PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

Marcelo Ribeiro Trierveiler

Caracterização de Instâncias por índice de discriminação para pré-processamento da região de competência em sistemas de seleção de múltiplos classificadores

DOUTORADO EM
CIÊNCIA DA COMPUTAÇÃO
PUCPR

CURITIBA
2025

Marcelo Ribeiro Trierveiler

Caracterização de Instâncias por índice de discriminação para preprocessamento da região de competência em sistemas de seleção de múltiplos classificadores

Curitiba - PR, Brasil
2025

Marcelo Ribeiro Trierveiler

Caracterização de Instâncias por índice de discriminação para preprocessamento da região de competência em sistemas de seleção de múltiplos classificadores

Tese apresentada ao Programa de Pós-Graduação em Informática aplicada (PPGIa) da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Doutor em Informática.

Orientador: Alceu de Souza Britto Jr.

Curitiba - PR, Brasil
2025

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

Trierveiler, Marcelo Ribeiro
T826c Caracterização de instâncias por índice de discriminação para processamento da
2025 região de competência em sistemas de seleção de múltiplos classificadores /
Marcelo Ribeiro Trierveiler ; orientador: Alceu de Souza Britto Jr. – 2025.
120 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2025.
Bibliografia: f. 104-120

1. Informática. 2. Aprendizado do computador. 3. Seleção dinâmica de classificadores. 4. Testes psicométricos. I. Britto Júnior, Alceu de Souza.
II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. III. Título.

CDD 20. ed. – 004

Bibliotecária: Luci Eduarda Wielganczuk – CRB 9/1118

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 09 de dezembro de 2025.

113-2025

DECLARAÇÃO

Declaro para os devidos fins, que **MARCELO RIBEIRO TRIERVEILER** defendeu a tese de Doutorado intitulada "**Caracterização de Instâncias por índice de discriminação para pré-processamento da região de competência em sistemas de seleção de múltiplos classificadores"**, na área de concentração Ciência da Computação no dia 07 de agosto de 2025, no qual foi aprovado.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade, firmo a presente declaração.

Documento assinado digitalmente
gov.br
JEAN PAUL BARDDAL
Data: 10/12/2025 17:06:14-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Jean Paul Barddal
Coordenador do Programa de Pós-Graduação em Informática

A todos que estiveram do meu lado, me incentivaram e ajudaram.

Agradecimentos

Agradeço a todos que estiveram do meu lado, a todos que incentivaram e ajudaram. Sinceramente prefiro não nomear tudo e todos que me ajudaram pois eu correria o risco de esquecer de algo ou de alguém. Digo isso, pois eu teria que considerar tudo e todas pessoas que de alguma forma cruzaram meu caminho acadêmico ou não, mas me deram subsídios para que eu chegasse onde estou. Realmente o agradecimento seria para toda a humanidade, todos animais e toda a matéria de nosso planeta de todas as épocas que puderam dar fundamento para tudo que temos à nossa volta. Mais uma vez prefiro não nomear mas agradecer ainda assim a tudo e a todos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O presente trabalho foi realizado com apoio do Instituto Federal do Paraná (IFPR).

Otimista: O copo de água está meio-cheio
Pessimista: O copo de água está meio-vazio
Desenvolvedor: O copo de água é duas vezes maior que o necessário!

Resumo

Em sistemas de seleção de múltiplos classificadores, a estimativa de competência dos classificadores para predizer uma instância desconhecida é uma questão de suma importância. Nestes sistemas, a região de competência fornece aos algoritmos uma maneira de estimar a competência de classificadores. Uma região de competência com baixa qualidade pode comprometer todo o sistema. Os algoritmos de pré-processamento podem contribuir substancialmente neste processo de seleção de instâncias, visando escolher aquelas com melhor qualidade. Entretanto, a maioria dos estudos se concentra em pré-processamento, utilizando somente informações estáticas de instâncias vizinhas. É necessária a obtenção de informações adicionais para estabelecer a relevância de uma instância que estará inserida em uma área de competência. Com este entendimento surgem várias perguntas, como: será que podemos melhorar a qualidade da região de competência através da poda dinâmica de instâncias de baixa qualidade? Como podemos conduzir esta poda? Como verificar instâncias de baixa qualidade? Nesta pesquisa, investigamos estas e outras perguntas relacionadas ao aprimoramento da região de competência de forma dinâmica. Foi explorada a adoção da estimativa do Índice de Discriminação proveniente da teoria dos testes psicométricos para caracterizar a importância de uma instância pertencente à região de competência. Pelos experimentos conduzidos, houve uma melhoria significativa dos algoritmos de seleção dinâmica estudados, indicando que a poda das instâncias a partir deste índice de discriminação pode ser uma alternativa viável nestes sistemas. Vários experimentos foram realizados para reconhecer o comportamento destas podas. Foi utilizado um protocolo experimental com 30 datasets e vários algoritmos de seleção dinâmica. Após a execução dos experimentos, observa-se que a metodologia proposta se configura como uma alternativa viável para oferecer aos algoritmos de seleção dinâmica uma nova perspectiva para a escolha de instâncias que irão compor a região de competência. Código disponível no link: https://github.com/marcelo-trier/DISi

Palavras-chave: Sistemas de Múltiplos Classificadores, Seleção Dinâmica de Classificadores, Testes Psicométricos.

Abstract

In multi-classifier selection systems, estimating the classifiers' competence in a POOL to predict an unknown instance is a significant issue. A low-quality region of competence can incorrectly evaluate the competence of classifiers and thus compromise the whole system. Preprocessing algorithms have a critical part in this process of selecting instances to choose those with better quality. However, most studies focus on preprocessing using only static information of the instances. More information is required to understand better the importance of an instance in the dynamic phase that can bring better quality to the region of competence (RoC). This way, it is possible to provide cases where dynamic selection systems benefit by correctly estimating the competence of their classifiers. From this understanding, several questions emerge, such as: Can we improve the quality of the competence region by dynamically pruning low-quality instances? How can we conduct this pruning? How can we verify low-quality samples? In the current work, we investigate these and other questions related to improving the competence region in a dynamic matter. We explored the adoption of discrimination index estimation from psychometric testing theory to characterize the importance of an instance belonging to the region of competence. After this estimation, pruning is done on cases with a lower discrimination index value. This paper also presents the proposal of a new measure to evaluate the quality of a region of competence called q-roc. Results show that this idea is promising for most of the dynamic selection methods analyzed. Through the experiments, there was a significant improvement in the studied dynamic selection methods, indicating that pruning the instances with a lower discrimination index can be a viable alternative in these systems. A case study was conducted that considers a dataset with a high difficulty level and six dynamic selection methods showing in detail the positive points of the proposed method.

Keywords: Multiple Classifier Systems, Dynamic Classifiers Selection, Psichometric Tests.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

$\alpha.k$	O parâmetro α é o multiplicador do tamanho da região de competência: k. Esta é uma maneira de estimar uma região local maior que o tamanho configurado: k.
acc	desempenho de um sistema de classificação (accuracy)
CD	estimativa de diferença crítica (critical difference)
CL	aprendizado baseado em currículo (curriculum learning)
`CTT`	teoria clássica dos testes (Classical Test Theory)
`DCS`	seleção dinâmica de classificador (dynamic classifier selection)
`DES`	seleção dinâmica de grupo de classificadores (dynamic ensemble selection)
`dRoC`	região de competência estimada por discriminação
`DS`	seleção dinâmica (dynamic selection)
`Dsel`	conjunto dados de validação a partir de um dataset (Dynamic Selection datataset)
DT	algorítmo de classificação baseado em árvore de decisão (decision tree)
ID	datasets desbalanceados (imbalanced datasets)
IH	dificuldade de uma instância (instance hardness)
`LI`	limite-inferior (inferior limmit), também visto com a menor acurácia de um sistema `MCS`, dado um `POOL` de classificadores
`IR`	estimativa de desbalanceamento das classes de um dataset (Imbalance Ratio)
IRT	teoria de resposta ao item (Item Response Theory)
`ITA`	Análise do Teste e Item (Item and Test Analysis)
`2-kNN`	outro nome para uma `RoC` formada pelo algoritmo knn de tamanho 2k

2kRoC	região de competência estimada por distância euclidiana contendo $2 * k$ instâncias
k-NN	k-vizinhos mais próxmos (k-nearest neighbors)
kRoC	região de competência estimada por distância euclidiana e $k = 7$.
LR	região local à $\mathtt{x}_q$ (local region)
MCS	sistemas de múltiplos classificadores (Multi-Classifier Systems)
MV	voto majoritário (majority vote)
OP	saída de todos classificadores do POOL (output profile)
Oráculo	sistema DES hipotético que seleciona sempre o classificador correto para predizer $\mathtt{x}_q$, caso este classificador exista.
PG	geração de protótipos (prototype generation)
POOL	conjunto de classificadores de um sistema de seleção dinâmica
PR	reconhecimento de padrões (pattern recognition)
PS	seleção de protótipos (prototype selection)
QR	estimativa de qualidade de todas RoC de um dataset (Quality RoC)
q-roc	estimativa de qualidade de uma RoC (quality of RoC)
RoC	região de competência (Region of Competence)
rRoC	região de competência estimada de maneira randômica
SB	seleção estática do melhor classificador de um POOL de classificadores (single best)
CC	seleção estática com a Combinação de todos Classificadores do POOL (all classifiers)
SS	seleção estática (static selection)
TN,FN	valores: verdadeiro-negativo e falso-negativo (true-negative, false-negative)
TP,FP	valores: verdadeiro-positivo, falso-positivo (true-positive, false-positive)

`U-L`	grupo dos classificadores mais competentes e grupo dos classificadores menos competentes pertencentes à um `POOL` (Upper/Lower)
`WTL`	estimativa de ganho, empate ou perda (win,tie,loss)
$\mathtt{x}_j$	instância que pertence ao conjunto `Dsel`, normalmente que participa da região de competência
$\mathtt{x}_q$	instância desconhecida, representando x_{query}. Outros trabalhos nomeiam a instância desconhecida como x_t ou x_{test} por ser proveniente do conjunto de teste

Sumário

1 Introdução

Classificação é uma tarefa presente em várias atividades humanas. Criar ou aprimorar um método que aumente a acurácia de sistemas de classificação pode contribuir para diferentes áreas científicas e sociais (BRITTO; SABOURIN; OLIVEIRA, 2014), como, por exemplo: o sequenciamento de DNA, o sensoriamento remoto, a biometria, entre outros.

Os sistemas de múltiplos classificadores (MCS) representam uma abordagem de classificação coletiva e configuram-se como uma alternativa aos sistemas de classificação monolíticos, especialmente quando estes não conseguem abranger toda a variabilidade do espaço de características de um determinado problema. Neste contexto, será necessário um método para reunir os votos de distintos classificadores, a fim de alcançar um resultado. Neste sentido, o agrupamento destes diversos votos pode ocorrer de diversas formas.

Um trabalho interessante sobre inteligência das multidões pode ser visto em (SUROWIECKI, 2004). O autor comenta que as decisões do coletivo são, em alguns casos, superiores às feitas por um único indivíduo. De acordo com o autor, a inteligência de uma multidão de indivíduos segue 4 pilares: (1) Diversidade de opinião; (2) Independência da escolha de opinião; (3) Descentralização, onde os indivíduos tiram conclusões a partir de seu conhecimento local; e (4) Agregação, onde a opinião de vários pode ser resumida em uma escolha. Estes pilares se repetem na área de seleção de múltiplos classificadores, como apontado por (ROKACH, 2010).

Estes trabalhos são somente um dos indícios de que os MCS também podem se beneficiar de conceitos e estatísticas já discutidos em outras áreas científicas. Esta pesquisa também engloba a investigação de trabalhos de diferentes áreas, visando assim a incorporação de conhecimentos já estabelecidos em MCS.

Diversas sugestões para a seleção de classificadores estão disponíveis, tanto de forma estática (Static Selection: SS) quanto de forma dinâmica (Dynamic Selection: DS). Em ambas metodologias, é possível selecionar: um único classificador (Dynamic Classifier Selection: DCS) ou um conjunto de classificadores (Dynamic Ensemble Selection: DES). Vários estudos apontam que a seleção dinâmica tem apresentado resultados promissores, conforme mencionado por diversos estudos (BRITTO; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018; BARBOZA et al., 2025), por abordar uma área específica da amostra de teste. Diante disso, esta pesquisa concentra-se na seleção dinâmica de classificadores DS.

Em um sistema de seleção dinâmica, será treinado um conjunto de classificadores (POOL), que poderá ser homogêneo ou heterogêneo. Existem várias maneiras de

construí-los, uma delas é por meio da utilização de diferentes instâncias do conjunto de treinamento. Desta forma, um dos desafios do DES consiste em determinar quais classificadores integrantes do POOL serão mais promissores na classificação de uma instância desconhecida ($\mathtt{x}_q$).

Nesses casos, surge a questão: Como escolher os classificadores mais eficientes para alcançar um desempenho mais parecido com o oráculo? Oráculo (KUNCHEVA, 2002; SOUZA et al., 2024) pode ser definido como um sistema hipotético que seleciona sempre um classificador correto (caso exista) para predizer uma instância desconhecida.

Cada algoritmo de seleção dinâmica possui sua métrica para estimar a competência dos classificadores para então serem selecionados. Esta competência é estimada de várias formas sendo uma das mais conhecidas sendo feita a partir de um conjunto de instâncias do dataset de validação, também chamado de região de competência: RoC.

De acordo com (LIMA; LUDERMIR, 2013; LIMA; SERGIO; LUDERMIR, 2014) é possível melhorar a performance dos métodos de seleção dinâmica trabalhando em melhores formas de extrair as regiões de competência. O trabalho de (SOUZA et al., 2024) comenta alguns métodos para definição da região de competência como por exemplo: vizinhança, agrupamento (clustering), espaço de decisão (decision space), particionamento recursivo (recursive partitioning), hipercaixa difusa (fuzzy hyperboxes), entre outros. Veremos a seguir, no estado da arte, mais detalhes sobre região de competência.

Entretanto, persiste na atualidade a falta de um acordo sobre a forma adequada de estimar uma região de competência, qual deve ser o seu tamanho, ou até mesmo se deveria ser ausente a RoC. As pesquisas mais atuais ainda mostram diversos trabalhos na área, como o caso (ZHANG; ZHU; LUO, 2025) que não são conclusivos em relação à RoC, indicando que ainda precisamos de mais estudos. De acordo com o entendimento presente em várias pesquisas é nítido que a seleção de instâncias para compor uma região de competência influencia o desempenho do método de seleção dinâmica e continua sendo uma das etapas de extrema importância. Mesmo em sistemas que não utilizam a contextualização de uma instância desconhecida ($\mathtt{x}_q$) como no trabalho de (PINTO; SOARES; MENDES-MOREIRA, 2016), os resultados não são conclusivos como também veremos no estado da arte a seguir.

1.1 Desafio

O principal desafio deste trabalho é encontrar uma maneira de estimar a região de competência que considere informações não apenas da vizinhança, mas também do POOL de classificadores. Neste trabalho perseguimos as seguintes dúvidas abaixo:

D_1 : Será que para cada instância desconhecida podemos extrair uma região de competência do dataset de validação que considere simultaneamente: (1) as instâncias mais próximas e também (2) os erros e acertos dos classificadores do `POOL`?

D_2 : Será que uma região de competência, definida em consideração por estes dois itens acima, terão um impacto significativamente positivo em algoritmos de seleção dinâmica de classificadores?

D_3 : Qual seria a maneira de conseguirmos ponderar as instâncias do dataset de validação a partir dos classificadores do `POOL`?

D_4 : Dentre as métricas existentes como poderemos quantificar uma região de competência para verificar quão bom/ruim seria este grupo de instâncias?

D_5 : Será que o `POOL` de classificadores conseguiria nos fornecer alguma estimativa sobre o impacto dos algoritmos de seleção dinâmica?

Foi pensando nestas perguntas que direcionamos as etapas da nossa pesquisa.

1.2 Objetivos

O principal objetivo deste estudo inclui o desenvolvimento de um novo método para escolher a área de competência para métodos de seleção dinâmica de classificadores, além de um novo método para avaliar essa área de competência estabelecida. Para tal, pretende-se empregar conceitos provenientes da teoria de testes psicométricos.

Como objetivos específicos, temos:

- construir um método que consiga extrair a importância de instâncias dentro de uma região de competência que sejam mais promissoras para os métodos de seleção dinâmica, considerando métricas de testes psicométricos;
- comparar de forma estatística o desempenho de diferentes sistemas de seleção dinâmica de classificadores, considerando uma região de competência estimada de maneira convencional com uma região de competência estimada por métricas provenientes das teorias de testes psicométricos;
- construir uma fórmula para que consigamos quantificar uma região de competência e assim podermos mensurar sua qualidade;
- construir uma nova maneira de avaliar o `POOL` de classificadores que nos permita enxergar melhor a sua eficácia ao ser utilizado em algoritmos de seleção dinâmica.

1.3 Proposta

Analisamos diferentes maneiras de medir a importância de uma instância dentro da região de competência que considere não apenas sua vizinhança, mas também o POOL de classificadores de um sistema de seleção dinâmica. Encontramos indícios positivos para esta investigação nos conceitos da **teoria de testes psicométricos**. A teoria de teste psicométricos possui maneiras de avaliar o conhecimento de indivíduos como: competência, habilidade, atitude, entre outras. Nosso interesse com esta pesquisa é a investigação dos cálculos presentes na teoria de testes psicométricos e sua transposição para o reconhecimento de importância das instâncias que irão compor uma região de competência em sistemas de seleção dinâmica.

Acreditamos que estas teorias possam ajudar a melhor caracterizar as instâncias de uma região de competência, e assim os métodos de seleção dinâmica consigam estimar de forma mais assertiva a competência de seus classificadores, trazendo uma melhora significativa no seu desempenho.

Além desta proposta inicial, este trabalho também define uma **nova métrica** capaz de quantificar a qualidade de uma região de competência para que assim consigamos verificar seu comportamento em diferentes métodos de seleção dinâmica. Ainda no sentido de avaliação, entendemos de que nada adiantaria um POOL onde todos classificadores possuem as mesmas predições dada uma instância qualquer. Pensando nisso, também propomos a investigação de uma nova maneira de avaliar de forma qualitativa um POOL de classificadores.

1.4 Hipóteses

Considerando as propostas anteriores, levantamos as seguintes hipóteses:

$\mathbf{H}_1$: A seleção adequada da região de competência com base em dados da teoria de testes psicométricos pode trazer um impacto positivo na acurácia. Esta hipótese possui em sua base a tentativa da escolha de instâncias que possuem maior poder de discriminação por permitir caracterizar uma região de competência através de duas variáveis: vizinhos mais próximos e também o conjunto de classificadores do POOL.

$\mathbf{H}_2$: O processo de escolha de uma região de competência é crucial nos algoritmos de seleção dinâmica. Definir uma métrica adequada que permita quantificar a qualidade de uma região de competência pode trazer uma forma de avaliar os métodos de seleção dinâmica não apenas pelos classificadores escolhidos, mas também pela forma com que a vizinhança é definida.

$\mathbf{H}_3$: Sabemos que o oráculo é considerado o limite-superior dos algoritmos de seleção dinâmica. No entanto ao compararmos estes algoritmos entre eles ou

com o próprio oráculo verifica-se a falta de uma outra medida que estime o limite-inferior através dos classificadores do POOL. Esta hipótese traz a ideia de que a diferença entre o limite-superior e limite-inferior pode ajudar a estimar a margem de impacto de um algoritmo de seleção dinâmica.

1.5 Contribuições

Neste trabalho apresentamos a investigação da importância de cada instância presente na região de competência estimada por uma fórmula da teoria de testes psicométricos. Depois de várias análises verificamos que a seleção de instâncias para uma região de competência estimada pelo índice de discriminação pode ter um impacto significativamente positivo no desempenho de sistemas de seleção dinâmica.

A grande diferença entre este método e os métodos existentes na literatura atual, se baseia em uma métrica de discriminação de uma instância que considera: (i) a vizinhança da instância desconhecida em conjunto com (ii) os classificadores presentes no POOL. Os resultados apresentados são promissores e nos fornecem indícios de que a estimativa da relevância de uma instância dentro da RoC através do índice de discriminação pode trazer benefícios no desempenho de sistemas de múltiplos classificadores.

Além desta contribuição principal, este trabalho também investigou a utilização de duas métricas de avaliação. A primeira métrica referente à avaliação qualitativa de uma região de competência. A segunda métrica, referente à avaliação do POOL de classificadores.

Apresentamos abaixo as contribuições científicas geradas a partir deste trabalho.

Aceito: Dynamic Ensemble Selection by K-Nearest Local Oracles with Discrimination Index (TRIERVEILER et al., 2018);

A ser submetivo: New insights for characterizing the region of competence on multiple classifier systems.

Código DISi: https://github.com/marcelo-trier/DISi

1.6 Organização

O Capítulo 2 mostra os principais conceitos envolvidos nesta pesquisa relacionados com a área de sistemas de múltiplos classificadores e testes psicométricos. O Capítulo 4 mostra a proposta deste trabalho, ou seja, a utilização dos conceitos de teorias de testes psicométricos para criação de um novo método para caracterizar

quais instâncias irão compor a região de competência. O Capítulo 5 mostra o protocolo experimental adotado, os experimentos conduzidos e seus resultados durante esta pesquisa. O Capítulo 6 apresenta a conclusão desta pesquisa e a finalização deste texto com os trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo abordaremos alguns conceitos existentes na área de sistemas de múltiplos classificadores, principalmente aqueles relacionados com os temas de: seleção dinâmica de classificadores e testes psicométricos.

De acordo com C. Bishop (BISHOP, 1995), o principal objetivo do Reconhecimento de Padrões (PR) é a produção de um sistema que faz predições para novos dados desconhecidos, i.e., um sistema que obtenha uma boa generalização.

A classificação é uma das tarefas bastante utilizadas no nosso dia-a-dia e, desta forma, esta tarefa tem recebido atenção da comunidade científica. Criar um método que aumente a acurácia de sistemas de classificação pode contribuir para diferentes áreas científicas e sociais (BRITTO; SABOURIN; OLIVEIRA, 2014) como por exemplo: sequenciamento de DNA, sensoriamento remoto, biometria, entre outros.

Várias são as propostas de indutores para executar a tarefa de classificação, e.g.: Perceptrons de Multi-Camada (MPL) (ROSENBLATT, 1958; RUMELHART; HINTON; WILLIAMS, 1986), Máquinas de Vetores de Suporte (SVM) (VAPNIK, 2013; JAIN; DUIN; MAO, 2000), k-vizinhos mais próximos (k-NN) (COVER; HART, 1967), Árvores de Decisão (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; BREIMAN et al., 1984), entre outros. No entanto o teorema No Free Lunch (WOLPERT, 1996) nos mostra que caso um algoritmo-A supere um algoritmo-B em algum problema, poderão existir outros problemas onde B supere A.

Os sistemas de múltiplos classificadores tiveram uma contribuição fundamental de Shapire (SCHAPIRE, 1990) que provou ser possível a construção de um método de classificação robusto a partir de um conjunto de classificadores *fracos* e *instáveis*. Classificadores fracos possuem uma baixa taxa de acerto. Modelos instáveis são aqueles que, com uma pequena alteração no conjunto de treinamento, terão uma alteração significativa no modelo treinado (SKURICHINA; DUIN, 2002). De acordo com (BREIMAN, 1996), entre as técnicas de amostragem do dataset para geração de modelos de classificação, o Bagging teve um melhor resultado para classificadores fracos.

Veremos neste capítulo o estudo sobre seleção de múltiplos classificadores e posteriormente a teoria de testes psicométricos e como esta teoria pode ajudar os métodos de seleção dinâmica.

2.1 Sistemas de Múltiplos Classificadores

Os MCS adotam um POOL, construído pelo treinamento de vários classificadores $\mathbf{c_i}$ a partir de amostras do conjunto de treino. Depois, para cada instância desconhecida

$\mathtt{x}_q$ a ser predita é feita uma combinação dos votos de cada classificador para então chegar a um resultado: $\hat{\mathtt{y}}$.

De acordo com (BRITTO; SABOURIN; OLIVEIRA, 2014), a criação de um indutor único (monolítico) para resolução de vários problemas, que tenha uma alta performance, pode ser impraticável e este foi um dos motivos para que pesquisadores investigassem melhor esta área. Os MCS podem acontecer em 3 fases: (1) Geração, (2) Seleção e (3) Integração, como mostrado na Figura 1.

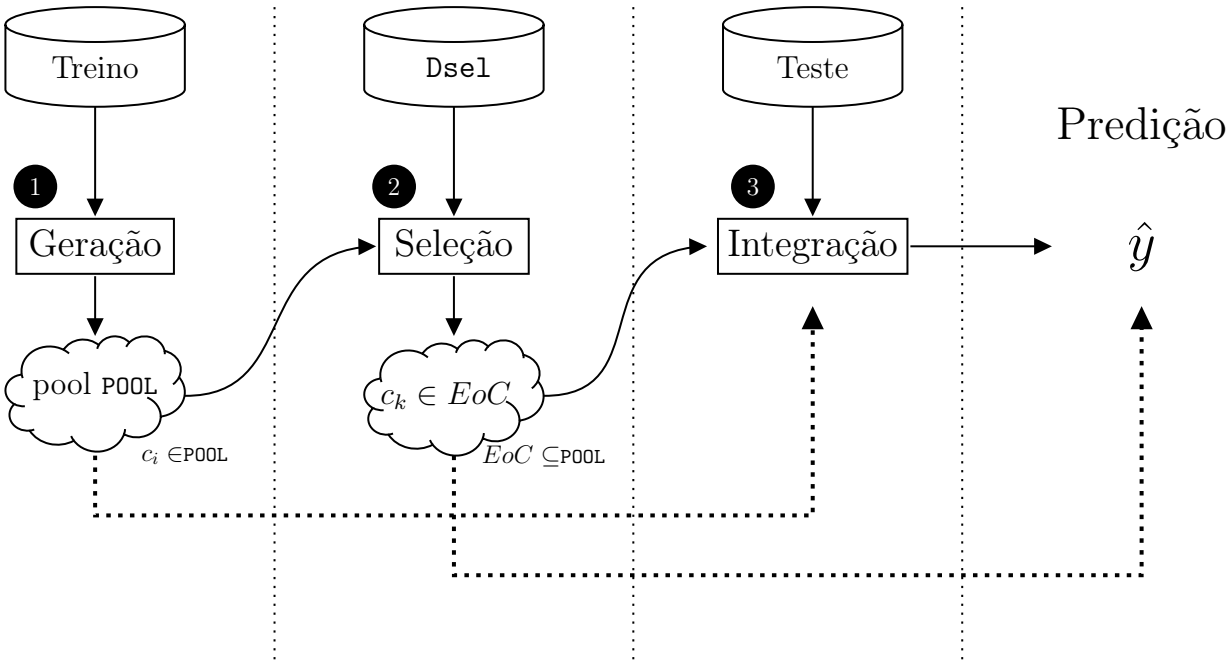


Figura 1 – As 3 Fases de um Sistema de Múltiplos Classificadores: (1) Geração, (2) Seleção, (3) Combinação. Adaptado de (BRITTO; SABOURIN; OLIVEIRA, 2014)

A primeira fase está relacionada com a construção de um POOL de classificadores. Na segunda etapa será selecionado um ou vários classificadores competentes para rotularem uma instância desconhecida ($\mathtt{x}_q$). A terceira etapa é responsável pela combinação dos votos de cada classificador, que será agrupado de alguma forma e uma predição final ($\hat{\mathtt{y}}$) será considerada como resultado. Alguns MCS não possuem a etapa de integração como o caso da seleção de classificador único, como será visto a seguir.

2.1.1 Geração

Espera-se que a geração de um POOL de classificadores seja feita de modo a permitir que eles sejam diversos entre si, pois de nada adiantaria vários classificadores que

fizessem as mesmas predições para cada $\mathtt{x}_q$ (CRUZ; SABOURIN; CAVALCANTI, 2018). Conforme comentado por (KUMAR; KUMAR, 2012; BRITTO; SABOURIN; OLIVEIRA, 2014) a diversidade pode ser explícita ou implícita.

A **diversidade implícita** está relacionada com a geração do classificador, ou seja, da indução de algum algoritmo por um determinado conjunto de dados. Esta diversidade implícita pode ser medida através de cálculos feitos no espaço de características das instâncias e pode ser estimada antes ou durante o treinamento dos classificadores através da variação de alguma opção. Muitos trabalhos como por exemplo: (WOŹNIAK; GRAÑA; CORCHADO, 2014), comentam a geração do `POOL` de forma homogênea ou heterogênea. Algumas técnicas consideram (re)amostragem de instâncias como por exemplo: Bagging (BREIMAN, 1996; EFRON; TIBSHIRANI, 1994) e Boosting, e as técnicas bastante citadas que fazem (re)amostragem no espaço de características são o Random Subspaces (HO, 1998) e Random Projection (FERN; BRODLEY, 2003; DANG et al., 2018). Outra característica explorada por (SOUZA et al., 2019) é a abordagem dinâmica para geração do `POOL` de classificadores. Nesta abordagem os classificadores são criados na etapa de generalização ao invés de uma forma estática na etapa inicial do sistema.

A **diversidade explícita** está relacionada com a medição de classificadores já treinados, ou seja, se eles são diversos entre si e por isso também podem receber o nome de medidas de diversidade (KUNCHEVA; WHITAKER, 2003). Esta medida é obtida no espaço de decisão (*decision space*) (saída) dos classificadores e calculada após os classificadores treinados. As equações de diversidade consideram classificadores aos pares ou em conjunto. Para as medidas de diversidade em pares é verificado quando classificadores acertam ou erram juntos uma determinada sequência de instâncias, como por exemplo: estatística-q, coeficiente de correlação, medida de desacordo, falha-dupla, entre outras. As medidas de diversidade para conjuntos de classificadores calculam uma média dos acertos em conjunto, e como exemplo temos: entropia, variância de Kohavi-Wolpert, concordância entre-classificadores, dificuldade, entre outras (KUNCHEVA; WHITAKER, 2003).

2.1.2 Seleção

A seleção de classificadores pode estar inserida na etapa de treinamento ou generalização, como vemos nas Figuras 2, 3 e 4. Quando a seleção é feita na etapa de treinamento, Figura 2, recebe o nome de seleção estática (Static Selection `SS`) e quando a seleção acontece na fase de generalização recebe o nome de seleção dinâmica (`DS`). Comentaremos brevemente a seleção estática para dar ênfase em seleção dinâmica.

A **seleção estática** recebe este nome pois é feita uma única vez na fase de treinamento e sempre o mesmo classificador (ou conjunto de classificadores) será

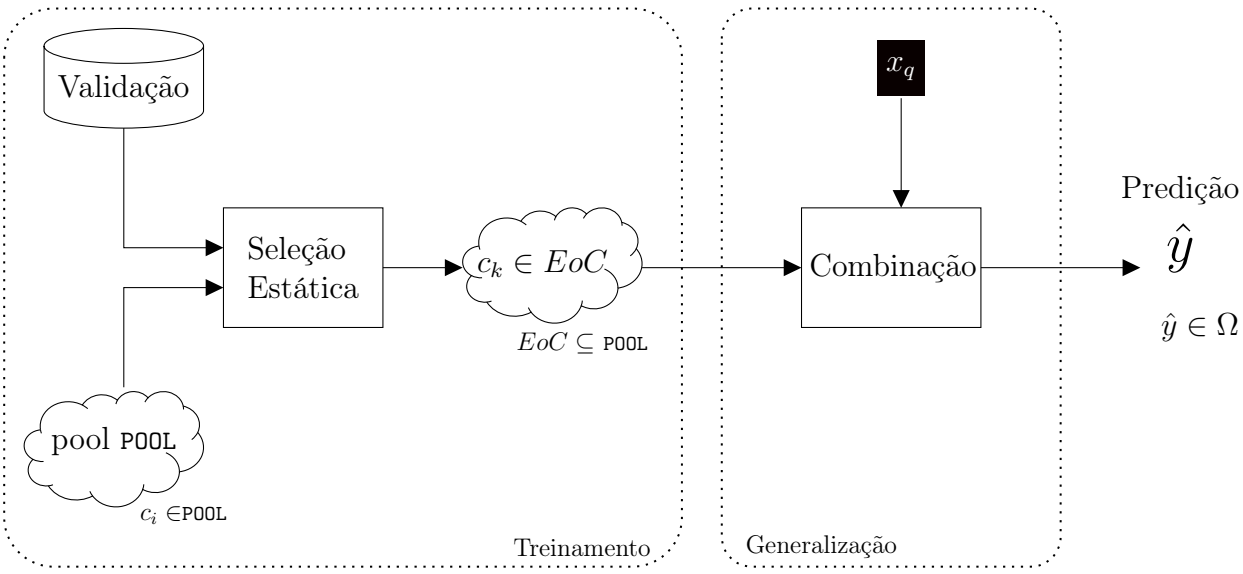


Figura 2 – Seleção estática de classificadores

utilizado para rotular as instâncias desconhecidas. Este tipo de sistema tem um gasto computacional alto na fase de treinamento, no entanto a fase de generalização existe uma tendência de ser menos custosa. Em alguns sistemas verifica-se que a combinação de votos, caso exista, tendem a ser mais sofisticados. De acordo com (CRUZ; SABOURIN; CAVALCANTI, 2018) o critério mais usado para selecionar um ensemble estático é diversidade e acurácia.

Considerando a **seleção dinâmica**, existe a possibilidade de manipular melhor as nuâncias da região local de cada instância desconhecida, i.e., sabendo que cada classificador $\mathbf{c_i}$ é um especialista em uma região do espaço de características, somente serão escolhidos aqueles que possuem competência para fazê-lo. Diferentemente da seleção estática, a seleção dinâmica possui esse nome porque na fase de generalização, cada instância desconhecida $\mathtt{x}_q$ possuirá um ou vários classificadores escolhidos para rotulá-la, conforme as Figuras 3 e 4.

Na seleção dinâmica de classificador único DCS (Dynamic Classifier Selection) será sempre escolhido um classificador mais competente enquanto na seleção dinâmica de ensembles DES (Dynamic Ensemble Selection) serão escolhidos um conjunto deles.

Os métodos de seleção dinâmica DCS e DES compartilham as etapas: (1) definição de competência e (2) escolha de classificadores. A primeira etapa pode ser subdivida em: (a) definição de uma fonte de informação (em alguns casos chamado de região de competência) e (b) cálculo de competência. Já a segunda etapa capta os classificadores mais aptos para estimar uma distância desconhecida.

As pesquisas em seleção dinâmica tem demonstrado superioridade em relação

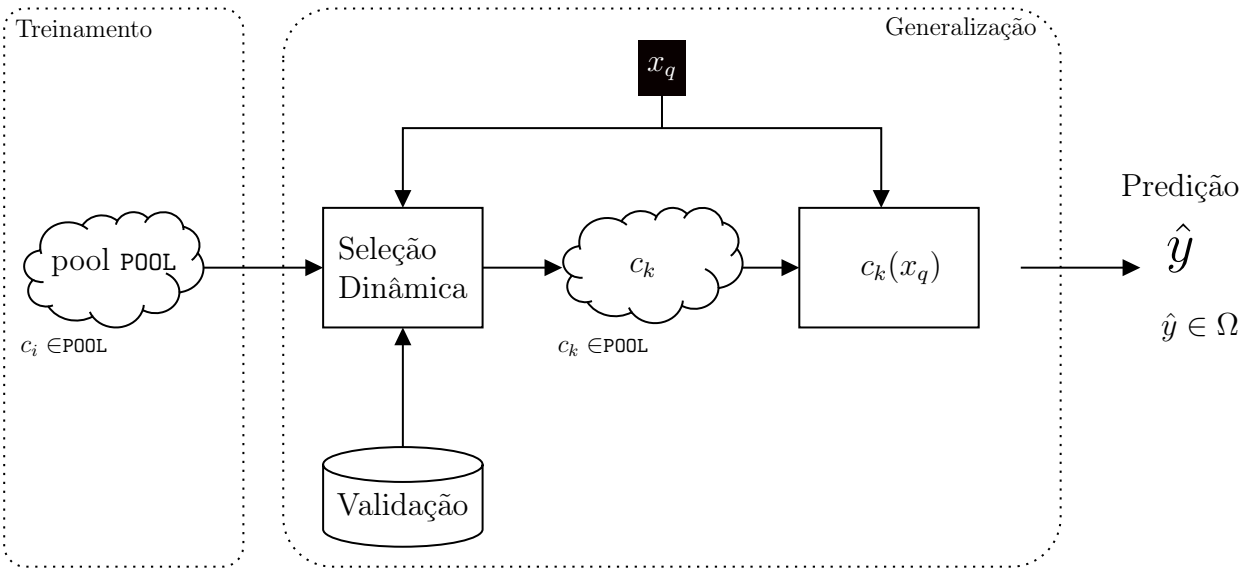


Figura 3 – Seleção dinâmica de classificadores (DCS)

à seleção estática. Em primeiro lugar, por realizar um esforço extra na escolha do(s) melhor(es) classificador(es) para cada nova instância $\mathtt{x}_q$. Em segundo lugar por fazer uma classificação considerando instâncias locais próximas à $\mathtt{x}_q$ ao invés de considerar todo conjunto de validação, conforme já relatado pelo trabalho de (CRUZ; SABOURIN; CAVALCANTI, 2018) e reforçado por (ZHANG; ZHU; LUO, 2025). Vemos também em outras pesquisas, como por exemplo (PINTO; SOARES; MENDES-MOREIRA, 2016; NARASSIGUIN; ELGHAZEL; AUSSEM, 2017) que existe a possibilidade da escolha de melhores classificadores sem a necessidade da região de competência.

Resultados mostram que a seleção dinâmica de ensembles tem obtido resultados interessantes em datasets pequenos (small dataset) e/ou desbalanceados (imbalanced ratio) (FERNÁNDEZ et al., 2008; CRUZ; SABOURIN; CAVALCANTI, 2018; CRUZ; SABOURIN; CAVALCANTI, 2017b; ROY et al., 2018). Ainda assim, vemos que as últimas pesquisas têm demonstrado resultados também superiores em datasets maiores como no caso dos trabalhos de (SOUZA et al., 2024; DAVTALAB; CRUZ; SABOURIN, 2023).

Selecionar um conjunto de classificadores, ao invés de apenas um, surgiu da dificuldade em encontrar a melhor medida de competência para os classificadores do POOL que serão selecionados. A ideia é a tentativa em diluir o risco de selecionar apenas um classificador conforme tem sido relatado pelo andamento das pesquisas (BARBOZA et al., 2025; ZHANG; ZHU; LUO, 2025; SOUZA et al., 2024).

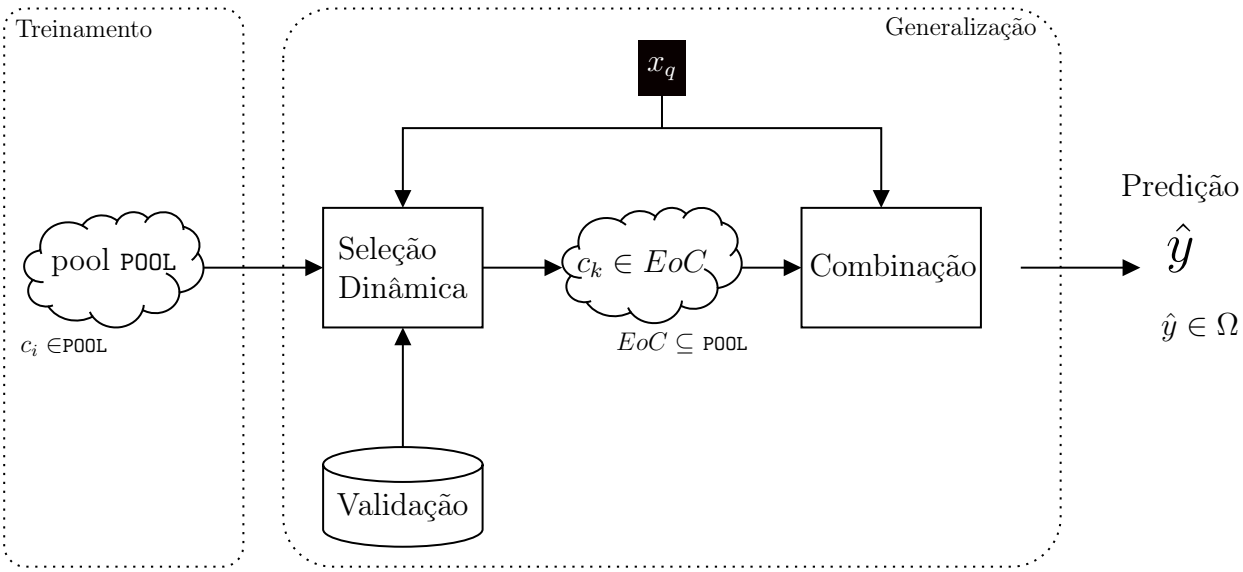


Figura 4 – Seleção dinâmica de ensembles (DES)

2.1.3 Combinação

Esta terceira parte de um sistema MCS tem a tarefa de combinar as predições de um conjunto de classificadores através de uma função. Caso o sistema MCS não possua um conjunto de classificadores para predizer a instância desconhecida, esta etapa torna-se desnecessária.

Podemos citar o voto majoritário como um dos exemplos bastante utilizados nesta área. A combinação pode ser feita de diferentes formas como comentado por (CRUZ; SABOURIN; CAVALCANTI, 2018), através de métodos: (1) treinados, (2) não-treinados e (3) dinamicamente ponderado.

Entre métodos (2) não-treinados estão incluídos o voto majoritário, soma, produto, máximo, mínimo, média, entre outros mostrados no trabalho de (KITTLER et al., 1998) e bastante utilizados até a data atual. Estes métodos são formados por uma equação que é estimada de forma estática para todos os problemas existentes.

Para tentar contornar as limitações dos métodos-estáticos, foram propostas as combinações treinadas (1). Neste caso, em cada problema é treinado um classificador que fará a combinação dos votos. O trabalho de (DUIN, 2002) discute estes métodos e inclusive comenta que apesar de existir uma tendência desta abordagem ser melhor, existe uma desvantagem: acrescentar problemas que antes não existiam como por exemplo um treinamento enviesado.

Ainda com estas desvantagens existem trabalhos que discutem que os métodos treinados possuem melhores desempenhos, como é o caso da mistura de especialistas (mixture of experts) (MASOUDNIA; EBRAHIMPOUR, 2014).

Considerando métodos (3) dinamicamente ponderados, existe uma agregação de votos de todos classificadores que foram selecionados para predição. Através da resposta de todos estes classificadores é então escolhido uma resposta final, ponderada pela competência local de cada classificador (CRUZ; SABOURIN; CAVALCANTI, 2018).

2.2 Teoria de Testes Psicométricos

O estudo científico de questionários aplicados à descoberta de informações sobre as pessoas tem uma longa história e é conhecido por Teste Psicométrico. De acordo com a (GAULT, 1907), os primeiros relatórios para descobrir informações sobre as pessoas datam da década de 1830. A informação é estimada através da coleta de respostas de um questionário aplicado à um grupo de pessoas.

Um questionário ou uma prova é composto por um conjunto de itens ou perguntas. O tipo mais comum de item é o chamado dicotômico, ou seja que tenha apenas dois tipos de respostas: certo ou errado. A psicometria é o estudo estatístico realizado através da aplicação de um questionário a um grupo de pessoas. Uma análise mais profunda deste questionário pode ser feita a partir das respostas dos indivíduos (respondentes) onde é extraído alguns valores de grande importância como é o caso da dificuldade e discriminação. Este processo nos ajuda a capturar o que chamamos de traço-latente (latent trait) de um indivíduo.

Os traços latentes podem ser entendidos como um comportamento pessoal, oculto ou não, que pode ser difícil de identificar externamente, mas que pode ser estimado a partir das respostas de um questionário. O traço latente pode representar: a aptidão para uma tarefa, a personalidade, o interesse, etc. É aplicado nos domínios da psicologia e também da educação. A dificuldade mede a probabilidade de êxito de cada item. A discriminação é uma medida que mostra até que ponto um item consegue separar os indivíduos que tenham conhecimento mais elevado dos indivíduos com conhecimento mais baixo.

Psicólogos, professores, estatísticos e matemáticos vêm tentado compreender os possíveis parâmetros das pessoas e dos itens através da análise das respostas a um questionário. Duas teorias importantes são amplamente discutidas: Teoria Clássica dos Testes (Classical Test Theory: `CTT`) (GULLIKSEN, 2013) também conhecida como Teoria da Pontuação Verdadeira; e Teoria da Resposta ao Item (Item Response Theory: IRT) (TUCKER, 1946; LORD, 1952).

É perfeitamente possível relacionar esta teoria com o reconhecimento de padrões. Para esta proposta iremos considerar os respondentes como classificadores de um `POOL` e assumiremos que as perguntas de um teste são instâncias em num dataset. Seguindo esta analogia, estamos interessados nas medidas que podem ser extraídas

a partir destes dois conjuntos: instâncias e classificadores. Consideraremos o estudo do Índice de Discriminação: Ð.

2.2.1 Como extrair medidas a partir de respostas

Iniciamos esta parte do texto com as palavras de (STONE, 1999), que em seu trecho considera difícil conseguir medir o conhecimento de uma pessoa e que mesmo não existindo fórmula para verificar o traço latente com perfeição, podemos nos ater em uma medida que seja próximo suficiente.

> *When we make measures, we do so with the intention of being accurate enough for our practical purpose. We do not expect absolute precision. Our notion of 'accurate' does not imply 'perfect'. Instead it implies 'close enough to be useful'. (STONE, 1999)*

De acordo com (WRIGHT BENJAMIN DRAKE; STONE, 1979), a teoria dos testes psicométricos deve prever que transformemos as respostas corretas de uma prova na habilidade mental de um indivíduo. Ou seja, é preciso que seja aplicado uma Prova para um indivíduo e assim conseguir estimar suas habilidades mentais sob determinado assunto.

Isso se torna necessário porque não se consegue enxergar o conhecimento dentro do cérebro de alguém. Pelo menos não até o momento. É através das respostas do indivíduo à um teste mental que conseguimos fazer medições de seu conhecimento sobre um assunto. Se pudéssemos visualizar um determinado conhecimento como uma linha, então poderíamos desenhar a figura 5 contendo o conhecimento de uma pessoa sobre aquele assunto.

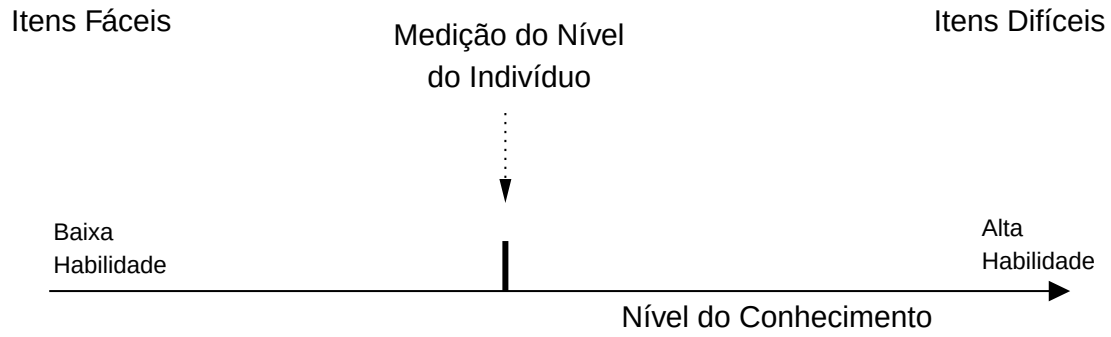


Figura 5 – Rasch Model Measures, adaptado de: (WRIGHT BENJAMIN DRAKE; STONE, 1979).

Na Figura 5 vemos que um teste mental pode ajudar a medir a posição do conhecimento de uma pessoa dentro de uma medida quantitativa, apenas na observação

de suas respostas. Nesta figura vemos uma régua que vai da esquerda (indicando baixa habilidade em um determinado assunto) para direita (alta habilidade). Ainda na figura, na parte de cima, vemos que nesta régua temos também a medição das questões de uma prova indo de perguntas fáceis à esquerda até difíceis à direita.

Nesta figura 5 também vemos que a medida apontada pela seta é a respresentação do conhecimento de alguém sobre o assunto em questão. Se pudermos avaliar várias respostas de um indivíduo à uma prova, podemos atribuir um nível de habilidade naquele assunto, como mostra a Figura 6. Consideramos que as perguntas foram posicionadas na figura de forma ordenada, do mais fácil ao mais difícil.

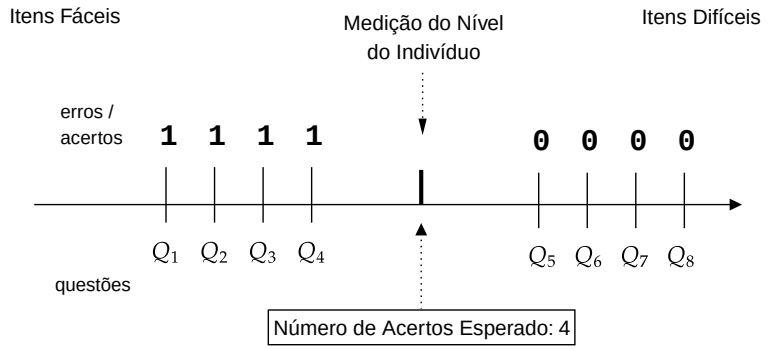


Figura 6 – Rasch Model Measures, adaptado de: (WRIGHT BENJAMIN DRAKE; STONE, 1979).

Nesta figura 6 vemos um teste com 8 questões. Neste exemplo, o conhecimento de um indivíduo foi medido verificando seus acertos, sendo que ele acertou as 4 primeiras perguntas. Desta forma podemos atribuir um score à pessoa e dizer que ela tem um determinado nível de habilidade para o conhecimento medido.

2.2.2 Viés da dificuldade

Existem alguns problemas com essa régua mostrada na Figura 6 para calcular a habilidade de uma pessoa. Supondo que quiséssemos medir o conhecimento de uma pessoa e que tivéssemos 5 provas para serem aplicadas, cada uma com 8 questões diferentes. Supondo que o nível de dificuldade destas questões também são diferentes, podemos ter um problema na medição conforme mostrado na Figura 7.

Nesta Figura 7 temos uma representação das respostas do mesmo indivíduo para os 5 testes em questão. As notas desta pessoa para cada teste foram respectivamente: 8, 0, 2, 6 e 4. Vemos nesta imagem que existem notas que são opostas como por exemplo o teste-1 e teste-2. Vemos claramente nesta imagem que o conhecimento absoluto do indivíduo não alterou em nenhum dos testes, mas foi alterado o nível de dificuldade das questões da prova.

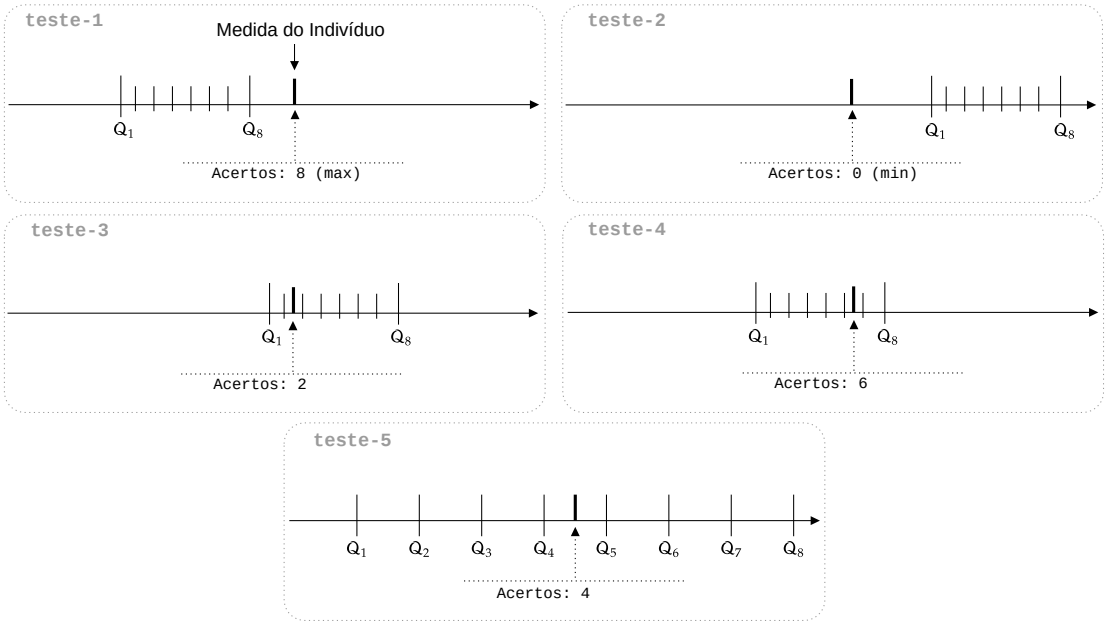


Figura 7 – Rasch Model Measures, adaptado de: (WRIGHT BENJAMIN DRAKE; STONE, 1979).

Neste exemplo podemos ver que a escolha dos ítens de uma prova pode tranquilamente influenciar positivamente ou negativamente na nota de um indivíduo, além de não conseguir medir corretamente suas habilidades mentais. Ainda nesta figura 7, vemos que ao final, no teste-5 vemos as perguntas igualmente espaçadas e distribuídas ao longo de todo o espectro da régua de dificuldade. Este teste-5 seria o único teste (dentre os demais) que conseguiria medir o conhecimento da pessoa.

Capturando o experimento acima onde vemos uma grande dispersão do nível de dificuldade das questões para medir o conhecimento de uma pessoa, gostaríamos de trazer também um cenário de cinco testes (com oito questões cada) sendo aplicadas simultaneamente para duas pessoas **P1** e **P2**, conforme Figura 8.

Podemos verificar já no teste-1 o problema onde temos que a pessoa1 (P1) não conseguiu acertar nenhuma questão e a pessoa2 (P2) acertou todas questões. No entanto podemos notar que a dificuldade das questões estão todas muito próximas não nos permitindo avaliar corretamente P1 nem P2.

No segundo, terceiro e quarto testes da figura 8 vemos que também não é possível confiar no score que foi atribuído às duas pessoas porque o nível de dificuldade das questões estão bastante enviesados. Este viés faz com que a diferença entre os scores trouxesse valor zero na diferença entre estas pessoas, o que não é verdade.

Avaliando também o teste-5 da figura 8, vemos que o nível de dificuldade entre

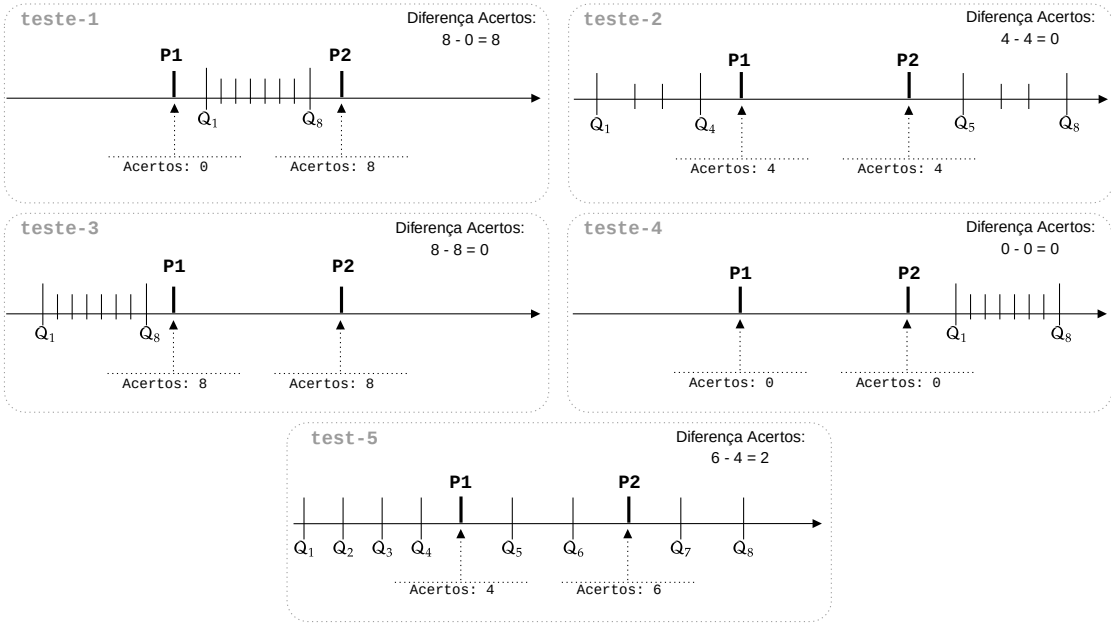


Figura 8 – Rasch Model Measures, adaptado de: (WRIGHT BENJAMIN DRAKE; STONE, 1979).

as perguntas está igualmente espaçado e sim, este seria um modelo correto para aplicar um teste de oito questões para um ou vários indivíduos.

Cabe aqui uma pequena análise que iremos comentar melhor mais à frente. Estamos vendo nestes exemplos que a atribuição de um ranking para um grupo de pessoas (que no nosso caso é um um grupo de classificadores) pode muito bem estar enviesado de acordo com a distribuição da dificuldade dos ítens. Nestes casos, se não sabemos ao certo qual a distribuição de dificuldade dos ítens em um teste como poderemos ter certeza se nosso ranking está correto? A resposta para esta e outras perguntas é investigado em nossa proposta Cap. 4.

2.2.3 Medidas de Psicometria: dificuldade, habilidade e discriminação

Considere as perguntas abaixo sobre duas pessoas que receberam o prêmio Nobel: qual das perguntas é mais difícil?

- Quem foi Madre Tereza de Calcutá?
- Quem foi Harry Markowitz?

A segunda pessoa é um pesquisador de ciências econômicas que criou a teoria de seleção de portifólio para diminuir o risco dos investidores e aumentar o retorno. A primeira pessoa não precisa ser comentada.

Colocamos as perguntas acima para identificar a dificuldade em alguma área do conhecimento. Saber ou não uma resposta não torna a pergunta fácil ou difícil. O fato de não sabermos uma questão faz com que seja difícil para nós (ou seja, algo bastante subjetivo), mas não para a questão propriamente dita.

Para a teoria de testes psicométricos, a dificuldade de uma questão está relacionada com um conjunto de indivíduos que conseguem respondê-la. Este conjunto pode ser formado por especialistas na área, por estudantes em uma sala de aula ou toda a população humana em um dado período da história. Um exemplo para isso é a teoria de abiogênese, onde a mais de um século se acreditava que era possível geração espontânea da vida a partir de outras matérias, que foram refutadas por Louis Pasteur (PASTEUR, 1860).

O fato de não sabermos a resposta de uma questão não a torna difícil. O que a torna difícil é se existe uma grande quantidade de pessoas que não sabem a resposta. Esta é uma das bases de testes psicométricos que calcula entre outras medidas a habilidade de pessoas e a dificuldade de questões.

Um ponto bastante interessante desta teoria é a caracterização da dificuldade das questões estarem intrinsecamente relacionadas com a habilidade das pessoas que a respondem. Este entendimento é de fundamental importância, pois nos próximos capítulos desta tese destacamos que: a dificuldade de instâncias em um dataset pode ser medida pela quantidade de classificadores que conseguem prevê-la, e não por uma métrica que considere somente o espaço de características.

Nos testes psicométricos temos algumas teorias reservadas para análise dos itens, utilizada para extrair medidas dos testes mentais. Por exemplo, os testes educacionais ou também conhecidos como: provas, são um exemplo de como pode ser feita a mensuração do conhecimento de estudantes na sala de aula.

Neste trabalho estamos considerando testes dicotômicos dentro da Análise dos Itens e não estamos considerando testes descritivos. As respostas dos testes descritivos, por exemplo: uma redação, possui uma quantidade de notas muito grande, podendo variar entre $\{0$ à $10\}$, além de ser também um pouco subjetivo. Por este motivo os testes descritivos não se encaixam neste trabalho.

Por sua vez, os testes dicotômicos se caracterizam por possuírem questões com somente uma resposta correta e uma ou várias respostas erradas, de forma que o indivíduo/aluno: acerta ou erra uma questão. Como exemplo de testes dicotômicos, temos questões do tipo verdadeiro/falso ou questões de múltipla escolha onde somente um dos itens estará correto.

2.2.4 Dificuldade de questões (b_j) e Habilidade de indivíduos (a_i)

Suponha uma prova respondida por N estudantes contendo M questões. A resposta (R) de cada i-ésimo estudante para cada j-ésima questão (R_{ij}) é dada pela Eq. 2.1, sendo $i = \{1..N\}, j = \{1..M\}$.

$$R_{ij} = \begin{cases} 1, & \text{caso acerto} \\ 0, & \text{caso errado} \end{cases} \tag{2.1}$$

De acordo com os pesquisadores (CROCKER; ALGINA, 1986), a dificuldade é a proporção de pessoas em uma população ou um grupo que consegue reponder corretamente uma pergunta. Assumindo R_{ij} como a matriz das respostas de i-ésimas pessoas e j-ésimas questões, temos que a difuldade pode ser estimada de acordo com a Eq. 2.2. Assim, a dificuldade de uma questão b_j pode ser representada pela proporção de estudantes que a acertaram.

$$b_j = \frac{\sum_j^M R_{ij}}{M} \tag{2.2}$$

A dificuldade de uma questão b_j varia entre o intervalo $[0, 1]$ e mostra que quanto mais próximo de 1 mais fácil foi a questão para aquele conjunto de pessoas. A habilidade de um indivíduo a_i pode ser representada pela Eq. 2.3 e varia entre o intervalo $[0, 1]$ e mostra que quanto maior o valor, melhor é a competência do estudante.

$$a_i = \frac{\sum_i^N R_{ij}}{N} \tag{2.3}$$

Trazemos um exemplo de estimativa, onde mostramos um exemplo de erros e acertos de dez estudantes em uma prova com sete questões, conforme mostrado na Tabela 1. Usaremos esta tabela para os exemplos e cálculos a seguir.

A partir do exemplo na Tabela 1 verifica-se que o estudante E7 é o mais competente para esta prova, enquanto o E4 é o menos competente. Também é possível fazer a mesma analogia às questões, onde a questão Q4 é a mais fácil e a questão Q6, mais difícil. Claro que a dificuldade de uma questão pode ser um conceito subjetivo, pois existirá diferença entre uma questão respondida por um especialista na área ou um estudante do ensino médio. Mas para nosso exemplo Tabela 1 consideramos uma sala com indivíduos de mesma idade.

2.2.5 Índice de Discriminação: Ð

A Análise do Teste e Item - ITA (Item and Test Analysis) - é feita para se ter uma visão mais clara de algumas variáveis, como: (1) do teste propriamente dito; (2) dos estudantes; e (3) das questões. A ITA está presente tanto na Teoria Clássica dos Testes, CTT (Classical Test Theory) como também em outras teorias, por exemplo: Item Response Theory (IRT). A Análise de Item prevê um índice de discriminação para cada questão, ou seja, uma medida para calcular o quanto uma

Tabela 1 – Erros e acertos dos estudantes (E) para cada uma das questões (Q). A última coluna mostra a soma das questões que cada estudante acertou e a última linha mostra o número de estudantes que acertou a questão.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	$\sum$
E1	0	0	1	1	1	0	0	3
E2	0	1	1	1	0	0	1	4
E3	1	0	0	1	1	0	1	4
E4	0	0	0	0	0	0	0	0
E5	0	0	0	1	0	0	0	1
E6	1	0	1	1	0	0	1	4
E7	1	1	1	1	1	1	1	7
E8	0	1	1	0	1	0	0	3
E9	0	0	0	1	1	0	0	2
E10	1	0	0	1	0	1	1	4
$\sum$	4	3	5	8	5	2	5	32

pergunta consegue separar os estudantes em dois grupos: os mais competentes e os menos competentes.

2.2.6 Porquê o valor 27% no Ð?

Muitas medidas de discriminação foram propostas ao longo dos anos e atualmente estas medidas ainda são discutidas como vemos no trabalho de (METSÄMUURONEN, 2019). Uma das teorias que o CTT incorporou em seu framework é conhecido como abordagem de grupos extremos, EGA (Extreme Groups Approach) ou também design de grupos extremos, EGD (Extreme Groups Design). De acordo com (PREACHER et al., 2005), a pesquisa mais antiga que utiliza a abordagem EGA foi feita em 1899 para medir a inteligência de crianças (BINET, 1899). Nesta pesquisa, um grupo de 32 crianças foram classificadas em dois grupos extremos: os mais inteligentes e os menos inteligentes e o estudo de (PREACHER et al., 2005) ainda comenta que EGA é um tipo de sobreamostragem (oversampling) bastante utilizada na estatística.

Nesta estratégia, os grupos extremos são nomeados por grupo superior e grupo inferior, ou então, grupos UL (Upper and Lower). A pesquisa de (FELDT, 1961) provou através de uma compração t-test que as médias dos grupos-UL é maximizada pelos grupos que correspondem a 27% de cada extremidade. A pesquisa de (FOWLER, 1992) também investigou estas medidas e concluiu que dentro de seus experimentos os grupos extremos de 27% possuem uma diferença estatística maior que a divisão em dois grupos de 50%.

O trabalho de (FELDT, 1961) foi complementar ao trabalho de (KELLEY, 1939), que capturou a ideia dos grupos UL para formular uma das medidas de discriminação mais conhecidas da teoria clássica dos testes: Índice de Discriminação. A pesquisa de (KELLEY, 1939) foi a primeira a introduzir o número de **27%** para estimar o índice de discriminação nos testes de pessoas. Após este trabalho muitas discussões surgiram nos anos seguintes e vários experimentos foram realizados estimando os grupos UL em: 50%, 33%, 25%, 20%, 16%, 12% e 10% que significa dividir o grupo de pessoas por respectivamente: {2, 3, 4, 5, 6, 8 e 10}. Ainda considerando outros trabalhos de exploração de diferentes proporções vemos que a pesquisa de (KELLEY, 1939) foi uma das poucas que matematicamente provou o poder da separação dos grupos UL em 27%.

A pesquisa de (WIERSMA; JURS, 1985) mostra que o número de 27% maximiza a diferença em uma distribuição normal enquanto provê dados suficientes para análise. A pesquisa de (MATLOCK-HETZEL, 1997) mostra que estes 27% são necessário pois precisa-se de alunos suficientes em cada grupo para prover estabilidade estatística enquanto que é preciso existir uma diferença de características entre os integrantes de cada grupo, para que a discriminação seja mais limpa.

Ao longo desta história de mais de um século, outros pesquisadores também confirmaram o poder da abordagem dos grupos UL em distribuições não-paramétricas. Em nossa revisão bibliográfica sobre o tema de EGA verificamos vários estatísticos que pesquisaram sobre os grupos UL, entre eles gostaríamos de citar como sendo de fundamental importância para esta área os trabalhos de: (FORLANO; PINTER, 1941; DOPPELT; POTTS, 1949; JOHNSON, 1951; FELDT, 1961; FELDT, 1963; BRIDGMAN, 1964; BRENNAN, 1972; BORICH; GODBOUT, 1974; ALF; ABRAHAMS, 1975; ABRAHAMS; ALF, 1978; GARG, 1983; FOWLER, 1992; PREACHER et al., 2005; PREACHER, 2014). Além destes trabalhos que estão relacionados à estatística, também podemos citar alguns trabalhos que contribuíram para a área de psicometria em suas análises de medidas de discriminação: (FINDLEY, 1956; ENGELHART, 1965; CURETON, 1966; ALEAMONI; SPENCER, 1969; WHITNEY; SABERS, 1971; OOSTERHOF, 1976; BURTON, 2001; LEI; DUNBAR; KOLEN, 2004; LIU, 2008). Baseado nesta revisão, iremos considerar o índice de discriminação através de seu cálculo através dos grupos UL com 27%.

2.2.7 Índice de Discriminação na Psicometria

A discriminação é uma medida que nos informa, dentro de uma prova, quanto uma questão consegue diferenciar efetivamente os estudantes com alto conhecimento daqueles outros estudantes com baixo conhecimento. Isso é importante porque para uma prova, o professor precisa saber qual foi a questão que mais consegue separar os alunos para então inferir alguma análise, como por exemplo um plano de

recuperação somente para um grupo de alunos. O cálculo de discriminante dentro da *Item and Test Analysis* do CTT é feito em 4 etapas:

1. ranquear os alunos pela sua habilidade calculado na Eq. 2.3 considerando as questões envolvidas,
2. separar uma parcela de estudantes em dois grupos: aqueles com alto conhecimento e os outros com baixo conhecimento – chamaremos de **grupo-competentes** e **grupo-não-competentes** respectamente,
3. para cada questão Q_j, computar o número de acertos que cada grupo obteve, sendo: $\mathcal{A}$ os acertos do grupo-competentes e $\bar{\mathcal{A}}$ os acertos do grupo-não-competentes,
4. calcular a diferença da quantidade de acertos dos dois grupos e dividir pelo tamanho (L). Este é o cálculo do índice de discriminação, como mostrado na Eq. 2.4.

$$\text{Đ} = \frac{\mathcal{A} - \bar{\mathcal{A}}}{L} \tag{2.4}$$

Na segunda etapa descrita anteriormente é previsto que seja feito dois grupos, sendo que o grupo-competentes será composto por $p\%$ dos alunos melhores ranqueados, e o grupo-não-competentes será composto por $p\%$ dos piores ranqueados. Na Análise do Item do CTT um dos valores mais adotados é $p = 27\%$.

O valor do índice de discriminação Đ varia entre [-1 e 1], sendo que quanto mais perto de 1, mais se consegue separar os estudantes com alto conhecimento daqueles com baixo conhecimento. Os valores negativos indicariam que os alunos com baixo conhecimento em uma prova acertaram uma questão em maior quantidade que os estudantes mais competentes. Como isso não deveria acontecer, normalmente em situações que o valor de Đ for negativo, a questão deve ser revista ou anulada.

Coninuando o exemplo da Tabela 1, ao separar os estudantes com maior habilidade daqueles com menor habilidade, teríamos os dois grupos: competentes e não-competentes conforme Eq. 2.5 mostrado abaixo:

$$\begin{aligned} \text{Competentes} &= \{E7, E2, E3, E6, E10\} \\ \tilde{\text{N}}\text{Competentes} &= \{E1, E8, E9, E5, E4\} \end{aligned} \tag{2.5}$$

A Tabela 2 mostra o cálculo de $\mathcal{A}$ e $\bar{\mathcal{A}}$ para cada uma das questões presentes no exemplo acima, Tabela 1. Com os acertos calculados para cada grupo Eq. 2.5 é possível então calcular o discriminante Đ Eq. 2.4 de cada uma das questões da prova, como mostrado na última linha da tabela.

Tabela 2 – Quantidade de alunos em cada grupo que acertaram cada uma das questões e estimativa do índice de Discriminação.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7
$\mathcal{A}$	4	2	3	5	2	3	5
$\bar{\mathcal{A}}$	0	1	2	3	3	0	0
Đ	0.80	0.20	0.20	0.40	-0.20	0.60	1.00

Neste exemplo, mostrado na Tabela 2, verificamos que Q7 possui um discriminante ideal, pois consegue separar os alunos dos dois grupos: competentes e não-competentes. Os discriminantes negativos devem ser considerados em uma análise futura e buscar uma maneira de eliminar aquela questão da prova ou reformulá-la, como é o caso de Q5.

2.2.8 Considerações sobre Psicometria

Com intuito de trazer um exemplo clássico já bastante visitado ao longo da história, vejamos um breve resumo do trabalho de (MATLOCK-HETZEL, 1997). Nesta pesquisa, a autora considera um grupo de 30 estudantes respondendo algumas questões, sendo que a análise é feita em cima destas questões individualmente. A Tabela 3 mostra os cálculos feitos.

Tabela 3 – Exemplos de cálculos de índice de discriminação proposto por (MATLOCK-HETZEL, 1997)

	Q1	Q2	Q3	Q4	Q5
$\mathcal{A}$	0	15	13	11	7
$\bar{\mathcal{A}}$	0	15	5	7	11
b_j	0.0	1.0	0.60	0.60	0.60
Đ	0.0	0.0	0.53	0.27	-0.27

Através do exemplo da autora, vemos que não somente a dificuldade b_j deve ser considerada, mas também o índice de discriminação Đ. Para as questões Q1 e Q2 verificamos que mesmo tendo uma grande diferença na dificuldade da questão, o índice de discriminação é o mesmo. O contrário acontece nas questões Q3 e Q4 onde temos a mesma dificuldade mas índice de discriminação diferentes. Quando comparamos Q4 e Q5 vemos que eles possuem um mesmo número de dificuldade mas discriminação oposta, ou seja, Q5 apresenta o mesmo valor, mas negativo.

Também podemos verificar na Tabela 3 acima, que apenas a medida de dificuldade de uma pergunta isoladamente não nos traz informação suficiente sobre

ela-própria nem sobre seus respondentes, pois verificando a dificuldade mínima ou máxima (b_j de Q1 e Q2), foi mostrado que a discriminação se manteve em $\texttt{Đ}= 0.0$.

Gostaríamos de enfatizar este ponto, porque esta é uma das principais motivações deste trabalho conforme comentado na introdução. Sabemos que na área de sistemas de multi-classificadores existem muitas considerações sobre dificuldade das instâncias (Instance Hardness: `IH`) mas não temos nenhum indício de pesquisa sobre a relação de dois conjuntos de classificadores: classficadores-competentes e classificadores-não-competentes, como é o caso do índice de discriminação que será visto no próximo capítulo.

2.3 Considerações Finais

Neste capítulo destacamos os principais conceitos e técnicas fundamentais para entendimento do método proposto. Primeiramente foi comentado as principais etapas dos sistemas de múltiplos classificadores com ênfase em seleção dinâmica. Em segundo lugar, mas não menos importante, foi relatado algumas pesquisas sobre testes psicométricos. Nesta segunda parte foi dado maior ênfase por ser um tipo de estudo bastante extenso e também porque não existem muitos trabalhos de intersecção entre Psicometria e `MCS`. No próximo capítulo serão apresentados os trabalhos relacionados com a nossa proposta.

3 Trabalhos Relacionados

O estudo do capítulo anterior apresentou os principais conceitos introdutórios da teoria de testes psicométricos que tem sido estudados há mais de um século por pesquisadores de diferentes áreas como: Matemáticos, Psicólogos, Estatísticos, entre outros. Claro que estas poucas páginas não conseguem sintetizar todas pesquisas sobre o tema, mas no mínimo dar uma visão geral desta teoria para que consigamos adotar partes deste conhecimento em nossa área de seleção de múltiplos classificadores. A seguir mostramos alguns estudos que enxergamos estar vinculados com este trabalho: Sistemas de Múltiplos Classificadores e Psicometria.

Importante comentar que dentro dos protocolos experimentais de sistemas de múltiplos classificadores (BRITTO; SABOURIN; OLIVEIRA, 2014), vemos constantemente a comparação com os métodos estáticos como: melhor classificador (Single Best: SB); a combinação de todos classificadores (CC); e Oráculo.

O método SB (RUTA; GABRYS, 2005) é considerado uma seleção estática onde o melhor classificador do POOL é escolhido na fase de treinamento para rotular todas instâncias desconhecidas $\mathtt{x}_q$. Já método CC (KITTLER et al., 1998) também é considerado uma seleção estática, mas todo o POOL é considerado na predição de uma instância desconhecida $\mathtt{x}_q$, onde a predição final é formada pelos votos de todos classificadores (*majority vote*). O Oráculo (KUNCHEVA, 2002) é uma medida que representa a quantidade de instâncias que conseguem ser corretamente preditas por pelo menos um classificador do POOL (esta medida é bastante importante em nosso trabalho e visto com mais detalhes à frente).

3.1 MCS & Psicometria

Uma das maneiras de melhorar os sistemas de aprendizagem foi proposta por (BENGIO et al., 2009) através de sua teoria de aprendizagem nomeada por: Curriculum Learning. De acordo com os autores, os humanos e animais aprendem melhor quando os exemplos são apresentados de uma forma ordenada e significativa, não quando os conceitos são apresentados de forma randômica. Este trabalho fornece indícios que existe uma relação entre a associação de teorias de aprendizagem e os sistemas de múltiplos classificadores.

A ideia de: "começar pequeno e ir avançando"; dentro da área do conhecimento e aprendizagem tem sido discutida entre psicólogos e educadores, inclusive Vygotsky (VYGOTSKY, 1987) criou o conceito de Zona de Desenvolvimento Proximal (Zone of Proximal Development: ZPD) que resumidamente significa fornecer apenas exemplos que indivíduos conseguem absorver dentro de sua capacidade, pois se

forem fornecidos exemplos de aprendizagem muito difíceis causará uma frustração e não terão o resultado esperado: a aprendizagem.

Mais uma vez, pelo fato deste trabalho estar associado à diferentes áreas do conhecimento, citamos a teoria do portifólio (MARKOWITZ, 1952) (prêmio Nobel em Economia). Esta teoria mostra que é possível reduzir o risco quando se seleciona mais de um ativo para uma carteira de investiments. Esta ideia de diluir o risco também é visto em MCS.

Esta ideia foi utilizada nos estudos de (SMITH; MARTINEZ, 2016) ao propor o seu conceito de *hardness ordering* que significa fornecer as instâncias de treinamento à um sistema de aprendizagem de forma ordenada pela sua dificuldade (*Instance Hardness*: IH). Desta forma, o sistema de aprendizagem poderá assim tirar proveito da maneira com que os humanos também aprendem: através do mais simples, ao mais complexo.

Na pesquisa de (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) é citado o aprendizado supervisionado relacionado à estudantes em uma sala de aula guiado por um professor e gostaríamos de incluir um trecho desta pesquisa abaixo:

> *This is called supervised learning or 'learning with a teacher'. Under this metaphor the 'student' presents an answer $\hat{y}$ for each x_i in the training sample, and the supervisor or 'teacher' provides either the correct answer and/or an error associated with the student's answer. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)*

Acreditamos que seja salutar a incorporação do índice de discriminação Ð (Discrimination index) na escolha de instâncias de uma região de competência para então podermos construir estratégias que tragam mais uma alternativa positiva ao estado-da-arte em sistemas de múltiplos classificadores. A partir desta pesquisa, verificamos que existe a busca entre a associação do: aprendizado-humano com os algoritmos-aprendizagem.

Estes registros acima foram tentativas de incorporar a área de psicometria com seleção de múltiplos classificadores como é o foco deste trabalho. Ainda assim, o conteúdo abaixo será dedicada às últimas pesquisas sobre regiões de competência para sistemas de seleção dinâmica que é um dos focos deste trabalho.

3.2 MCS & Regiões de Competência

Iniciaremos esta parte do trabalho sobre a manipulação da região de competência através da seleção de protótipos (Prototype Selection, PS). Foram feitos alguns estudos interessantes sobre este assunto e por este motivo já iniciaremos por ele, vejamos. A manipulação da região de competência através de seleção de protótipos obtiveram resultados interessantes, como é o caso de (CRUZ; SABOURIN; CAVALCANTI, 2016). Sabemos que o estudo de Seleção de Protótipos é bastante

vasto (OLVERA-LÓPEZ et al., 2010), no entanto vemos que existem trabalhos que contenham a interseção entre seleção de protótipos e região de competência. Chamamos de protótipos como um subconjunto dos dados que tentam capturar a essência da distribuição dos dados dentro de uma região de competência, podendo esta região ser formada por apenas poucos vizinhos ou então todo um dataset. Várias foram as estratégias para construir sistemas de seleção de protótipos e de acordo com vários autores (GARCIA et al., 2011; TRIGUERO et al., 2012; ZUBEK; KUNCHEVA, 2018) podemos agrupá-las em: condensação, edição e híbridas. Vários são os trabalhos que propõem a manipulação de dados através destas técnicas, por exemplo: (CRUZ; SABOURIN; CAVALCANTI, 2016) mostra uma comparação entre seis técnicas de PS adaptadas para dynamic ensemble selection, comparadas com o k-NN. De acordo com os autores, foram escolhidas técnicas de edição e híbridas devido ao interesse na performance do sistema. As técnicas avaliadas pelo autor foram: Random Mutation Hill Climbins (SKALAK, 1994); Relative Neighborhood Graph (SÁNCHEZ; PLA; FERRI, 1997) (RNG); Edited Nearest Neighbor (WILSON, 1972) (ENN); Steady-State Memetic Algorithm (GARCÍA; CANO; HERRERA, 2008); Adaptive Search Algorithm (ESHELMAN, 1991); e General Genetic Algorithm (KUNCHEVA, 1995).

Outro método bastante interessante que consiste na edição do `Dsel` foi proposto por (CRUZ; CAVALCANTI; Ing Ren, 2011) e seu algoritmo de filtro adaptativo: `fa-kNN` (Filter Adaptive Distance KNN). Neste trabalho o autor realiza a remoção de instâncias em duas etapas, sendo a primeira na fase offline: treinamento do `POOL` de classificadores utilizando `e-kNN`; e a segunda etapa realizada na fase de teste: classificação das instâncias desconhecidas, utilizando o método Adaptive distance `k-NN` (`a-kNN` (WANG; NESKOVIC; COOPER, 2007)) .

O trabalho proposto por (CRUZ; SABOURIN; CAVALCANTI, 2017b) trás uma maneira única de trabalhar com os dados em duas fases. Em primeiro é aplicado uma seleção de protótipos ao conjunto de validação durante a etapa de aprendizagem, com o intuito de diminuir a sobreposição entre classes. O método utilizado nesta primeira etapa é k-NN editado (Edited Nearest Neighbor, ENN). Na segunda etapa o artigo cita o emprego do algoritmo k-NN adaptativo (AKNN) na fase de captura da `RoC`, utilizado para reduzir o impacto de instâncias ruidosas. Este trabalho se parece com o trabalho `fa-kNN`, mas traz diferentes abordagens, testes, cenários e resultados.

A pesquisa de (SABOURIN et al., 1993) propôs a competência dos classificadores baseado em ordenação (rank) o que deu nome ao algoritmo: Classifier Rank. Após este trabalho, foi proposto um novo método de ordenação modificada (WOODS; KEGELMEYER; BOWYER, 1997) (Modified Classifier Ranking, `MCR`). Esta ordenação é feita através das instâncias presentes na região de competência. Este método faz uma ordenação das instâncias da RoC pela distância com $\mathtt{x}_q$, ou

seja, a RoC é estimada pelo algoritmo k-NN. Após a ordenação pela distância, a competência é calculada como o número de acertos consecutivos.

No trabalho de (WOODS; KEGELMEYER; BOWYER, 1997) foi proposto também os dois métodos baseados na acurácia local: `OLA` (Overall Local Accuracy) e `LCA` (Local Classifier Accuracy). O primeiro algoritmo estima a competência do classificador presente no pool baseado em sua acurácia na região de competência, e o segundo (`LCA`) verifica qual instância dentro da RoC possui a mesma classe de predição do classificador por $\mathtt{x}_q$. Basicamente a diferença entre `OLA` e `LCA` é a região de competência porque logo após este cálculo, será escolhido o classificador com maior competência.

O trabalho de (SMITS, 2002) propõe uma região de competência onde cada instância $\mathbf{x_j}$ é ponderada pela sua distância à $\mathtt{x}_q$. Por este motivo este método considera a acurácia local modificada que dá nome ao algoritmo `MLA` (Modified Local Accuracy). De acordo com este trabalho (SMITS, 2002), o autor considera que instâncias mais próximas de $\mathtt{x}_q$ tenham uma maior influência que as mais distantes.

No trabalho de (GIACINTO; ROLI, 1999) é considerada uma saída probabilística de cada classificador para cada classe do problema e com isso foram criados os métodos aPriori e aPosteriori. Estes métodos são análogos ao `OLA` e `LCA`, mas consideram uma saída probabilística.

O trabalho de (GIACINTO; ROLI, 2001) utiliza o espaço de decisão (Behavior-Knowledge Space, `BKS` (HUANG; SUEN, 1995)) para compor o comportamento de seu sistema `DCS` denominado `MCB` (Multiple Classifier Behaviour). Neste trabalho, para cada $\mathbf{x_j}$ é estimado um um perfil de saída (output profile) dos classificadores presentes no `POOL`. Com isso é construído um espaço pelas decisões dos classificadores. Este método explora a predição de todos classificadores do `POOL` como fonte de informação (BRITTO; SABOURIN; OLIVEIRA, 2014). Neste método a `RoC` é estimada pelas instâncias que estiverem mais relacionadas com $\mathtt{x}_q$ no espaço de decisão `BKS`.

Vemos também pelo trabalho de (BRUN et al., 2016) que as medidas de complexidade e acurácia local podem ajudar a compor as instâncias $\mathbf{x_j}$ da região local. O sistema chamado de seleção dinâmica baseado em complexidade (`DSOC`) considera a similaridade da complexidade entre região local e o conjunto de treino; e também considera a distância entre a classe predita e o centróide da classe no conjunto de validação. Por fim, a acurácia local é estimada conforme realizado no algoritmo `LCA`.

No trabalho de (SOARES et al., 2006; SANTANA et al., 2006) temos dois sistemas interessantes que utilizam conceitos de acurácia e diversidade para construção de listas de prioridade. Estas listas são construídas a partir do `POOL`, sendo que uma lista ordenada a partir de acurácia e outra ordenada pela diversidade entre pares de

classificadores. Os sistemas que se utilizam desta ideia são DES-Clustering `DES-CLU` e `DES-DKNN`.

O trabalho de (KO; SABOURIN; BRITTO, 2008) utiliza o conceito de oráculo local para estimativa de classificadores promissores considerando uma região de competência. Os métodos que fazem parte desta abordagem são conhecidos como Knora-Eliminate (`KNE`) e Knora-Union (`KNU`). Para o cálculo de competência é considerado o oráculo dentro de uma região local. O `KNU` seleciona os classificadores que acertam pelo menos uma instância na região de competência e o `KNE` considera somente classificadores que acertam todas.

Proposto por (WOLOSZYNSKI; KURZYNSKI, 2011), seu sistema de seleção dinâmica utiliza a medida de um classificador randômico como limiar (threshold) para selecionar os classificadores e compor o ensemble. Em seu algoritmo (Randomized Reference Classifier, `RRC`) primeiramente é calculado a fonte de competência (source of competence) (WOLOSZYNSKI; KURZYNSKI, 2010) para cada instância do conjunto de validação. Neste trabalho é considerado que um classificador randômico possui uma probabilidade de acertar uma instância a partir do número total de classes do problema. Os classificadores são agrupados e caso obtenham uma competência maior ou igual à um classificador randômico, serão selecionados para predizer a instância desconhecida.

O trabalho proposto por (CAVALIN; SABOURIN; SUEN, 2012) é parecido com `KNU` onde a competência é calculada como sendo o número de instâncias que cada classificador acerta dentro da RoC. Este método chamado k-vizinhos com perfil de saída (k-Nearest Output Profile, `KNOP`), o espaço de decisão é utilizado como fonte de informação, ou seja, é medido a proximidade do perfil-de-saída (output profile, `OP`) com as outras instâncias do conjunto de validação. O `OP` de uma instância é a predição de todos classificadores do `POOL` para esta dada instância. Ao final é feito uma seleção de instâncias que que possuam um `OP` mais próximo do `OP` de $\mathtt{x}_q$.

O trabalho proposto por (CRUZ et al., 2015) chamado de Meta-DES (`MDES`) se baseia em calcular meta-características para o treinamento de um meta-classificador que dirá quais serão os classificadores competentes. O mesmo autor também propôs (CRUZ; SABOURIN; CAVALCANTI, 2017a), sendo a competência predita por um meta-classificador e sua resposta variando entre [0-1]. Ambos frameworks se baseiam em extrair meta-características (meta-features) de um conjunto de fontes de informação: região local e perfil de saída. O segundo framework ainda faz uma seleção de meta-características diferente para cada problema, realizado uma vez no início do sistema. O primeiro trabalho (no início deste parágrafo) realiza combinação dos classificadores com voto majoritário, sendo que o segundo trabalho (CRUZ; SABOURIN; CAVALCANTI, 2017a) também faz combinação por voto majoritário, mas ponderado pela competência.

O trabalho proposto por (TRIERVEILER et al., 2018) também chamado de

DISi (discriminate index) trás uma proposta diferente para estimar a região de competência. Neste trabalho a região de competência é estimada em cojunto com o POOL de classificadores. Esta união é feita a partir de teorias provindas de CTT (Classical Test Theory) utilizada para ranquear estudantes de uma sala de aula a partir das respostas à um teste. Esta teoria vem sendo construída há mais de um século, significando que passou por um bom período de amadurecimento como é o caso do índice de discriminação. Este índice traz à tona a eficácia da seleção de perguntas em uma prova para conseguir melhor separar os estudantes mais competentes daqueles não-competentes. Este índice de discriminação foi trazido para a seleção de instâncias em uma região de competência com intuito de podar instâncias que supostamente não contribuiriam em separar bons classificadores dos demais, facilitando assim para o algoritmo DES ter uma melhor performance.

Alguns trabalhos focaram em considerar a seleção de classificadores mais promissores sem a ajuda da região de competência, vejamos: A proposta de CHADE (PINTO; SOARES; MENDES-MOREIRA, 2016) prevê diretamente quais classificadores de um ensemble são competentes para cada instância, utilizando uma cadeia de classificadores que captura os acertos/erros entre eles. Para isso, o algoritmo cria um dataset Dsel' com características da saída dos classificadores do POOL, e com isso transforma o problema em multi-rótulos (multi-labels). A decisão final resulta da combinação apenas dos classificadores previstos como competentes, dispensando o uso de regiões de competência baseadas em vizinhança. O mesmo acontece para a pesquisa de (NARASSIGUIN; ELGHAZEL; AUSSEM, 2017) que também prevê os classificadores mais competentes sem a necessidade da RoC, mas executa esta tarefa de uma maneira a tentar corrigir a limitação do CHADE introduzindo uma modelagem probabilística.

Convém comentar que outros trabalhos também pesquisaram sobre este assunto de seleção de classificador(es) mais promissor(es) sem ajuda de uma região de competência. O trabalho de (SANTOS; SABOURIN; MAUPIN, 2008) tenta a não-utilização de RoC ao criar vários conjuntos de classificadores e com isso para cada $\mathbf{x}_q$ é selecionado um grupo com maior confiança. O trabalho de (ELMI et al., 2023) trata a classificação como um problema multi-rótulo (multi-label), conhecido como DES-ML (DES Multi-Label). O trabalho de (ZHANG; ZHU; LUO, 2025) propõe uma medida de sinergia, relacionada à Teoria dos Jogos (John von Neumann; Oscar Morgenstern, 1945) com uma variante do valor de Shapley (SHAPLEY et al., 1953).

Temos alguns trabalhos que tiveram seu foco em grafos, como o caso do trabalho de (LI et al., 2019). Neste método é utilizado uma abordagem em grafo para escolha de instâncias para compor a região de competência. Neste caso a vizinhança de $\mathbf{x}_q$ é estimada utilizando dois grafos: must-link e cannot-link. Estes grafos possuem pesos em suas arestas de conexão e por fim a competência dos classificadores é

estimada pelos pesos nestes dois grafos (DAVTALAB, 2023). Outra abordagem de grafos utilizada como região de competência pode ser visto no trabalho (SOUZA et al., 2024) e posteriormente (SOUZA et al., 2023). Nestes últimos dois trabalhos, os autores propõem uma nova abordagem de seleção dinâmica de ensembles que busca superar as limitações associadas à definição da RoC em cenários com dados esparsos e sobrepostos, frequentes em problemas de alta dimensionalidade e poucos exemplos (high dimensional small sample sized, HDSSS). Em vez de depender exclusivamente da suposição de localidade para construir a RoC, o GNN-DES modela a estrutura local dos dados em forma de grafo e as relações entre classificadores por meio de um conjunto de meta-rótulos (meta-label). As conexões entre instâncias são ponderadas de acordo com sua proximidade no espaço de decisão e sua relação de classe. Desta forma são formados vínculos fortes ou fracos no espaço de decisão. As dependências entre classificadores são codificadas como meta-rótulos multiclasse e o processo de seleção é aprendido de forma fim-a-fim por uma GNN (Graph Neural Network).

No estudo desenvolvido por (OLIVEIRA; CAVALCANTI; SABOURIN, 2017) - FIRE-DES, os autores propõem uma estrutura para gerenciar dois tipos de regiões de competência: regiões seguras (safe region) e regiões de indecisão (indecision region). De acordo com os autores, $\mathbf{x}_q$ pode se encontrar em uma região de indecisão caso sua RoC contenha instâncias de diferentes classes. Nestas ocasiões o algoritmo criou uma maneira de operar nestas incertezas e nomeou esta poda de classificadores como: DFP (Dynamic Frienemy Prunning) e chamou de frienemies (friend or enemy) as instâncias vizinhas. Este método DFP captura apenas classificadores que possuam suas fronteiras de decisão dentro da RoC. O segundo método desenvolvido pelo mesmo grupo de pesquisadores (CRUZ et al., 2019), apelidado de FIRE-DES++, aprimora esta maneira de definir a RoC. Primeiro é feito uma filtragem do Dsel para eliminação de supostos ruídos e em segundo lugar é utilizado k-NN-Equality que em regiões indecisas irá manter no mínimo um classificador de cada classe.

O trabalho proposto por (SOUZA et al., 2019) tem uma abordagem de geração dinâmica de POOLs de classificadores voltada para o contexto DCS. Esta geração dinâmica acontece quando a instância se encontra em regiões difíceis do espaço de características (sobreposição de diferentes classes) identificado pela medida de complexidade de desacordo de vizinhos (k-Disagreeing Neighbors, kDN). Nestas situações, o método OLP (Online Local Pool) constrói um POOL especializado à instância $\mathbf{x}_q$. Caso a região não tenha sobreposição de diferentes classes o método utilizado é o de vizinhança mais próxima k-NN. Neste sentido este método associa a RoC à medidas de dificuldade de instâncias como kDN, para gerar pools especializados sob demanda. Outro trabalho realizada pelos mesmos pesquisadores (SOUZA et al., 2022) expande o trabalho para o cenário de alta dimensionalidade onde: os problemas com RoC devido à escassez de vizinhos; e sobreposição de classes,

é uma realidade. Este estudo chamado de OLP++ incorporou métodos como vizinhança adaptativa, redução de espaço de busca e também geração incremental de classificadores.

O artigo (ELMI; EFTEKHARI, 2020) propõe um novo método de seleção dinâmica de ensembles: DES-hesitant, que integra conjuntos-fuzzy-indecisos (hesitant fuzzy sets, HFS) ao processo de classificação. O método considera que a competência de um classificador deve ser avaliada sob múltiplos critérios locais, como: acurácia, probabilidade, fração, ranqueamento e função potencial - calculados sobre diferentes regiões de competência, definidas no espaço-de-características e no espaço-de-decisão. Esses critérios são organizados em uma matriz de decisão fuzzy-hesitante, que permite representar a incerteza sobre os graus de competência de cada classificador em diferentes regiões. A seleção final dos classificadores é realizada através de um processo de agregação HDMCDM (hesitant fuzzy multiple criteria decision making), onde um limiar é definido e somente classificadores com valor acima deste limiar são selecionados.

O trabalho de (DAVTALAB; CRUZ; SABOURIN, 2022) propõe um *framework* denominado de seleção dinâmica baseado em hipercaixas fuzzy (FH-DES, Fuzzy hyperbox, Dynamic Ensemble Selection), que modela as regiões de competência, e também incompetência, de cada classificador por meio de hipercaixas fuzzy. Cada hipercaixa representa um sub-conjunto de amostras delimitado por seus pontos extremos (mínimo e máximo), na tentativa de reduzir a complexidade de armazenamento e processamento em comparação à abordagens baseadas em KNN. Desta forma, o método cria vários sub-conjuntos a partir do DSEL que de acordo com os autores será menor que o número de instâncias, trazendo benefício na hora da busca por instâncias mais próximas na `RoC`. O método cria também duas variantes FH-DES-C e FH-DES-M que significa rodar o algoritmo com uma RoC estimada através de instâncias corretamente-classificadas ou com instâncias mal-classificadas, respectivamente. Por este motivo é utilizado os caracteres 'C' e 'M' ao final. Os autores comentam que em testes com 900 mil instâncias o método permitiu diminuir para aproximadamente 4 mil hipercaixas. Este trabalho possui uma outra versão (DAVTALAB; CRUZ; SABOURIN, 2023) que aprimora a escalabilidade e reduz a complexidade computacional ao realizar todo o cálculo das regiões de competência na fase off-line de treinamento; assim, a classificação de uma instância envolve apenas computações de pertinência fuzzy aos *hyperboxes*, eliminando buscas por vizinhos na fase de teste. Estes métodos são descritos também na Tese (DAVTALAB, 2023), onde também informa que o método fuzzy hyperbox tem sido estudado nas últimas décadas, por exemplo o trabalho de (GABRYS; BARGIELA, 2000).

No artigo de (CHOI; LIM, 2021) é apresentado um framework `DES` que aprimora a definição das regiões de competência de cada $\mathtt{x}_q$. Para superar os limites de uma

RoC tradicional, que possui pontos sensíveis a ruído e sobreposição de classes, o DDES propõe duas variantes para RoC: (i) DDES-I: fará uma escolha de região de competência elipsoidal com seus eixos (axes) correspondendo às variâncias de $\mathbf{x}_q$. (ii) DDES-M: também de forma elipsoidal, mas com a distância sendo considerada pela medida Mahalanobis. De acordo com os autores, estas regiões de competência elipsoidais permitem selecionar de forma mais precisa os classificadores competentes para cada exemplo.

O trabalho de (BARBOZA et al., 2025) possui foco em mudança-comportamento (concept drift) que possui uma característica de alteração do conjunto de instâncias Dsel ao longo do tempo. Devido à esta realidade foi criado DES incremental e adaptativo (INCA-DES, INCremental and Adaptative DES). Neste cenário, o algoritmo necessita de uma RoC que tenha um tempo de resposta rápido para fluxo de dados (data setream) e foi utilizado o algoritmo de árvores K-D (K-Dimensional Tree). De acordo como os autores isso diminui o tempo de busca de instâncias em uma árvore e também permite a remoção e/ou alteração de instâncias com alta velocidade, dado a necessidade do ambiente de *concept drift*.

Nosso último estudo (FARHANGIAN et al., 2025) comenta sobre combinação de estruturas para detecção de notícias falsas (fake news): DRES. Por ser um sistema que trabalha com textos, sua primeira etapa é a transformação do texto em características que consigam ser interpretadas pelos algoritmos de seleção dinâmica. Como existe mais de uma maneira de transformação de textos, o método cria conjuntos diferentes de características para cada texto. Ao tentar classificar uma instância desconhecida $\mathbf{x}_q$, é buscado em vários destes conjuntos quais são as regiões de competência que possuam o menor índice de dificuldade (instance hardness). Após esta análise, é escolhido o POOL de classificadores que pertence ao mesmo conjunto e então um algoritmo DES é utilizado. A inovação neste caso para nosso estudo está na multiplicidade de regiões de competência para serem escolhidas.

Vemos que dentro de um sistema de seleção de classificadores várias são as formas para que esta escolha aconteça. Uma das maneiras que vemos é a partir de instâncias mais próximas à $\mathbf{x}_q$ que serão utilizadas para definir a competência de classificadores. No entando, existem várias outras formas de captura de instâncias e até mesmo sistemas que se moldam para a não-captura, como o caso (PINTO; SOARES; MENDES-MOREIRA, 2016). Para facilitar esta visualização mostramos na Tabela 4 os sistemas aqui listados e a maneira com que a RoC é estimada.

Tabela 4 – Lista dos metodos de seleção dinâmica e suas respectivas regiões de competência.

Ano	abrev	Nome Completo	definição RoC	Espaço Caract	Referência
1993	RANK	Classifier Rank algorithm	k-NN	FS	(SABOURIN et al., 1993)
1997	LCA	Local Class Accuracy	k-NN	FS	(WOODS; KEGELMEYER; BOWYER, 1997)
1997	OLA	Overall Local Accuracy	k-NN	FS	(WOODS; KEGELMEYER; BOWYER, 1997)
1999	API	A Priori	k-NN	FS	(GIACINTO; ROLI, 1999)
1999	APO	A Posteriori	k-NN	FS	(GIACINTO; ROLI, 1999)
2001	MCB	Multiple Classifier Behaviour	k-NN	FS	(GIACINTO; ROLI, 2001)
2002	MLA	Modified Local Accuracy	k-NN	FS	(SMITS, 2002)
2006	DES-CLU	Dynamic Selection Clustering	agrupamento	FS	(SOARES et al., 2006; SANTANA et al., 2006)
2006	DES-KNN	Dynamic Selection k-NN	k-NN	FS	(SOARES et al., 2006; SANTANA et al., 2006)
2008	DOCS	Dynamic overproduce-and-choose strategy	no-roc	–	(SANTOS; SABOURIN; MAUPIN, 2008)
2008	KNE	KNORA Eliminate	k-NN	FS	(KO; SABOURIN; BRITTO, 2008)
2008	KNU	KNORA Union	k-NN	FS	(KO; SABOURIN; BRITTO, 2008)
2011	FA-KNN	Filter-Adaptive Region of Competence	k-NN	FS	(CRUZ; SABOURIN; CAVALCANTI, 2017b)
2011	RRC	Randomized Reference Classifier	função potencial	FS	(WOLOSZYNSKI; KURZYNSKI, 2011)
2012	KNOP	K-nearest Output Profiles	k-NN	DS	(CAVALIN; SABOURIN; SUEN, 2012)
2015	META-DES	DES framework using meta-learning	k-NN	FS + DS	(CRUZ et al., 2015)
2016	CHADE	Chained Dynamic Ensemble	no-roc	–	(PINTO; SOARES; MENDES-MOREIRA, 2016)
2016	DSOC	Dynamic Selection Based on Complexity	k-NN	FS + DS	(BRUN et al., 2016)
2017	FIRE-DES	Frienemy Indecision Region DES	k-NN	FS	(OLIVEIRA; CAVALCANTI; SABOURIN, 2017)
2017	META-DES.O	META-DES Oracle	k-NN	FS + DS	(CRUZ; SABOURIN; CAVALCANTI, 2017a)
2017	PPC	Probabilistic Classifier Chains	no-roc	–	(NARASSIGUIN; ELGHAZEL; AUSSEM, 2017)
2018	DISi	RoC by Discriminant Index	k-NN	FS	(TRIERVEILER et al., 2018)
2019	FIRE-DES++	Fire-DES improved	k-NN	FS	(CRUZ et al., 2019)
2019	GDEP	Graph-based Dynamic Ensemble Pruning	grafo	DS	(LI et al., 2019)
2019	OLP	Online Local Pool	part. recursivo	DS	(SOUZA et al., 2019)
2020	DES-HFS	DES Hesitant fuzzy multiple criteria	multiplo k-NN	FS + DS	(ELMI; EFTEKHARI, 2020)
2021	DDES	Distribution-Based DES Framework	k-NN	FS	(CHOI; LIM, 2021)
2022	OLP++	OLP improved	part. recursivo	DS	(SOUZA et al., 2022)
2023	DES-ML	DES Multi-label	no-roc	–	(ELMI et al., 2023)
2023	FH-DES	Fuzzy hyperboxes DES	hip. fuzzy	FS	(DAVTALAB; CRUZ; SABOURIN, 2023)
2024	GNN-DES	Graph Neural Network DES	grafo	DS	(SOUZA et al., 2024; SOUZA et al., 2023)
2025	DES-AS	DES based on Algorithm Shapley	no-roc	–	(ZHANG; ZHU; LUO, 2025)
2025	DRES	Dynamic Representation and Ensemble Selection	k-NN (k-d tree)	FS	(FARHANGIAN et al., 2025)
2025	INCA-DES	Incremental and Adaptative DES	multiplo k-NN	FS	(BARBOZA et al., 2025)

Obs: k-NN: k-vizinhos mais próximos; part. recursivo: particionamento recursivo; no-roc: ausência de região de competência; hip. fuzzy: hipercaixas fuzzy; FS: espaço de características (feature-space); DS: espaço de decisão (decision space).

Também trazemos a taxonomia adotada pela comunidade científica onde define o tipo de cada RoC (ou sua ausência). Esta taxonomia também foi utilizada na Tabela 4 e também pode ser encontrada em outros trabalhos como (DAVTALAB, 2023).

- k-vizinhos mais próximos (k-nearest neighbors);
- agrupamento (clustering);
- função potencial (potential function);
- particionamento recursivo (recursive partitioning);
- hipercaixas fuzzy (fuzzy hyperboxes);
- grafo (graph based);
- – ausência – (no-roc).

3.3 Considerações Finais

Neste capítulo relatamos os principais trabalhos relacionados ao nosso método principal: extração de região de competência para sistemas de seleção dinâmica. Vemos que o método proposto `DISi` traz como novidade para esta área uma abordagem de seleção dinâmica e uma alternativa para aprimoramento da região de competência baseado em psicometria. No intuito de organizar melhor este trabalho trouxemos a Tabela 4 com várias contribuições para extração da região de competência ao longo das últimas três décadas. Em segundo lugar foi trazido a taxonomia atual utilizada como métodos para constituição de uma região de competência (inclusive sua ausência).

4 Método Proposto

A principal ideia deste estudo é a investigação da possibilidade em incorporar métricas procedentes das teorias de testes psicométricos em sistemas de seleção de múltiplos classificadores para melhor caracterizar as instâncias em uma região de competência.

Considerando que um método de seleção dinâmica escolhe os classificadores mais promissores em tempo de execução, enfatizamos a importância de realizar mais pesquisas e análises considerando as hipóteses levantadas na introdução deste trabalho.

1. índice de discriminação para selecionar dinamicamente uma região de competência que possua maior poder de informação;

2. investigação de uma nova métrica de qualidade para região de competência que não seja influenciada pelo viés dos algoritmos de seleção dinâmica;

3. investigação de uma nova métrica para o `POOL` de classificadores que considere o score mínimo de um algoritmo de seleção dinâmica.

4.1 Considerações Iniciais

A área da psicometria possui várias medidas, dentre elas: habilidade de estudantes, dificuldade das questões, índice de discriminação, entre outras. Foi analisada a incorporação destas primeiras três medidas para construção de um novo algoritmo de seleção de vizinhos mais póximos, mas antes precisamos fazer algumas associações entre estas duas áreas de pesquisa: Psicometria e Aprendizagem de Máquina, como mostrado na Tabela 5.

Pensando nestas associações podemos construir um sistema de seleção dinâmica que filtre as instâncias da região de competência pelo índice de discriminação, ou seja, na diferença entre os acertos do grupo de classificadores competentes e o grupo dos classificadores não-competentes, conforme Eq. 2.4.

Na psicometria, o grupo de indivíduos que mais acerta questões: grupo-competentes, pode ser visualizado na seleção dinâmica como o grupo de classificadores selecionados. No outro caso, o grupo dos indivíduos não-competentes, chamaremos de ensemble de classificadores com menor acurácia dentro de uma região de competência.

Tabela 5 – Associações da área de sistemas de múltiplos classificadores e testes psicométricos

Machine Learning	Representação	Testes Psicométricos
instância desconhecida	$\mathtt{x}_q$	pergunta de uma prova, ou seja, uma informação desconhecida que se queira resolver
classificador	$\mathbf{c_i}$	estudante, ou seja, um objeto ativo no sistema que provê respostas para questões
pool de classificadores	`POOL`	uma sala de aula com vários estudantes
competência do classificador	$\not\subset$	habilidade de um aluno em reponder uma questão
dataset de validação	`Dsel`	um banco de dados de questões que o professor irá utilizar para fazer uma prova
região de competência	`RoC`	uma prova feita pelo professor para validar um conhecimento na turma, ou seja, serão escolhidos vários perguntas x_i próximas de $\mathtt{x}_q$ para avalidar um determinado assunto/conhecimento na turma

4.2 Proposta de uma nova Região de Competência

Uma das maneiras em estimar a região de competência é através da seleção de k-vizinhos mais próximos à instância desconhecida $\mathtt{x}_q$. A proposta deste trabalho é adicionar uma pré-tarefa para tentar ponderar quem são os k-vizinhos mais promissores baseado em métricas de distância e competência dos classificadores. Esta pré-tarefa consiste em selecionar primeiramente um conjunto maior que o valor de $k-$vizinhos pré configurado, ou seja, será selecionado $(\alpha * k)$-vizinhos, sendo α explicado a seguir.

A Figura 9 mostra como a região de competência pode ser estimada neste método. A Figura 9(a) mostra uma região de competência estimada apenas pela distância de $\mathtt{x}_q$. A Figura 9(b) mostra como uma região de competência pode ser estimada, considerando o índice de discriminação. Neste exemplo, a Figura 9(b) é

estimada uma região local com $\alpha = 2$ e $k = 7$. Isto significa que o círculo exterior contém $\alpha * k = 14$ instâncias.

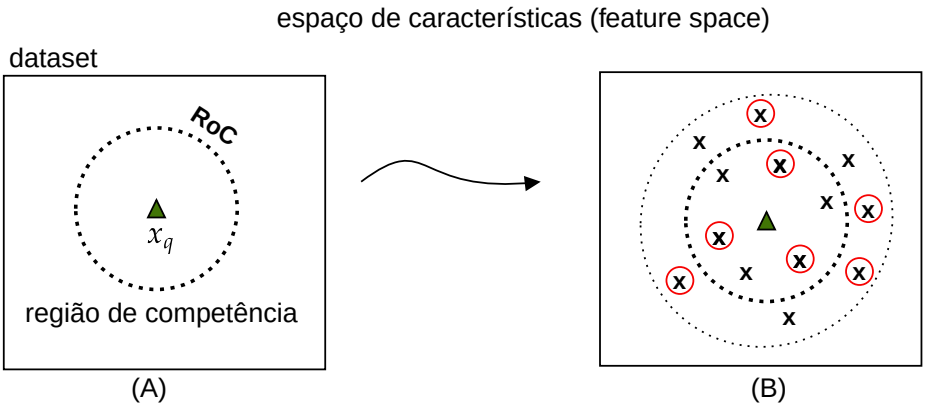


Figura 9 – Aumento de uma região local para estimar a região de competência.

Depois desta etapa serão escolhidas as instâncias que possuem uma maior qualidade que neste projeto está sendo considerado as instâncias com maior Índice de Discriminação. A Figura 9(b) mostra um exemplo de instâncias com círculo vermelho representando os k-vizinhos mais discriminantes.

4.2.1 Algoritmo Proposto

O Algoritmo 1 inicia estimando a região local através de alguma função de vizinhança. Na primeira linha, o conjunto S recebe as instâncias vizinhas de $\mathtt{x}_q$ dentro de `Dsel`, com o tamanho de $\alpha.k$. O parâmetro α é o multiplicador do tamanho da região de competência. Esta tarefa é necessária para estimar uma região local maior que a `RoC` para então escolher instâncias com maior qualidade.

Em seguida estima-se a acurácia de cada classificador do `POOL` considerando a região S. Os classificadores são ordenados e os grupos extremos Upper e Lower (UL) são capturados (linhas 3 e 4). A linha 5 mostra o loop para cada instância. A dificuldade da instância x_i é então estimada para cada grupo UL nas linhas 6 e 7. O valor `Đ` é estimado na linha 8, calculando a diferença de dificuldade entre os grupos *Upper* e *Lower*. No final, são selecionadas k-instâncias com o maior valor `Đ`. O código com o Algoritmo proposto pode ser acessado no link: https://github.com/marcelo-trier/DISi.

Algoritmo 1 Região de Competência estimada por Índice de Discriminação - dRoC.

Input: pool of classifiers: `POOL`
Input: Dynamic selection dataset: `Dsel`
Input: Size of nearest neighbors: k, and factor of size: α
Input: unknown instance: $\mathtt{x}_q$
Output: `RoC` based on discriminant measure

```
S ← k-NN(x_q, Dsel, α.k)
POOL^r ← rank POOL by accuracy on S
U ← get Upper group of classifiers on POOL^r
L ← get Lower group of classifiers on POOL^r
for each x_i in S do
    b_u ← difficulty of x_i for U
    b_l ← difficulty of x_i for L
    Đ_i ← b_u − b_l (CROCKER; ALGINA, 1986)
end for
RoC ← select k instances with best Đ
return RoC
```

4.2.2 Alternativa ao Algoritmo Proposto

O algoritmo proposto primeiro aumenta o número de instâncias na região de competência para depois capturar aquelas mais promissoras. Com isso, vemos claramente duas etapas: (1) verificar a vizinhança de $\mathtt{x}_q$ com uma quantidade maior que uma região de competência tradicional, e (2) dentro deste novo conjunto, selecionar somente $k-$vizinhos mais promissores.

Mas com essa nova proposta surge então algumas questões: Que abordagem alternativa poderíamos adotar para comparar uma região de competência que selecione $\alpha*k$-vizinhos mais próximos? Que abordagem alternativa poderia capturar os $\alpha * k$-vizinhos mais próximos e depois reduzidar a região para k-vizinhos? Foi pensando nestas perguntas que foram feitas as implementações: (1) seleção de vizinhança de tamanho: $\alpha * k$-vizinhos que chamamos de `2kRoC`; e (2) seleção de vizinhança randômica, chamado de: `rRoC`. Veremos estas duas regiões de competência a seguir.

Nossa primeira implementação (1) `2kRoC` simplesmente captura uma região de competência com um tamanho $2 * k$ maior que as outras `RoC`. Esta primeira implementação pode ser visualizada na Figura 9(b), considerando que não serão podadas nenhuma das instâncias que pertencem ao círculo maior.

Nesta segunda implementação (2) `rRoC` realiza uma poda das instâncias de maneira randômica. A Figura 9 também pode ajudar a explicar este conceito. A

principal diferença entre dRoC e rRoC é que esta segunda selecionará aleatoriamente as instâncias da Figura 9(b). Segue abaixo o algoritmo com poda randômica:

1. Selecionar $\alpha * k$-vizinhos mais próximos,
2. Escolher aleatoriamente k-vizinhos.

Algoritmo 2 Região de Competência selecionada randomicamente - rRoC.

Input: pool of classifiers: `POOL`
Input: Dynamic selection dataset: `Dsel`
Input: Size of nearest neighbors: k, and factor of size: α
Input: unknown instance: $\mathtt{x}_q$
Output: `RoC` based on random instances from local region
1: S $\leftarrow$ k-NN($\mathtt{x}_q$, `Dsel`, $\alpha.k$)
2: `RoC` $\leftarrow$ Random select k instances from S
3: **return** `RoC`

A ideia principal deste tipo de análise é mostrar que a escolha de instâncias baseadas em discriminação `dRoC` não segue a mesma tendência de uma região de competência estimada com o dobro de instâncias ou uma `RoC` com poda aleatória.

4.2.3 Análise do método DES-DISi

Em nosso primeiro artigo foram reportados os resultados da adaptação da teoria de testes psicométricos para um algoritmo de seleção dinâmica, chamado de "Dynamic Ensemble Selection based on Discrimination Index"(DES-DISi) (TRIERVEILER et al., 2018).

O algoritmo de discriminação faz um pré-processamento das instâncias na região de competência, para então selecionar aquelas mais promissoras para esta análise. Para fazer testes e comparar com os outros algoritmos de `DS`, foi implementada esta estimativa da `RoC` sobre um algoritmo de seleção dinâmica já conhecido: Knora-Union (`KNU`). Desta forma, podemos assumir que esta versão do DISi pode ser comparado ao `KNU`, mas com um preprocessamento na região de competência pelo índice de discriminação.

A análise reportada no artigo (TRIERVEILER et al., 2018) mostrou que o algoritmo DISi foi superior a outros 16 algoritmos de seleção dinâmica e outros dois algoritmos de seleção estática. No entanto, considerando que o DISi é um algoritmo para estimar a região de competência e considerando que foi utilizado o Knora-Union como método de seleção dinâmica, gostaríamos de propor como segunda análise: a continuação do referido artigo. Esta seção trata justamente

da comparação entre a execução do KNU sob duas regiões de competência: uma RoC estimada por k-vizinhos mais próximos (chamaremos de: kRoC) e uma RoC estimada por índice de discriminação (dRoC).

Para realização desta análise, a seguinte questão norteou a condução do teste: O método DISi produz resultados diferentes do método KNU original? A proposta desta análise é investigar se o KNU-kRoC é diferente estatisticamente do método DISi (KNU-dRoC).

Para esta proposta executamos duas vezes o mesmo experimento, sendo a única diferença entre eles o algoritmo para a estimativa da região de competência: kRoC e dRoC. Os datasets utilizados nesta análise foram os mesmos utilizados no trabalho inicial (TRIERVEILER et al., 2018). A primeira análise foi baseada no valor de Wilcoxon entre essas duas abordagens, e abaixo é reportado o p-value desta comparação.

$$\text{p-value} = 0.002887 \tag{4.1}$$

Este resultado da Análise 4.1 mostra um $\alpha < 0.01$, que significa que podemos descartar a hipótese H_0 indicando que estes dois métodos são estatisticamente diferentes. Este resultado foi obtido através dos valores mostrados ao final deste trabalho, nas Tabelas 18 e 17.

Tabela 6 – Teste estatístico de *Friedman* para todos algoritmos. As duas primeiras colunas consideram o algoritmo DISi enquanto que as duas últimas colunas representam o valor do *rank Test* sem o algoritmo DISi.

ALG	Rank Test	ALG	Rank Test
DISi	4.43 (2.62)	KNN	4.57 (3.24)
KNN	5.17 (3.47)	KNOP	5.17 (3.27)
KNOP	5.8 (3.57)	MTD	5.73 (3.5)
MTD	6.43 (3.85)	KNU	5.87 (2.76)
KNU	6.57 (2.97)	CLU	6.03 (4.06)
CLU	6.67 (4.36)	API	6.33 (2.87)
API	7.03 (3.24)	KNE	6.93 (4.14)
KNE	7.6 (4.52)	OLA	7.1 (3.77)
OLA	7.87 (4.03)	MCB	7.9 (4.3)
MCB	8.67 (4.57)	RRC	8.8 (4.54)
RRC	9.63 (4.74)	CC	9.97 (4.45)
CC	10.83 (4.65)	DSOC	11.13 (3.6)
DSOC	12.1 (3.66)	Rank	12.37 (4.06)
Rank	13.27 (4.23)	APO	13.9 (3.19)
APO	14.9 (3.19)	MLA	13.97 (2.3)
MLA	14.97 (2.3)	LCA	14.37 (1.96)
LCA	15.37 (1.96)	SB	15.03 (1.79)
SB	16.03 (1.79)		

Também foi estimado o teste de postos sinalizados de Friedman (Friedman rank test) (FRIEDMAN, 1937) comparando o `DISi` e `KNU-kRoC`. A comparação exibida na Tabela 6 mostra quatro colunas, sendo que as primeiras duas à esquerda mostram os *ranks* dos métodos e seus valores, enquanto que as duas últimas colunas à direita mostram o mesmo só que sem o método `DISi`. Vemos que o comportamento do `KNU` não se alterou em nenhuma das colunas. Vemos também na tabela que o experimento executado com `DISi` ficou em um patamar diferente do `KNU`.

Através dos resultados dos experimentos desta seção, podemos responder a questão:

- O método DISi é diferente do método original KNU?
- **Resposta**: Encontramos indícios que mostram que estes dois métodos são diferentes entre sí, através do comportamento dos resultados dos métodos motrados acima. Esta análise foi importante para podermos continuar com os próximos passos desta pesquisa. Agora que foi feito o levantamento da diferença entre os métodos quando executados com kRoC e dRoC, continuaremos a investigação de uma nova medida de qualidade de uma região de competência.

4.3 Qualidade da região de competência: **q-roc**

Considerando o índice de acertos na região local como um critério crítico para os métodos de seleção dinâmica estimar a competência dos classificadores, é proposto uma medida de qualidade para a região de competência que considere a acurácia dos classificadores. Esta medida iremos chamar de `q-roc`, que será explicada a seguir. Antes, gostaríamos de salientar que esta é uma medida realizada após o experimento, nos mesmos moldes que a acurácia é estimada. Ou seja, esta medida irá mostrar um valor de qualidade após finalizado um teste. Isso se torna importante porque iremos comparar dois conjuntos: o conjunto de classificadores-considerados-competentes e o conjunto de classificadores-realmente-competentes para predizer uma instância desconhecida. Vejamos a seguir.

Pensando em investigar o poder de informação de uma região de competência, consideramos a seguinte motivação:

- Como poderíamos verificar se uma região de competência foi apropriada para estimar a competência de classificadores?
- Como medir a qualidade do conjunto de instâncias que foram selecionadas para compor a `RoC`?

Para responder a estas perguntas, utilizaremos a definição de oráculo a partir de diferentes pontos de vista. De acordo com (BRITTO; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018), **Oráculo** pode ser considerado como uma medida que representa a proporção de instâncias em um conjunto de dados que podem ser corretamente prevista por pelo menos um classificador no `POOL` e também indica o possível **limite superior** de um sistema de seleção dinâmico.

Com intuito de exemplo, vamos pegar uma instância qualquer $\mathtt{x}_q$ e uma região de competência ao seu redor: `RoC`. Com isso, consideramos como: oráculos locais, os classificadores do `POOL` que conseguem predizer corretamente ao menos uma instância dentro da `RoC`. Dentro desta análise *apriori*, chamaremos este grupo de classificadores de **conjunto-A**. Consideramos esta análise como sendo: *apriori*, porque temos uma informação antecipada sobre a classe das instâncias que pertencem à `RoC`.

De maneira análoga, após a instância $\mathtt{x}_q$ ser predita, podemos realizar uma análise que consideramos como sendo *aposteriori*, ou seja, uma análise feita após sabermos a classificação correta de $\mathtt{x}_q$. Nesta segunda análise, também podemos investigar quais foram os classificadores do `POOL` que contribuíram para a real-classificação de $\mathtt{x}_q$. Chamaremos este segundo conjunto de classificadores de: **conjunto-B**. Consideramos este segundo grupo como *aposteriori* porque ele possui informações após identificar a classe correta de $\mathtt{x}_q$.

Com isso, esperamos analizar dois pontos de vista simultaneamente: (A) o conjunto de classificadores que conseguem predizer ao menos 1 instância da `RoC` e (B) como sendo o conjunto de classificadores que conseguem predizer $\mathtt{x}_q$ corretamente. A Figura 10 mostra um exemplo destes grupos: conjunto-A e conjunto-B.

A Figura 10 mostra $\mathtt{x}_q$ representado pelo triângulo e o círculo a sua volta mostra a região de competência. O cojunto-A indica os classificadores competentes para predizer corretamente as instâncias da `RoC` e o conjunto-B mostra os classificadores competentes para predizer $\mathtt{x}_q$. A intersecção entre as regiões $A \cap B$ indica a probabilidade de um algoritmo de seleção dinâmica encontrar um classificador correto para $\mathtt{x}_q$ a partir do conjunto-A.

Considerando esta intersecção dos conjuntos *aPriori* e *aPosteriori*, propomos uma nova medida: qualidade-RoC, ou apenas `q-roc`. A Equação 4.2 mostra esta medida, onde: o número de classificadores que participam da intersecção $|A \cap B|$, dividido pelo número de classificadores no conjunto-A. Desta forma, `q-roc` indica qualidade média de uma região de competência e varia entre: $[0, 1]$.

$$q_{roc} = \frac{|A \cap B|}{|A|}, \quad [0 \leq q_{roc} \leq 1] \tag{4.2}$$

A principal ideia desta medida `q-roc` é avaliar a possível qualidade da `RoC` a partir das intersecções de outras duas análises. Com isso, poderemos avaliar a qualidade de diferentes regiões de competência fornecidas por diferentes algoritmos.

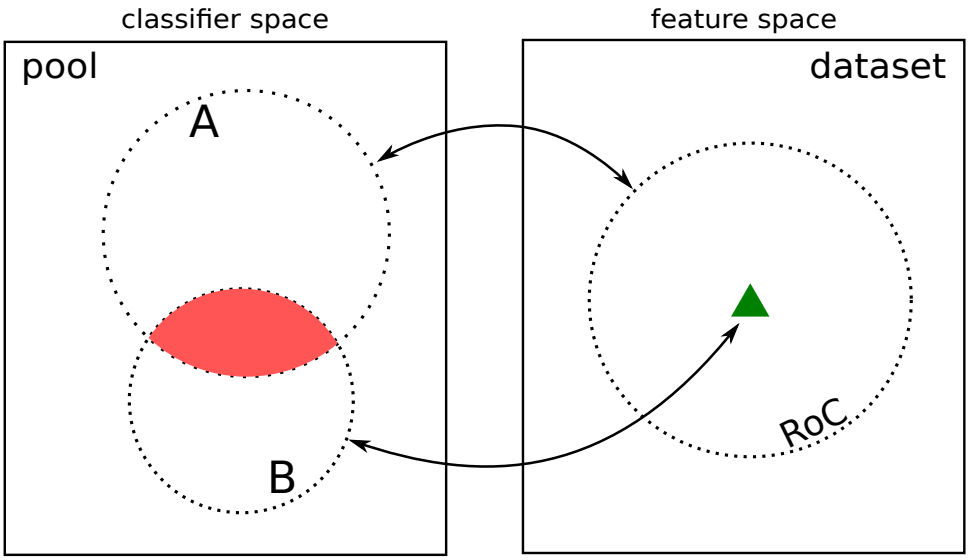


Figura 10 – Análise dos conjuntos A e B no espaço de classificação relacionados com instâncias da região de competência e instância desconhecida ($\mathtt{x}_q$).

Quanto maior o valor da interseção: $|A \cap B|$, maior o valor de `q-roc` e consequentemente maior o número de classificadores realmente-competentes que conseguem predizer $\mathtt{x}_q$. Chamaremos de classificadores realmente-competentes o conjunto de classificadores que conseguem predizer corretamente $\mathtt{x}_q$.

Nosso argumento que justifica a adoção de uma métrica de qualidade é enfatizado abaixo:

- Considerando duas regiões de competências estimadas a partir da mesma $\mathtt{x}_q$, um algoritmo de seleção dinâmica terá uma maior chance de predizer corretamente esta instância desconhecida se tiver um `q-roc` superior que outra região de competência em questão. Ou seja, quanto maior o conjunto de classificadores pertencentes à intersecção de $|A \cap B|$, maior a chance de um algoritmo de seleção dinâmica ter uma performance superior em relação à ele próprio se considerasse uma `RoC` com `q-roc` mais baixo.

Vejamos a seguir uma análise da interseção entre os conjuntos A e B.

4.3.1 Insersecção entre os conjuntos A e B = ∅

Caso a intersecção seja vazia: $|A \cap B| = \emptyset$, como mostrado na Figura 11, sabemos que o valor `q-roc` $= \mathbf{0}$, e nestes casos os algoritmos de seleção dinâmica não serão capazes de encontrar um classificador competente para predizer corretamente $\mathtt{x}_q$ (a menos que faça isso randomicamente). Este valor nulo de `q-roc` pode acontecer

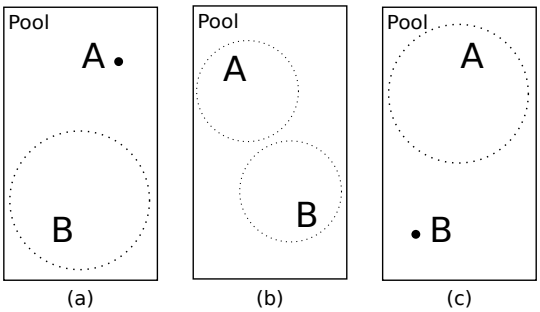


Figura 11 – Exemplos onde não existe intersecção entre conjuntos A e B: (A∩B)=∅, ou seja: q-roc=0. Na Fig(a) temos um exemplo do conjunto A vazio: $A = \emptyset$ and $0 < |B| < |Pool|$. Na Fig(c) temos um exemplo do conjunto B vazio, ou seja, $B = \emptyset$ and $0 < |A| < |Pool|$.

porque os métodos DS estimam a competência dos classificadores associadas com alguma métrica, e.g. acurácia. Vejamos os casos abaixo:

1. $|A| = 0$, Figura 11(a), ou seja, não existem classificadores que consigam predizer corretamente uma instância que pertença à RoC. Neste caso, mesmo que existam classificadores realmente-competentes (conjunto-B), eles não conseguirão ser selecionados pelos métodos de seleção dinâmica;

2. $|A| \neq 0$, $|B| \neq 0$, $|A \cap B| = 0$, Figura 11(b), ou seja, existem classificadores competentes que conseguem estimar as instâncias da RoC e também existem classificadores competentes que conseguem estimar $\mathbf{x}_q$, mas não se consegue estimar nenhum classificador do conjunto-B a partir do conjunto-A. Em outras palavras, não existe intersecção entre o conjunto-A e conjunto-B.

3. $|B| = 0$, Figura 11(c), ou seja, mesmo havendo classificadores que consigam predizer corretamente alguma instância da RoC, não existem classificadores competentes que consigam predizer $\mathbf{x}_q$. Este seria o caso onde uma instância $\mathbf{x}_q$ não consiga ser predita independente do que seja feito, a menos que seja gerado classificadores de maneira dinâmica, como é o caso do trabalho de: (SOUZA et al., 2019).

4.3.2 Insersecção entre os conjuntos A e B $\neq \emptyset$

Por outro lado, se a intersecção for máxima: $\frac{|A \cap B|}{|A|} = 1$, teremos como valor $\texttt{q-roc} = \mathbf{1}$ como mostrado na Figura 12. Isso significa que a maioria (ou talvez todos) métodos de seleção dinâmica (que utilizam RoC no espaço de características) irão classificar corretamente $\mathbf{x}_q$, independente do seu algoritmo de estimativa de competência.

Isso significa que a medida `q-roc` pode nos fornecer alguns *insights* sobre a qualidade de uma região de competência. A Figura 12 nos mostra 3 situações distintas que podem acontecer a partir da situação proposta na Figura 10. Em duas destas situações, o valor de `q-roc` $= \mathbf{1}$, vejamos:

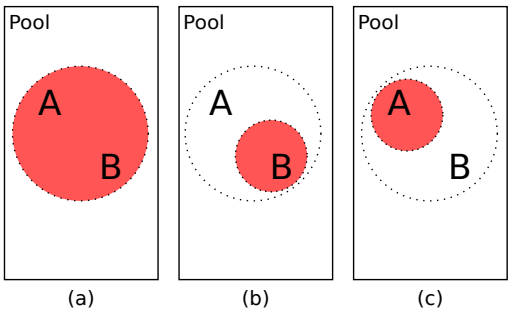


Figura 12 – Intersecção dos conjuntos *apriori* e *aposteriori* (conjuntos A e B). O score é máximo nas figuras (a) e (c), ou seja, q-roc=1. Na figura (b) o número do score varia entre: $0 < qroc < 1$.

1. $|A| = |B|$, Figura 12(a) mostra uma situação onde o conjunto-A é igual ao conjunto-B, ou seja, todas as instâncias que pertencem à `RoC` fornecem informações suficientes para que classificadores sejam escolhidos e consigam classificar corretamente $\mathtt{x}_q$. Este é o exemplo ideal que é buscado pelos algoritmos de seleção dinâmica;

2. $|A \cap B| = |B|$, $B \subset A$, Figura 12(b), mostra uma situação onde o conjunto-B é um sub-conjunto de A, ou seja, o conjunto-B está contido no conjunto-A. Nestes casos um sistema de seleção dinâmica terá uma máxima probabilidade de escolher um classificador correto igual ao tamanho do conjunto-B dividido pelo conjunto-A. Esta situação é a mesma do que já foi comentado na Equação 4.2. No entanto, sabendo que o conjunto-B é um subconjunto de A, a equação seria simplesmente: $q_{roc} = \frac{|B|}{|A|}$.

3. $|A \cap B| = |A|$, $A \subset B$, Figura 12(c) encontramos uma situação onde o conjunto-B é maior que o conjunto-A, ou seja, o número de classificadores realmente-competentes para classificar $\mathtt{x}_q$ é maior do que o número de classificadores que acertam as instâncias da região de competência. Nestes casos, da mesma forma que a Figura 12(a), qualquer classificador que acerte ao menos uma instância da `RoC` será competente para predizer $\mathtt{x}_q$.

4.3.3 Discussão

Foi levantado nos primeiros três exemplos da Figura 11, situações onde a intersecção entre as análises *aPriori* e *aPosteriori* é nula. Considerando que muitos dos algoritmos de seleção dinâmica possuem uma dependência da região de competência para escolher os classificadores mais promissores, como mostrado na Tabela 4, de nada adiantaria o melhor dos algoritmos DS nestes casos da figura 11.

Em todos os seis casos levantados anteriormente, o único caso em que poderíamos verificar o desempenho de um algoritmo de seleção dinâmica sobre outro seria na situação da figura 12(b). Além deste caso, no início desta seção, figura 10, é mostrada uma imagem onde possa haver melhorias de desempenho a partir de algoritmos de seleção dinâmica. Nesta figura 10, vemos que o conjunto-A e conjunto-B possuem uma intersecção e o aumento desta intersecção é uma das investigações que os algoritmos de seleção dinâmica buscam.

Gostaríamos de deixar aqui registrado que o **desempenho de um algoritmo de seleção dinâmica depende em vários casos desta intersecção entre os conjuntos** $A \cap B$. Este é um entendimento importante pois grande parte das comparações entre algoritmos de seleção dinâmica que foram estudados no capítulo 3 vem sendo feitas sem estas devidas considerações.

Ao comparar um algoritmo de seleção dinâmica com outro, muitas métricas já foram propostas, sendo a pesquisa (DEMŠAR, 2006) bastante citada. Este estudo investiga a comparação de algoritmos de classificação sobre múltiplos datasets, no entanto o artigo não propõe nenhuma medida para métricas de sistemas de seleção dinâmica que possuem um `POOL` de classificadores construídos. Esta é uma das ideias que deu origem ao estudo da seção 4.4 que veremos adiante.

4.3.4 Qualidade de todas `RoC`: Score QR

Esta medida foi introduzida para mostrar uma visão geral de todas q-roc. A Eq. 4.3 nos mostra a média de todas q-roc estimadas para um determinado dataset.

$$QR = \frac{\sum \text{q-roc}}{|Te|} \tag{4.3}$$

Onde a $\sum$ q-roc representa a soma de todas medidas q-roc do conjunto de teste, e $|Te|$ representa a cardinalidade do conjunto de Teste, ou seja, o tamanho do Test set. A medida QR varia entre $[0, 1]$, e quanto maior este valor, maior é a intersecção entre os conjuntos: A e B e com isso, melhor a chance do algoritmo de seleção dinâmica obter uma boa performance. Este score reporta a probabilidade de um método de seleção dinâmica selecionar corretamente um classificador, considerando que existe classificadores realmente-competentes na `RoC`. A seção de experimentos 5 mostra a estimativa desta medida para os vários métodos de seleção dinâmica.

4.4 **Limite-Inferior** de um POOL de classificadores

De acordo com (BRITTO; SABOURIN; OLIVEIRA, 2014), o limite superior não fornece toda informação necessária para estimar a qualidade do POOL de classificadores. Nesta seção estamos interessados em conhecer o poder de informação que um POOL de classificadores possui.

O desempenho de um algoritmo de seleção dinâmica depende em vários casos da escolha de instâncias para compor a região de competência que consigam fornecer uma maior confiabilidade na escolha dos classificadores presentes no POOL.

Como já mostrado anteriormente na Figura 10, queremos aumentar a intersecção do conjunto de classificadores-considerados-competentes (descrito anteriormente como conj-A) com o conjunto dos classificadores-realmente-competentes (descrito como conj-B). Veremos nesta seção uma análise do **limite inferior** como sendo o conjunto de instâncias que são corretamente preditas por todos classificadores presentes no POOL.

Entre várias das medidas utilizadas em sistemas de múltiplos classificadores, uma bastante importante é o Oráculo. Este Oráculo pode ser considerado como uma medida de um modelo hipotético de seleção de classificadores ideal. A ideia foi proposta por (KUNCHEVA, 2002) e adotada por vários pesquisadores que considera o limite superior (também chamado de: upper limit) de um POOL de classificadores. Apesar do desempenho de um MCS poder ser maior que o Oráculo em teoria, ele ainda representa o alvo a ser perseguido por um sistema de múltiplos classificadores. Abaixo apresentamos a Figura 13 que mostra um esquemático do limite superior. Esta imagem representa dois POOLs formados a partir do mesmo dataset. A diferença no limite-superior mostrado para cada exemplo é diferente devido a natureza randomica com que os classificadores são criados.

Dado um determinado POOL de classificadores, o Oráculo vem sendo utilizado para avaliar o máximo desempenho de um MCS sob este POOL. Esta informação nos mostra uma lacuna que pode ser melhorada entre os sistemas atuais e o Oráculo.

Recentemente o trabalho de (SOUZA et al., 2017) fez uma série de experimentos onde garantia um POOL que contivesse o Oráculo com acurácia de 100%. No entanto, foi constatado que mesmo assim não houveram significativas melhoras no desempenho nos algoritmos MCS utilizados em cima dos datasets testados. A conclusão deste trabalho foi que talvez o Oráculo com desempenho de 100% não seja a melhor tarefa de pesquisa a ser perseguida na busca de um melhor POOL de classificadores.

Sabendo que o Oráculo vem sendo citado como como limite-superior do POOL de classificadores, gostaríamos de propor uma nova métrica para avaliação do POOL que seja responsável por fornecer um limite-inferior. Também estamos interessados em como esta métrica pode ajudar os sistemas de múltiplos classificadores. O limite-inferior pode nos fornecer informações sobre a acurácia mínima de um método

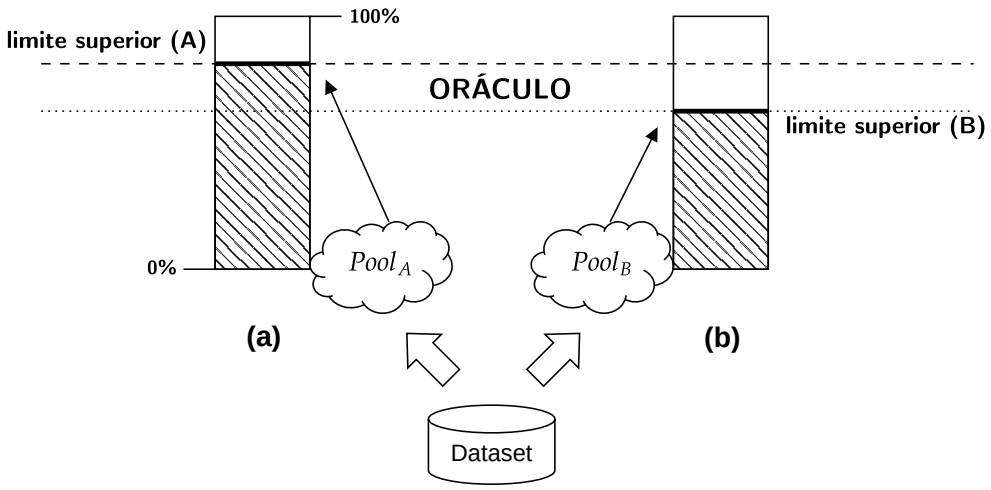


Figura 13 – Limite-Superior.

de seleção dinâmica. Este valor nos informa uma medida de complexidade dos dados em relação ao POOL de classificadores construído, ou seja, é uma medida que está relacionada com a caracterização da qualidade do POOL de classificadores.

Olhando mais a fundo para esta medida, podemos enxergar que não existe uma maneira de um um método de seleção dinâmica ter uma performance pior que o limite-inferior. Ou seja, se todos classificadores no POOL podem predizer corretamente uma instância, então todos os métodos de seleção dinâmica terão a mesma resposta em relação à instância em questão. Nestes casos vemos que não existe nenhum mérito na escolha de classificadores pelo método de seleção dinâmica. Isso pode dificultar o rankeamento de algoritmos dado um problema, ou então, pode dificultar a estimativa de qualidade de um sistema em relação à outros.

Refletindo melhor sobre esta medida, trouxemos um exemplo para ilustrar o caso. A Imagem 14 mostra um esquemático de dois POOLs de classificadores sendo construídos a partir do mesmo dataset, conforme apresentado anteriormente na Figura 13. No entanto nesta figura 14 destacamos o **poder de impacto** de um POOL de classificadores estimado a partir da diferença entre limite-superior e limite-inferior, representado na figura 14 por Delta: Δ.

Vemos que em cada um dos POOLs temos duas linhas representando a porcentagem de instâncias máximas e mínimas que conseguem ser preditas pelos classificadores. No POOL_A vemos uma diferença pequena entre o limite-superior e o limite-inferior. Já no POOL_B vemos que esta diferença é maior, mesmo possuindo um limite-superior mais baixo.

Gostaríamos de mostrar com esta imagem 14 que mesmo o Oráculo nos mostrando um limite-superior de um POOL para os algoritmos de seleção dinâmica, ainda

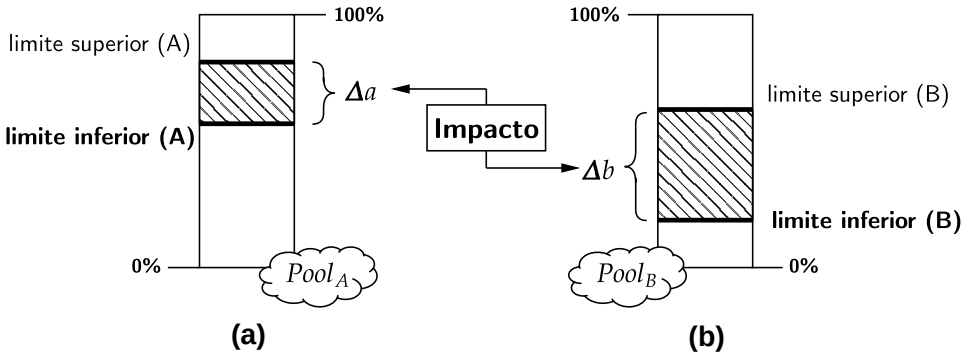


Figura 14 – Limite-Inferior.

assim precisamos avaliar qual o mínimo de instâncias que este POOL conseguirá predizer.

Quanto mais próximo estes dois limites, menor será a diferença de desempenho de um algoritmo de seleção dinâmica em comparação à outro. Com isso menor a possibilidade de comparar diferentes algoritmos. Ou em outras palavras, menor a confiança nesta comparação.

Isto acontece porque a margem de diferença entre dois algoritmos de seleção dinâmica será estimada dentro da faixa de valores Δ. Será este Δ que mostrará o poder de impacto de um método de seleção dinâmica. Caso obtenhamos uma margem pequena deste impacto Δ, menor as chances de conseguirmos diferenciar os métodos de seleção dinâmica.

Esta ideia pode ajudar a conseguirmos métricas mais fiéis do que apenas o número de acertos (acurácia) ou outra medida de desempenho principalmente para sistemas de seleção de múltiplos classificadores que tem uma forte dependência de um POOL de classificadores. Sabendo das medidas de oráculo $\mathcal{O}$ e limite-inferior LI, poderemos estimar a faixa do oráculo Δ conforme Eq. 4.4 abaixo.

$$\Delta = \mathcal{O} - \texttt{LI} \tag{4.4}$$

4.4.1 Exemplo

Se pegarmos como exemplo, um POOL de classificadores onde o $\mathcal{O}$= 70% e LI= 70%, significa que Δ= 0 e desta forma todos os algoritmos de seleção dinâmica terão a mesma performance de 70%. Isto pode acontecer em casos onde existe alguma região no dataset que pode ser considerada fácil e todos classificadores do POOL conseguem acertar qualquer instância daquela região. No entanto em outras regiões consideradas muito difíceis, não existe classificador que consiga

predizer corretamente uma instância. Isto pode acontecer também em problemas desbalanceados e/ou em sistemas que possuam o número de instâncias muito pequeno, um dos focos deste trabalho.

Mesmo esta situação do exemplo acima podendo acontecer, ainda assim é um exemplo hipotético. Não vimos em nossos experimentos nenhum exemplo onde o limite-superior é igual ao limite-inferior. No entanto, verificamos que em alguns momentos o limite-inferior é muito próximo do Oráculo, fazendo com que a performance dos sistemas de seleção dinâmica sejam muito próximas como mostraremos na investigação prática a seguir Seção 5.3.

4.4.2 Conjunto de **N**-classificadores: $\mathbb{C}_N$

O Oráculo pode ser entendido como sendo um sistema *ideal* que sempre irá selecionar o classificador correto para predizer uma instância desconhecida, caso exista. Gostaríamos de analisar melhor a medida do oráculo e também propor maneiras de enxergá-lo.

Vemos que para cada instância existe um conjunto de classificadores no `POOL` que conseguem respondê-la corretamente. Desta forma, chamaremos de $\mathbb{C}_{\mathbf{1}}$: o número de instâncias que conseguem ser classificadas corretamente por 1% dos classificadores do `POOL`; $\mathbb{C}_{\mathbf{2}}$: o número de instâncias que conseguem ser preditas por 2% dos classificadores, e assim por diante. Desta forma, teremos que $\mathbb{C}_{\mathbf{100}}$ será o número de instâncias que conseguem ser preditas por todos classificadores do `POOL` e $\mathbb{C}_{\mathbf{0}}$ como número de instâncias que não conseguem ser preditas por nenhum classificador deste `POOL`.

Cabe aqui lembrar que estamos realizando experimentos com um `POOL` de tamanho igual a 100 classificadores e desta forma, os valores aqui representados por $\{\mathbb{C}_{\mathbf{0}}, \mathbb{C}_{\mathbf{1}}, ..., \mathbb{C}_{\mathbf{100}}\}$ corresponde ao número de instâncias que são preditas corretamente por cada número de classificadores do `POOL`.

Esta interpretação será importante para nós porque iremos estimar nossos cálculos de Oráculo e limite-inferior baseado nestes valores de: $\mathbb{C}_{\mathbf{0}}$, $\mathbb{C}_{\mathbf{1}}$ e $\mathbb{C}_{\mathbf{100}}$.

Considerando que o oráculo é uma medida que mostra quantas instâncias conseguem ser preditas por pelo menos um classificador do `POOL`, entendemos que: o oráculo $\boldsymbol{\mathcal{O}}$ é na verdade uma medida estimada entre a diferença do número de instâncias que não conseguem ser preditas por nenhum classificador ($\mathbb{C}_{\mathbf{0}}$) e o número total de todas instâncias. A Eq. 4.5 a seguir mostra esta estimativa.

$$\boldsymbol{\mathcal{O}} = \mathbf{1} - \mathbb{C}_{\mathbf{0}} \tag{4.5}$$

Que significa dizer que o Oráculo é igual ao total de instâncias 100% (representado na função pelo número: 1) menos a proporção de instâncias que não conseguem ser corretamente classificadas por nenhum classificador do `POOL`: $\mathbb{C}_{\mathbf{0}}$.

Uma outra forma de escrever o oráculo dentro desta análise seria considerá-lo como sendo o somatório de todos valores de $\mathbb{C}_j$ sendo: $\mathbf{1 <= j <= 100}$, ou seja: $\boldsymbol{\mathcal{O}} = \sum_{j=1}^{100} \mathbb{C}_j$.

Esta forma de considerar o cálculo foi motivado pelo limite-inferior `LI`, onde enxergamos que sua estimativa pode ser feita através da Eq. 4.6 ou da Eq. 4.7. Ou seja, o limite-inferior é a proporção de instâncias onde todos classificadores irão fornecer a mesma resposta. Nestes casos, podemos obter como resposta o valor de $\mathbb{C}_{\mathbf{0}}$ e/ou $\mathbb{C}_{\mathbf{100}}$.

$$\texttt{LI} = \mathbb{C}_{\mathbf{0}} + \mathbb{C}_{\mathbf{100}} \tag{4.6}$$

$$\texttt{LI} = \mathbb{C}_{\mathbf{100}} \tag{4.7}$$

Enxergamos que estas duas funções podem ser utilizadas para o cálculo do limite-inferior, sendo que a Eq. 4.6 considera as instâncias que não são preditas por nenhum classificador do `POOL` ($\mathbb{C}_{\mathbf{0}}$) e a Eq. 4.7 considera apenas as instâncias que conseguem ser preditas por todos classificadores ($\mathbb{C}_{\mathbf{100}}$).

A principal ideia destas estimativas e também da consideração dos acertos $\mathbb{C}_{\boldsymbol{N}}$, está na possibilidade de conseguirmos enxergar o número de classificadores que conseguem acertar um determinado conjunto de instâncias.

Como exemplo, trazemos uma interpretação a partir de nossos registros de log. Nestes registros encontramos repetidas vezes problemas difíceis de serem resolvidos, como por exemplo o dataset Haberman, onde em algumas situações obtivemos um número muito alto de $\mathbb{C}_{\mathbf{1}}$. Isso indica que em várias situações teríamos apenas um classificador que conseguiria acertar uma instância desconhecida.

Conforme comentado anteriormente, vemos que já existiram outras pesquisas, como é o caso de (SOUZA et al., 2017), onde garantiram um `POOL` com Oráculo de 100%, mas que tiveram o mesmo resultado considerando um `POOL` gerado randomicamente com Oráculo menor que 100%. Isso mostra que muito provavelmente o `POOL` gerado com Oráculo de 100% obteve um valor muito alto de instâncias que são preditas por poucos classificadores. Esta poderia ser uma das respostas para o caso do trabalho citado.

Na investigação prática a seguir iremos mostrar mais informações que foram coletadas dos experimentos com o dataset Haberman.

4.5 Considerações Finais

Dentro desta seção foram feitas várias reflexões a partir das informações coletadas do método proposto. Em um primeiro momento foram feitas associações que traçaram um paralelo entre as áreas de aprendizagem de máquina e psicometria.

Depois foi apresentado o método proposto e seu algoritmo. Para que pudéssemos obter informações iniciais, foi implementado método proposto e mais outros dois métodos para extração da RoC que considera um tamanho de região de competência parecidos com esta proposta.

Foi proposto também uma nova medida para avaliar a qualidade da região de competência: q-roc. Ainda neste captítulo foi mostrado a proposta do limite-inferior que fornece uma métrica complementar ao Oráculo em sistemas de seleção dinâmica de classificadores. Nesta parte do texto também foi apresentado uma nova maneira de enxergar o cálculo do oráculo e também do limite-inferior.

5 Experimentos

De acordo com as hipóteses levantadas, acreditamos que a teoria de testes psicométricos podem nos fornecer insights para avaliação e seleção de instâncias em uma região de competência a fim de possibilitar com que os algoritmos de seleção dinâmica possuam uma `RoC` de melhor qualidade. Desta forma os methodos DS podem estimar competência dos classificadores de forma mais confiável, trazendo um melhor score na predição de instâncias desconhecidas.

O primeiro experimento proposto foi a união de (1) um sistema de seleção dinâmica já bastante consolidado (KNORA-Union (KO; SABOURIN; BRITTO, 2008)) com (2) a proposta principal desta pesquisa: uma região de competência baseada em índice de discriminação: dRoC. Na ocasião, a união destes algoritmos foi chamado de: Dynamic Ensemble Selection by K-Nearest Local Oracles with Discrimination Index (DES-DISi) (TRIERVEILER et al., 2018). Ou seja, podemos resumir que o metodo DISi é a união do KNU com um método de pré-processamento da região de competência.

Este primeiro experimento foi realizado por verificarmos bons resultados fornecidos pelo KNORA-Union e por estar presente em várias das pesquisas realizadas com seleção dinâmica como por exemplo o trabalho de (VRIESMANN et al., 2010).

Foram obtidas análises interessantes a partir deste primeiro trabalho e com ele foram levantados outras questões:

1. Como medir a qualidade de uma região de competência?
2. Sabendo que o DISi captura $\boldsymbol{n * 2}$ vizinhos mais próximos e depois seleciona as instâncias com maior potencial de discriminação, como seria a comparação com um sistema de captura de região de competência que faça algo parecido de maneira randômica?
3. Como poderíamos enxergar de fato o que acontece com uma região de competência ao selecionarmos as instâncias com maior discriminação?
4. O índice de discriminação funciona na maioria dos casos?

5.1 Protocolo Experimental e Parâmetros

Nesta seção encontramos uma descrição do protocolo experimental e seus parâmetros. Os experimentos foram realizados seguido o mesmo protocolo experimental que já adotado em outros trabalhos: (CRUZ; SABOURIN; CAVALCANTI, 2018;

BRUN et al., 2016; SOUZA et al., 2017). Alguns trabalhos não utilizam todos os parâmetros aqui listados devido à natureza intrínseca de cada problema. A seguir, os parâmetros de nossos experimentos.

Os métodos de seleção dinâmica utilizados foram selecionados especificamente para cada experimento. No entanto é possível conferir os métodos e sua abreviação na Tabela 4, no capítulo 3. Além dos algoritmos listados, também propomos a comparação do método proposto com a performance do oráculo, considerado-o como limite-superior de cada simulação. Também foram utilizados dois métodos de seleção estática: Combinação de todos Classificadores (CC) e *Single Best* (SB) ambos recomendados no trabalho (BRITTO; SABOURIN; OLIVEIRA, 2014). A biblioteca DesLib (CRUZ et al., 2018) foi usada como implementação da maioria dos métodos listados.

Os experimentos foram conduzidos com **30 datasets** a partir de diferentes fontes: UC Irvine Machine Learning Repository (UCI) (DHEERU; TANISKIDOU, 2017), Knowledge Extraction Evolutionary Learning (KEEL) (ALCALÁ-FDEZ et al., 2011), Ludmila Kuncheva Collection (LKC) sobre dados médicos (KUNCHEVA, 2004) e também dados da biblioteca Matlab (PRTOOLS) (DUIN et al., 2000). Estes datasets também foram utilizados em vários outros experimentos: (CRUZ; SABOURIN; CAVALCANTI, 2018; TRIERVEILER et al., 2018) e estão sendo mostrados na Tabela 7. Estes datasets foram utilizados por serem considerados de problemas difíceis, com diferentes números de características, diferentes números de classes e diferentes desbalanceamentos (`IR`).

A Tabela 8 mostra os parâmetros utilizados para cada simulação. Para cada dataset, foram feitas 20 replicações. Para cada replicação, cada dataset foi particionado em Treinamento, Dsel (validação) e Teste, com as respectivas proporções: Tr=0.5; Dsel=0.25; Te=0.25. Cada partição foi feita randomicamente e estratificada, ou seja, respeitando as proporções de classes.

Os classificadores foram treinados por amostragem tipo bootstrap do conjunto de Treinamento. Cada amostra do bootstrap possui uma proporção de 50% do conjunto de Treinamento que significa um total de 25% do dataset total. Cada bootstrap foi feito de acordo com o processo de reamostragem, ou seja, uma instância pode participar mais de uma vez na amostra.

Cada uma das bases de dados utilizadas foram normalizadas entre os valores [0, 1] por uma função de MinMax, como mostra a Equação: $\boldsymbol{x} = \frac{(\boldsymbol{x}-\boldsymbol{x}_{min})}{(\boldsymbol{x}_{max}-\boldsymbol{x}_{min})}$.

Para os métodos que utilizaram região de competência, foi aplicado k-vizinhos mais próximos (`k-NN`) com distância Euclidiana. Caso o dataset pertença a um problema multi-classe, foi feita a decomposição um-versus-todos (one-versus-all: OvA), proposto por (LORENA; CARVALHO; GAMA, 2008).

O classificador utilizado foi o Perceptron em sua implementação feita por *Stocastic Gradient Descendent* (SGD) implementado na biblioteca numpy (OLIPHANT,

Tabela 7 – Lista de datasets selecionados

Dataset	Abbrev.	#Instances	#Features	#Classes	IR	Source
Adult	AD	690	14	2	1.25	UCI
Banana	BN	2000	2	2	1.00	PRTools
Blood	BL	748	4	2	3.20	UCI
CTG	CT	2126	21	3	9.40	UCI
Diabetes	DB	766	8	2	1.86	UCI
Ecoli	EC	336	7	8	71.50	UCI
Faults	FA	1941	27	7	12.24	UCI
German	GE	1000	24	2	2.33	STATLOG
Glass	GL	214	9	6	8.44	UCI
Haberman	HA	306	3	2	2.78	UCI
Heart	HE	270	13	2	1.25	UCI
ILPD	IL	583	10	2	2.49	UCI
Ionosphere	IO	351	34	2	1.79	UCI
Laryngeal1	L1	213	16	2	1.63	LKC
Laryngeal3	L3	353	16	3	4.11	LKC
Lithuanian	LT	2000	2	2	1.00	PRTools
Liver	LV	345	6	2	1.38	UCI
Magic	MG	19020	10	2	1.84	KEEL
Mammo	MM	830	5	2	1.06	KEEL
Monk	MO	432	6	2	1.12	KEEL
Phoneme	PH	5404	5	2	2.41	KEEL
Segmentation	SE	2310	19	7	1.00	KEEL
Sonar	SO	208	60	2	1.14	KEEL
Thyroid	TH	692	16	2	12.06	LKC
Vehicle	VH	846	18	4	1.10	STATLOG
Vertebral	VT	300	6	2	2.13	UCI
WDBC	WB	569	30	2	1.68	KEEL
WDVG	WV	5000	21	3	1.03	UCI
Weaning	WE	302	17	2	1.00	LKC
Wine	WI	178	13	3	1.48	UCI

Considerações – Abbrev: Abreviação do nome do método; IR Desbalanceamento do dataset (Imbalance ratio).

2007). Este indutor foi escolhido por ser considerado um classificador fraco e instável (SCHAPIRE, 1990) com um número de épocas máximo igual a 100. O indutor Perceptron ainda possui o reforço de outras pesquisas, como o caso de (NOURBAKHSH; HOSEINPOUR, 2017): *"Schapire proved that a strong classification can be obtained by weak classifiers combination"*.

O POOL foi gerado de forma homogênea e seu tamanho igual a 100 classificadores do tipo Perceptron (já comentado acima). Para cada replicação e para cada dataset foi construído um POOL diferente e todos algoritmos testados foram submetidos ao mesmo POOL para que seja possível uma comparação estatística ao final.

Tabela 8 – Lista de parâmetros utilizados.

Parâmetro	Descrição
Ambiente desenvolvimento	Python (ROSSUM, 1995) versão-3.8; Numpy (OLIPHANT, 2007); SciPy (JONES et al., 2001–); Scikit Learn (JONES et al., 2001–) (PEDREGOSA et al., 2011); Matplotlib (HUNTER, 2007) Estatística: PyCM (HAGHIGHI et al., 2018)
Métodos Seleção Dinâmica	DesLib (CRUZ et al., 2018) (version: v0.1-Jan / 2018)
Número Datasets	30 datasets (diferentes fontes, veja Tab: 7)
Número replicações	20 (para cada dataset)
Divisão do Dataset	Estratificado: Treino (50%), Validação (25%) e Teste (25%)
Instâncias de Treino	Algoritmo: Bootstrap; 50%; Stratified
Tamanho `POOL` Classificadores	100 classificadores
Tipo do `POOL`	Homogêneo
Tipo do Classificador	Perceptron - Stocastic Gradient Descendent (SGD)
Parâmetros do Classificador	Número de épocas: 100
Função de distância	k-vizinhos mais próximos, distância Euclideana
Número de vizinhos mais próximos	$\boldsymbol{k = 7}$
Normalização dos Dados	Equação MinMax

O tamanho da região de competência escolhida foi $\boldsymbol{k = 7}$ devido à pesquisas anteriores: (KO; SABOURIN; BRITTO, 2008; BRUN et al., 2016; CRUZ et al., 2015). Por fim, os métodos de seleção dinâmica foram configurados com voto majoritário como algoritmo de fusão. A Tabela 8 mostra um resumo dos parâmetros utilizados nos experimentos.

5.2 Estatísticas

De acordo com (GARCÍA et al., 2010) a análise de desempenho de vários indutores pode ser realizada a partir de dois pontos de vista: para apenas um dataset ou uma análise multi-datasets. Um estudo extenso feito por (DIETTERICH, 1998) mostra várias possibilidades para um dataset, no enanto seguiremos o estudo de (DEMŠAR, 2006) pela sua ênfase no estudo estatístico considerando múlti-classificadores e multi-datasets e também por sua importância em vários outros

trabalhos de seleção dinâmica (BRITTO; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018). Verifica-se que em seleção dinâmica a proposta de um novo algoritmo é normalmente validada em um conjunto de datasets em relação à seus pares, inclusive a teoria No-Free-Lunch (WOLPERT, 1996) que instiga os testes em mais de um problema.

Convém lembrar que a comparação de resultados estatísticos requer um intervalo de significância $\boldsymbol{\alpha}$ e que dentro deste trabalho adotaremos os níveis $\boldsymbol{\alpha} = 0.1, 0.05, 0.01$, com seguintes valores: $\boldsymbol{z_\alpha} = 1{,}28, 1{,}64, 2{,}32$.

Os testes estatísticos tentam provar, para um determinado nível de significância, a veracidade da hipótese nula $\boldsymbol{H_0}$. Caso o resultado estatístico não consiga comprovar a veracidade de $\boldsymbol{H_0}$, rejeita-se esta hipótese, ou seja, $\boldsymbol{H_0}$ é considerado falso para um determinado $\boldsymbol{\alpha}$. Outra medida bastante utilizada são os graus-de-liberdade (degrees of freedom, df), ou seja, o número de observações que são livres para variar quando estiver sendo feito uma estimativa estatística. Em alguns testes estatísticos com $\boldsymbol{n}$ observações, podemos dizer que 1 grau de liberdade $\boldsymbol{df = (n-1)}$ significaria que temos a liberdade para variar $\boldsymbol{(n-1)}$ amostras.

O estudo de (DEMŠAR, 2006) comenta que devem ser feitos experimentos suficientes para cada dataset e que os algoritmos sejam avaliados usando as mesmas amostras randômicas. O autor comenta dois tipos de comparação: (1) par-a-par, e (2) em grupo. Autor também comenta que não existe garantia de que os resultados de desempenho entre vários sistemas seguirão uma curva normal. Por este motivo iremos ivestigar os resultados através de testes não-paramétricos. Para os testes estatísticos pareados, abordaremos (I) postos sinalizados de Wilcoxon e (II) gráfico de ganho/empate/perda (`WTL`). Para os testes estatísticos em grupo, abordaremos (III) Friedman-rank-test e (IV) o gráfico Diferença Crítica (`CD`).

I **Teste de Wilcoxon**: Este é um teste não-paramétrico (WILCOXON, 1945) e cosidera seus parâmetros desconhecidos. Neste teste é feito um rank do desempenho de dois métodos. Depois é feito uma média destes *ranks* e o valor $\boldsymbol{z_w}$ é calculado. A hipótese nula $\boldsymbol{H_0}$ assume que os resultados são equivalentes. Esta hipótese é aceita se o valor estimado p-value$\geq \boldsymbol{\alpha}$.

II **Teste de Sinal** (`WTL`): Este teste estatístico (Sign test) é feito entre dois métodos (SHESKIN, 2003; SALZBERG, 1997). A hipótese nula $\boldsymbol{H_0}$ assume que dois classificadores possuem desempenho equivalente. Este cálculo `WTL` também mostra através de um número crítico $\boldsymbol{n_c}$ se a diferença no resultado de duas comparações é estatisticamente significante. O autor sugere o cálculo do valor crítico $\boldsymbol{n_c}$, dado pela Eq. 5.1 que significa a improbabilidade de $\boldsymbol{H_0}$, sendo o valor $\boldsymbol{n_{exp}}$ igual ao número de experimentos realizados.

$$\boldsymbol{n_c = \frac{n_{exp}}{2} + z_\alpha \frac{\sqrt{n_{exp}}}{2}} \tag{5.1}$$

III **Teste de Friedman** (Friedman-rank-test (FRIEDMAN, 1937; DEMŠAR, 2006)): Diferente do teste de sinal, este teste faz uma análise de todos-com-todos, ao invés de um-a-um como acima. A hipótese nula $\boldsymbol{H_0}$ assume que todos algoritmos possuem mesma média de desempenho para um determinado $\boldsymbol{\alpha}$. A equação original de Friedman faz um ranking dos algoritmos para cada dataset, sendo o melhor com rank: 1, o segundo rank: 2, e assim por diante. Caso haja empate é feito uma média dos *ranks*. Ao final, todos algoritmos terão um número de ordenação para cada dataset e então é calculado a média deste número para cada algoritmo em questão.

IV **Diferença Crítica** (`CD`): A partir da tabela de *distribuição-F* é possível extrair o valor *F-crítico* ($\boldsymbol{F_c}$), enquanto o valor *F-calculado* ($\boldsymbol{F_F}$) é utilizado para inferir $\boldsymbol{H_0}$. Desta forma, caso o $\boldsymbol{F_F} > \boldsymbol{F_c}$ então rejeita-se $\boldsymbol{H_0}$: as observações não possuem mesma média. A `CD` mostra vários procedimentos para verificar quais algoritmos possuem desempenho estatisticamente semelhantes, inclusive o autor (DEMŠAR, 2006) comenta sobre o gráfico `CD`, que mostra os procedimentos que são estatisticamente semelhantes conectados por uma barra.

5.3 Investigação prática

A partir dos questionamentos feitos durante este capítulo, gostaríamos de acrescentar esta investigação prática. Nesta análise verificamos o comportamento da seleção de instâncias por índice de Discriminação em relação à uma região de competência onde a vizinhança estimada apenas por distância.

Neste experimento os mesmos parâmetros foram utilizados e a única diferença foi a estimativa da `RoC`. Verificamos um impacto positivo no método `LCA` com `dRoC` comparado com o mesmo método com `kRoC`. Por esta razão escolhemos esta análise para uma investigação mais profunda do método `LCA`.

O método `LCA` foi adotado por dois motivos: (1) ele obteve um dos piores desempenhos em nosso trabalho anterior: (TRIERVEILER et al., 2018), e (2) durante a fase de escolha de classificador, caso haja mais de um classificador competente, ele escolhe randomicamente um destes. O problema neste tipo de abordagem é que durante a estimativa de competência, onde muitos classificadores que não classificariam corretametne $\mathtt{x}_{\boldsymbol{q}}$ podem receber uma estimativa alta e assim podem ser selecionados randomicamente. Isso faz com que o `LCA` seja frágil em situações onde existam muitos classificadores com estimativa de máxima competencia.

Outro aspeto interessante é que o `LCA` define a competência do classificador pelo número de respostas corretas nas instâncias da `RoC` que possuem a mesma classe

atribuída para $\mathtt{x}_{\boldsymbol{q}}$. Assim sendo, o LCA verifica se o classificador tem competência suficiente para classificar corretamente as instâncias da RoC que são da mesma classe.

Tabela 9 – Datasets listados pela diferença entre o algoritmo LCA considerando duas regiões de competência: uma contendo o algoritmo de índice de discriminação dLCA e outra estimada de maneira convencional kLCA. Estão sendo mostrados apenas os 10 primeiros resultados (de um total de 30 datasets) considerando a maior diferença entre as RoCs.

DATASET	ORA	LCA	dLCA	Diff
Haberman	74.29	52.57(21.95)	**63.22(14.66)**	10.658
Liver	97.91	52.91 (5.42)	**61.63 (4.73)**	8.721
Vertebral	99.60	73.73 (5.51)	**81.87 (4.38)**	8.133
Monk	98.57	70.79 (4.54)	**78.66 (3.81)**	7.870
German	99.34	64.92 (5.84)	**71.92 (1.77)**	7.000
Magic	97.99	76.23 (1.51)	**82.27 (0.47)**	6.038
Phoneme	97.59	75.60 (0.86)	**81.43 (1.01)**	5.829
WDVG	99.29	77.56 (1.89)	**82.50 (1.28)**	4.94
Faults	97.49	60.72 (3.04)	**64.85 (2.03)**	4.124
Mammo	98.97	75.46 (4.17)	**79.40 (2.62)**	3.937

Note: A coluna **Diff** mostra a diferença entre o método LCA estimado para duas regiões de competência. O teste de significância de Wilcoxon foi aplicado para 20 replicações. Os valores que estão sublinhados indica uma diferença significativa de pvalue < 0.05 e os números em negrito significam **pvalue < 0.01**. A lista com todos datasets e contendo mais informações pode ser conferida na Tabela 7.

Após a escolha do método, investigamos qual foi o dataset que obteve a maior diferença entre os resultados do LCA considerando as regiões de competência kRoC e dRoC. A Tabela 9 apresenta os 10 primeiros resultados ranqueados pela maior diferença entre estes métodos, mostrado na última coluna: Diff. A partir desta análise o dataset Haberman foi escolhido para realizar os experimentos desta seção. Vejamos abaixo algumas informações sobre este dataset.

Haberman é um dataset pequeno com pouco mais de 300 instâncias, 3 features e 2 classes, sendo o número de instâncias de cada classe igual a $\{225, 81\}$. Este dataset mostra os casos de pacientes em um hospital que sofreram intervenção cirúrgica. O dataset Haberman também foi escolhido por verificarmos que: é um problema difícil, com poucas instâncias, poucas características e moderadamente desbalanceado, sendo a razão de desbalanceamento igual a: IR= **2.78**. Baseado em nosso protocolo de experimentos, foram feitas 20 replicações para nosso teste.

Após a escolha do dataset Haberman, iniciamos a investigação de nossas variáveis já mencionadas anteriormente. A Tabela 10 mostra os resultados obtidos nestas 20 simulações. Foram coletadas informações do dataset `Dsel` e também do dataset `TEST` para verificarmos o comportamento de cada uma das variáveis.

Tabela 10 – Estimativa das informações de Oráculo, limite-inferior (IL), $\mathbb{C}_0$, $\mathbb{C}_1$ e $\mathbb{C}_{100}$ para os datasets `Dsel` e `TEST`.

	`Dsel`					`TEST`				
Rep	ORA	IL	$\mathbb{C}_0$	$\mathbb{C}_1$	$\mathbb{C}_{100}$	ORA	IL	$\mathbb{C}_0$	$\mathbb{C}_1$	$\mathbb{C}_{100}$
01	75.32	38.96	24.68	29.87	14.29	75.00	43.42	25.00	27.63	18.42
02	77.92	87.01	22.08	00.00	64.94	85.53	81.58	14.47	02.63	67.11
03	77.92	33.77	22.08	25.97	11.69	73.68	44.74	26.32	22.37	18.42
04	77.92	84.42	22.08	01.30	62.34	84.21	84.21	15.79	00.00	68.42
05	63.64	49.35	36.36	01.30	12.99	68.42	50.00	31.58	02.63	18.42
06	68.83	50.65	31.17	33.77	19.48	68.42	48.68	31.58	22.37	17.11
07	68.83	50.65	31.17	25.97	19.48	60.53	56.58	39.47	11.84	17.11
08	63.64	53.25	36.36	20.78	16.88	21.05	48.68	00.00	51.32	11.84
09	59.74	50.65	40.26	11.69	10.39	50.00	68.42	50.00	03.95	18.42
10	81.82	85.71	18.18	01.30	67.53	81.58	77.63	18.42	01.32	59.21
11	77.92	84.42	22.08	00.00	62.34	85.53	80.26	14.47	00.00	65.79
12	77.92	33.77	22.08	25.97	11.69	88.16	27.63	11.84	28.95	15.79
13	80.52	83.12	19.48	02.60	63.64	78.95	90.79	21.05	00.00	69.74
14	83.12	77.92	16.88	01.30	61.04	80.26	90.79	19.74	00.00	71.05
15	81.82	80.52	18.18	05.19	62.34	84.21	82.89	15.79	06.58	67.11
16	79.22	85.71	20.78	00.00	64.94	81.58	78.95	18.42	02.63	60.53
17	77.92	89.61	22.08	01.30	67.53	80.26	84.21	19.74	01.32	64.47
18	45.45	76.62	54.55	15.58	22.08	39.47	80.26	60.53	09.21	19.74
19	77.92	92.21	22.08	03.90	70.13	80.26	88.16	19.74	03.95	68.42
20	88.31	23.38	11.69	20.78	11.69	81.58	34.21	18.42	18.42	15.79

Nesta investigação verificamos o comportamento das medidas: Oráculo, limite-inferior, $\mathbb{C}_0$, $\mathbb{C}_1$ e $\mathbb{C}_{100}$. Vejamos que conforme já comentado anteriormente o Oráculo é estimado como sendo $\mathbf{100\%} - \mathbb{C}_0$ e o limite-inferior como sendo $\mathbb{C}_0 + \mathbb{C}_{100}$. A coluna $\mathbb{C}_1$ nos mostra as instâncias que conseguem ser preditas por apenas um classificador do `POOL`, consideradas difíceis.

Vemos que em algumas situações como por exemplo na replicação 02, o valor de $\texttt{Dsel}|\mathbb{C}_1$ é igual a zero. Isso mostra que não existe nenhuma instância que seja predita por apenas um classificador. Vemos também outras informações interessantes na replicação 10 onde o limite-inferior é maior que o oráculo. Sim, isso é perfeitamente possível acontecer dado o cálculo com que é feito o limite-inferior.

Neste caso da replicação 10, vemos 67.53% das instâncias conseguem ser preditas por todos classificadores do `POOL` $\mathbb{C}_{100}$. Além disto 18.18% das instâncias não conseguem ser preditas por nenhum classificador do `POOL`. A soma destes dois

valores é sim um dos cálculos de limite-inferior proposto, como mostrado na Eq. 4.6. Este foi um dos motivos que decidimos escolher o limite-inferior como sendo apenas $\mathbb{C}_{\mathbf{100}}$, ou seja, Eq. 4.7 para que tenhamos valores que não ultrapassem o oráculo nesta fase de investigação.

Após estas pequenas análises e com as informações da execução desta investigação, foi criada a Tabela 11 que nos mostra as acurácias do método `LCA` sendo executado de maneira convencional e `dLCA` com cálculo de discriminante. Para termos uma comparação, também capturamos as informações do Oráculo e limite-inferior para termos uma baliza dos métodos. Também capturamos informações de dois outros algoritmos de seleção estática para comparação: todos classificadores `CC` e Single Best `SB`.

O método `CC` foi implementado para considerar todo o `POOL` de classificadores para cada instância desconhecida, utilizando o método de voto majoritário para decidir a resposta final. Já o método Single Best foi implementado para considerar apenas o melhor classificador do `POOL` para predizer todas as instâncias.

Considerando todas estas informações, escolhemos a execução **número-12** da Tabela 11, porque ela possui a maior diferença entre kRoC e dRoC. Como podemos verificar, esta simulação-12 possui uma diferença de mais de 40 pontos percentuais entre LCA e dLCA. O Oráculo obteve valor: 77.92% e isto nos mostra a porcentagem de instâncias do dataset que pode ser corretamente predita por pelo menos um classificador do `POOL`. Nesta simulação-12 podemos notar também um pequeno **limite-inferior**, igual a 11.69%. Desta forma vemos que $\boldsymbol{\Delta}$ Eq. 4.4 possui um valor de 66.23. Isso significa que cada algoritmo de seleção dinâmica terá uma faixa de mais de 50% para se diferenciar de seus pares. Gostaríamos de enfatizar novamente que esta é uma parte extremamente importante e que justifica a utilização do limite-inferior como medida de complexidade e nos permite estimar essa diferença com o limite-superior mostrando a faixa de instâncias viáveis para estudo a partir de um `POOL` de classificadores.

5.3.1 Investigação do método `LCA`

Sendo o método `LCA` do tipo `DCS`, sabemos que ele seleciona apenas uma hipótese do `POOL` de classificadores para predizer uma instância desconhecida. No entanto, quando existem empates entre os classificadores mais competentes, os algoritmos do tipo `DCS` escolhem apenas uma hipótese randomicamente entre os empates. Notamos que em muitos casos, devido à função de estimativa de competência dos classificadores, o `LCA` acaba fornecendo valores máximos para classificadores que não são verdadeiramente-competentes.

Desta forma, observamos que o `LCA` ao selecionar randomicamente algum destes classificadores pode fazer a escolha errada. Uma das maneiras de aprimorar o resultado do `LCA` seria fornecer uma região de competência onde as instâncias

Tabela 11 – Valores de acurácia do método LCA rodando com uma região de competência estimada apenas por k-nn. O algoritmo dLCA representa o mesmo método só que rodando com uma região de competência estimada por índice de discriminação. Dataset utilizado: Haberman.

Rep	LCA	dLCA	Diff	ORA	LI	CC	SB
01	35.53	64.47	28.95	75.32	14.29	26.32	27.63
02	73.68	73.68	00.00	77.92	64.94	75.00	73.68
03	34.21	64.47	30.26	77.92	11.69	26.32	26.32
04	73.68	73.68	00.00	77.92	62.34	75.00	73.68
05	28.95	44.74	15.79	63.64	12.99	26.32	26.32
06	42.11	57.90	15.79	68.83	19.48	26.32	26.32
07	30.26	47.37	17.11	68.83	19.48	26.32	26.32
08	26.32	36.84	10.53	63.64	16.88	26.32	00.00
09	26.32	39.47	13.16	59.74	10.39	26.32	26.32
10	73.68	73.68	00.00	81.82	67.53	73.68	73.68
11	73.68	73.68	00.00	77.92	62.34	73.68	73.68
12	30.26	71.05	40.79	77.92	11.69	26.32	26.32
13	73.68	73.68	00.00	80.52	63.64	73.68	72.37
14	73.68	73.68	00.00	83.12	61.04	72.37	78.95
15	73.68	73.68	00.00	81.82	62.34	73.68	73.68
16	73.68	73.68	00.00	79.22	64.94	72.37	73.68
17	73.68	73.68	00.00	77.92	67.53	73.68	73.68
18	26.32	31.58	05.26	45.45	22.08	26.32	26.32
19	73.68	73.68	00.00	77.92	70.13	73.68	73.68
20	34.21	69.73	35.53	88.31	11.69	26.32	39.47

Rep: número de replicações. IL: limite-inferior - mostra a porcentagem de instâncias que todos classificadores do POOL conseguem predizer a instância desconhecida corretamente. Diff: mostra a diferença entre os métodos LCA e dLCA.

pudessem contribuir com o ranqueamento dos classificadores verdadeiramente-competentes. Considerando a estimativa de competência de classificadores e seus empates, possuímos as perguntas abaixo que nos motivam nesta exploração.

(i) Qual o número de classificadores no POOL que irão acertar cada instância do dataset de TEST?

(ii) Qual o número de empates de classificadores com competência máxima?

(iii) Dentre as hipóteses escolhidas, qual a frequência de escolha de cada classificador no POOL?

5.3.2 Classificadores realmente-competentes

De acordo com a pergunta **(i)**, foram capturados os dados da simulação-12 e gerado a Tabela 12. A ideia neste pequeno experimento é fazer uma análise *aposteriori* para mostrar qual o número de classificadores que realmente acertam cada instância do dataset de TEST. Podemos ver na Tabela 12 que aprox. 11.80% das instâncias não podem ser preditas por nenhum classificador do POOL, representando $\mathbb{C}_{\mathbf{0}}$. A maioria das instâncias: 61.80% conseguem ser corretamente preditas por entre 1 e 5 classificadores do POOL. Um porcentagem considerável de aproximadamente 26.4% das instâncias conseguem ser preditas corretamente por um número de 95 a 100 classificadores, o que nos mostra quão fácil são estas instâncias para os algoritmos de seleção dinâmica.

Se analisarmos somente $\mathbb{C}_{\mathbf{0}}$ e $\mathbb{C}_{\mathbf{100}}$, podemos ver que $\mathbf{27,63\%}$ das instâncias obterão o mesmo resultado a partir de qualquer classificador. Isto representa mais que um quarto do dataset que não conseguirá ser utilizado para ranquear a habilidade de nenhum classificador, tornando nestes momentos os algoritmos de MCS completamente indiferentes.

Tabela 12 – Número de classificadores verdadeiramente-competentes.

#Classifiers	% of Instances
00	11.84%
01	28.95%
02	10.53%
03	03.95%
04	06.58%
05	11.84%
95	01.32%
97	01.32%
98	05.26%
99	02.63%
100	15.79%

Analisando a Tabela 12, entendemos que o **limite-inferior** pode ser enxergado como a primeira e/ou última linha da tabela. Se pegarmos o caso da Tabela 11, execução-12 - vemos que o LI é igual a 11.69% indicando que este é a porcentagem de instâncias no dataset Dsel que conseguem ser preditas por todos classificadores do POOL.

O limite-inferior, do ponto de vista de exploração dos dados, pode ser visualizado sob o prisma do dataset Dsel ou do dataset TEST. Na Tabela 11 é mostrado o IL do conjunto Dsel, enquanto que a Tabela 12 é mostrado o IL do conjunto de TEST, sendo LI= $\mathbf{15,79\%}$.

5.3.3 Comportamento da região de competência estimada por índice de discriminação

Outra informação importante para nossa análise é saber como se comporta a uma região de competência estimada de duas formas: k-vizinhos mais próximos kRoC e

discriminante `dRoC`. Neste sentido, gostaríamos de verificar qual a porcentagem de instâncias da dRoC possui intersecção com as instâncias de uma kRoC, ou seja, qual probabilidade de uma instância dRoC estar em uma kRoC. Consideraremos os mesmos parâmetros da simulação anterior: dataset Haberman, simulação-12.

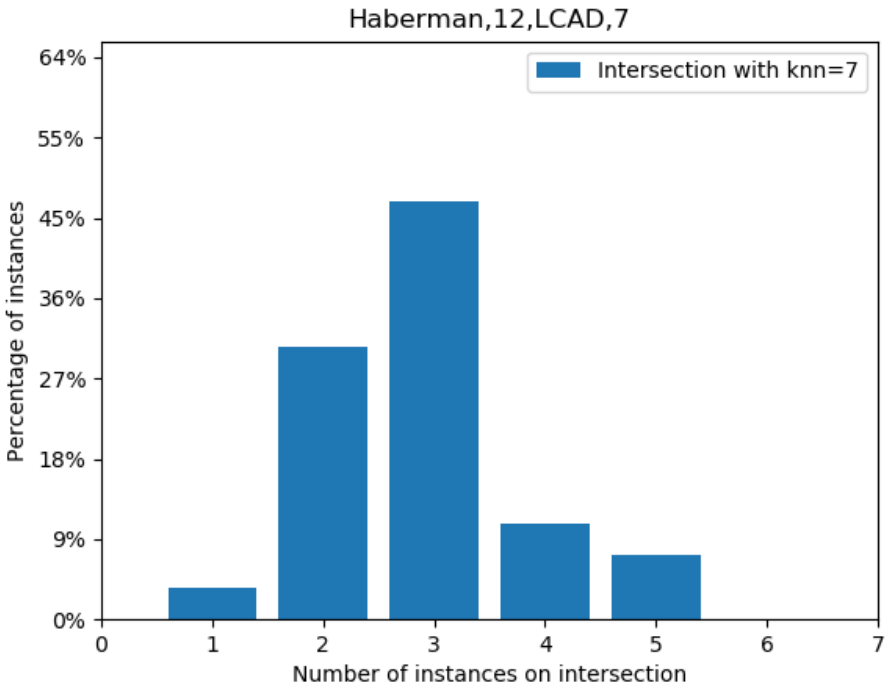


Figura 15 – Quantas instâncias existem entre a intersecção de uma região de competência kRoC e outra dRoC?

Vemos pela Figura 15 que a intersecção mais comum entre `RoC` vs. `dRoC` para os experimentos realizados com o dataset Habberman - simulação:12, são de 3 instâncias ocupando quase 50% das ocasiões. Outra informação importante para nossa análise é a verificação do nível de dificuldade (Instance Hardness - `IH`) de uma instância (SMITH; MARTINEZ; GIRAUD-CARRIER, 2013) em relação à um conjunto de dados, por exemplo uma `dRoC`. Neste trabalho de (SMITH; MARTINEZ; GIRAUD-CARRIER, 2013), `IH` é definido conforme algumas medidas de complexidade, sendo uma bastante utilizada: `kDN` (k-Disagreeing Neighbors) conforme comentado por (LORENA et al., 2019). Esta medida `kDN` estima o número de desacordos que uma instância possui com sua vizinhança, e pode assumir valores entre $[\mathbf{0-1}]$, sendo que quanto maior este número maior o número de instâncias em desacordo. Podemos assumir "desacordo"como: instâncias vizinhas que possuem uma classe diferente da instância em questão.

Com isso, vemos pela Figura 16 que uma dRoC possui um grande número de instâncias com baixo `IH`, ou seja, aprox. 92% das instâncias tem um `IH` entre $[\mathbf{0; 0.1}]$, para este caso. Vemos também que uma `kRoC` possui instâncias em diferentes níveis de dificuldade.

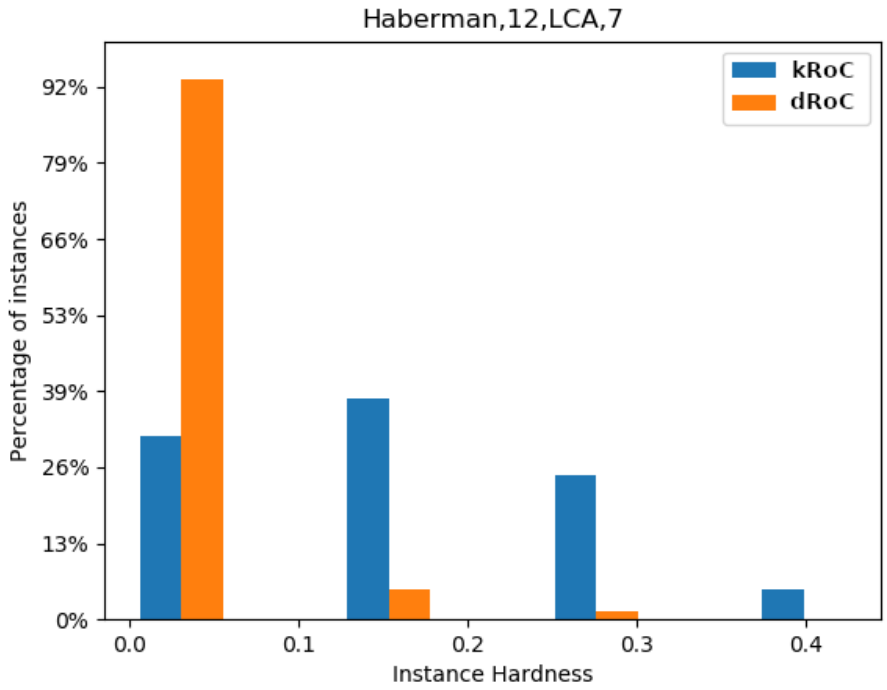


Figura 16 – Verificação da dificuldade de uma instância (Instance Hardness - IH)

5.3.4 Comportamento dos classificadores

Após uma visão geral, iremos investigar qual o comportamento dos classificadores no `POOL` dentro de nosso exemplo: Habberman, 12; a partir de duas regiões de competência kRoC e dRoC. A Figura 17 nos mostra um histograma do número de classificadores que pertencem a uma região de competência que acertariam uma instância desconhecida $\mathbf{x}_{\boldsymbol{q}}$, ou seja, a contagem da intersecção entre o conjunto-A e conjunto-B (veja: 4.3) também referenciados neste texto respectivamente como classificadores-considerados-competentes e classificadores-realmente-competentes.

Considerando `kRoC` : existem diferentes números de classificadores que acertam as instâncias. Neste caso vemos que uma porcentagem pequena de classificadores entre 1 e 10 acertam as instâncias da `RoC`. Também podemos verificar

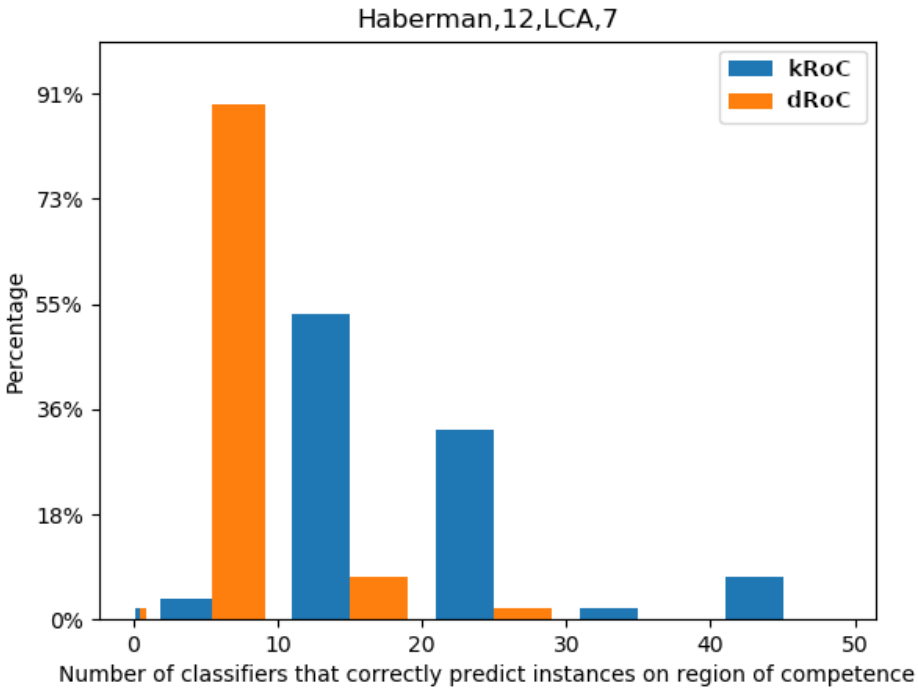


Figura 17 – Histograma da frequência do número de classificadores que conseguem predizer corretamente uma instância que pertence à região de competência (RoC)

neste caso que em aprox. 50% dos casos temos um número de classificadores entre 10 e 20 que acertam as instâncias da RoC. Também temos que em pouco mais de 30% dos casos temos entre 20 e 30 classificadores.

Considerando dRoC : vemos que em aprox. 90% dos casos existem 1 e 10 classificadores que acertam as instâncias da região. Considerando o LCA como um algoritmo DCS, sabemos que quanto menor o número de empates de classificadores-considerados-competentes mais positivo será o desempenho do método. Podemos considerar que dado um método dRoC que consiga reduzir o número dos classificadores-considerados-competentes, reduzirá também os **classificadores-fake** reduzindo assim o risco do método de escolher um classificador que não contribui com o oráculo. Consideramos **classificadores-fake** como aqueles que pertencem ao conjunto-A, mas que não pertencem ao conjunto-B, ou seja, são considerados competentes, mas que não conseguirão predizer corretamente $\mathbf{x}_{q}$. Em outras palavras, classificadores que serão selecionados para predizer $\mathbf{x}_{q}$ mas que não contribuem para a acurácia do sistema.

A Figura 18 mostra uma visão da porcentagem de empates dos classificadores. Já comentamos que no método LCA, a grande quantidade de empates de classificadores com máxima competência dificultava a escolha dentre eles pelo método de seleção dinâmica.

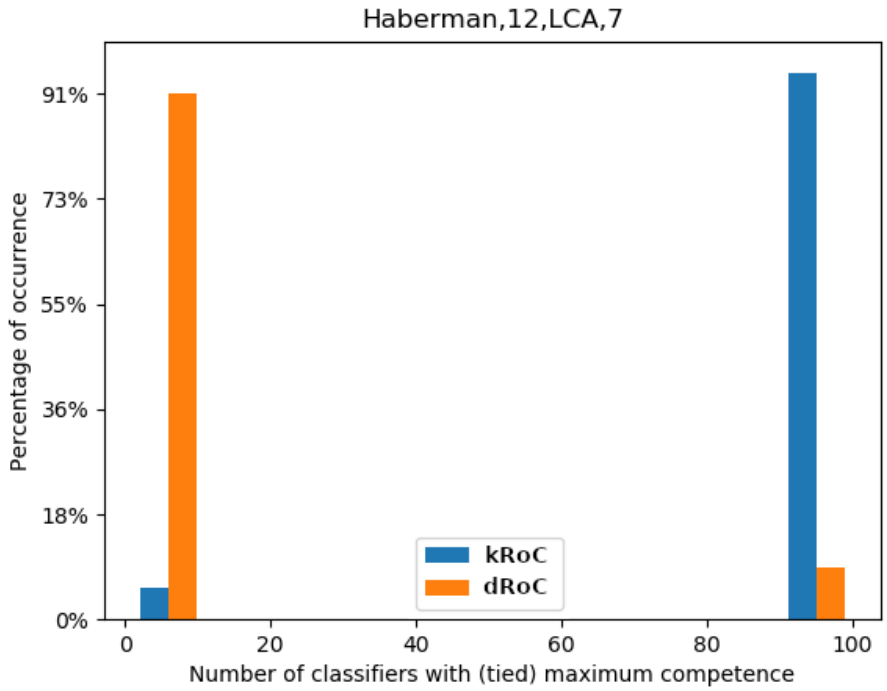


Figura 18 – Histograma mostrando o empate de classificadores que atingiram a máxima competência.

Este histograma da Figura 18 nos mostra que a região de competência dRoC possui em aprox. 90% dos casos um empate entre 0 e 10 classificadores mais competentes. O oposto acontece quando selecionamos uma kRoC, onde verificamos que mais de 90% dos casos existe empates entre 90 e 100 classificadores. Esta figura mostra para este caso que uma região de competência estimada por índice de discriminação pode reduzir o número de empates de classificadores mais competentes, podendo assim trazer melhores resultados, principalmente para esta nossa análise do LCA.

5.3.5 Comportamento dos valores do índice de discriminação

A partir de nossos resultados, gostaríamos de investigar qual seria o comportamento dos valores de discriminação para cada uma das regiões de competência

capturadas durante a simulação em questão: Habberman, 12. Nesta análise, verificamos através da Figura 19 que em uma kRoC convencional com $\boldsymbol{k = 7}$, uma parte considerável de quase 60% das instâncias possui um índice de discriminação igual a 1.0 (valor máximo de discriminação), mostrado do lado direito do gráfico a barra em laranja.

Ainda assim uma grande parcela de quase 40% possui como índice de discriminação um valor igual a $\text{Đ}= \mathbf{0.2}$. Ainda nesta figura, vemos que uma região de competência estimada como dRoC, grande parte das instâncias possui discriminação máxima como era de se esperar.

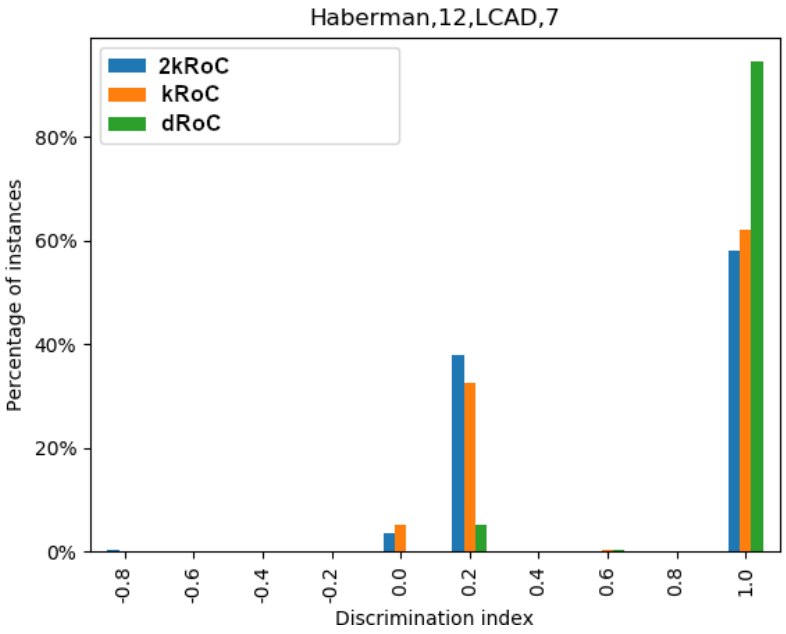


Figura 19 – Índice de discriminação para instâncias da região de competência.

5.3.6 Comportamento de predições

Após as investigações a respeito dos valores do indice de discriminação e do comportamento dos classificadores e suas competências, iremos verificar também o comportamento dos erros e acertos feitos pelo algoritmo LCA, considerando as duas regiões de competência de interesse: kRoC vs. dRoC. O histograma da Figura 20 mostra a relação das predições corretas/incorretas no eixo-x e sua porcentagem de ocorrências no eixo-y.

O label **all-agree** indica que todos classificadores do `POOL` concordam com a predição, ou seja, não existe o que deve ser estimado neste caso. Este valor

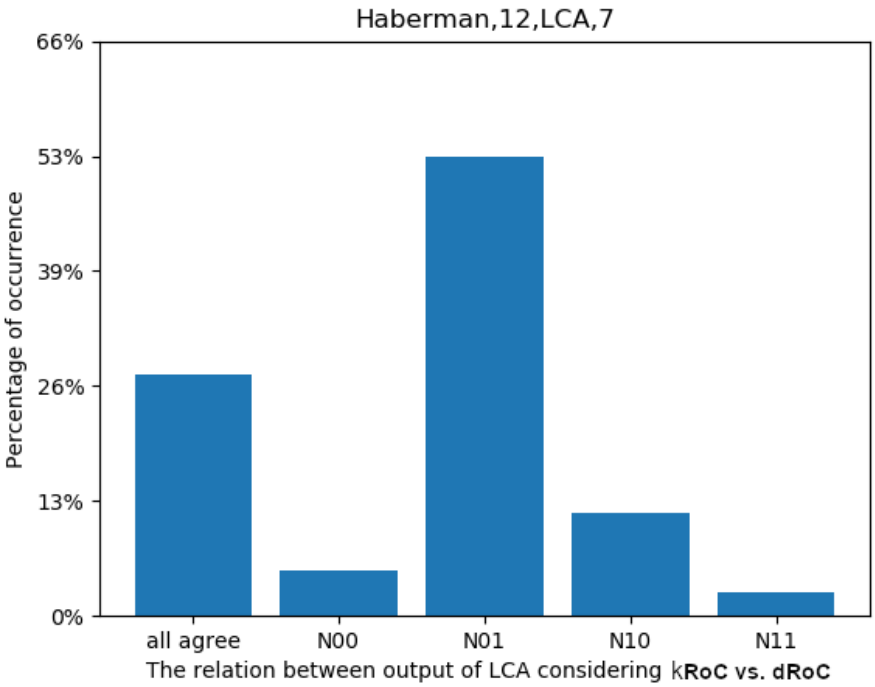


Figura 20 – Histograma de acertos para o algoritmo LCA considerando regiões de competência: convencional (kRoC) e discriminante (dRoC).

indica duas situações bem distintas, mas que todos classificadores concordam: quando 0% dos classificadores conseguem acertar $\mathtt{x}_{\boldsymbol{q}}$, e também quando 100% concordam. Fizemos questão de adicionar esta informação neste histograma para deixar registrada a existência desta medida. Esta medida nos mostra que em aprox. 26% das instâncias, os classificadores do POOL possuem unanimidade de respostas, nas situações em que todos ou ou nenhum deles acerta $\mathtt{x}_{\boldsymbol{q}}$.

Os indices: N00, N01, N10 e N11 mostram a diversidade de respostas dos considerando as duas RoCs analisadas, conforme comenta o artigo (KUNCHEVA; WHITAKER, 2003). Neste índice consideramos os valores **0** como erro de predição e **1** indicando o acerto. Chamaremos de kLCA as respostas do algoritmo executado com kRoC, e chamaremos de dLCA o algoritmo rodado com dRoC. Desta forma, N00 indica quando o LCA erra em ambos os casos, N01 indica quando o kLCA erra e o dLCA acerta; N10 indica quando o kLCA acerta e dLCA erra; por fim, N11 indica quando o LCA prediz corretamente uma instâncias nas duas RoCs. Percebemos que pela Figura 20 que o número de acertos com uma dRoC (N01) é superior a 51% dos casos, indicando mais da metade das instâncias testadas.

Tabela 13 – Valores da medida q-roc para 15 instâncias do dataset Haberman, execução do experimento número-12.

	kRoC				dRoC			
id	|A|	$\|\boldsymbol{A} \cap \boldsymbol{B}\|$	|B|	q-roc	|A|	$\|\boldsymbol{A} \cap \boldsymbol{B}\|$	|B|	q-roc
0	2	0	0	0.000	2	0	0	0.000
1	100	5	5	0.050	8	5	5	0.620
2	2	2	100	1.000	2	2	100	1.000
3	100	2	2	0.020	2	2	2	1.000
4	100	97	97	0.970	5	2	97	0.400
5	100	1	1	0.010	1	1	1	1.000
6	100	1	1	0.010	3	1	1	0.330
7	100	3	3	0.030	97	2	3	0.020
8	100	95	95	0.950	5	0	95	0.000
9	100	1	1	0.010	3	1	1	0.330
10	100	4	4	0.040	5	4	4	0.800
11	100	1	1	0.010	3	1	1	0.330
12	20	1	1	0.050	21	1	1	0.050
13	2	2	100	1.000	2	2	100	1.000
14	100	98	98	0.980	3	1	98	0.330

5.4 Desempenho entre diferentes tipos de RoC

Os experimentos desta seção foram realizados para avaliar o impacto no desempenho de três tipos de região de competência: kRoC, dRoC e rRoC. Também estamos interessados em avaliar a qualidade das regiões de competência através da proposta: **q-roc**, Eq. 4.3. Os experimentos a seguir tiveram os mesmos parâmetros descritos no início do Capítulo 5.

Considerando as análises anteriores, foram mantidos os testes com dataset Haberman, simulação: 12. A tabela 13 mostra a qualidade da região de competência **q-roc** e o número de classificadores dentro de três conjuntos: A, B e $\boldsymbol{A} \cap \boldsymbol{B}$; para 15 instâncias do conjunto de dados Haberman considerando `kRoC`, e do outro: `dRoC`.

A tabela 13 está dividida em duas regiões de competência: kRoC e dRoC e mostra algumas informações interessantes nas 15 instâncias (id), numeradas de 0 a 14. As colunas $|\boldsymbol{A}|$, $|\boldsymbol{A} \cap \boldsymbol{B}|$ e $|\boldsymbol{B}|$ expressam a cardinalidade desses conjuntos, ou seja, o número de classificadores dentro de cada conjunto. A coluna q-roc mostra o score estimado de acordo com a Eq. 4.2.

id=0 , na primeira linha do kRoC, temos os valores $|\boldsymbol{A}| = \mathbf{2}$, $|\boldsymbol{B}| = \mathbf{0}$. Neste tipo de situação não é possível para os classificadores do `POOL` predizer corretamente $\mathbf{x}_{\boldsymbol{q}}$ porque não há nenhum classificador no conjunto dos Oráculos, conjunto-B. Esta situação se repete para dRoC.

id=1 , vemos um cenário diferente: a coluna kRoC mostra que o número de classificadores no conjunto-A é igual a 100, no entanto, considerando dRoC esse número diminuiu para 8 fazendo com que q-roc seja 0,05 no primeiro caso e 0,62 no segundo. Notamos que a diminuição do número de classificadores no conjunto-A trouxe uma melhoria no score q-roc deste exemplo.

id=3 , temos um cenário ideal para o funcionamento do dRoC. kRoC possui q-roc=0,02 e a mesma instância no dRoC possui q-roc=1,0. Considerando este caso, o kRoC assume que todos os classificadores do `POOL` são capazes de prever corretamente as instâncias da `RoC`, mas apenas dois classificadores são competentes para prever $\mathtt{x}_{\boldsymbol{q}}$. Neste mesmo exemplo considerando uma dRoC, vemos que apenas dois classificadores do conjunto são capazes de prever as instâncias `RoC` e são exatamente estes dois classificadores que são competentes para prever corretamente $\mathtt{x}_{\boldsymbol{q}}$. Neste caso, q-roc(kRoC)=0.02, enquanto q-roc(dRoC)=1.

id=4 , temos outro cenário: o número de classificadores no conjunto |B|=97, ou seja, podemos considerar uma instância que é fácil de ser predita e seu q-roc é aprox. 1. No entanto, na a região dRoC, o q-roc é inferior a 0,5.

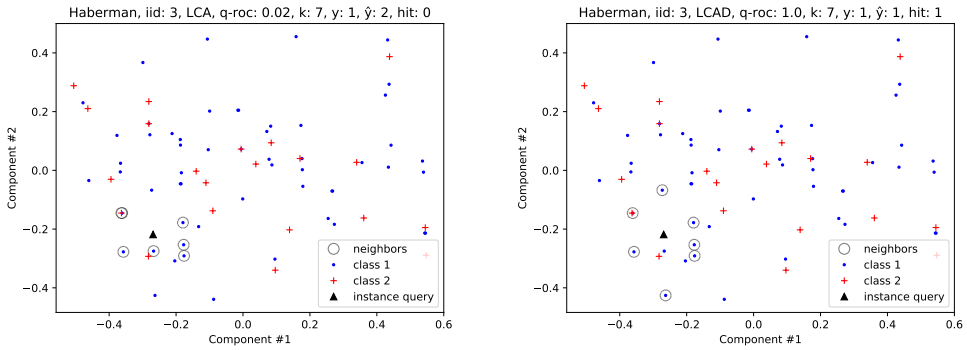
id=5 , temos outra instância interessante, onde 100 classificadores no conjunto A e apenas 1 classificador no conjunto B, o que faz com que q-roc(kRoC)=0,01. No entanto, vemos que q-roc(dRoC)=1.0.

id=8 , temos uma situação diferente, onde o número de classificadores no conjunto-B=95 mostrando que é uma instância fácil. No entanto não acontece na região dRoC onde não há classificador na intersecção $\boldsymbol{A} \cap \boldsymbol{B}$.

Fizemos uma transformação linear PCA do dataset `Dsel` para mostrar uma maneira visual e investigar melhor o dataset Haberman, simulação-12. Nesta análise foram avaliadas duas instâncias $\mathtt{x}_{\boldsymbol{q}}$: no primeiro exemplo (**E1**) uma instância onde `dRoC` obtivesse sucesso e o segundo exemplo (**E2**), o oposto. As Figuras 21 e 22 mostram uma visão do dataset `Dsel` através de seus componentes principais (PCA). Cada imagem mostra uma instância desconhecida e a sua região de competência.

A tabela14 mostra os experimentos e os valores estimados para cada uma das regiões de competência analisadas, discutidas abaixo:

E1: Neste exemplo, Tabela14, a instância $\mathtt{x}_{\boldsymbol{q}}$ pertence à classe 1 e a maioria dos vizinhos são da mesma classe, como visto na Figura 21. Neste exemplo o número de classificadores no conjunto |A| para as linhas `kRoC` e `rRoC` é igual a 100, ou seja, todos classificadores do `POOL` conseguem acertar ao menos uma instância na região de competência. O número de classificadores que realmente acertarão $\mathtt{x}_{\boldsymbol{q}}$ é igual a 2 como visto no conjunto |B|.



(a) (b)

Figura 21 – Exemplo1: Análise PCA (Principal Component Analysis) do algoritmo LCA, executado no dataset Haberman, experimento número-12, instância número 3, vizinhança k=7. (a) Verificamos que o método `kRoC` não foi possível responder corretamente e (b) o método `dRoC` conseguiu responder corretamente à instância desconhecida $\mathbf{x}_{\boldsymbol{q}}$.

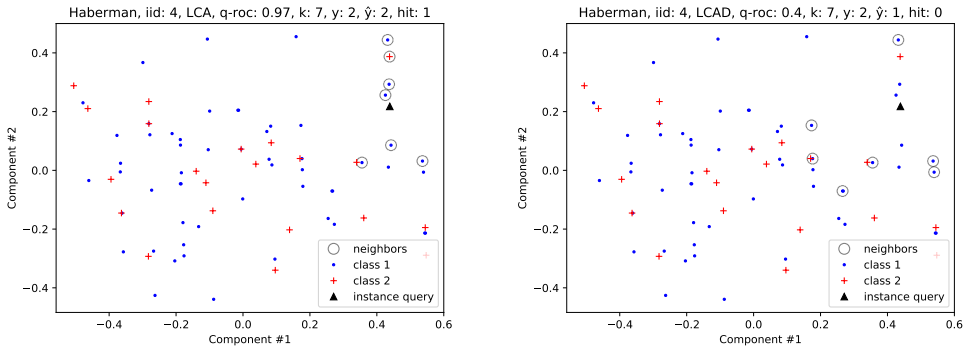
Tabela 14 – Estudo de caso de duas instâncias para o dataset Haberman, considerando três tipos de região de competência.

	type	class	|A|	|B|	|A∩B|	q-roc	predicted
E1	kRoC	1	100	2	2	0.02	2
	dRoC	1	2	2	2	1	1
	rRoC	1	100	2	2	0.02	2
E2	kRoC	2	100	97	97	0.97	2
	dRoC	2	5	97	2	0.4	1
	rRoC	2	100	97	97	0.97	2

Colunas: (1) tipo da região de competência; (2) classe correta da instância; (3) |A|: número de classificadores que pertencem ao conjunto-A; (4) |B|: número de classificadores que pertencem ao conjunto-B; (5) q-roc: medida de qualidade da RoC baseada na equação 4.2; (6) esta última coluna mostra o resultado do LCA que foi predito para cada uma das RoCs apresentadas.

O valor de q-roc indica a proporção de quantos classificadores no conjunto |A| irão classificar corretamente $\mathbf{x}_{\boldsymbol{q}}$. No caso de uma `kRoC` e `rRoC` vemos que o valor de q-roc é baixo. Neste primeiro exemplo, o LCA consegue responder corretamente a instância somente com a região de competência `dRoC`.

E2: Neste exemplo, a instância $\mathbf{x}_{\boldsymbol{q}}$ pertence à classe 2 e a maioria dos vizinhos são da classe oposta, como pode ser visto na Figura 22(b). Na Table 14, podemos ver que o número de classificadores que conseguirão responder corretamente $\mathbf{x}_{\boldsymbol{q}}$ é



(a) (b)

Figura 22 – Exemplo2: Análise PCA do algoritmo LCA, executado no dataset Haberman, experimento número-12, instância número 4, vizinhança k=7. (a) Verificamos que o método kRoC obteve uma predição correta; e (b) o método dRoC não foi possível responder corretamente instância desconhecida $\mathbf{x}_{\boldsymbol{q}}$.

igual a 97 mostrado em |B|, e o número de classificadores no conjunto A é grande para kRoC e rRoC. Com um q-roc alto, existe uma maior chance dos algoritmos de seleção dinâmica conseguirem encontrar um classificador competente para $\mathbf{x}_{\boldsymbol{q}}$. Vemos também na Tabela 14 que o valor de q-roc para uma região discriminante é menor, significando que existe uma intersecção muito baixa entre A e B neste caso.

Percebemos neste estudo de caso que através da Tabela 14 e pelas Figuras 21 e 22, o índice de discriminação terá alguns problemas pois ele sempre tenta encontrar as instâncias que são mais discriminantes. Em situações onde a intersecção dos conjuntos A e B é próximo do tamanho do POOL de classificadores, existe um risco do índice de discriminação podar instâncias que ajudariam a estimar boas competências aos classificadores, como é o caso exemplo **E2**. Para continuar com o estudo de caso, veremos como se comportou 20 replicações no dataset Haberman a seguir.

Foram executadas 20 replicações no dataset Haberman para obter uma melhor análise estatística. A tabela 15 mostra a média destas 20 execuções para cada uma das regiões de competência analisadas: kRoC, dRoC e rRoC.

A Tabela 15 mostra que o número de classificadores no conjunto A é grande quando consideramos uma kRoC e rRoC. O mesmo não acontece para uma região estimada por discriminante, em que obteve uma média de 18.92 classificadores.

O valor do número de classificadores no conjunto |B| não mudou para nenhum caso porque a instância $\mathbf{x}_{\boldsymbol{q}}$ utilizada é sempre a mesma. Vemos uma grande diferença entre os números de intersecção |A∩B|, sendo que a região de discriminação obteve um valor mais baixo.

Esta situação nos indica que para este caso o índice de discriminação diminuiu

Tabela 15 – Informação da média de 20 execuções para o método LCA usando o dataset Haberman.

Type	$\|A\|^{20}$	$\|B\|^{20}$	$\|A \cap B\|^{20}$	QR^{20}	Acc (%)
kRoC	95.10	27.55	26.23	29.85	30.26
dRoC	18.92	27.55	03.59	54.01	71.05
rRoC	90.01	27.55	26.25	30.64	34.21

a intersecção dos classificadores. No entanto vemos que a medida de qualidade QR teve um aumento significativo em relação aos outros dois algoritmos. Isto pode indicar que neste caso, o índice de discriminação diminui a intersecção dos classificadores considerados competentes. Apesar da discriminação diminuir o número de |A| e |A∩B|, ele aumenta o q-roc, porque a Eq. 4.2 é formada pela divisão do número de classificadores no conjunto-A. Desta forma, aumentando o a qualidade da região de competância, vemos que a acurácia também aumentou chegando à **71.05%** no método executado com dRoC o que significa um aumento de mais de 100% em relação aos outros dois métodos kRoC e rRoC.

Verificamos neste exemplo também que uma região de competência randômica `rRoC` seguiu uma tendência de acompanhar os valores estimados de `kRoC`, e ainda assim teve uma performance melhor neste dataset.

Para complementar esta análise, coletamos o valor de QR para as 20 replicações. A média é mostrado na Figura 23. Esta figura mostra que para o dataset Haberman, em metade dos casos, uma `dRoC` possui uma maior qualidade da região de competência, Eq.4.3.

Investigando um pouco melhor a Figura 23, vemos que em 50% dos casos os dois métodos dRoC e kRoC tiveram um empate com uma acurácia alta, acima de 70 pontos. Se investigássemos esses valores apenas pela métrica de acurácia, veríamos que simplesmente os dois métodos possuem supostamente o mesmo desempenho. No entanto, vemos que o limite-inferior nestes métodos, mostrado na Tabela 11 é muito alto, o que pelos nossos comentários da Figura 14 mostra que não há instâncias sufientes disponíveis para trazer melhores estimativas para os métodos.

5.5 Investigação de diferentes regiões de competência com o método proposto

Dentro da área de seleção dinâmica de classificadores, a região de competência desempenha um importante papel na obtenção do conjunto de instâncias que será utilizada para estimar a competência de um classificador. Esta seleção de instâncias

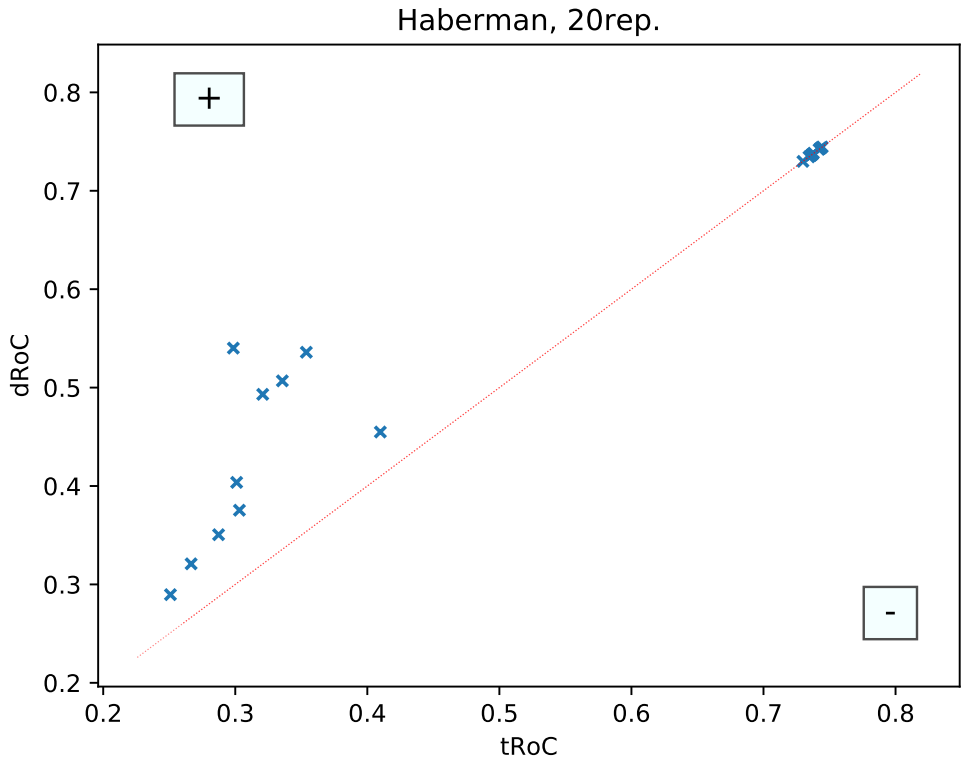


Figura 23 – Valores da medida QR para o dataset Haberman para cada uma das execuções, considerando LCA e duas RoC.

mais promissoras do conjunto DSEL trará impacto significativo no desempenho dos métodos.

Após os resultados iniciais do capítulo anterior obtivemos indícios de que nossa proposta é promissora e com este entendimento faremos nesta seção uma investigação de diferentes regiões de competência com nosso método **dRoC**.

5.5.1 Proposta para o experimento

Faremos experimentos contendo a seguinte lista de regiões de competência: `2-kNN` descrito neste trabalho em Figura 9(b); `r-kNN` também descrito no Algoritmo 2; `e-kNN` (ENN) por ser um algoritmo bastante citado na literatura; `g-kNN` (RNG) por obter o melhor desempenho nas simulações do autor em (CRUZ; SABOURIN; CAVALCANTI, 2016); `fa-kNN` pela diferença entre outros métodos e por último o algoritmo proposto por este trabaho: `d-kNN` que seleciona uma região de competência pelo índice de discriminação.

Tabela 16 – Lista com as regiões de competência selecionadas para este estudo.

Abbrev.	INFO
2-KNN	Região local que contém o dobro de instâncias
r-KNN	Região local extraída de forma randômica dentro de 2KNN
e-KNN	Região local capturada do DSEL', após poda de instâncias no DSEL pelo algoritmo Edited Nearest Neighbor
g-KNN	Região local capturada de DSEL', após poda de instâncias no DSEL pelo algoritmo Relative Neighborhood Graph
d-KNN	Região local capturada através do algoritmo proposto, ou seja, 2KNN com poda pelo menor discriminante
fa-KNN	Região local estimada por duas podas: DSEL' realizada no início e DSEL'' com poda dinâmica

A tabela 16 mostra um resumo destes métodos de região de competência. As simulações realizadas com estes métodos descritos utilizaram os mesmos valores já listados na Tabela 8. Foi optado por comparar o dRoC proposto com outros métodos que também utilizam uma região de competência de tamanho fixo, já listados na Tabela 4. Desta Tabela 4 foram utilizados os algoritmos para teste em um total de 11: `RNK`, `OLA`, `LCA`, `KNE`, `KNU`, `KNOP`, `MCB`, `DESP`, `MLA`, `MDES`, `DES-DKNN`.Foram considerados métodos de seleção estática e seleção dinâmica de classificadores.

5.5.2 Resultados

Após a execução das simulações obtivemos um total de 330 resultados (30 datasets $\times$ 11 métodos). Após os experimentos os resultados foram coletados e a primeira figura deste experimento foi gerada: Figura 24.

Esta figura mostra os resultados do teste WTL para cada um dos métodos analisados em comparação com o método base: KNN. A linha vertical sinaliza o valor-crítico $\boldsymbol{n_c}$ Eq.5.1, ou seja, mostra quando o teste da hipótese nula $\boldsymbol{H_0}$ pode ser rejeitada. O valor de alfa utilizado foi de: $\boldsymbol{\alpha = 0.05}$ e desta forma, caso o método possua um número de vitórias maior que $\boldsymbol{n_c}$, assumimos a hipótese $\boldsymbol{H_1}$ indicando que os métodos são diferentes com 95% de chance. Neste exemplo a linha vertical sinaliza $\boldsymbol{n_c \approx 180}$. De todas as `RoC` analisadas, vemos que g-knn e d-knn obtiveram resultados significativamente melhor que o método knn original.

Após esta análise, foi conduzido o teste Friedman rank tests, conforme comentado em: 5.2-III, contendo os mesmos resultados que o teste acima: 30 datasets e 11 métodos de seleção dinâmica ($\mathbf{330 = 30 \times 11}$). Após estimar o ranking de cada item, aplicamos o teste post-hoc de Bonferroni-Dunn comparando os ranks de cada algoritmo aqui analisado. A condução deste teste também está descrito em 5.2-IV. A figura 25 contendo a diferença crítica (*Critical Difference* - `CD`) mostra o resultado deste teste.

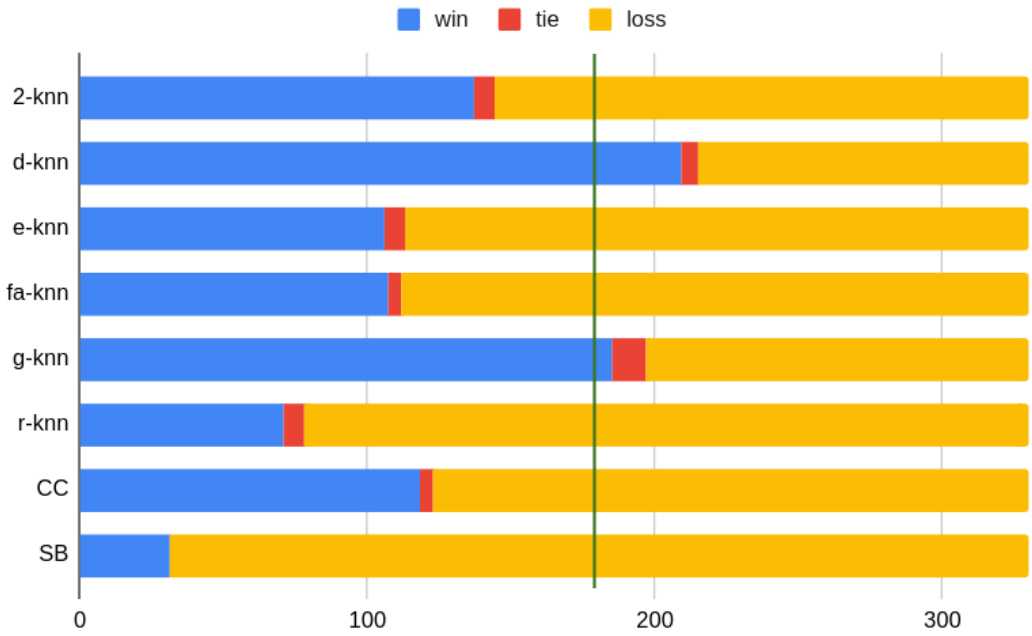


Figura 24 – Gráfico contendo informações de ganho, empate e perda (wtl) de 330 resultados, sendo 30 datasets e 11 métodos para cada um dos métodos de região de competência mostrados na tabela 16.

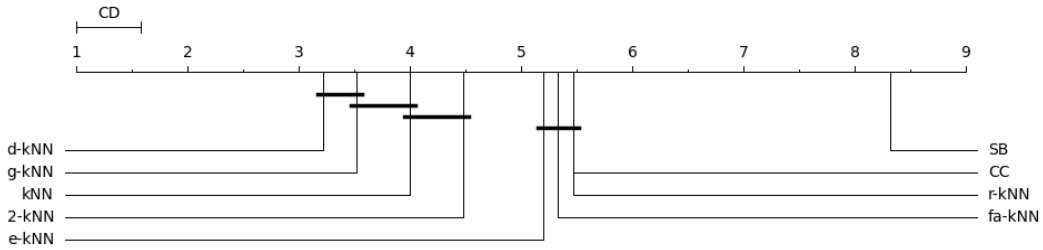


Figura 25 – Gráfico de diferença crítica (CD) que mostra $\boldsymbol{CD = 0.58075}$ para os métodos de região de competência mostrados na tabela 16.

A Figura 25 mostra um valor crítico de CD= **0.58075** para os métodos analisados na Tabela 16, sendo os valores de cada método: `d-kNN`: 3.22; `g-kNN`: 3.52; `k-NN`: 4.0; `2-kNN`: 4.48; `e-kNN`: 5.2; `fa-kNN`: 5.33; `r-kNN`: 5.47; CD: 5.47; SB: 8.32.

Podemos verificar por estes valores que `d-kNN` foi o método com menor valor, seguido pelo método `g-kNN`. Estes dois métodos estão conectados no gráfico CD porque eles representam o mesmo nível de performance.

Vemos também na Figura 25 que o método `k-NN` aparece em terceiro, seguido do método `2-kNN` que por sua vez captura duas vezes mais instâncias para sua

região de competência. Podemos ver que os métodos estáticos foram ranqueados por último sendo o método SB aquele que teve a performance mais baixa.

Analisando todos os 330 resultados compostos por 30 datasetes e 11 métodos cada, podemos enxergar que a região de competência estimada por índice de discriminação `d-kNN` pode ser uma alternativa para sistemas de múltiplos classificadores.

5.6 Considerações Finais

Aqui destacamos o motivo pelo qual acreditamos que o índice de discriminação pode contribuir com os futuros trabalhos relacionados com a seleção da região de competência em sistemas de múltiplos classificadores.

Os experimentos realizados neste capítulo foram de grande importância para compreender em maior profundidade como o índice de discriminação pode ser uma maneira de fornecer aos algoritmos de seleção dinâmica uma maneira alternativa de ponderar as instâncias que farão parte de uma região de competência. Desta forma, os métodos de seleção dinâmica podem contar com mais uma maneira de capturar a região de competência através do método: `dRoC`, e com isso podendo obter uma maior chance de selecionar um classificador correto para predizer uma instância desconhecida.

Um dos grande achados desta pesquisa está na ideia por trás do índice de discriminação que precisa capturar ao mesmo tempo: informações das instâncias pertencentes à RoC em conjunto com os classificadores existentes no `POOL`, provendo informações que antes não eram estimadas. Além disso, este método provê na fase de estimativa da RoC que os classificadores menos competentes sejam considerados no cálculo, através dos grupos-UL provindo da teoria dos testes psicométricos.

Os experimentos realizados durante esta pesquisa nos forneceram ideias de como se comportam as estimativas de competência de classificadores para uma dada região de competência baseada no índice de discriminação.

Sistemas de seleção dinâmica estimam a competência de classificadores de acordo com suas heurísticas em uma região local da instância desconhecida $\mathtt{x}_{\boldsymbol{q}}$. No entanto verificamos que, caso uma instância $\boldsymbol{x_i}$ que pertença à RoC onde todos classificadores do `POOL` possuem a mesma resposta, pode não contribuir significativamente para a escolha correta do classificador(es) mais competente(s). Nestes casos, $\boldsymbol{x_i}$ não provê informações para os métodos DS sobre quais classificadores são realmente-competentes para predizer corretamente $\mathtt{x}_{\boldsymbol{q}}$.

Da mesma maneira, podem existir instâncias $\boldsymbol{x_i}$ que pertencem à RoC onde não há nenhum classicador que consegue predizer corretamente sua classe, fazendo com que ela não provenha nenhuma informação importante para a escolha do melhor classificador de $\mathtt{x}_{\boldsymbol{q}}$. Em ambos os casos, a unanimidade de respostas de todos

classificadores do POOL, não nos trás informações que os diferenciem entre os bons e os ruins, pois todos possuem a mesma resposta.

Desta forma, instâncias $\boldsymbol{x_i}$ que pertencem à uma região de competência estimada apenas considerando métricas de distância podem não prover informações suficientes para os sistemas de seleção dinâmica escolherem os classificadores mais promissores. Seguindo um dos exemplos anteriores, em casos onde todos classificadores são unânimes na predição de $\boldsymbol{x_i}$, sabemos que esta análise pode gerar classificadores-fake, ou seja, classificadores que serão selecionados para predizer $\mathtt{x}_{\boldsymbol{q}}$ mas que não contribuem para a acurácia do sistema. Nestes casos, estes classificadores-fake poderão fazer com que o método de seleção dinâmica falhe ao tentar escolher o classificador mais competente.

Após todos insights, a pergunta que sempre esteve latente durante todo o desenvolvimento desta pesquisa foi por que afinal o índice de discriminação pode ser uma alternativa promissora em sistemas de seleção dinâmica.

Por que afinal utilizar o índice de discriminação? O índice de Discriminação pode prover uma estimativa das instâncias que são mais propensas a ajudar na separação dos classificadores mais competentes e dos menos competentes. Este método pode remover os classificadores-fake (considerados-competentes) na fase de escolha de instâncias para a região de competência. Esta é a hipótese por trás do índice de discriminação.

6 Conclusão

Incorporamos uma medida importante provinda da Teoria Clássica dos Testes: índice de discriminação, na seleção de instâncias para compor a região de competência. Esta região ainda é crucial em sistemas de seleção dinâmica de *ensembles* principalmente em situações onde o dataset é pequeno e desbalanceado. Dentro da incorporação desta métrica, destacamos como uma das principais contribuições deste trabalho uma possível resposta para a pergunta: "Por que o índice de discriminação funciona?", descrito na seção: 5.6. Alguns resultados já foram divulgados junto a comunidade científica através do artigo (TRIERVEILER et al., 2018). Este artigo mostra que não somente o índice de discriminação é uma alternativa viável, como também foi o primeiro artigo escrito dentro de nossa pesquisa que consegue extrair informações importantes a partir de classificadores considerados não-competentes (necessários para estimativa do índice de discriminação).

Revisitando as perguntas iniciais deste trabalho, podemos então fazer algumas conclusões como disposto abaixo.

Desafio-1 : Será que para cada instância desconhecida podemos extrair uma região de competência do dataset de validação que considere simultaneamente: (1) as instâncias mais próximas e também (2) os erros e acertos dos classificadores do `POOL`? Qual seria a maneira de conseguirmos ponderar as instâncias do dataset de validação a partir dos classificadores do `POOL`? Será que uma região de competência, definida em consideração por estes dois itens terão um impacto significativamente positivo em algoritmos de seleção dinâmica de classificadores?

- Hipótese-1: Construir um método que consiga extrair a importância de instâncias dentro de uma região de competência que sejam mais promissoras para os métodos de seleção dinâmica, considerando métricas de testes psicométricos; e comparar de forma estatística o desempenho de diferentes sistemas de seleção dinâmica de classificadores, considerando uma região de competência estimada de maneira convencional com uma região de competência estimada por métricas provenientes das teorias de testes psicométricos.
- Conclusão-1: Foi encontrado na teoria de testes psicométricos a métrica de índice de discriminação, que consegue diferenciar instâncias através de erros e acertos dos classificadores do `POOL`.

Desafio-2 : Dentre as métricas existentes como poderemos quantificar uma região de competência para verificar quão bom/ruim seria este grupo de instâncias?

- Hipótese-2: Construir uma fórmula para que consigamos quantificar uma região de competência e assim podermos mensurar sua qualidade;
- Conclusão-2: Criação de uma métrica de qualidade da região de competência `q-roc` pode nos fornecer uma maneira de mensurar sua qualidade.

Desafio-3 : Será que o `POOL` de classificadores conseguiria nos fornecer alguma estimativa sobre o impacto dos algoritmos de seleção dinâmica?

- Hipótese-3: Construir uma nova maneira de avaliar o `POOL` de classificadores que nos permita enxergar melhor a sua eficácia ao ser utilizado em algoritmos de seleção dinâmica.
- Conclusão-3: Criação e definição do Limite-Inferior que pode nos ajudar a encontrar o $\mathbf{\Delta}$-`POOL`, ou seja, uma faixa de valores que o `POOL` pode fornecer para os algoritmos de seleção dinâmica.

Como trabalho futuro, vemos que existe um vasto campo de pesquisa na adoção de outras métricas da psicometria para escolha de instâncias. Isto abre um leque com várias opções para ser explorado. Uma das ideias futuras é conseguir extrair as regiões de competência de tamanho variado, que maximize o índice de discriminação de uma região. Além disto, capturar outras medidas de psicometria e ponderar as instâncias de uma região de competência com estas medidas parecem também ser um campo fértil e promissor a ser explorado.

Apêndices

Tabela 17 – Resultados reportados na conferência internacional ICTAI'18. Cada linha representa os resultados de um dataset e cada coluna representa um método de seleção dinâmica. Os valores nesta tabela representam a acurácia média e o desvio padrão para 20 replicações para cada dataset.

	API	APO	LCA	MCB	MLA	OLA	ORA	RANK
AD	83.92 (1.95)	83.78 (2.41)	82.12 (4.06)	82.47 (2.66)	82.12 (4.06)	83.26 (3.31)	99.13 (0.74)	78.43 (3.81)
BN	97.38 (0.60)	95.87 (0.81)	95.77 (1.04)	97.22 (0.64)	95.77 (1.04)	97.29 (0.53)	99.93 (0.12)	91.11 (1.92)
BL	76.52 (2.08)	74.01 (6.25)	72.83 (9.95)	76.74 (2.54)	72.75 (9.41)	76.63 (2.51)	94.41 (3.76)	73.13 (4.28)
CT	89.90 (0.78)	85.57 (0.88)	86.36 (1.13)	89.07 (1.01)	86.37 (1.12)	89.10 (0.97)	99.12 (0.44)	88.25 (1.24)
DB	74.82 (3.19)	71.22 (2.43)	69.51 (4.30)	73.70 (2.39)	69.51 (4.30)	74.19 (2.50)	99.38 (0.52)	70.89 (2.78)
EC	82.86 (3.33)	79.40 (3.48)	80.18 (2.68)	80.54 (4.69)	80.18 (2.68)	81.96 (3.71)	97.50 (1.09)	78.21 (5.53)
FA	69.42 (1.36)	60.98 (2.48)	61.35 (2.82)	67.41 (1.69)	61.35 (2.82)	68.47 (2.04)	97.31 (1.03)	65.82 (2.52)
GE	72.66 (2.86)	65.36 (5.01)	64.70 (6.51)	70.52 (3.12)	64.70 (6.51)	71.74 (3.12)	99.46 (0.57)	69.40 (3.59)
GL	58.30 (5.19)	48.21 (5.10)	49.34 (5.98)	61.79 (5.54)	49.34 (5.98)	61.42 (5.25)	97.17 (2.63)	59.34 (5.42)
HA	55.86 (1.94)	52.70 (2.18)	52.70 (2.18)	63.88 (1.43)	53.03 (2.14)	63.22 (1.47)	73.82 (1.38)	60.92 (1.50)
HE	80.37 (3.22)	79.33 (4.15)	76.79 (3.92)	76.12 (3.32)	76.79 (3.92)	76.27 (4.11)	98.88 (1.74)	75.07 (4.67)
IL	68.41 (2.67)	68.21 (3.87)	67.10 (6.89)	70.03 (2.71)	67.45 (6.93)	69.48 (3.02)	99.48 (0.63)	67.66 (2.98)
IO	85.45 (3.93)	80.17 (3.93)	81.36 (4.63)	85.23 (3.69)	81.36 (4.63)	85.45 (3.64)	99.72 (0.63)	82.16 (3.94)
L1	80.09 (4.31)	78.02 (5.62)	78.11 (6.31)	79.06 (3.77)	78.11 (6.31)	79.72 (5.12)	100.00 (0.00)	77.36 (5.34)
L3	72.90 (3.28)	70.63 (2.72)	70.62 (2.84)	68.64 (4.98)	70.62 (2.84)	69.49 (3.89)	97.61 (2.90)	67.78 (5.35)
LT	96.16 (0.76)	93.79 (1.34)	93.39 (1.46)	96.20 (0.77)	93.39 (1.46)	96.27 (0.77)	99.76 (0.31)	90.59 (1.53)
LV	63.20 (5.11)	50.87 (5.92)	54.01 (6.84)	66.28 (3.56)	54.24 (7.06)	66.57 (4.46)	97.44 (3.45)	65.52 (4.84)
MG	82.17 (0.55)	76.49 (1.14)	76.23 (3.88)	82.14 (0.64)	76.34 (3.91)	82.19 (0.62)	97.99 (1.23)	76.39 (2.39)
MM	78.84 (2.83)	76.81 (4.08)	74.35 (6.27)	78.91 (2.50)	74.93 (4.50)	79.06 (2.43)	98.86 (0.98)	75.82 (2.46)
MO	84.07 (2.52)	67.92 (3.78)	71.71 (5.70)	83.66 (3.82)	71.71 (5.70)	84.21 (3.93)	99.58 (0.56)	84.77 (3.41)
PH	83.08 (1.12)	75.16 (0.83)	75.61 (1.27)	81.99 (0.80)	75.75 (1.26)	82.03 (0.82)	97.65 (1.47)	80.89 (1.12)
SE	93.80 (1.01)	87.89 (1.34)	89.50 (1.48)	93.93 (1.00)	89.90 (1.49)	93.71 (1.10)	99.20 (0.65)	88.91 (1.76)
SO	78.75 (6.22)	68.08 (6.22)	67.69 (6.88)	78.08 (6.22)	67.69 (6.88)	76.15 (6.87)	99.71 (0.70)	77.60 (5.62)
TH	96.10 (1.56)	94.88 (1.16)	94.54 (1.32)	95.87 (1.52)	94.54 (1.32)	95.72 (1.69)	99.88 (0.24)	93.35 (1.48)
VH	75.90 (1.85)	66.14 (2.56)	68.91 (3.68)	75.26 (1.74)	68.91 (3.68)	76.11 (2.29)	99.08 (0.82)	73.96 (3.64)
VT	82.27 (5.25)	71.07 (7.03)	75.00 (5.64)	85.00 (2.87)	75.00 (5.64)	84.60 (3.68)	99.67 (0.59)	80.93 (5.58)
WB	96.90 (1.32)	94.72 (1.85)	95.00 (2.50)	95.67 (1.54)	95.00 (2.50)	95.77 (1.53)	99.86 (0.37)	93.87 (2.48)
WV	83.69 (0.91)	76.38 (2.09)	77.84 (3.89)	83.05 (0.98)	77.84 (3.89)	83.73 (0.92)	99.36 (0.32)	80.32 (1.27)
WE	80.33 (4.49)	72.20 (6.01)	73.00 (8.68)	76.47 (5.09)	73.00 (8.68)	80.13 (4.47)	99.73 (0.55)	78.60 (3.31)
WI	96.93 (2.25)	93.30 (2.99)	95.57 (2.81)	94.55 (3.64)	95.57 (2.81)	97.05 (2.46)	100.00 (0.00)	93.64 (3.00)

Tabela 18 – Resultados reportados na conferência internacional ICTAI'18. Cada linha representa os resultados de um dataset e cada coluna representa um método de seleção dinâmica. Os valores nesta tabela representam a acurácia média e o desvio padrão para 20 replicações para cada dataset.

	CLU	DISi	KNE	k-NN	KNOP	KNU	MTD	ORA	RRC
AD	85.35 (2.31)	86.02 (2.76)	81.92 (3.21)	84.91 (2.25)	86.66 (2.57)	86.16 (2.90)	84.36 (2.29)	99.13 (0.74)	86.51 (3.05)
BN	92.04 (1.94)	97.07 (0.73)	96.85 (0.63)	95.75 (1.31)	93.17 (1.53)	96.47 (0.64)	96.56 (0.76)	99.93 (0.12)	84.87 (1.34)
BL	77.83 (2.21)	77.89 (1.89)	73.85 (3.63)	77.62 (1.51)	76.15 (3.09)	76.15 (2.63)	74.22 (3.24)	94.41 (3.76)	73.66 (9.26)
CT	88.98 (0.80)	89.21 (0.76)	89.63 (1.31)	89.67 (0.94)	88.98 (1.02)	88.80 (1.07)	89.97 (1.74)	99.12 (0.44)	87.73 (1.45)
DB	76.43 (2.46)	76.48 (2.34)	73.52 (3.15)	75.23 (2.70)	76.87 (2.76)	77.03 (2.47)	74.27 (2.28)	99.38 (0.52)	76.64 (2.63)
EC	84.64 (2.99)	83.87 (2.95)	82.08 (3.87)	84.70 (3.21)	84.35 (3.26)	83.93 (2.83)	82.68 (2.55)	97.50 (1.09)	83.57 (2.88)
FA	68.21 (1.98)	69.27 (1.70)	68.79 (1.96)	69.39 (1.86)	68.84 (2.00)	68.54 (2.78)	70.12 (1.56)	97.31 (1.03)	64.76 (4.58)
GE	75.38 (2.63)	74.64 (2.48)	71.72 (2.99)	74.58 (2.51)	73.28 (3.24)	72.76 (3.00)	72.18 (3.52)	99.46 (0.57)	73.48 (2.72)
GL	56.04 (3.78)	57.36 (5.59)	62.17 (4.04)	59.62 (5.39)	54.43 (4.63)	55.66 (4.95)	59.53 (5.92)	97.17 (2.63)	49.91 (5.98)
HA	50.39 (2.39)	61.25 (1.54)	62.76 (1.47)	50.39 (2.39)	53.49 (2.15)	52.76 (2.19)	56.45 (2.04)	73.82 (1.38)	50.33 (2.40)
HE	83.36 (2.29)	82.91 (1.97)	77.31 (2.35)	82.09 (1.88)	83.13 (2.70)	82.91 (2.35)	80.30 (3.95)	98.88 (1.74)	82.69 (3.12)
IL	70.72 (1.94)	70.14 (2.17)	69.24 (3.95)	69.24 (2.90)	71.66 (3.02)	71.34 (2.66)	67.17 (3.80)	99.48 (0.63)	71.79 (3.40)
IO	86.65 (2.76)	86.25 (3.15)	87.05 (3.66)	87.27 (2.49)	87.10 (2.89)	86.31 (2.96)	87.33 (2.81)	99.72 (0.63)	86.08 (3.10)
L1	81.98 (5.10)	83.02 (5.05)	78.68 (5.38)	81.89 (5.10)	83.11 (4.26)	82.64 (4.69)	80.38 (3.99)	100.00 (0.00)	82.92 (4.22)
L3	72.39 (2.76)	72.67 (2.75)	69.43 (4.42)	72.61 (3.53)	72.33 (3.36)	72.73 (3.26)	70.85 (3.75)	97.61 (2.90)	71.36 (3.18)
LT	89.65 (1.69)	94.86 (1.11)	95.76 (1.05)	94.48 (1.04)	91.59 (1.48)	95.22 (1.01)	95.46 (0.95)	99.76 (0.31)	82.08 (1.74)
LV	65.35 (4.13)	67.73 (3.99)	66.05 (5.24)	66.28 (3.52)	62.44 (3.50)	58.43 (3.29)	62.67 (5.70)	97.44 (3.45)	62.21 (4.04)
MG	79.40 (0.72)	82.46 (0.47)	80.59 (0.78)	81.84 (0.74)	81.21 (0.95)	80.91 (0.69)	82.33 (0.57)	97.99 (1.23)	76.69 (2.07)
MM	80.31 (3.11)	80.63 (2.71)	76.04 (2.66)	80.48 (2.98)	80.48 (2.54)	80.24 (2.23)	77.49 (3.01)	98.86 (0.98)	79.35 (2.33)
MO	82.31 (2.97)	84.12 (3.85)	88.06 (3.93)	85.23 (3.03)	84.35 (3.85)	82.36 (4.12)	90.74 (3.76)	99.58 (0.56)	80.46 (3.40)
PH	75.24 (1.17)	81.18 (1.05)	83.25 (0.82)	78.98 (1.97)	77.39 (1.28)	77.69 (1.35)	83.43 (0.84)	97.65 (1.47)	71.84 (0.54)
SE	91.69 (1.17)	92.52 (0.94)	94.97 (1.01)	93.47 (1.08)	93.67 (1.01)	92.52 (1.14)	95.03 (0.97)	99.20 (0.65)	90.64 (1.28)
SO	78.17 (5.59)	77.31 (5.57)	80.19 (4.63)	79.23 (4.78)	77.31 (5.28)	77.50 (5.86)	79.71 (4.85)	99.71 (0.70)	78.08 (5.45)
TH	96.82 (1.25)	96.71 (1.01)	95.84 (1.40)	96.82 (1.18)	96.99 (0.93)	96.99 (0.95)	96.45 (1.07)	99.88 (0.24)	96.82 (0.97)
VH	77.30 (2.23)	76.97 (2.54)	77.39 (2.58)	77.51 (1.95)	76.61 (2.62)	75.66 (2.75)	76.80 (2.11)	99.08 (0.82)	74.31 (2.67)
VT	86.00 (2.72)	85.20 (4.10)	84.13 (3.79)	85.93 (2.95)	84.60 (3.76)	83.80 (4.04)	83.80 (3.60)	99.67 (0.59)	81.80 (4.58)
WB	97.32 (1.28)	97.54 (1.34)	96.87 (1.30)	97.50 (1.22)	97.43 (1.41)	97.46 (1.30)	97.29 (1.52)	99.86 (0.37)	97.29 (1.28)
WV	86.04 (0.90)	85.17 (1.02)	83.48 (0.94)	84.95 (0.85)	84.77 (1.31)	83.78 (1.50)	84.33 (1.14)	99.36 (0.32)	82.44 (1.86)
WE	82.00 (4.16)	81.40 (4.65)	80.80 (5.00)	82.20 (4.04)	81.93 (4.27)	81.27 (4.23)	80.27 (3.34)	99.73 (0.55)	80.47 (3.90)
WI	97.84 (2.02)	98.30 (1.93)	97.73 (2.33)	97.84 (1.88)	98.30 (1.93)	98.30 (1.93)	98.07 (2.12)	100.00 (0.00)	98.18 (1.89)

Anexos

Referências

ABRAHAMS, Norman M; ALF, Edward F. Relative costs and statistical power in the extreme groups approach. *Psychometrika*, Springer, v. 43, n. 1, p. 11–17, 1978. Citado na página 38.

ALCALÁ-FDEZ, Jesús; FERNÁNDEZ, Alberto; LUENGO, Julián; DERRAC, Joaquín; GARCÍA, Salvador; SÁNCHEZ, Luciano; HERRERA, Francisco. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, Citeseer, v. 17, 2011. Citado na página 72.

ALEAMONI, Lawrence M; SPENCER, Richard E. A comparison of biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 29, n. 2, p. 353–358, 1969. Citado na página 38.

ALF, Edward F; ABRAHAMS, Norman M. The use of extreme groups in assessing relationships. *Psychometrika*, Springer, v. 40, n. 4, p. 563–572, 1975. Citado na página 38.

BARBOZA, Eduardo V. L.; ALMEIDA, Paulo R. Lisboa De; JUNIOR, Alceu De Souza Britto; SABOURIN, Robert; CRUZ, Rafael M. O. Inca-des: An incremental and adaptive dynamic ensemble selection approach using online k-d tree neighborhood search for data streams with concept drift. *Information Fusion*, abr. 2025. Citado 4 vezes nas páginas 18, 28, 50 e 51.

BENGIO, Yoshua; LOURADOUR, Jerome; COLLOBERT, Ronan; WESTON, Jason. Curriculum learning. *Icml*, 2009. ISSN 0022510X. Citado na página 42.

BINET, Alfred. Attention et adaptation. *L'annee psychologique*, Université René Descartes, Paris 5, v. 6, n. 1, p. 248–404, 1899. Citado na página 37.

BISHOP, Christopher M. *Neural Networks for Pattern Recognition*. [s.n.], 1995. 482 p. ISSN 01621459. ISBN 0198538642. Disponível em: <https://www.jstor.org/stable/2965437?origin=crossref>. Citado na página 24.

BORICH, Gary D; GODBOUT, Robert C. Extreme groups designs and the calculation of statistical power. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 34, n. 3, p. 663–675, 1974. Citado na página 38.

BREIMAN, Leo. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, aug 1996. ISSN 0885-6125. Disponível em: <http://link.springer.com/10.1007/BF00058655>. Citado 2 vezes nas páginas 24 e 26.

BREIMAN, Leo; FRIEDMAN, Jerome; STONE, Charles J.; OLSHEN, R.A. *Classification and Regression Trees*. New York: Chapman & Hall, 1984. 368 p. ISBN 9780412048418. Citado na página 24.

BRENNAN, Robert L. A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 32, n. 2, p. 289–303, 1972. Citado na página 38.

BRIDGMAN, CS. The relation of the upper-lower item discrimination index, d, to the bivariate normal correlation coefficient. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 24, n. 1, p. 85–90, 1964. Citado na página 38.

BRITTO, Alceu de Souza; SABOURIN, Robert; OLIVEIRA, Luiz E. S. Dynamic selection of classifiers - a comprehensive review. *Pattern Recognition*, v. 47, n. 11, p. 3665–3680, 2014. ISSN 00313203. Citado 11 vezes nas páginas 9, 18, 24, 25, 26, 42, 45, 60, 65, 72 e 75.

BRUN, Andre Luiz; BRITTO, Alceu S.; OLIVEIRA, Luiz S.; ENEMBRECK, Fabricio; SABOURIN, Robert. Contribution of data complexity features on dynamic classifier selection. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. [S.l.]: IEEE, 2016. v. 2016-Octob, p. 4396–4403. ISBN 978-1-5090-0620-5. Citado 5 vezes nas páginas 45, 51, 71, 72 e 74.

BURTON, Richard F. Do item-discrimination indices really help us to improve our tests? *Assessment & Evaluation in Higher Education*, Taylor & Francis Group, v. 26, n. 3, p. 213–220, 2001. Citado na página 38.

CAVALIN, Paulo R.; SABOURIN, Robert; SUEN, Ching 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, sep 2012. ISSN 00313203. Citado 2 vezes nas páginas 46 e 51.

CHOI, Ye-Rim; LIM, Dong-Joon. Ddes: A distribution-based dynamic ensemble selection framework. *IEEE Access*, v. 9, p. 40743–40754, 2021. Citado 2 vezes nas páginas 49 e 51.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, jan 1967. ISSN 0018-9448. Disponível

em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1053964http://ieeexplore.ieee.org/document/1053964/>. Citado na página 24.

CROCKER, Linda; ALGINA, James. *Introduction to classical and modern test theory*. [S.l.]: ERIC, 1986. Citado 2 vezes nas páginas 36 e 56.

CRUZ, Rafael M. O.; CAVALCANTI, George D. C.; Ing Ren, Tsang. A method for dynamic ensemble selection based on a filter and an adaptive distance to improve the quality of the regions of competence. In: *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011. p. 1126–1133. ISBN 978-1-4244-9635-8. ISSN 2161-4393. Disponível em: <http://ieeexplore.ieee.org/document/6033350/>. Citado na página 44.

CRUZ, Rafael M. O.; HAFEMANN, Luiz G.; SABOURIN, Robert; CAVALCANTI, George D. C. *DESlib: A Dynamic ensemble selection library in Python*. 2018. Citado 2 vezes nas páginas 72 e 74.

CRUZ, Rafael M. O.; OLIVEIRA, Dayvid VR; CAVALCANTI, George DC; SABOURIN, Robert. Fire-des++: Enhanced online pruning of base classifiers for dynamic ensemble selection. *Pattern Recognition*, Elsevier, v. 85, p. 149–160, 2019. Citado 2 vezes nas páginas 48 e 51.

CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George DC. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, v. 41, p. 195–216, 2018. ISSN 15662535. Citado 10 vezes nas páginas 18, 26, 27, 28, 29, 30, 60, 71, 72 e 75.

CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George D. C. Analyzing different prototype selection techniques for dynamic classifier and ensemble selection. *Neural Computing and Applications*, 2016. ISSN 0941-0643. Citado 3 vezes nas páginas 43, 44 e 93.

CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George D. C. Meta-des.oracle: Meta-learning and feature selection for dynamic ensemble selection. *Information Fusion*, Elsevier B.V., v. 38, p. 84–103, nov 2017. ISSN 15662535. Citado 2 vezes nas páginas 46 e 51.

CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George D. C. Prototype selection techniques for dynamic classifier and ensemble selection. *Proceedings of the International Joint Conference on Neural Networks*, Springer London, v. 2017-May, n. July 2016, p. 3959–3966, 2017. ISSN 09410643. Citado 3 vezes nas páginas 28, 44 e 51.

CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George D. C.; Ing Ren, Tsang. Meta-des: A dynamic ensemble selection framework using meta-learning. *Pattern Recognition*, Elsevier, v. 48, n. 5, p. 1925–1935, may 2015. ISSN 00313203. Citado 3 vezes nas páginas 46, 51 e 74.

CURETON, Edward E. Simplified formulas for item analysis. *Journal of Educational Measurement*, JSTOR, v. 3, n. 2, p. 187–189, 1966. Citado na página 38.

DANG, Manh Truong; LUONG, Anh Vu; VU, Tuyet-Trinh; NGUYEN, Quoc Viet Hung; NGUYEN, Tien Thanh; STANTIC, Bela. An ensemble system with random projection and dynamic ensemble selection. In: NGUYEN, Ngoc Thanh; HOANG, Duong Hung; HONG, Tzung-Pei; PHAM, Hoang; TRAWIŃSKI, Bogdan (Ed.). Dong Hoi City, Vietnam: Springer International Publishing, 2018, (Lecture Notes in Computer Science, v. 10751). p. 576–586. Citado na página 26.

DAVTALAB, Reza. *Dynamic ensemble selection using fuzzy min-max hyperboxes*. Tese (Doutorado) — École de technologie supérieure, 2023. Citado 3 vezes nas páginas 48, 49 e 52.

DAVTALAB, Reza; CRUZ, Rafael M. O.; SABOURIN, Robert. Dynamic ensemble selection using fuzzy hyperboxes. *IEEE International Joint Conference on Neural Network*, jul. 2022. Citado na página 49.

DAVTALAB, Reza; CRUZ, Rafael M. O.; SABOURIN, Robert. A scalable dynamic ensemble selection using fuzzy hyperboxes. v. 102, p. 102036–102036, 2023. Citado 3 vezes nas páginas 28, 49 e 51.

DEMŠAR, Janez. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, JMLR. org, v. 7, p. 1–30, 2006. Citado 4 vezes nas páginas 64, 74, 75 e 76.

DHEERU, Dua; TANISKIDOU, Efi Karra. *UCI Machine Learning Repository*. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. Citado na página 72.

DIETTERICH, Thomas G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, v. 10, n. 7, p. 1895–1923, oct 1998. ISSN 0899-7667. Disponível em: <http://www.mitpressjournals.org/doi/10.1162/089976698300017197>. Citado na página 74.

DOPPELT, Jerome E; POTTS, Edith M. The constancy of item-test correlation coefficients computed from upper and lower groups. *Journal of Educational Psychology*, Warwick & York, v. 40, n. 6, p. 378, 1949. Citado na página 38.

DUIN, RPW; JUSZCZAK, P; PACLIK, P; PEKALSKA, E; RIDDER, D De; TAX, DMJ; VERZAKOV, S. A matlab toolbox for pattern recognition. *PRTools version*, Citeseer, v. 3, p. 109–111, 2000. Citado na página 72.

DUIN, Robert P. W. The combining classifier: to train or not to train? In: *Object recognition supported by user interaction for service robots*. IEEE Comput. Soc, 2002. v. 2, p. 765–770. ISBN 0-7695-1695-X. ISSN 1051-4651. Disponível em: <http://ieeexplore.ieee.org/document/1048415/>. Citado na página 29.

EFRON, Bradley; TIBSHIRANI, Robert J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994. Citado na página 26.

ELMI, Javad; EFTEKHARI, Mahdi. Dynamic ensemble selection based on hesitant fuzzy multiple criteria decision making. *Studies in fuzziness and soft computing*, p. 1–13, jan. 2020. Citado 2 vezes nas páginas 49 e 51.

ELMI, Javad; EFTEKHARI, Mahdi; MEHRPOOYA, Adel; RAVARI, Mohammad Rezaei. A novel framework based on the multi-label classification for dynamic selection of classifiers. *International Journal of Machine Learning and Cybernetics*, Springer, v. 14, n. 6, p. 2137–2154, 2023. Citado 2 vezes nas páginas 47 e 51.

ENGELHART, Max D. A comparison of several item discrimination indices 1. *Journal of Educational Measurement*, Wiley Online Library, v. 2, n. 1, p. 69–76, 1965. Citado na página 38.

ESHELMAN, Larry J. The chc adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. In: . Morgan Kaufmann Publishers, Inc., 1991. v. 1, n. 1975, p. 265–283. Disponível em: <http://dx.doi.org/10.1016/B978-0-08-050684-5.50020-3>. Citado na página 44.

FARHANGIAN, Faramarz; ENSINA, Leandro A.; CAVALCANTI, George D. C.; CRUZ, Rafael M. O. Dres: Fake news detection by dynamic representation and ensemble selection. set. 2025. Citado 2 vezes nas páginas 50 e 51.

FELDT, Leonard S. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, Springer, v. 26, n. 3, p. 307–316, 1961. Citado 2 vezes nas páginas 37 e 38.

FELDT, Leonard S. Note on use of extreme criterion groups in item discrimination analysis. *Psychometrika*, Springer, v. 28, n. 1, p. 97–104, 1963. Citado na página 38.

FERN, Xiaoli Z; BRODLEY, Carla E. Random projection for high dimensional data clustering: A cluster ensemble approach. In: *Twentieth International Conference on International Conference on Machine Learning*. Washington, DC, USA: AAAI Press, 2003. p. 186–193. ISBN 1-57735-189-4. Citado na página 26.

FERNÁNDEZ, Alberto; GARCÍA, Salvador; JESUS, María José del; HERRERA, Francisco. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, v. 159, n. 18, p. 2378–2398, sep 2008. ISSN 01650114. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0165011407005660>. Citado na página 28.

FINDLEY, Warren G. A rationale for evaluation of item discrimination statistics. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 16, n. 2, p. 175–180, 1956. Citado na página 38.

FORLANO, George; PINTER, Rudolf. Selection of upper and lower groups for item validation. *Journal of Educational Psychology*, Warwick & York, v. 32, n. 7, p. 544, 1941. Citado na página 38.

FOWLER, Robert L. Using the extreme groups strategy when measures are not normally distributed. *Applied Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 16, n. 3, p. 249–259, 1992. Citado 2 vezes nas páginas 37 e 38.

FRIEDMAN, Milton. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, v. 32, n. 200, p. 675–701, dec 1937. ISSN 0162-1459. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522>. Citado 2 vezes nas páginas 59 e 76.

GABRYS, B.; BARGIELA, A. General fuzzy min-max neural network for clustering and classification. *IEEE Transactions on Neural Networks*, v. 11, n. 3, p. 769–783, maio 2000. ISSN 1045-9227. Citado na página 49.

GARCÍA, Salvador; CANO, José Ramón; HERRERA, Francisco. A memetic algorithm for evolutionary prototype selection: A scaling up approach. *Pattern Recognit.*, v. 41, p. 2693–2709, 2008. ISSN 00313203. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0031320308000770>. Citado na página 44.

GARCIA, Salvador; DERRAC, Joaquín; CANO, José Ramón; HERRERA, Francisco. Prototype selection for nearest neighbor classification: Taxonomy and

empirical study. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE, p. 417–435, 2011. Citado na página 44.

GARCÍA, Salvador; FERNÁNDEZ, Alberto; LUENGO, Julián; HERRERA, Francisco. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, Elsevier Inc., v. 180, n. 10, p. 2044–2064, may 2010. ISSN 00200255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2009.12.010http://linkinghub.elsevier.com/retrieve/pii/S0020025509005404>. Citado na página 74.

GARG, Rashmi. An empirical comparison of three strategies used in extreme group designs. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 43, n. 2, p. 359–371, 1983. Citado na página 38.

GAULT, Robert H. A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, Taylor & Francis, v. 14, n. 3, p. 366–383, 1907. Citado na página 30.

GIACINTO, Giorgio; ROLI, Fabio. Methods for dynamic classifier selection. *Proceedings - International Conference on Image Analysis and Processing, ICIAP 1999*, p. 659–664, 1999. ISSN 03029743. Citado 2 vezes nas páginas 45 e 51.

GIACINTO, Giorgio; ROLI, Fabio. Dynamic classifier selection based on multiple classifier behaviour. *Pattern Recognition*, v. 34, n. 9, p. 1879–1881, sep 2001. ISSN 00313203. Citado 2 vezes nas páginas 45 e 51.

GULLIKSEN, Harold. *Theory of Mental Tests*. Routledge, 2013. Disponível em: <https://doi.org/10.4324/9780203052150>. Citado na página 30.

HAGHIGHI, Sepand; JASEMI, Masoomeh; HESSABI, Shaahin; ZOLANVARI, Alireza. Pycm: Multiclass confusion matrix library in python. *Journal of Open Source Software*, The Open Journal, v. 3, n. 25, p. 729, may 2018. Disponível em: <https://doi.org/10.21105/joss.00729>. Citado na página 74.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning*. 2. ed. New York, NY: Springer New York, 2009. v. 1. 745 p. (Springer Series in Statistics, v. 1). ISSN 0172-7397. ISBN 978-0-387-84857-0. Disponível em: <http://www.springerlink.com/index/10.1007/b94608http://link.springer.com/10.1007/978-0-387-84858-7>. Citado 2 vezes nas páginas 24 e 43.

HO, Tin Kam. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

v. 20, n. 8, p. 832–844, 1998. ISSN 01628828. Disponível em: <http://ieeexplore.ieee.org/document/709601/>. Citado na página 26.

HUANG, Y.S.; SUEN, C.Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 17, n. 1, p. 90–94, 1995. ISSN 01628828. Disponível em: <http://ieeexplore.ieee.org/document/368145/>. Citado na página 45.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 74.

JAIN, Ak; DUIN, Rpw; MAO, J. Statistical pattern recognition: A review. *IEEE transactions on pattern analysis and machine intelligence*, v. 22, n. 1, p. 4–37, 2000. ISSN 01628828. Citado na página 24.

John von Neumann; Oscar Morgenstern. *Theory of games and economic behavior*. Princeton University, 1945. Disponível em: <https://doi.org/10.1038/157172a0>. Citado na página 47.

JOHNSON, A Pemberton. Notes on a suggested index of item validity: The ul index. *Journal of Educational Psychology*, Warwick & York, v. 42, n. 8, p. 499, 1951. Citado na página 38.

JONES, Eric; OLIPHANT, Travis; PETERSON, Pearu et al. *SciPy: Open source scientific tools for Python*. 2001–. [Online; accessed 09-Mai-2018]. Disponível em: <http://www.scipy.org/>. Citado na página 74.

KELLEY, Truman L. The selection of upper and lower groups for the validation of test items. *Journal of educational psychology*, Warwick & York, v. 30, n. 1, p. 17, 1939. Citado na página 38.

KITTLER, J.; HATEF, M.; DUIN, R.P.W.; MATAS, J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 20, n. 3, p. 226–239, mar 1998. ISSN 01628828. Disponível em: <http://ieeexplore.ieee.org/document/547205/http://ieeexplore.ieee.org/document/667881/>. Citado 2 vezes nas páginas 29 e 42.

KO, Albert Hung Ren; SABOURIN, Robert; BRITTO, Alceu de Souza. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, v. 41, n. 5, p. 1718–1731, may 2008. Citado 4 vezes nas páginas 46, 51, 71 e 74.

KUMAR, Gulshan; KUMAR, Krishan. The use of artificial-intelligence-based ensembles for intrusion detection: A review. *Applied Computational Intelligence and Soft Computing*, v. 2012, p. 1–20, 2012. ISSN 1687-9724. Disponível em: <http://www.hindawi.com/journals/acisc/2012/850160/>. Citado na página 26.

KUNCHEVA, L.I. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 2, p. 281–286, 2002. ISSN 01628828. Citado 3 vezes nas páginas 19, 42 e 65.

KUNCHEVA, L. *Ludmila kuncheva collection of real medical data sets*. 2004. Accessed: 12-April-2018. Disponível em: <http://pages.bangor.ac.uk/~mas00a/activities/real_data.htm>. Citado na página 72.

KUNCHEVA, Ludmila; WHITAKER, Christopher. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. In: *IEE Workshop Intelligent Sensor Processing*. [S.l.]: Springer, 2003. v. 51, n. 2, p. 181–207. Citado 2 vezes nas páginas 26 e 87.

KUNCHEVA, Ludmila I. Editing for the k-nearest neighbors rule by a genetic algorithm. *Pattern recognition letters*, Elsevier, v. 16, n. 8, p. 809–814, 1995. Citado na página 44.

LEI, Pui-Wa; DUNBAR, Stephen B; KOLEN, Michael J. A comparison of parametric and nonparametric approaches to item analysis for multiple-choice tests. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 64, n. 4, p. 565–587, 2004. Citado na página 38.

LI, Danyang; WEN, Guihua; LI, Xu; CAI, Xianfa. Graph-based dynamic ensemble pruning for facial expression recognition. *Applied Intelligence*, Springer, p. 1–19, 2019. Citado 2 vezes nas páginas 47 e 51.

LIMA, Tiago Pessoa Ferreira De; LUDERMIR, Teresa Bernarda. Optimizing dynamic ensemble selection procedure by evolutionary extreme learning machines and a noise reduction filter. In: *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*. IEEE, 2013. p. 546–552. ISBN 978-1-4799-2972-6. ISSN 10823409. Disponível em: <http://ieeexplore.ieee.org/document/6735298/>. Citado na página 19.

LIMA, Tiago Pessoa Ferreira de; SERGIO, Anderson Tenorio; LUDERMIR, Teresa Bernarda. Improving classifiers and regions of competence in dynamic ensemble selection. In: *2014 Brazilian Conference on Intelligent Systems*. IEEE, 2014. p. 13–18. ISBN 9781479956180. Disponível em: <http://ieeexplore.ieee.org/document/6984800/>. Citado na página 19.

LIU, Fu. *Comparison of several popular discrimination indices based on different criteria and their application in item analysis*. Athens, GA: [s.n.], 2008. (Master Thesis). Citado na página 38.

LORD, Frederic M. A theory of test scores. In: . [s.n.], 1952. Disponível em: <https://api.semanticscholar.org/CorpusID:118536487>. Citado na página 30.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de; GAMA, João M. P. A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, v. 30, n. 1-4, p. 19–37, dec 2008. ISSN 0269-2821. Disponível em: <http://link.springer.com/10.1007/s10462-009-9114-9>. Citado na página 72.

LORENA, Ana C.; GARCIA, Luís P. F.; LEHMANN, Jens; SOUTO, Marcilio C. P.; HO, Tin Kam. How complex is your classification problem?: A survey on measuring classification complexity. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 52, n. 5, p. 107:1–107:34, 2019. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/3347711>. Citado na página 82.

MARKOWITZ, Harry. Portfolio selection. *The Journal of Finance*, Wiley, v. 7, n. 1, p. 77–91, mar 1952. Disponível em: <https://doi.org/10.1111/j.1540-6261.1952.tb01525.x>. Citado na página 43.

MASOUDNIA, Saeed; EBRAHIMPOUR, Reza. Mixture of experts: A literature survey. *Artificial Intelligence Review*, v. 42, n. 2, p. 275–293, 2014. ISSN 02692821. Citado na página 29.

MATLOCK-HETZEL, Susan. Basic concepts in item and test analysis. In: *Annual Meeting of the Southwest Educational Research Association*. Austin: [s.n.], 1997. p. 1–7. Citado 3 vezes nas páginas 11, 38 e 40.

METSÄMUURONEN, Jari. Underestimation of the item discrimination power and somers' d as an alternative for the item–total-and item–rest correlations. *Preprint. DOI: http://dx. doi. org/10.13140/RG*, v. 2, n. 18487.16803, 2019. Citado na página 37.

NARASSIGUIN, Anil; ELGHAZEL, Haytham; AUSSEM, Alex. Dynamic ensemble selection with probabilistic classifier chains. In: *Machine Learning and Knowledge Discovery in Databases*. [S.l.]: Springer International Publishing, 2017. p. 169–186. ISBN 978-3-319-71249-9. Citado 3 vezes nas páginas 28, 47 e 51.

NOURBAKHSH, Azamossadat; HOSEINPOUR, Mohaddeseh Mohammad. Multiple feature extraction and multiple classifier systems in face recognition. In:

SPRINGER. *Proceedings of the Computational Methods in Systems and Software*. [S.l.], 2017. p. 111–122. Citado na página 73.

OLIPHANT, T. E. Python for scientific computing. *Computing in Science Engineering*, v. 9, n. 3, p. 10–20, May 2007. ISSN 1521-9615. Citado 2 vezes nas páginas 73 e 74.

OLIVEIRA, Dayvid VR; CAVALCANTI, George DC; SABOURIN, Robert. Online pruning of base classifiers for dynamic ensemble selection. *Pattern Recognition*, Elsevier, v. 72, p. 44–58, 2017. Citado 2 vezes nas páginas 48 e 51.

OLVERA-LÓPEZ, J. Arturo; CARRASCO-OCHOA, J. Ariel; MARTÍNEZ-TRINIDAD, J. Francisco; KITTLER, Josef. A review of instance selection methods. v. 34, n. 2, p. 133–143, 2010. ISSN 0269-2821. Disponível em: <http://link.springer.com/10.1007/s10462-010-9165-y>. Citado na página 44.

OOSTERHOF, Albert C. Similarity of various item discrimination indices. *Journal of Educational Measurement*, JSTOR, p. 145–150, 1976. Citado na página 38.

PASTEUR, Louis. *Nouvelles expériences relatives aux générations dites spontanées*. [S.l.]: Mallet-Bachelier, 1860. Citado na página 35.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 74.

PINTO, Fábio; SOARES, Carlos; MENDES-MOREIRA, João. Chade: Metalearning with classifier chains for dynamic combination of classifiers. In: *Machine Learning and Knowledge Discovery in Databases*. [S.l.]: Springer International Publishing, 2016. p. 410–425. ISBN 978-3-319-46128-1. Citado 5 vezes nas páginas 19, 28, 47, 50 e 51.

PREACHER, Kristopher J. Extreme groups designs. *The encyclopedia of clinical psychology*, John Wiley & Sons, Inc. Hoboken, NJ, USA, p. 1–4, 2014. Citado na página 38.

PREACHER, Kristopher J; RUCKER, Derek D; MACCALLUM, Robert C; NICEWANDER, W Alan. Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological methods*, American Psychological Association, v. 10, n. 2, p. 178, 2005. Citado 2 vezes nas páginas 37 e 38.

ROKACH, Lior. Ensemble-based classifiers. *Artificial Intelligence Review*, v. 33, n. 1-2, p. 1–39, feb 2010. ISSN 0269-2821. Disponível em: <http://link.springer.com/10.1007/s10462-009-9124-7>. Citado na página 18.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, American Psychological Association (APA), v. 65, n. 6, p. 386–408, 1958. Disponível em: <https://doi.org/10.1037/h0042519>. Citado na página 24.

ROSSUM, Guido van. *Python Library Reference*. National Research Institute for Mathematics and Computer Science, Amsterdam, 1995. [Online; accessed 09-May-2018]. Disponível em: <https://ir.cwi.nl/pub/5009/05009D.pdf>. Citado na página 74.

ROY, Anandarup; CRUZ, Rafael M. O.; SABOURIN, Robert; CAVALCANTI, George D. C. A study on combining dynamic selection and data preprocessing for imbalance learning. *Neurocomputing*, Elsevier B.V., n. February, 2018. ISSN 09252312. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0925231218300936>. Citado na página 28.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, oct 1986. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/323533a0>. Citado na página 24.

RUTA, Dymitr; GABRYS, Bogdan. Classifier selection for majority voting. *Information Fusion*, v. 6, n. 1, p. 63–81, mar 2005. ISSN 15662535. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1566253504000417>. Citado na página 42.

SABOURIN, M.; MITICHE, A.; THOMAS, D.; NAGY, G. Classifier combination for hand-printed digit recognition. In: *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*. IEEE Comput. Soc. Press, 1993. p. 163–166. ISBN 0-8186-4960-7. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=395758http://ieeexplore.ieee.org/document/395758/>. Citado 2 vezes nas páginas 44 e 51.

SALZBERG, Steven L. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, v. 1, n. 3, p. 317–328, 1997. ISSN 13845810. Disponível em: <http://link.springer.com/10.1023/A:1009752403260>. Citado na página 75.

SÁNCHEZ, J. S.; PLA, F.; FERRI, F. J. Prototype selection for the nearest neighbour rule through proximity graphs. *Pattern Recognition Letters*, v. 18, n. 6, p. 507–513, 1997. ISSN 01678655. Citado na página 44.

SANTANA, Alixandre; SOARES, Rodrigo G.F.; CANUTO, Anne M. P.; De Souto, Marcilio C. P. A dynamic classifier selection method to build ensembles using accuracy and diversity. *Proceedings of the Ninth Brazilian Symposium on Neural Networks, SBRN'06*, p. 2–7, 2006. Citado 2 vezes nas páginas 45 e 51.

SANTOS, Eulanda M.; SABOURIN, Robert; MAUPIN, Patrick. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition*, v. 41, n. 10, p. 2993–3009, 2008. ISSN 00313203. Citado 2 vezes nas páginas 47 e 51.

SCHAPIRE, Robert E. The strength of weak learnability. *Machine Learning*, v. 5, n. 2, p. 197–227, jun 1990. ISSN 0885-6125. Disponível em: <http://link.springer.com/article/10.1007/BF00116037http://link.springer.com/10.1007/BF00116037>. Citado 2 vezes nas páginas 24 e 73.

SHAPLEY, Lloyd S et al. A value for n-person games. Princeton University Press Princeton, 1953. Citado na página 47.

SHESKIN, David J. *Handbook of parametric and nonparametric statistical procedures*. [S.l.]: crc Press, 2003. Citado na página 75.

SKALAK, David B. Prototype and feature selection by sampling and random mutation hill climbing algorithms. In: . Elsevier, 1994. p. 293–301. ISBN 1558603352. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B978155860335650043X>. Citado na página 44.

SKURICHINA, Marina; DUIN, Robert P W. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, v. 5, n. 2, p. 121–135, jun 2002. ISSN 1433-7541. Disponível em: <http://link.springer.com/10.1007/s100440200011>. Citado na página 24.

SMITH, Michael R.; MARTINEZ, Tony. A comparative evaluation of curriculum learning with filtering and boosting in supervised classification problems. *Computational Intelligence*, v. 32, n. 2, p. 167–195, 2016. ISSN 14678640. Citado na página 43.

SMITH, Michael R.; MARTINEZ, Tony; GIRAUD-CARRIER, Christophe. An instance level analysis of data complexity. *Machine Learning*, Springer Nature, v. 95, n. 2, p. 225–256, nov 2013. Citado na página 82.

SMITS, P.C. Multiple classifier systems for supervised remote sensing image classification based on dynamic classifier selection. *IEEE Transactions on Geoscience and Remote Sensing*, v. 40, n. 4, p. 801–813, apr 2002. ISSN 0196-2892. Citado 2 vezes nas páginas 45 e 51.

SOARES, Rodrigo G. F.; SANTANA, Alixandre; CANUTO, Anne M. P.; SOUTO, Marcilio C. P. de. Using accuracy and diversity to select classifiers to build ensembles. In: *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. [S.l.]: IEEE, 2006. p. 1310–1316. ISBN 0-7803-9490-9. ISSN 10987576. Citado 2 vezes nas páginas 45 e 51.

SOUZA, Mariana A.; CAVALCANTI, George D.C.; CRUZ, Rafael M.O.; SABOURIN, Robert. Online local pool generation for dynamic classifier selection. *Pattern Recognition*, Elsevier Ltd, v. 85, p. 132–148, jan 2019. ISSN 00313203. Disponível em: <https://doi.org/10.1016/j.patcog.2018.08.004https://linkinghub.elsevier.com/retrieve/pii/S0031320318302802>. Citado 4 vezes nas páginas 26, 48, 51 e 62.

SOUZA, Mariana A.; CAVALCANTI, George D. C.; CRUZ, Rafael M. O.; SABOURIN, Robert. On the characterization of the oracle for dynamic classifier selection. *Proceedings of the International Joint Conference on Neural Networks*, v. 2017-May, n. October, p. 332–339, 2017. Citado 4 vezes nas páginas 65, 69, 71 e 72.

SOUZA, Mariana A.; SABOURIN, Robert; CAVALCANTI, George D.C.; CRUZ, Rafael M.O. A dynamic multiple classifier system using graph neural network for high dimensional overlapped data. *Information Fusion*, v. 103, p. 102145, 2024. ISSN 1566-2535. Citado 4 vezes nas páginas 19, 28, 48 e 51.

SOUZA, Mariana Assunção de; SABOURIN, Robert; CAVALCANTI, George D. C.; CRUZ, Rafael M. O. Olp++: An online local classifier for high dimensional data. *Information Fusion*, 2022. Citado 2 vezes nas páginas 48 e 51.

SOUZA, Mariana de Araújo; SABOURIN, Robert; CAVALCANTI, George Darmiton da Cunha; CRUZ, Rafael M. O. Gnn-des: A new end-to-end dynamic ensemble selection method based on multi-label graph neural network. p. 59–69, ago. 2023. Citado 2 vezes nas páginas 48 e 51.

STONE, Benjamin Wright; Mark. *Measurement Essentials*. Wilmington, Delaware: Wide Range, Inc, 1999. Citado na página 31.

SUROWIECKI, James. *The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and*

Nations. reprint. [S.l.]: Doubleday, 2004. 296 p. ISBN 0385503865, 9780385503860. Citado na página 18.

TRIERVEILER, Marcelo Ribeiro; BRITTO, Alceu de Souza; OLIVEIRA, Luiz; SABOURIN, Robert. Dynamic ensemble selection by k-nearest local oracles with discrimination index. In: *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE Computer Society, 2018. ISBN 978-1-5386-7449-9. ISSN 2375-0197. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/ICTAI.2018.00120>. Citado 9 vezes nas páginas 22, 46, 51, 57, 58, 71, 72, 76 e 98.

TRIGUERO, Isaac; DERRAC, Joaquín; GARCIA, Salvador; HERRERA, Francisco. A taxonomy and experimental study on prototype generation for nearest neighbor classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 42, n. 1, p. 86–100, jan 2012. ISSN 1094-6977. Disponível em: <http://ieeexplore.ieee.org/document/5709998/>. Citado na página 44.

TUCKER, L.R. Maximum validity of a test with equivalent items. In: . [s.n.], 1946. Disponível em: <https://doi.org/10.1007/BF02288894>. Citado na página 30.

VAPNIK, Vladimir. *The Nature of Statistical Learning Theory*. [S.l.]: Springer New York, 2013. (Information Science and Statistics). ISBN 9781475732641. Citado na página 24.

VRIESMANN, Leila Maria; BRITTO, Alceu S.; Soares Oliveira, Luiz Eduardo; SABOURIN, Robert. Using additional neighborhood information in a dynamic ensemble selection method: improving the knora approach. *17th International Conference on Systems, Signals and Image Processing Using*, p. 6–9, 2010. Citado na página 71.

VYGOTSKY, Lev. Zone of proximal development. *Mind in society: The development of higher psychological processes*, Harvard University Press Cambridge, MA, v. 5291, p. 157, 1987. Citado na página 42.

WANG, Jigang; NESKOVIC, Predrag; COOPER, Leon N. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters*, v. 28, n. 2, p. 207–213, jan 2007. ISSN 01678655. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167865506001917>. Citado na página 44.

WHITNEY, Douglas R; SABERS, Darrell L. Two generalizations of the item discrimination index to multi-score items. *The Journal of Experimental Education*, Taylor & Francis, v. 39, n. 3, p. 88–92, 1971. Citado na página 38.

WIERSMA, W; JURS, S G. *Educational measurement and testing*. [S.l.]: Allyn and Bacon, 1985. Citado na página 38.

WILCOXON, Frank. Individual comparisons by ranking methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80, dec 1945. ISSN 00994987. Disponível em: <http://www.jstor.org/stable/10.2307/3001968?origin=crossrefhttps://www.jstor.org/stable/10.2307/3001968?origin=crossref>. Citado na página 75.

WILSON, Dennis L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2, n. 3, p. 408–421, jul 1972. ISSN 0018-9472. Disponível em: <http://ieeexplore.ieee.org/document/4309137/>. Citado na página 44.

WOLOSZYNSKI, Tomasz; KURZYNSKI, Marek. A measure of competence based on randomized reference classifier for dynamic ensemble selection. In: *2010 20th International Conference on Pattern Recognition*. IEEE, 2010. p. 4194–4197. ISBN 978-1-4244-7542-1. Disponível em: <http://ieeexplore.ieee.org/document/5597745/>. Citado na página 46.

WOLOSZYNSKI, Tomasz; KURZYNSKI, Marek. A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognition*, Elsevier BV, v. 44, n. 10-11, p. 2656–2668, oct 2011. Citado 2 vezes nas páginas 46 e 51.

WOLPERT, David H. The lack of a priori distinctions between learning algorithms. *Neural Computation*, v. 8, n. 7, p. 1391–1420, 1996. ISSN 0899-7667. Citado 2 vezes nas páginas 24 e 75.

WOODS, Kevin; KEGELMEYER, W.P.; BOWYER, Kevin. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 4, p. 405–410, apr 1997. ISSN 01628828. Citado 3 vezes nas páginas 44, 45 e 51.

WOŹNIAK, Michał; GRAÑA, Manuel; CORCHADO, Emilio. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, v. 16, n. 1, p. 3–17, 2014. ISSN 15662535. Citado na página 26.

WRIGHT BENJAMIN DRAKE; STONE, Mark H. *Best Test Design (Rasch Measurement Series)*. 1. ed. [S.l.]: MESA Press, 1979. ISBN 978-0941938006. Citado 5 vezes nas páginas 9, 31, 32, 33 e 34.

ZHANG, ZL; ZHU, YH; LUO, XG. Des-as: Dynamic ensemble selection based on algorithm shapley. *Pattern Recognition*, jan. 2025. Citado 4 vezes nas páginas 19, 28, 47 e 51.

ZUBEK, Julian; KUNCHEVA, Ludmila. *Learning from Exemplars and Prototypes in Machine Learning and Psychology*. [S.l.], 2018. Disponível em: <http://arxiv.org/abs/1806.01130>. Citado na página 44.