendix of terms used in vacuum technology). *International journal of mass spectrometry and ion processes*, Elsevier Science Publishing Company, Inc., v. 142, n. 3, p. 209–240, 1995.

TOMACHEWSKI, D. Utilização de aprendizado de máquina para classificação de bactérias através de proteínas ribossomais. Universidade Estadual de Ponta Grossa, 2017.

TOMACHEWSKI, D. *et al.* Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. *Bioinformatics*, Oxford University Press, v. 1, p. 3, 2018b.

TOMACHEWSKI, D. *et al.* Pukyu - banco de dados de massa molecular de proteínas ribossomais. Universidade Estadual de Ponta Grossa, 2018a.

URBANOWICZ, R. J. *et al.* Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, Elsevier, v. 85, p. 189–203, 2018.

VRIEZE, J. de. *The littlest farmhands*. [S.l.]: American Association for the Advancement of Science, 2015.

WARDHANI, N. W. S. *et al.* Cross-validation metrics for evaluating classification performance on imbalanced data. In: IEEE. *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. [S.l.], 2019. p. 14–18.

WEI, L. *et al.* A novel hierarchical selective ensemble classifier with bioinformatics application. *Artificial intelligence in medicine*, Elsevier, v. 83, p. 82–90, 2017.

WICHARD, J. D. Model selection in an ensemble framework. In: IEEE. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. [S.l.], 2006. p. 2187–2192.

WIESER, A. *et al.* Maldi-tof ms in microbiological diagnostics—identification of microorganisms and beyond (mini review). *Applied microbiology and biotechnology*, Springer, v. 93, n. 3, p. 965–974, 2012.

WOLOSZYNSKI, T.; KURZYNSKI, M. A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognition*, Elsevier, v. 44, n. 10-11, p. 2656–2668, 2011.

WOODS, K.; KEGELMEYER, W. P.; BOWYER, K. Combination of multiple classifiers using local accuracy estimates. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 19, n. 4, p. 405–410, 1997.

ZHANG, L.; JIANG, L.; LI, C. A discriminative model selection approach and its application to text classification. *Neural Computing and Applications*, Springer, p. 1–15, 2017.

ZHU, X.; WU, X.; YANG, Y. Dynamic classifier selection for effective mining from noisy data streams. In: IEEE. *Fourth IEEE International Conference on Data Mining (ICDM'04)*. [S.l.], 2004. p. 305–312.