Análise da relação entre medidas de complexidade de dados e o desempenho de classificadores

Vitor Lisboa Nogueira
Universidade Estadual do Oeste do Paraná - UNIOESTE
Cascavel-PR, Brasil
vitorln10@hotmail.com

André Luiz Brun
Universidade Estadual do Oeste do Paraná - UNIOESTE
Cascavel-PR, Brasil
andre.brun@unioeste.br

***Resumo*—This paper examines the relationship between the descriptors for the complexity of the dataset and the performance of the classifier. The main objective was to analyze whether classifiers trained on data of similar complexity perform similarly in terms of accuracy. The complexity measures were divided into three categories: Overlap, Class Separability and Data Geometry, Topology and Density. The experiment, conducted on 26 datasets, revealed that some metrics show a significant correlation with classification performance. Among them, L1, T1, F2 and L2 stand out, showing a behavior more consistent with accuracy. In contrast, F3, N1, F4 and N3 were the metrics with the least similarity. These findings contribute to a deeper understanding of how intrinsic data characteristics affect the classification task, indicating that prior knowledge of complexity descriptors can help estimate the accuracy rate of classification models.**

***Keywords*—complexity descriptors; classifier performance; data influence.**

***Resumo*—Este trabalho investiga a relação entre descritores de complexidade de conjuntos de dados e o desempenho de classificadores. O objetivo central foi analisar se classificadores treinados com dados de complexidade similar apresentam desempenhos semelhantes em termos de acurácia. Foram utilizadas medidas de complexidade agrupadas em três categorias: sobreposição, separabilidade das classes e características de geometria, topologia e densidade dos dados. O experimento, que foi conduzido sobre 26 bases de dados, indicou que algumas métricas apresentam relação significativa com o desempenho dos classificadores. Dentre elas, destacam-se L1, T1, F2 e L2, que mostraram comportamento mais alinhado à acurácia. Em contrapartida, F3, N1, F4 e N3 foram as métricas com menor similaridade. Esses achados contribuem para uma compreensão mais profunda sobre o impacto das características intrínsecas dos dados na tarefa de classificação indicando que um conhecimento prévio dos descritores de complexidade pode auxiliar na estimação da taxa de acertos dos modelos de classificação.**

***Palavras-chave*—descritores de complexidade; desempenho de classificadores; influência dos dados.**

I. Introdução

O reconhecimento de padrões é uma área da Ciência da Computação e da Inteligência Artificial que tem como objetivo identificar regularidades ou estruturas em dados, permitindo que sistemas computacionais interpretem e tomem decisões com base em informações complexas. Historicamente fundamentado em métodos estatísticos e matemáticos, esse campo evoluiu significativamente com o avanço das tecnologias recentes, especialmente com o surgimento de técnicas inteligentes como o aprendizado de máquina e redes neurais profundas. Essas inovações permitiram o desenvolvimento de sistemas mais precisos e adaptativos, capazes de lidar com grandes volumes de dados e contextos altamente variáveis.

Entre as diversas aplicações do reconhecimento de padrões, destaca-se a tarefa de classificação, que trata-se de um problema supervisionado e consiste em atribuir categorias ou rótulos a objetos, sinais ou eventos com base em suas características — sendo amplamente utilizada em áreas como diagnóstico médico, visão computacional, segurança e análise de dados.

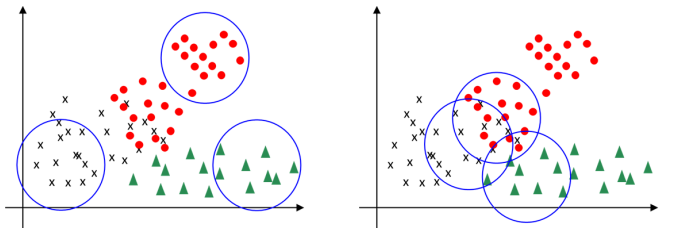
A estratégia de classificação mais comum é a monolítica, que consiste em um único modelo responsável por atribuir a classe a todos os objetos do conjunto. Contudo, tais modelos podem ter problemas para lidar com conjuntos que apresentam maior variabilidade. Visando suplantar tal dificuldade surgem, como alternativa, os Sistemas de Múltiplos Classificadores (SMCs) – sistema com diversos modelos distintos que partem da premissa que classificadores diferentes geram erros diferentes – que visam responder melhor à variabilidade de certos problemas [1].

Os SMCs podem ser divididos em três fases: geração, seleção e integração. Na primeira, um *pool* de classificadores é gerado, podendo variar entre a estratégia homogênea – classificadores com a mesma técnica de aprendizagem, porém treinados em conjuntos distintos (ou com parâmetros distintos) – ou heterogênea – classificadores com técnicas de aprendizagem distintas, porém treinados no mesmo conjunto. Na segunda fase, um subconjunto destes classificadores é selecionado segundo um critério pré estabelecido, enquanto na última fase, uma decisão final é tomada com base nas predições dos classificadores selecionados [2].

O comportamento dos classificadores é dependente do con-

junto de dados em que foram treinados, por isso tem-se a preocupação em como gerar subconjuntos de aprendizado mais adequados. Em problemas mais simples, as instâncias têm classes com atributos mais distintos, sendo mais fáceis diferenciá-las. Em conjuntos mais complexos, as classes têm atributos sobrepostos, tornando difícil a distinção das instâncias de classes diferentes.

Considere, por exemplo, o conjunto apresentado na Figura 1. Na ilustração foram escolhidas 14 instâncias das classes "X" e "◯" e oito da classe "△" para se efetuar o treinamento. No primeiro caso, o *Bagging* escolheu as instâncias destacadas em azul (Figura 1(a)). Percebe-se que as amostras selecionadas são bastante distintas entre os dois grupos, tendo suas características nos eixos X e Y com valores bem diferentes. Neste caso, o problema apresentará uma complexidade reduzida. No segundo exemplo (Figura 1(b)), nota-se que as instâncias selecionadas estão bem mais próximas, tendo valores nos dois eixos sobrepostos. Os índices de complexidade deste grupo mostrarão valores mais elevados, caracterizando o problema como mais difícil.



(a) Seleção facilmente separável (b) Seleção intrincada de exemplares

Figura 1: Exemplo de estimação de complexidade

Para tentar detalhar os conjuntos de dados foram propostas as medidas de complexidade, que tentam, através de índices, inferir o grau de complexidade do problema a ser resolvido. Tais medidas são comumente separadas em três categorias [3]: sobreposição; separabilidade das classes e medidas de geometria, topologia e densidade.

Estudos indicam que a análise da complexidade dos conjuntos de treino pode contribuir com as etapas de geração e seleção de SMCs [2]–[4]. Visto isso, neste trabalho busca-se compreender melhor o funcionamento das medidas de complexidade, e qual sua relação com a acurácia dos classificadores gerados, tendo como hipótese que classificadores treinados com conjuntos de complexidade semelhantes têm acurácia semelhante.

Para que a combinação dos classificadores presentes em um *pool* seja vantajosa é necessário que haja certa diversidade entre os elementos, a qual depende que os classificadores comentam erros complementares. Além disso, é importante que os membros do conjunto não sejam muito robustos, o que pode fazer com que a diversidade diminua. Sabendo que há trabalhos na literatura que adotam medidas de complexidade como critério de aptidão para selecionar ou gerar subconjuntos do pool de classificadores [2], [5], [6], é importante entender se os classificadores que treinam em conjuntos de dificuldade similares apresentam comportamentos similares em termos de taxa de acertos. Essa informação pode ajudar na compreensão e definição de medidas de complexidade adequadas na etapa da seleção dos classificadores.

Neste trabalho buscou-se analisar e correlacionar descritores de complexidade dos dados usados no treinamento de classificadores frente ao seu desempenho em termos de acurácia. O objetivo foi entender se classificadores treinados em conjuntos de dados com complexidade similares também terão taxas de acertos semelhantes. Além disso, almejou-se identificar quais medidas de complexidade têm comportamento mais parecido com a acurácia dos classificadores.

II. Fundamentação

É de conhecimento geral que o comportamento dos classificadores está diretamente ligado à qualidade do conjunto de dados utilizado em seu treinamento, por isso a preocupação em gerar subconjuntos mais adequados. Em problemas mais simples, as instâncias têm classes com atributos discrepantes, sendo mais fáceis diferenciá-las. Em conjuntos mais complexos, todavia, as classes têm atributos intrincados, tornando difícil a distinção das instâncias de classes diferentes.

Uma estratégia para descrever tais conjuntos é adotando as medidas de complexidade que tentam, através de índices, descrever o grau de complexidade do conjunto [7].

As medidas de complexidade comumente são separadas em três categorias [8] [3] [9]: sobreposição (F1, F2, F3, F4); separabilidade das classes (L1, L2, N1, N2, N3) e medidas de geometria, topologia e densidade (L3, N4, T1, T2, D1, D2, D3). Outra forma de categorizar uma medida de complexidade é pelo foco da análise que está sendo efetuada no conjunto de dados.

A. Medidas de sobreposição das classes

O objetivo das medidas de sobreposição é estimar o quão sobrepostas duas classes estão no espaço de características. Para tal, examina-se o intervalo e a dispersão dos atributos de instâncias de classes diferentes.

A Relação Máxima do Discriminante de Fischer (F1) estima o quão separadas são duas classes de acordo com alguma característica específica [9]. É possível considerar F1 como sendo a distância entre os centros de duas classes. Para tal, calcula-se F1 através de um índice cujo valor aumenta à medida

que mais afastados os centros das classes sejam um do outro [10].

A medida F2 (Sobreposição de Atributos por Classe) estima a sobreposição dos valores de um atributo de instâncias de duas classes diferentes [3]. É possível calcular esta sobreposição de um atributo encontrando seus valores máximos e mínimos e, em seguida, calculando-se a razão entre a intersecção da região das duas classes e amplitude total do atributo. O valor total de F2 é dado pelo produto entre os valores F2 individuais de cada um dos atributos do conjunto.

A Eficiência Máxima por Atributo Individual (F3) calcula a capacidade individual que cada atributo tem de classificar as instâncias. A eficiência de cada característica é calculada pela razão entre o número de instâncias que podem ser separadas por esta característica em relação ao número total de instâncias do conjunto [11]. Sendo assim, espera-se que quanto maior o valor de F3, mais discriminante seja o atributo. A eficiência máxima individual de um atributo será definida pelo maior valor obtido entre todos os atributos considerados [3].

A medida F4 (Eficiência Coletiva dos Atributos), assim como F3, calcula a propriedade discriminante dos atributos, porém, ao contrário de F3, nesta medida todos os atributos são considerados para o valor do índice.

Para calcular a propriedade discriminante coletiva, primeiramente, seleciona-se o atributo que consegue separar o maior número de instâncias (maior poder discriminativo). Em seguida, todas as instâncias que puderam ser discriminadas são removidas do conjunto de dados. O próximo atributo mais discriminativo é selecionado, repetindo o processo de eliminar as instâncias discriminadas por tal atributo. Este procedimento é repetido até que todas as instâncias sejam discriminadas, ou até que todos os atributos tenham sidos analisados [11]. O valor de F4 é dado pela razão de instâncias que foram discriminadas, pelo número total de elementos do conjunto.

B. Medidas de separabilidade das classes

As medidas de separabilidade avaliam o quão complexo é o comportamento dos conjuntos nas regiões de fronteira entre as classes, geralmente adotando estratégias de vizinhança.

A medida L1 (Soma Minimizada da Distância de Erro de um Classificador Linear) é calculada através da soma das distâncias euclidianas (δ) entre fronteira linear formada por um classificador linear ótimo e cada uma das amostras classificadas erroneamente [8]. A medida evidencia o quanto os dados do conjunto de treino são linearmente separáveis [3].

A Taxa de Erro de um Classificador Linear sobre o Treino (L2) representa a medida da taxa de erro obtida através da utilização de um classificador linear ótimo sobre as instâncias de treino [3], [5] [7]. A ideia é utilizar o mesmo classificador construído em L1 para verificar quantas amostras estão posicionadas no lado oposto da fronteira linear que o lado correspondente à sua classe. O valor de L2 é calculado pela razão entre o número de elementos classificados erroneamente e número total de instâncias analisadas.

A Fração de Pontos na Região de Fronteiras (N1) se baseia na construção de uma árvore de cobertura mínima (MST - *Minimum Spanning Tree*), na qual todos os pontos do conjunto de dados são conectados de forma a minimizar a soma das distâncias. Os pontos que estiverem conectados com exemplares de outra classe são ditos pontos de fronteira. Calcula-se N1 através da razão entre a contagem de elementos de fronteira pelo número total de instâncias do conjunto.

A medida N2 (Proporção das Distâncias Intra/Inter classes até o vizinho mais próximo) consiste em comparar a distância de cada elemento e seu vizinho mais próximo dentro da classe, com a distância até o vizinho mais próximo pertencente a outra classe [2]. O valor de N2 se dá pela razão entre o somatório de todas as distâncias euclidianas dos elementos para seu vizinho mais próximo em sua classe, e o somatório de todas as distâncias entre os mesmos elementos e os vizinhos mais próximos de outra classe [11].

A Taxa de erro do classificador KNN pela abordagem *Leave-One-Out* (N3) é estimada pela taxa de erro de um classificador KNN com o valor de k sendo configurado como $k = 1$ [11]. A taxa de erros é estimada pelo método *Leave-One-Out*, que consiste em testar cada instância em um classificador treinado com os demais elementos do conjunto [7]. Valores próximos de 0 para N3 indicam espaçamento entre os elementos da região de fronteira das classes, enquanto valores próximos de 1 indicam sobreposição entre tais classes [5].

C. Medidas de geometria, topologia e densidade

As medidas de Geometria, Topologia e Densidade buscam apresentar uma compreensão mais espacial do relacionamento das classes, por meio da descrição da geometria das mesmas.

A Fração de Esferas de Cobertura Máxima (T1) representa o número de círculos necessários para cobrir cada classe. Inicialmente, é criado um círculo centralizado em cada instância do conjunto de dados, o qual cresce até tocar uma instância de outra classe. As circunferências que estão completamente situadas dentro de outras circunferências são ditas redundantes, sendo assim descartadas [7].

O Número médio de pontos por dimensão (T2) descreve a densidade da distribuição espacial de amostras através da razão entre número de instâncias no conjunto perante o número de atributos [11]. Esta medida é apontada pelos autores como uma medida que não traz informações pertinentes a respeito da separabilidade das classes quando utilizado um classificador linear. Contudo, T2 fornece informações relevantes em casos

onde os classificadores são não lineares, como o KNN, por exemplo [5].

A Não-Linearidade de um Classificador Linear (L3) analisa a não linearidade de um classificador em relação ao conjunto de dados. Este método consiste em criar um conjunto de teste através da interpolação linear, com coeficientes gerados aleatoriamente, entre duas instâncias da mesma classe randomicamente selecionadas. Então L3 corresponderá ao valor da taxa de erro quando avalia-se o novo conjunto de testes utilizando um classificador linear similar ao de L1 [3].

III. Metodologia

O fluxograma da análise desenvolvida é ilustrado na Figura 2. Inicialmente, cada base é dividida em conjuntos de treino, validação e teste (Figura 2-A). Em seguida, a partir do conjunto de treino, utilizando o algoritmo *Bagging* (Figura 2-B), são gerados 100 subconjuntos para treino de 100 modelos de classificação. Tais conjuntos são gravados em arquivos do tipo ARFF [1] para análise posterior.

Os subconjuntos gerados são então usados para a construção dos modelos de classificação (Figura 2-C). Os classificadores construídos são então submetidos à estimação de sua acurácia. Esse processo será feito sobre o conjunto de teste. Tais valores são enviados à etapa seguinte do processo (Figura 2-E).

Além de servirem de base para o treinamento dos modelos os subconjuntos são utilizados para estimação dos índices de complexidade (Figura 2-D) empregando-se a biblioteca ECoL (*Extended Complexity Library*) [12]. Para cada conjunto de treino fornecido, a biblioteca retorna um vetor contendo os índices de complexidade requisitados.

Após a conclusão da estimação dos índices de complexidade e da acurácia dos classificadores gerados, é realizada a análise da relação entre tais índices (Figura 2-E). Nesta fase, é estimado o comportamento das variações entre pares de classificadores para a acurácia e complexidade.

Visando avaliar a robustez dos experimentos realizados, todo o processo foi executado dez vezes, desde a formação dos conjuntos de treino, teste e validação a partir da base original, até a análise da relação entre acurácia e medidas de complexidade.

A. Conjuntos de dados (A)

O processo de desenvolvimento adotou o mesmo conjunto de bases de dados utilizado no trabalho de [7], composto por 26 bases, as quais são formadas apenas de atributos numéricos e não apresentam valores faltantes. Além disso, tais bases são frequentemente usadas na literatura como base para avaliação de métodos de classificação.

Destas bases, quatorze são originárias do repositório da Universidade da Califórnia em Irvine (UCI) [13], duas são procedentes do repositório KEEL (*Knowledge Extraction based on Evolutionary Learning*) [14], outras quatro pertencentes à LKC (*Ludmila Kuncheva Collection of Real Medical Data*) [15], quatro provenientes do projeto StatLog [16] e duas bases geradas artificialmente com o *toolbox* PRTools do Matlab.

A Tabela I apresenta cada uma das bases de dados analisadas nesse trabalho juntamente com sua respectiva origem. Além disso, são apresentadas as principais informações de cada base, as quais: quantidade total de instâncias do conjunto, o número de atributos de cada instância de base, e o número de classes em que um objeto dessa base pode ser classificado.

Tabela I: Principais características das bases usadas nos experimentos

Base de dados	Instâncias	Atributos	Classes	Fonte
Adult	690	14	2	UCI
Banana	2000	2	2	PRTools
Blood	748	4	2	UCI
CTG	2126	21	3	UCI
Diabetes	766	8	2	UCI
Faults	1941	27	7	UCI
German	1000	24	2	STATLOG
Haberman	306	3	2	UCI
Heart	270	13	2	STATLOG
ILPD	583	10	2	UCI
Ionosphere	350	34	2	UCI
Laryngeal1	213	16	2	LKC
Laryngeal3	353	16	3	LKC
Lithuanian	2000	2	2	PRTools
Liver	345	6	2	UCI
Mammo	830	5	2	KEEL
Monk	432	6	2	KEEL
Phoneme	5404	5	2	STATLOG
Segmentation	2310	19	7	UCI
Sonar	208	60	2	UCI
Thyroid	692	16	2	LKC
Vehicle	847	18	4	STATLOG
Vertebral	300	6	2	UCI
WBC	569	30	2	UCI
WDVG	5000	21	3	UCI
Weaning	302	17	2	LKC

O processo se inicia com a divisão das bases em três novos conjuntos (Figura 2-A), sendo eles treino, validação, teste. Esta divisão é feita de forma aleatória e sem reposição e, além disso, são mantidas as proporções (estratificação) das classes da base de dados original em cada novo conjunto.

O primeiro conjunto, Treino, tem como função servir de base para o aprendizado do classificador, sendo este composto de 50% das amostras da base de dados original. O segundo conjunto, Validação, é usado no processo de calibragem de parâmetros do classificador, tendo 25% das instâncias do con-

[1] *Attribute-Relation File Format* (ARFF) é um formato de arquivo que visa representar, de forma padronizada, conjuntos de dados com instâncias independentes e não ordenadas, e que não envolvem relacionamentos entre as mesmas.

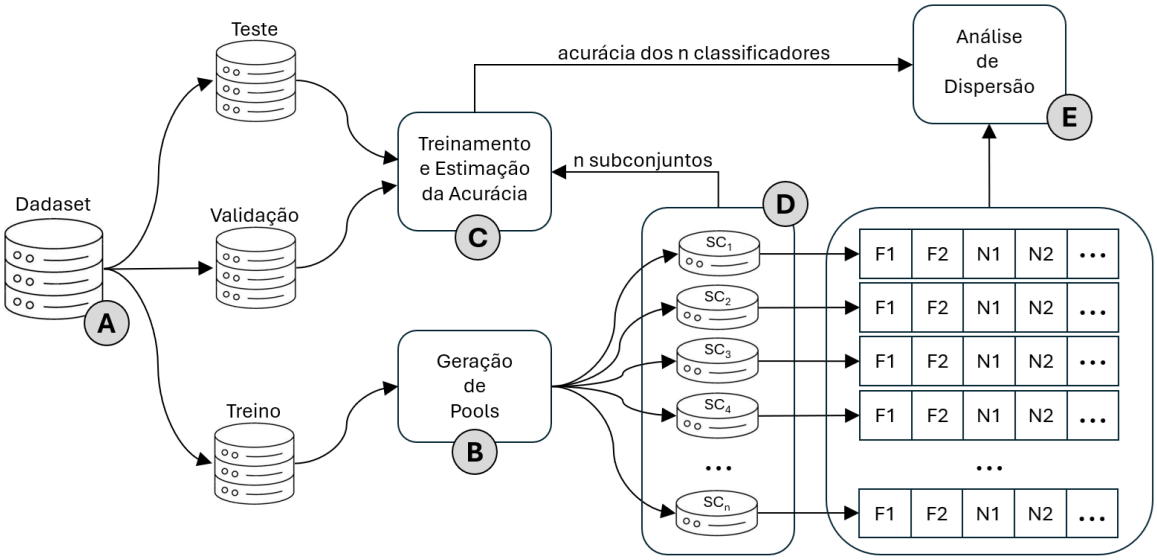


Figura 2: Fluxograma das etapas realizadas do trabalho

junto. Já o conjunto de Teste é utilizado na estimação da acurácia do classificador, contendo os 25% dos dados restantes.

B. Geração dos subconjuntos de treino (B)

A partir do conjunto Treino, são gerados 100 subconjuntos de treino (SC_1, SC_2, SC_3, ..., SC_{100}), por meio do algoritmo de *Bagging*. Cada subconjunto contém 20% do tamanho do conjunto de Treino, podendo repetir instâncias entre os conjuntos ou mesmo dentro de um mesmo conjunto. O *Bagging* foi escolhido como algoritmo de geração de *pools* por ser simples e amplamente utilizado na literatura.

C. Treinamento dos modelos e estimação da acurácia (C)

Buscando maior variabilidade entre as acurácias dos classificadores treinados por cada subconjunto, foi implementado um classificador fraco do tipo *perceptron*, de modo que pequenas alterações nos conjuntos de treino resultem em uma maior variação entre os classificadores em termos de opinião e taxas de acerto. Tal configuração é adequada para um SMC de *pool* homogêneo pois permite que os classificadores tenham erros complementares.

Durante o treinamento optou-se por manter uma configuração de parâmetros que menos interfira na acurácia dos classificadores, nesse sentido, foi adotada a configuração padrão do *perceptron* da biblioteca *scikit-learn*.

Nesta etapa, após cada um dos subconjuntos gerados pelo *Bagging* terem sido usados para construir modelos de classificação, estes são submetidos à estimação de sua acurácia.

O processo consiste em empregar cada um dos 100 modelos gerados na classificação do conjunto de teste e estimar o percentual de instâncias rotuladas corretamente. O valor resultante é um índice pertencente ao intervalo $[0, 1]$ e corresponde à relação do número de acertos perante o total de instâncias.

D. Estimação das Métricas de Complexidade

Assim que os subconjuntos são construídos é realizada a fase de estimação da assinatura de complexidade. Para tal, foi empregada a biblioteca ECoL (*Extended Complexity Library*) [12] que é uma biblioteca de aprendizado de máquina, implementada na linguagem R, que dispõe de um total de vinte e nove descritores de complexidade para problemas de classificação e regressão, dos quais, por serem os descritores para problemas de classificação mais comuns na literatura, apenas doze serão empregados neste trabalho: F1, F2, F3, F4, L1, L2, L3, N1, N2, N3, N4 e T1.

A biblioteca ECoL por padrão retorna todos os índices de complexidades normalizados no intervalo $[0, 1]$, evitando assim que alguma das métricas tenha mais peso do que outra, dada sua faixa de variação.

E. Análise da Relação entre Acurácia e Complexidade

Após estimar as métricas de complexidade de cada subconjunto e a acurácia dos cem classificadores, o passo seguinte consiste em analisar a relação entre cada métrica de complexidade e a acurácia dos classificadores treinados com cada um dos subconjuntos.

Visando ilustrar como o processo é realizado, considere um conjunto composto de cinco classificadores fictícios (C_1, C_2, C_3, C_4, C_5). A Tabela II apresenta as acurácias do conjunto de classificadores e a Tabela III, as medidas da complexidade F2 estimada sobre os conjuntos de treino dos cinco classificadores. Pode-se perceber que todos os índices levantados encontram-se no intervalo $[0, 1]$.

Tabela II: Exemplo de acurácia de cinco classificadores fictícios

Classificador	Acurácia
C_1	0,8000
C_2	0,8500
C_3	0,9000
C_4	0,8700
C_5	0,8300

Tabela III: Valor da medida F2 referente aos conjuntos usados para treinar os cinco classificadores fictícios

Conjunto de treino	F2
CT_1	0,7300
CT_2	0,7000
CT_3	0,6500
CT_4	0,5700
CT_5	0,7000

A fim de estimar a relação entre as acurácias dos classificadores e as métricas de complexidade usadas para descrever a complexidade dos conjuntos sobre os quais tais classificadores foram treinados realizou-se a combinação de todos para todos os elementos do conjunto, formando assim uma combinação de n classificadores combinados dois a dois. O número total de combinações, NTC, pode ser estimado por $(n^2 - n)/2$. Dessa forma, para um conjunto composto por 100 classificadores, como o adotado neste trabalho, obtém-se um total de 4950 combinações possíveis.

A partir dos valores de acurácia e da medida F2 apresentados nas Tabelas II e III, respectivamente, é possível calcular a discrepância dos comportamentos de dois classificadores. Tais valores estão consolidados na Tabela IV em que a segunda coluna contém os valores das diferenças em termos de acurácia e na terceira coluna são apresentadas as diferenças entre os pares de conjuntos de treino usados nos mesmos classificadores, mas agora no espaço de complexidade da medida F2. Dado o cenário ilustrado em que o *pool* era composto por cinco classificadores, a Tabela IV é composta por dez linhas, descrevendo a relação entre todos os pares de classificadores.

De forma a melhor representar a diferença em termos de acurácia de complexidade dos classificadores comparados, os valores apresentados na Tabela IV podem ser representados em um plano cartesiano em que o eixo das abscissas corresponde às diferenças em termos de acurácia e o eixo das ordenadas

Tabela IV: Exemplo de tabela de pontos do gráfico

	Dissimilaridade da acurácia	Dissimilaridade de F2
C_1 vs C_2	-0,0500	0,0300
C_1 vs C_3	-0,1000	0,0800
C_1 vs C_4	-0,0700	0,1600
C_1 vs C_5	-0,0300	0,0300
C_2 vs C_3	-0,0500	0,0500
C_2 vs C_4	-0,0200	0,1300
C_2 vs C_5	0,0200	0,0000
C_3 vs C_4	0,0300	0,0800
C_3 vs C_5	0,0700	-0,0500
C_4 vs C_5	0,0400	0,1300

refere-se às diferenças em termos da medida de complexidade. Uma vez que a diferença máxima em termos de acurácia e complexidade estará dentro do intervalo $[-1, 1]$, o gráfico explora apenas esse espectro do plano cartesiano.

A dispersão dos pontos no espaço de acurácia e complexidade reflete o grau de similaridade entre os classificadores. Caso o comportamento destes seja parecido, a tendência é que o ponto que representa a sua relação esteja próximo da origem. Por outro lado, se a diferença for grande, em qualquer dos eixos, a tendência é que o ponto esteja posicionado mais longe da origem.

Para quantificar a relação entre a acurácia e a métrica de complexidade optou-se em adotar a Distância Euclidiana Média (DEM) entre todos os pontos. A Equação 1 detalha como é calculada esta medida. Na equação, NTC representa o número total de pontos, p corresponde a um ponto do conjunto e $\delta(p_i, p_j)$ refere-se à distância euclidiana entre os pontos i e j. Espera-se que quanto menor o DEM, maior a similaridade entre a acurácia e a complexidade dos classificadores do conjunto.

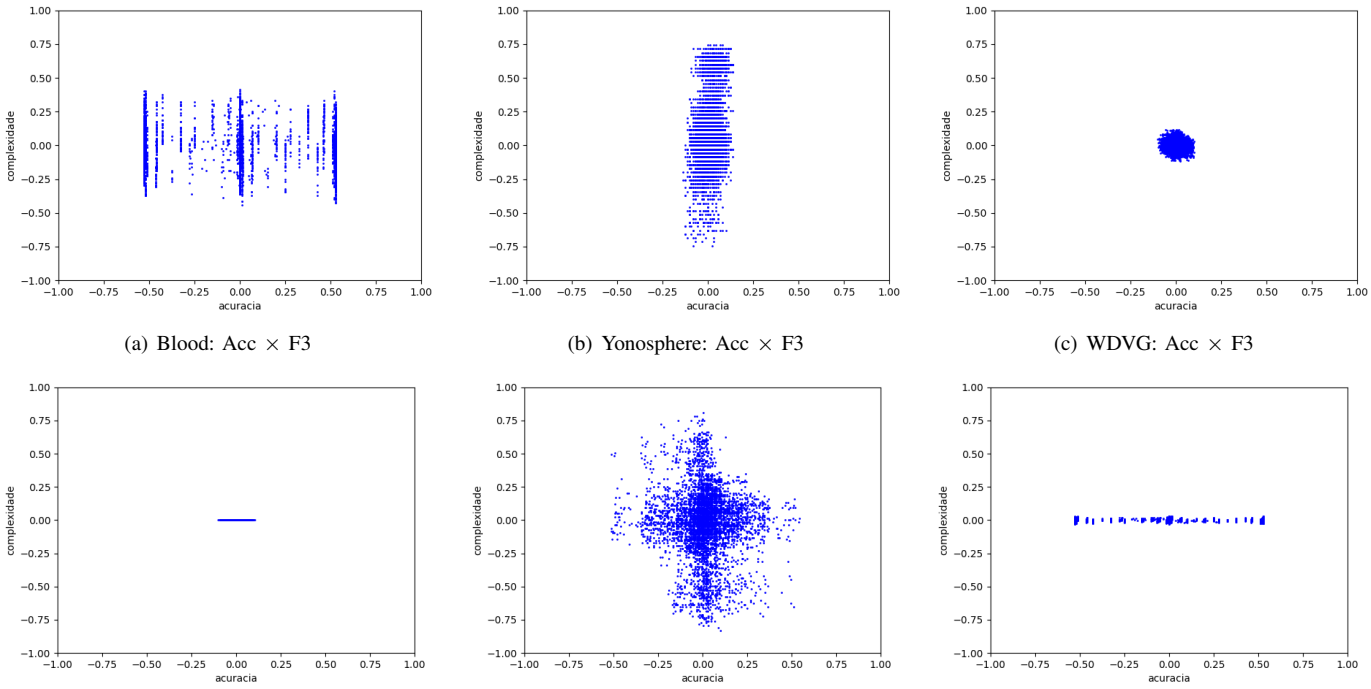
$$DEM = \frac{\sum_{i=1}^{NTC-1} \sum_{j=i+1}^{NTC} \delta(p_i, p_j)}{NTC} \quad (1)$$

IV. Resultados e Discussão

A. Análise da dispersão

Após a realização dos testes descritos na Seção III-E, é possível analisar o comportamento de cada gráfico de dispersão em diferentes bases de dados. Tal análise torna evidente grandes diferenças na relação Acurácia-Complexidade de diferentes bases de dados.

Na análise da relação Acurácia-F3 (Figura 3(a)) podemos ver uma variação mediana nos eixos da acurácia e no eixo da métrica F3 para a base Blood. É interessante observar que os classificadores apresentam "faixas" de acurácia que consistem em classificadores que apresentam comportamentos muito parecidos perante o conjunto de teste.



(a) Blood: Acc × F3 (b) Yonosphere: Acc × F3 (c) WDVG: Acc × F3

(d) Adult: Acc × F2 (e) Mammo: Acc × F4 (f) Blood: Acc × T1

Figura 3: Comportamentos observados para a relação entre acurácia e várias medidas de complexidade

Na Figura 3(b) pode-se perceber uma alta variação no eixo da métrica F3 para a base Ionosphere e, ao mesmo tempo, pouca variação no eixo das acurácias, indicando que classificadores tendem a ter número de acertos próximos apesar das variações na medida F3.

Já a Figura 3(c) apresenta uma baixa variação em ambos os eixos (variação na medida F3 e acurácia) para a base WDVG. Percebe-se que praticamente toda a concentração dos pontos ocorre no espaço [-0,15,0,15] para as duas variáveis estimadas.

Tais comportamentos, por sua vez, ficam evidenciados ao se comparar a Distância Euclidiana Média de cada uma das bases. O valor de dispersão para a base Blood (no exemplo ilustrado) foi de 0,4394. A base Ionosphere, na repetição apresentada, alcançou dispersão de 0,2322. Já a base WDVG, que mostrou-se muito mais concentrada, indicando grande similaridade entre a acurácia e a métrica F3, com uma DEM de 0,0661.

Outro apontamento observado diz respeito às métricas de complexidade que apresentaram valor semelhante para todos os conjuntos, gerando assim cenários em que variou-se apenas na acurácia. A Figura 3(d) ilustra um exemplo de métrica que não varia para todos os conjuntos. No cenário, a medida F2 não apresentou variações entre os 100 conjuntos usados no treinamento dos classificadores, apenas na acurácia dos mesmos, variando no intervalo $[-0,2, +0,2]$. O valor observado para DEM neste caso foi de 0,0492, um valor baixo em comparação aos demais.

Ao contrário do observado na Figura 3(d), há casos onde o gráfico apresenta dispersão mais expressiva, como observado na Figura 3(e). No entanto, apesar da dispersão cobrir uma região mais abrangente do espaço, o valor para DEM foi de 0,3384, o que indica que ainda há concentração dos elementos próximos da origem do plano cartesiano.

Outro comportamento interessante observado ao longo das repetições é visto na Figura 3(f). Na ilustração é possível perceber que os classificadores apresentaram faixas de acurácia e, ao longo dessas, pequenas variações na medida T1. Neste caso o valor obtido para DEM foi de 0,3890, confirmando certa dissimilaridade entre o comportamento em termos de acurácia frente aos valores da medida T1 estimada sobre os conjuntos em que foram treinados.

B. Análise da Distância Euclidiana Média

Na Tabela V são detalhados os valores médios da DEM para as 10 execuções do experimento para cada uma das bases testadas (linhas) frente às 12 métricas de complexidade (colunas). Destaca-se em vermelho (negritados) a métrica que obteve

Tabela V: Valor médio das dispersões para cada medida de complexidade considerando as 10 repetições executadas

Bases	F1	F2	F3	F4	L1	L2	L3	N1	N2	N3	N4	T1
Adult	0,085	**0,072**	0,177	0,151	0,086	0,089	0,088	0,145	0,089	0,094	0,077	0,073
Banana	0,142	0,154	0,164	0,201	0,133	0,144	0,147	0,137	0,131	**0,130**	0,146	0,133
Blood	0,400	0,456	0,439	0,470	0,396	0,396	0,420	0,426	0,400	0,405	0,415	**0,384**
CTG	0,051	**0,029**	0,127	0,048	0,032	0,035	0,034	0,063	0,040	0,042	0,041	0,030
Diabetes	0,159	0,164	0,215	0,275	0,162	0,179	0,188	0,204	0,162	0,172	0,166	**0,148**
Faults	0,071	**0,051**	0,070	0,056	0,052	0,052	0,052	0,090	0,063	0,071	0,064	**0,051**
German	0,057	0,111	0,144	0,400	0,079	0,078	0,081	0,120	0,071	0,080	0,058	**0,053**
Haberman	0,238	0,287	0,318	0,393	0,232	0,242	0,287	0,298	0,236	0,256	0,263	**0,214**
Heart	0,119	**0,092**	0,286	0,125	**0,092**	0,093	**0,092**	0,229	0,124	0,137	0,096	0,101
ILPD	0,074	**0,064**	0,147	0,100	0,088	0,116	0,133	0,152	0,089	0,107	0,102	0,065
Ionosphere	0,084	**0,066**	0,232	0,099	**0,066**	0,069	0,068	0,162	0,093	0,099	0,096	0,071
Laryngeal1	0,155	**0,103**	0,255	0,156	0,112	0,124	0,121	0,258	0,147	0,169	0,138	0,131
Laryngeal3	0,136	**0,078**	0,125	0,096	0,081	0,084	0,085	0,191	0,112	0,132	0,113	0,083
Lithuanian	0,144	0,155	0,171	0,175	0,137	0,146	0,148	0,141	0,134	**0,133**	0,141	0,134
Liver	0,133	0,135	0,199	0,181	0,150	0,186	0,206	0,228	0,150	0,175	0,166	**0,120**
Mammo	0,150	0,187	0,197	0,332	0,126	0,138	0,149	0,177	0,125	0,129	0,146	**0,111**
Monk	0,094	0,212	0,152	0,121	0,101	0,121	0,126	0,181	0,102	0,122	0,096	**0,077**
Phoneme	0,101	0,126	0,113	0,121	0,101	0,110	0,115	0,111	0,100	0,102	0,107	**0,097**
Segmentation	0,136	**0,133**	0,138	**0,133**	**0,133**	**0,133**	0,137	0,148	0,136	0,137	0,137	**0,133**
Sonar	0,136	**0,115**	0,294	0,156	**0,115**	**0,115**	**0,115**	0,257	0,155	0,177	0,120	0,118
Thyroid	0,056	**0,044**	0,094	0,052	0,046	0,050	0,051	0,082	0,064	0,058	0,049	0,144
Vehicle	0,124	0,094	0,133	0,106	0,095	0,097	0,098	0,151	0,109	0,122	0,124	**0,093**
Vertebral	0,149	**0,090**	0,218	0,138	0,106	0,123	0,123	0,198	0,121	0,128	0,125	0,106
WBC	0,258	**0,238**	0,281	0,239	**0,238**	0,239	0,239	0,267	0,250	0,246	0,240	0,251
WDVG	0,042	**0,038**	0,066	0,040	0,041	0,042	0,041	0,064	0,045	0,047	0,040	**0,038**
Weaning	0,109	**0,093**	0,236	0,151	**0,093**	0,095	0,094	0,202	0,129	0,145	0,109	0,095

o menor valor de dispersão para cada base, caracterizando-a como a métrica com maior relação complexidade-acurácia. Por outro lado, em azul (sublinhados) são representadas as métricas com maior valor de DEM. Tais valores são as combinações de acurácia e medida de complexidade que apresentam as maiores variações.

A partir dos valores médios das DEMs pode-se analisar quais métricas obtiveram os menores e maiores valores para cada base de dados, sendo esta informação inversamente proporcional à relação da métrica com acurácia. Percebe-se que a métrica que conseguiu mais vezes o menor valor de DEM é F2, sendo este o menor em 14 das 26 bases de dados. A medida L1 apresentou a maior similaridade (menor DEM) em 6 bases. Já as medidas L2, L3 e N3 apresentaram a menor dispersão para duas bases presentes no conjunto. Por fim, a F4 teve a menor dispersão para a base Segmentation.

Com relação aos comportamentos de maior dispersão, podemos destacar a medida F3 que obteve o maior valor para DEM em 12 das 26 bases testadas. Além de F3, outra medida que apresentou valores altos para DEM foi F4, tendo alcançado o maior valor em 7 bases. A medida N1 apresentou a maior dispersão em quatro bases de dados, enquanto F2 obteve os maiores valores de DEM em duas bases. Já a medida T1 apresentou a maior dispersão apenas para a base Thyroid.

Além da análise de similaridade das medidas de complexidade perante o comportamento da acurácia dos classificadores, analisou-se também as variações das métricas de uma mesma base de dados. Neste sentido destacou-se a base Blood ao apresentar os maiores valores de DEM em todas as métricas, variando de 0,470 em F4 e 0,384 em T1 (Figura 4(a)).

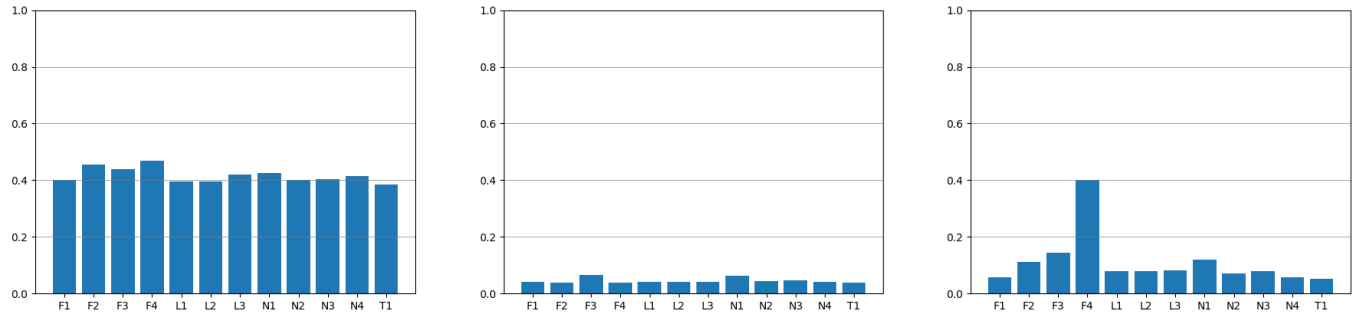
A base WDVG, por sua vez, apresentou os menores valores de dispersão média, tendo como maior valor F3 = 0,066 e menor valor F2 = 0,038 (Figura 4(b)).

Em outros cenários, como para a base German, observou-se um comportamento mais discrepante entre as métricas, como representado na Figura 4(c). Pode-se perceber que o valor de F4 mostrou-se consideravelmente maior aos demais índices de complexidade. Ela apresentou um DEM de 0,4000 enquanto a média das outras onze métricas foi de 0,0847.

Visando analisar se as medidas apresentaram comportamento estatisticamente significativo ao longo das 26 bases de dados aplicou-se o teste de Kruskal-Wallis com 5% de significância. Os resultados indicaram rejeição da hipótese nula (H_0), indicando que há pelo menos uma métrica de complexidade com comportamento distinto das demais.

Para ranquear as métricas de complexidade (Tabela VI) de forma a identificar aquelas que mais influenciam a acurácia de um classificador utilizou-se o teste *post-hoc* de Nemenyi, tendo como entrada as médias disponíveis na Tabela V. Os dados evidenciam que as métricas mais fortemente relacionadas com a acurácia foram L1, T1 e F2 enquanto F4, N1 e F3 indicaram a menor influência sobre a acurácia dos classificadores testados.

Com base nos resultados do teste de Nemenyi, obteve-se o valor para determinar se existe, ou não, uma diferença



(a) Blood (b) WDVG (c) German

Figura 4: Exemplo de estimação de complexidade

Tabela VI: Ranking das métricas de complexidade

ID	Medida	Ranking Médio
0	L1	3,1346
1	T1	3,1923
2	F2	4,2885
3	L2	5,3269
4	N2	5,6923
5	L3	6,0777
6	N4	6,1538
7	F1	6,3846
8	N3	7,5769
9	F4	8,8077
10	N1	10,3846
11	F3	11,0000

estatisticamente significativa entre cada par de medidas de complexidade. A Figura 5 apresenta uma matriz de diferença todos-com-todos em que quanto mais escuro é a intersecção das métricas, mais distintos são os conjuntos. Pode-se perceber que em grande parte das combinações não há diferença significativa. No entanto, em diversos casos observou-se que as métricas apresentaram variações significativas ao longo das 26 bases de dados. Nesse sentido as métricas F1 e F2 foram aquelas que mostraram mais diferenças significativas frente às demais métricas. A primeira divergiu estatisticamente de 9 métricas enquanto a segunda apresentou diferença significativa para 8 das 11 métricas comparadas.

Dado os resultados do teste de Nemenyi, também é possível construir um gráfico de diferença crítica (Figura 6), onde as medidas são ordenados das melhores (mais influentes) para as piores (menos influentes), e as barras horizontais correspondem à distância crítica (DC). Caso a diferença de rankings médios entre as medidas seja maior que DC considera-se que a diferença entre elas é significativa.

Os valores dos rankings médios de cada medida de complexidade, para as 26 bases de dados são apresentados na Tabela VI. Com base nos valores tabelados e na Figura 6 pode-se notar que a métrica com maior relação com a acurácia é a L1, tendo, porém, diferença crítica apenas com as medidas F3, N1, F4, N3 e F1. Já F3 mostrou a menor relação com a acurácia, tendo

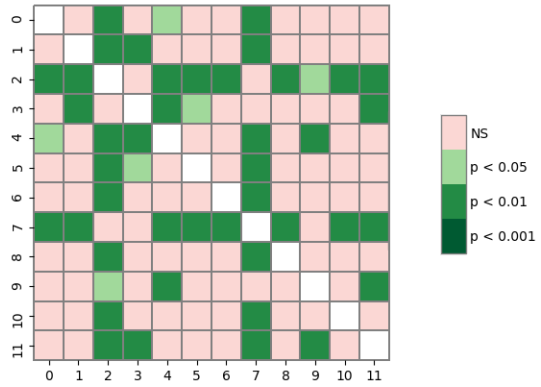


Figura 5: Relação de diferença entre métricas

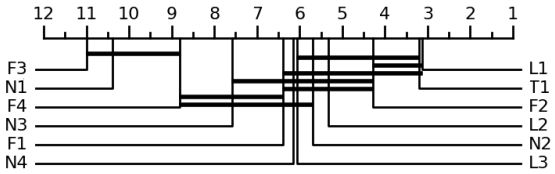


Figura 6: Gráfico de diferença crítica das métricas de complexidade

uma diferença não significativa apenas com N1 e F4.

Notou-se que, apesar de ter o maior número de menores resultados, F2 apresentou apenas o terceira maior relação com a acurácia dos classificadores. Isso pode indicar maior instabilidade da métrica perante diferentes conjuntos. Já a métrica F3, que havia sido a métrica com os maiores valores de DEM, teve este resultado confirmado pelo teste de Nemenyi, sendo a métrica com menor relação com a acurácia dos classificadores.

V. Conclusão

Com base no fato de que há relação entre o comportamento dos classificadores e os conjuntos sobre os quais eles foram treinados, neste trabalho se propôs elencar quais métricas de complexidade retiradas do conjunto de treino de um classificador melhor se relacionam com a acurácia do mesmo, no sentido de que ao treinar classificadores com conjuntos de complexidade semelhante, obtém-se resultados em termos de acurácia semelhante.

Para realizar tal análise, desenvolveu-se um protocolo para a geração de subconjuntos através do algoritmo *Bagging* que eram, então, usados no treinamento de perceptrons. Além disso, a partir de tais conjuntos foram obtidos seus descritores de complexidade. Com o intuito de maior robustez na análise, foram estudadas 26 bases de dados, cada uma gerando 100 subconjuntos por repetição, em um total de 10 repetições.

Percebeu-se que existe relação entre algumas métricas de complexidade provenientes dos conjuntos de treino e a acurácia de classificadores. Dentre as métricas que apresentaram comportamento mais similar entre a acurácia e a complexidade dos conjuntos de treino podemos destacar L1, T1, F2 e L2. Por outro lado, as medidas que mostraram menor similaridade em termos de complexidade foram F3, N1, F4 e N3.

Porém, como todos os testes foram feitos com classificadores lineares do tipo *perceptron*, percebeu-se então que aqueles que se utilizam de técnicas de classificação linear, como L1 e L2, obtiveram bons resultados, enquanto técnicas que utilizam estratégias de vizinhança, como N1 e N3, e técnicas inspiradas no funcionamento de árvores de decisão, como F3 e F4, obtiveram resultados ruins. Nesse sentido, é interessante que sejam experimentados modelos de classificação baseados em outros critérios, como o KNN, SVM ou MLP.

Além disso, pode ser pertinente avaliar se os comportamentos aqui observados se repetiriam em aplicações de regressão ao invés de classificação. Tendo em vista que a biblioteca ECoL apresenta métricas de complexidade para problemas de regressão, acredita-se ser útil uma análise da relação dessas métricas com a acurácia dos algoritmos de regressão.

Referências

[1] A. S. Britto Jr., R. Sabourin, and L. E. S. Oliveira, "Dynamic selection of classifiers — a comprehensive review," *Pattern Recognition*, vol. 47, p. 3665–3680, 2014. [Online]. Available: https://doi.org/10.1016/j.patcog.2014.05.003

[2] A. L. Brun, A. S. Britto Jr., L. S. Oliveira, and F. Enembreck, "A framework for dynamic classifier selection oriented by the classification problem difficulty," *Pattern Recognition*, vol. 76, pp. 175–190, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.10.038

[3] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 289–300, Mar 2002. [Online]. Available: https://doi.org/10.1109/34.990132

[4] N. Macià, E. Bernadó-Mansilla, A. Orriols-Puig, and T. K. Ho, "Learner excellence biased by data set selection: A case for data characterisation and artificial data sets," *Pattern Recognition*, vol. 46, no. 3, p. 1054–1066, 2013. [Online]. Available: https://doi.org/10.1016/j.patcog.2012.09.022

[5] A. L. Brun, A. S. Britto Jr., L. S. Oliveira, F. Enembreck, and R. Sabourin, "Contribution of data complexity features on dynamic classifier selection," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 4396–4403. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2016.7727774

[6] M. Monteiro, A. S. Britto Jr., J. P. Barddal, L. S. Oliveira, and R. Sabourin, "Classifier pool generation based on a two-level diversity approach," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2414–2421. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICPR48806.2021.9412137

[7] M. J. dos Santos, A. L. Brun, and R. A. Silva, "Um framework para análise da relação entre tamanho e complexidade de conjuntos de dados," *Revista Brasileira de Computação Aplicada*, vol. 13, no. 2, p. 1–15, May 2021. [Online]. Available: http://dx.doi.org/10.5335/rbca.v13i2.10898

[8] T. K. Ho and M. Basu, "Measuring the complexity of classification problems," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, ser. ICPR-00, vol. 2. IEEE Comput. Soc, p. 43–47. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2000.906015

[9] J. S. Sánchez, R. A. Mollineda, and J. M. Sotoca, "An analysis of how training data complexity affects the nearest neighbor classifiers," *Springer-Verlag London Limited*, vol. 10, no. 3, pp. 189–201, Sep 2007. [Online]. Available: https://doi.org/10.1007/s10044-007-0061-2

[10] A. I. Landeros, "Data complexity and classifier selection," Tese de Doutorado, University of Alabama, Tuscaloosa, 2008.

[11] A. Orriols-Puig, N. Macià, and T. K. Ho, *Documentation for the Data Complexity Library in C++*, Grup de Recerca en Sistemes Intel - ligents La Salle - Universitat Ramon Llull, Barcelona, Dec 2010.

[12] A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto, and T. K. Ho, "How complex is your classification problem? a survey on measuring classification complexity," *ACM Computing Surveys (CSUR)*, vol. 52, pp. 1–34, 2019. [Online]. Available: http://dx.doi.org/10.1145/3347711

[13] D. Dua and C. Graff. (2017) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. [Online]. Available: http://archive.ics.uci.edu/ml

[14] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.

[15] L. I. Kuncheva, *Combining Pattern Classifiers*, 1st ed. New Jersey: John Wiley & Sons, Inc., 2004.

[16] R. D. King, C. Feng, and A. Sutherland, "StatLog: comparison of classification algorithms on large real-world problems," *Applied Artificial Intelligence*, vol. 9, no. 3, pp. 289–333, 1995. [Online]. Available: https://doi.org/10.1080/08839519508945477