Revista Brasileira de Computação Aplicada, Julho, 2021

DOI: 10.5335/rbca.v13i2.10898
Vol. 13, Nº 2, pp. 1–15
Homepage: seer.upf.br/index.php/rbca/index

RBCA
ISSN 2176-6649

ARTIGO ORIGINAL

Um framework para análise da relação entre tamanho e complexidade de conjuntos de dados

A framework for analyzing the relationship between size and complexity of data sets

Mateus José dos Santos[1], André Luiz Brun [1], Ronan Assumpção Silva [2]

[1]Universidade Estadual do Oeste do Paraná (UNIOSTE), [2]Instituto Federal do Paraná (IFPR)

mateusjose19@hotmail.com,*andre.brun@unioeste.br; ronan.silva@ifpr.edu.br

Recebido: 26/04/2021. **Revisado:** 24/11/2020. **Aceito:** 18/05/2021.

Resumo

Na área de Reconhecimento de Padrões, um problema de classificação é dito complexo quando as observações de classes diferentes apresentam elevada semelhança. Ao reconhecer a estimativa de complexidade como um fator importante na obtenção de acurácia, a literatura propôs uma variedade de descritores de complexidade. Porém, não se conhece a sensibilidade desses descritores quanto à variação do tamanho dos conjuntos de dados. Neste trabalho, este comportamento foi analisado a partir da construção e execução de um *framework*. Os descritores foram medidos em 20.800 subconjuntos criados a partir de: i) 26 problemas de classificação, ii) 2 geradores e iii) 4 tamanhos. Os resultados comprovaram que a sensibilidade dos descritores ao tamanho é uma realidade. Se mostrou menos perceptível nas medidas F1, F2, L2, N4, L3, T1, D2 e D3. Entretanto, as medidas F3, F4, N1, N2 e N3 se mostraram mais afetadas por variações no número de instâncias presentes nos conjuntos de dados.

Palavras-Chave: Bagging; Boosting; Medidas de Complexidade; Tamanho do Conjunto de Dados.

Abstract

In the Pattern Recognition field, a classification problem is complex when the samples of different classes are highly similar. Consequently, the literature proposed a variety of complexity descriptors, considering the importance of complexity as a promising factor to obtain accuracy. However, the sensitivity of these descriptors is unknown, considering the size of the dataset. In this work, the goal is to analyze this behavior. For that reason, a variety of descriptors were estimated in 20,800 subsets created from: i) 26 classification problems, ii) 2 generators, and iii) 4 sizes. The results proved that the descriptors' sensitivity to size is a reality. It proved less noticeable in some metrics such as F1, F2, L2, N4, L3, T1, D2, and D3. However, metrics F3, F4, N1, N2, and N3 are the most affected by variations in the number of instances present in the dataset.

Keywords: Bagging; Boosting; Complexity Measures; Dataset Size.

1 Introdução

A área de Reconhecimento de Padrões (RP) tem como uma de suas principais aplicações a Classificação que é um processo que consiste na atribuição de uma classe a um determinado objeto cuja classe ainda não foi determinada. Para isso, é necessário o uso de um ou mais classificadores, que são os elementos responsáveis por rotular novos

elementos a partir da análise de um conjunto de objetos, também conhecidos por exemplos, observações ou instâncias. Estes apresentam as características de um problema em estudo e cabe ao classificador analisá-las e aprendê-las, pois assim terá maiores chances de atribuir a classe correta a um objeto em particular.

Para que um classificador possa realizar o reconhecimento de objetos, ele precisa ter contato prévio com uma variedade de exemplares de todas as classes do problema, na intenção de identificar os fatores que caracterizam cada uma das classes. Esse processo é chamado de aprendizagem ou treinamento. O aprendizado de máquina é uma abordagem de análise de dados que, através de algoritmos, que visa simular o conhecimento humano, ou seja, permitir que sistemas se adaptem de forma independente levando em consideração cenários anteriores, assim reproduzindo decisões e resultados confiáveis (Tan et al., 2009).

Sabe-se que a capacidade de acerto de um sistema de reconhecimento é sensível ao classificador utilizado. Outro fator que influencia nesta capacidade é referente ao comportamento dos classificadores, pois são dependentes do conjunto de dados no qual utilizado para treinamento. Este conjunto pode apresentar uma variedade de fatores que impactam na precisão do reconhecimento, tais como o tamanho do conjunto de treino, a representatividade dos exemplares (objetos ou instâncias) e o balanceamento entre as classes, apenas para citar alguns. Além desses fatores, o problema submetido a classificação em si pode influenciar nos resultados, pois a aquisição dos dados também levanta outros problemas de pesquisa específicos. Por exemplo, em cenários em que as características são muito similares a ponto de dificultar diferenciá-las, a acurácia dos métodos de classificação podem ser inferiores em relação àqueles onde os atributos de cada classe apresentam maior variabilidade entre si.

Diante desse contexto, estimar a complexidade dos dados para determinar quão difícil será a tarefa do classificador pode apresentar vantagens. Na literatura estão propostas várias medidas de complexidades (Ho, 1998, Sánchez et al., 2007). Estas medidas tentam, através de índices, descrever o grau de complexidade do conjunto usado para o treinamento do classificador ou do conjunto de classificadores. Espera-se então que, a análise da complexidade do problema impacte no sucesso de um sistema de classificação, pois dada a estimativa de complexidade, estratégias mais apropriadas para a classificação podem ser consideradas. Portanto, a escolha do algoritmo de classificação, a divisão da base, a reamostragem dos dados de treino, são apenas alguns dos fatores a serem considerados a partir do conhecimento prévio do quão complexo é o problema em considerado.

Muitas vezes a tarefa de classificação envolve muita variabilidade ou complexidade, fazendo com que um classificador individual não seja capaz de aprender efetivamente sobre todo o espaço de busca ou ser capaz de identificar as classes com precisão. Uma estratégia para tentar mitigar tal problema é a adoção de um sistema de múltiplos classificadores (SMC).

Os SMCs podem ser considerados uma das técnicas de aprendizado mais robustas e precisas. São utilizados para melhorar a performance de classificadores não tão robustos através da combinação das opiniões de diversos indutores esperando que o resultado obtido seja mais acurado (Ponti Jr., 2011).

Uma boa estratégia para a tentativa de construção de classificadores, que sejam bons em diferentes regiões do espaço de busca, é treinar os classificadores em conjuntos distintos e de tamanhos variados. Por exemplo, um classificador pode ser treinado para identificar uma flor de acordo com o tamanho de sua pétala, pela cor da planta ou mesmo com base no formato do caule. Dentro do conjunto de classificadores treinados, haverá aqueles que serão mais hábeis em identificar a cor, diferenciar melhor o tamanho da pétala e outros o formato do caule. Espera-se então que, ao combinar-se os três tipos de especialidade, o resultado alcançado seja melhor.

Um SMC pode ser de dois tipos: i) homogêneo ou ii) heterogêneo. Nos sistemas homogêneos, as mesmas técnicas de indução são utilizadas para modelar os classificadores, podendo variar o conjunto de características ou parâmetros. Já nos sistemas de múltiplos classificadores heterogêneos, as técnicas usadas para treinar os classificadores são diferentes, mas os dados treinamento são os mesmos. Dentre as abordagens homogêneas, há as técnicas clássicas como o *Bagging*, *Boosting* e *Randon Subspace*.

Os subconjuntos gerados pelo processo de *Bagging* e *Boosting* afetarão o desempenho do classificador pois têm a incumbência de escolher quais elementos formarão o conjunto no qual o classificador será treinado. Ao fazer o processo de construção dos subconjuntos para aprendizado podem ser escolhidas instâncias de classes distintas mas com características muito similares, que faz com que elas estejam muito próximas no espaço de características. Devido a essa similaridade, a dificuldade do classificador em identificar essas instâncias será maior, ou seja, possivelmente implicará em diminuição da sua acurácia.

Esta dificuldade na classificação pode ser estimada pelas medidas de complexidade (Ho, 1998, Sánchez et al., 2007), através das quais tenta-se determinar o grau de dificuldade de uma solução para um determinado problema. Tais descritores, no entanto, são sensíveis às alterações nos conjuntos. Um ponto importante, que ainda carece de estudos, é tentar identificar quais índices de complexidade são mais robustos às alterações na dimensão do conjunto de treino.

O objetivo principal consiste na análise da relação entre o tamanho do conjunto de treino com sua assinatura de complexidade. Para tanto, criou-se um protocolo robusto capaz de avaliar o comportamento das medidas de complexidade de acordo com a variação no tamanho dos conjuntos usados para o treino dos classificadores. Ao projetar esse protocolo experimental, revelou-se um *framework*, necessário para realizar a análise.

Para responder ao objetivo principal, o trabalho foi dividido em três etapas distintas. A primeira etapa consiste na geração de subconjuntos de diferentes tamanhos. Na etapa seguinte, se faz necessária a estimativa das medidas de complexidade relativas aos diferentes tamanhos de conjuntos de treino. Por fim, na última etapa o objetivo é identificar as medidas que estão mais sensíveis à variação do tamanho do conjunto de dados.

Este trabalho está dividido em 5 seções. A Seção 2 trata dos conceitos mais importantes para entendimento sobre

o trabalho, começando sobre reconhecimento de padrões e indo até o funcionamento das medidas de complexidade, apresentando a ideia básica e funcionamento dos mesmos. Na Seção 3 é detalhado o *framework* desenvolvido para a realização do trabalho e alcance da proposta da pesquisa. Em seguida, na Seção 4 serão apresentados os experimentos realizados e resultados obtidos e na Seção 5, as conclusões obtidas durante o desenvolvimento dos experimentos.

2 Fundamentos

Para problemas de classificação considerados simples, normalmente emprega-se apenas um classificador (dito monolítico). Este é responsável, portanto, por rotular todas as instâncias do conjunto. No entanto, para cenários mais complexos o classificador monolítico pode apresentar desempenho insuficiente, uma vez que ele pode não ser robusto suficiente para absorver toda a variabilidade do problema. Diante dessa limitação, a adoção de vários classificadores no processo surge como alternativa para a eficácia do sistema classificatório (Kittler et al., 1998, Jain et al., 2000, Skurichina and Duin, 2002, Kuncheva and Whitaker, 2003, Ko et al., 2008).

A efetividade em se adotar vários classificadores depende de um critério fundamental: os classificadores devem apresentar diversidade entre si. Assim, cometendo erros não relacionados, aumenta-se a chance de que padrões com características distintas sejam classificados corretamente. Logo, grande parte da eficiência de um SMC depende da tarefa de formação dos classificadores que comporão o conjunto.

O algoritmo de *Bagging* proposto por Breiman (1996) consiste em sortear aleatoriamente, e com reposição, elementos do conjunto de treino para a geração de subconjuntos de elementos distintos, tomando como base o conjunto original de dados. Estes subconjuntos são usados para o treinamento dos classificadores. A partir de seu treinamento diante destes conjuntos, espera-se obter diversidade, pois o aprendizado do classificador acontece a partir de exemplos distintos. Tanto o *Boosting* como o *Bagging* baseiam-se na ideia de sorteio considerando-se o conjunto de treinamento. Entretanto, no *Boosting* a escolha dos exemplos considera pesos para cada instância, aumentando a chance dos subconjuntos conterem instâncias com maiores dificuldades de serem classificadas corretamente. Este processo inicia com o sorteio aleatório de um conjunto de elementos, portanto, todos tem a mesma chance de serem selecionados. Em seguida, é feita a classificação das amostras sorteadas. Aquelas que forem classificadas erroneamente terão seus pesos aumentados, fazendo com que, em um sorteio seguinte, tenham mais chances de serem selecionadas a compor o novo subconjunto juntamente com outras instâncias sorteadas. Considera-se que as instâncias rotuladas indevidamente são consideradas difíceis de serem classificadas (Freund and Schapire, 1996).

Proposto por Ho (1998), o *Random Subspace* constrói o novo conjunto de treino por meio do sorteio de subespaços do conjunto de atributos da base de treinamento. A ideia é que, dentre um conjunto de n características para cada instância, sejam selecionados k atributos aleatoriamente (em que $k < n$) para compor cada conjunto de treino.

A complexidade de um problema está relacionada à dificuldade de classificação do mesmo, baseado nas instâncias e atributos que o compõem. Dessa maneira, ela tem relação com o desempenho da classificação, pois caso os dados tenham uma complexidade grande tem-se a possibilidade de que os classificadores não consigam classificar os novos padrões de forma efetiva (Brun et al., 2016).

Por exemplo, considere que na Fig. 1-a é ilustrado um conjunto de dados de treinamento, a partir do qual é necessário escolher 5 instâncias de cada classe para realizar a aprendizagem de um classificador. O processo de construção desse subconjunto de treino terá a incumbência (considerando o *Bagging* ou *Boosting*) de escolher quais instâncias serão usadas no treinamento.

Digamos que o método de geração de subconjuntos escolheu as instâncias circuladas, como na Fig. 1-b, para formar um subconjunto de treino. Neste caso, o problema será simples e o classificador terá boa acurácia uma vez que pode-se perceber que os atributos comprimento e peso das instâncias selecionadas são bastante distintos entre os dois grupos. Tal discrepância será evidenciada através de um descritor de complexidade.

Em outro cenário, representado na Fig. 1-c, são selecionadas instâncias que estão bem mais próximas no espaço de características (Comprimento x Peso). Este problema é mais complexo e, consequentemente, a acurácia tende a ser menor do que no exemplo anterior. O índice de complexidade deste grupo mostrará valores mais acentuados, caracterizando o problema como mais difícil de ser classificado.

Diante do cenário exposto na Fig. 1, nos cenários b e c, percebe-se evidentemente que a estimativa da complexidade é promissora para a construção de classificadores mais precisos. Sendo assim, visando estimar a complexidade de um problema foram propostas as medidas de complexidade, as quais são divididas em três categorias (Ho and Basu, 2002, Sánchez et al., 2007, Brun et al., 2016, Brun et al., 2018, Lorena et al., 2019), que são: Sobreposição (*overlap*) das classes, Separabilidade das classes e Medidas de geometria, topologia e densidade.

As medidas de sobreposição focam na efetividade de uma característica individual na identificação das classes. Já as medidas de separabilidade buscam estimar quão separáveis são as classes do problema examinando a existência e as formas das regiões de fronteira (Sotoca et al., 2005, Cavalcanti et al., 2012). As medidas de geometria, topologia e densidade visam descrever a geometria ou a forma das variações abrangidas por cada classe visando oferecer compreensão mais superficial do relacionamento das classes (Ho and Basu, 2002, Sotoca et al., 2005, Ho et al., 2006, Luengo and Herrera, 2010, Cavalcanti et al., 2012).

Na Tabela 1 são apresentadas as medidas adotadas neste trabalho divididas de acordo com sua categoria. São listados os nomes das medidas seguidos da sua respectiva descrição, na qual é possível diferenciá-las quanto à atuação. Descrições mais aprofundadas sobre as medidas de complexidade podem ser encontradas nos trabalhos de Ho et al. (2006) e Lorena et al. (2019) onde são detalhados e discutidos cada um dos índices.

Considerando a importância dos descritores de complexidade para o avanço dos SMCs, em Lorena et al. (2019) encontra-se uma série de trabalhos relacionados. Den-

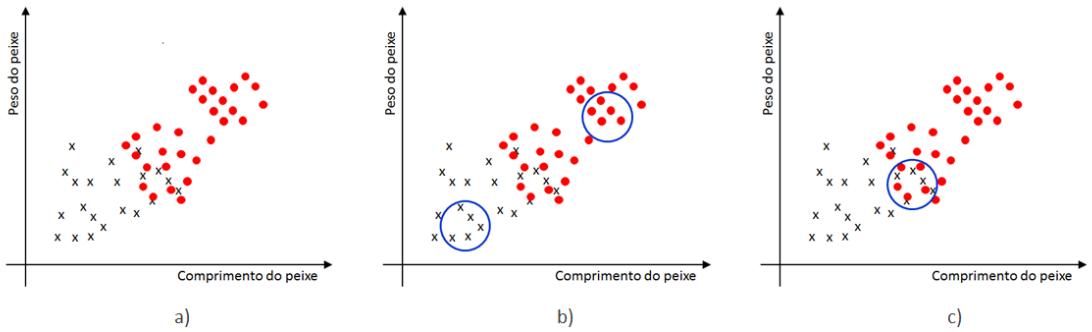

Figura 1: Exemplo de estimação de complexidade

tre os desafios observados pelos autores, sugere-se que o comportamento dos descritores de complexidade precisa ser mais bem entendido para a construção de melhores sistemas de classificação. Eles afirmam que aplicando tal conhecimento poderia auxiliar, por exemplo, em etapas de pré-processamento dos dados e também verificar a relação das medidas quanto a classificadores que consideram a separabilidade linear dos dados. Portanto, dada a necessidade, o presente estudo se propõe a análise de SMC que emprega o classificador Perceptron para a separabilidade linear dos dados, na pretensão de compreender o comportamento dos descritores perante a variação no tamanho do conjunto usado para o treinamento dos classificadores.

3 Metodologia

Neste trabalho objetivou-se analisar se há existência de relação entre o tamanho do conjunto de treino com sua assinatura de complexidade. Para tanto, foi necessária a construção de um *framework* que fosse capaz de obter, com base em um conjunto genérico de entrada, suas informações de complexidade. A estrutura implementada é ilustrada na Fig. 2.

Inicialmente, foi necessária a divisão das bases de dados em conjuntos de treino, teste e validação (Fig. 2-A). Após a divisão foram gerados subconjuntos para o treinamento utilizando o *Bagging* e *Boosting* (Fig. 2-B). Com base nos subconjuntos formados foi possível efetuar a análise de complexidade de cada um (Fig. 2-C). Estimadas as medidas de complexidade pôde-se efetuar a análise de correlação entre tamanho e assinatura de complexidade de cada subconjunto (Fig. 2-D). Cada uma destas etapas é melhor detalhada nas seções seguintes.

Os métodos implementados neste trabalho foram desenvolvidos em linguagem Python, em conjunto com as bibliotecas Pandas e *scikit-learn*. Pandas é uma biblioteca de software livre, que permite a análise de dados e estrutura de dados em alta performance (Augspurger et al., 2018). Em Python, para facilitar o uso de algoritmos de classificação, regressão e agrupamento, utilizou-se a biblioteca *scikit-learn* (Pedregosa et al., 2011).

3.1 Base de Dados (Processos A-B)

A primeira parte do processo consiste em realizar a divisão das bases em três novos subconjuntos: treino, teste e validação. Esta divisão é feita de forma aleatória de maneira que as proporções das classes do conjunto de dados original sejam mantidas (estratificadas). O primeiro conjunto é empregado no aprendizado do classificador. O segundo conjunto, Validação, é usado na definição de parâmetros do classificador. Já o conjunto de teste é utilizado na avaliação de acurácia do classificador.

Com a realização da divisão das bases de dados, é possível obter-se os conjuntos de treinos a partir dos quais serão gerados subconjuntos. Considere por exemplo, a base WBC, composta por 569 instâncias distribuídas em duas classes B (357 elementos) e M (212 registros). Ao se fazer a divisão estratificada do conjunto original, a base de treino terá 178 e 106 elementos para as classes B e M, respectivamente. Esse conjunto é utilizado para formar os subconjuntos usados para treinar os classificadores. Os métodos responsáveis pela formação de tais subconjuntos, em nosso trabalho, são o *Bagging* e o *Boosting*.

O processo de amostragem para a construção dos grupos de treino adota um limite de instâncias, normalmente uma proporção do tamanho do conjunto de teste. Neste trabalho foram adotados subconjuntos com as proporções de 10%, 33%, 50% e 66%. Para critério de ilustração, considere que o percentual adotado foi de 66%, ou seja, cada subconjunto gerado, terá 66% do tamanho do conjunto de treino. No cenário ilustrado anteriormente, cada subconjunto gerado por *Bagging* ou *Boosting* serão compostos de 117 instâncias da classe B e 70 exemplares da classe M.

3.2 Estimação das medidas de Complexidade (Processo C)

Uma vez construídos os subconjuntos fez-se necessária a estimação de sua assinatura de complexidade. Esta parte do processo efetua a estimação de complexidade utilizando a DCoL (*Data Complexity Library*) (Orriols-Puig et al., 2010) que consiste de uma biblioteca de aprendizado de máquina, implementada em C++, que provém o cálculo de quatorze descritores de complexidade, dos quais doze serão empregados neste trabalho. As medidas disponíveis na biblioteca estão assinaladas na última coluna da Tabela 1.

Para a utilização da DCoL é necessário que todos os

Tabela 1: Conjuntos de medidas de complexidade divididas de acordo com sua categoria

medida	Descrição	DCoL
	Medidas de Sobreposição	
F1	Pode ser interpretado como a distância entre o centro de duas classes, de forma que, quanto maior o valor do índice, maior a separação entre as classes	✓
F2	Corresponde ao produto das sobreposições de duas classes para cada um dos atributos presentes no conjunto de dados	✓
F3	Refere-se à máxima separabilidade dos atributos de um conjunto de dados, analisando todos os atributos um a um	✓
F4	Corresponde à separabilidade conjunta dos atributos. Exprime quão bem o conjunto de atributos consegue separar os elementos das duas classes	✓
	Medidas de Separabilidade	
L1	Exprime a soma das distâncias das instâncias classificadas erroneamente até a fronteira construída por um classificador linear	✓
L2	Indica a taxa de erros cometidas por um classificador linear sobre o conjunto em análise	✓
N1	Corresponde à fração de elementos que estão mais próximos de instâncias de classes diferentes da deles	✓
N2	Exprime a razão entre as somas das distâncias observadas entre elementos da mesma classe perante a soma das distâncias para elementos de classes distintas	✓
N3	Refere-se à taxa de erro de um classificador KNN adotando-se a abordagem *leave one out*	✓
	Medidas de geometria, topologia e densidade	
L3	Indica a taxa de erros cometidas por um classificador linear sobre um conjunto construído pela interpolação entre elementos escolhidos aleatoriamente do conjunto	✓
N4	Indica a taxa de erros cometidas por um classificador KNN sobre um conjunto construído pela interpolação entre elementos escolhidos aleatoriamente do conjunto	✓
T1	Corresponde ao número de circunferências necessárias para cobrir cada uma das classes presentes no conjunto	✓
T2	corresponde à razão entre o número de elementos do conjunto de dados pelo número de atributos que formam a base	✓
D1	Refere-se ao número médio de amostras por unidade de volume onde os elementos estão distribuídos	✗
D2	Representa o volume médio ocupado pelos k vizinhos mais próximos de cada instância de treino	✗
D3	Consiste na razão do número de elementos que a maioria dos seus vizinhos é de classes diferentes perante o total de instâncias pertencentes à classe	✗
C1	Corresponde ao balanceamento das classes no conjunto de dados estimando-se a entropia normalizada da distribuição dos tamanhos das classes	✗

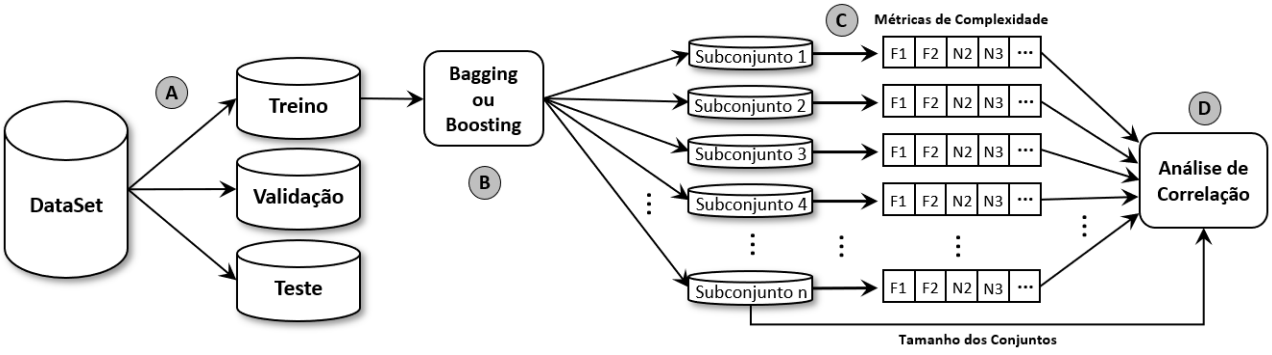

Figura 2: Estrutura geral do *framework* construído

arquivos dos subconjuntos estejam em formato ".arff" (*Attribute-Relation Format File*), que são usados como entrada para a biblioteca. A saída consiste de um vetor com atributos numéricos gravados em arquivo ".txt" para cada uma das entradas.

Além das medidas disponíveis na DCoL (vide Tabela 1) foram empregados outros dois descritores de implementação própria: Volume de Vizinhança Local (D2) e Densidade da Classe na Região de Sobreposição (D3) propostos por Sánchez et al. (2007).

Dentre as medidas apresentadas na Tabela 1, há três que não foram empregadas neste trabalho: D1, T2 e C1. As duas primeiras, por dependerem totalmente do número de instâncias, são facilmente influenciadas pelo tamanho do

conjunto. Já a medida C1, por representar o balanço entre as classes do problema, não apresentará variação de acordo com o tamanho dos conjuntos uma vez que mantemos a estratificação original das classes.

3.3 Análise de Correlação (Processo D)

Estimadas as assinaturas de complexidade de cada subconjunto, o passo seguinte consiste em analisar o comportamento de cada medida de complexidade perante a variação do tamanho do subconjunto, ou seja, o tamanho do subconjunto em relação a cada medida calculada. Este processo foi realizado para todos os subconjuntos gerados por *Bagging* e *Boosting*. Para a análise da correlação foi utilizado o coeficiente de correlação de Pearson, que é definido na Eq. (1).

$$r = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad (1)$$

onde x_i corresponde a cada amostra do primeiro conjunto (medida de complexidade), y_i refere-se a cada amostra do segundo conjunto (tamanho do subconjunto gerado por *Bagging* ou *Boosting*) e n é o número de elementos de cada conjunto.

O coeficiente de correlação de Pearson (r), é utilizado para realizar a medição do grau da correlação linear entre duas variáveis quantitativas. Seu valor está contido entre -1 e 1, onde -1 corresponde a uma relação negativa, enquanto o valor 1 indica uma relação positiva. Se o valor for 0, entende-se que não há relação (Vieira, 2015).

Dentro do limite da correlação pode se identificar, segundo Vieira (2015), alguns níveis intermediários, conforme apresentado na Tabela 2. Valores inferiores a 0,25 indicam que que correlação entre os valores é nula ou pequena. Já valores entre 0,25 e 0,5 indicam uma correlação mais intensa, porém ainda considerada fraca. A partir de 0,5 até 0,75 a correlação é dada como moderada para os conjuntos analisados. Caso a correlação observada seja superior a 0,75 entende-se que um conjunto tem forte correlação com o outro. Valores muito próximos ou iguais a 1 indicam uma correlação perfeita.

Tabela 2: Faixas de interpretação da correlação de Pearson

Valor de *r* (+ ou -)	Interpretação
0.00 a 0.25	Uma correlação pequena ou nula
0.25 a 0.50	Uma correlação fraca
0.50 a 0.75	Uma correlação moderada
0.75 a 1.00	Uma correlação forte ou perfeita

4 Resultados Experimentais

Enquanto a seção anterior teve como objetivo explanar, de uma forma genérica, quais são os métodos utilizados na pesquisa, nesta seção são detalhadas quais os parâmetros aplicados ao experimento, bem como realizou-se a validação do protocolo e análise dos resultados alcançados.

4.1 Bases de dados

Visando obter resultados mais consistentes na avaliação entre o tamanho do conjunto de treino com a sua assinatura de complexidade, optou-se em utilizar um conjunto composto de vinte e seis bases bases de dados, as quais são apresentadas na Tabela 3. Quatorze são originárias do repositório da UCI (Dua and Graff, 2017), duas são procedentes do repositório KEEL (*Knowledge Extraction based on Evolutionary Learning*) (Alcalá-Fdez et al., 2011), outras quatro pertencentes à LKC (*Ludmila Kuncheva Collection of Real Medical Data*) (Kuncheva, 2004), quatro provenientes do projeto STATLOG (King et al., 2000) e duas bases geradas artificialmente com o *toolbox PRTools* do Matlab.

A primeira parte do processo consiste em realizar a divisão das bases. Para a tarefa, cada uma das vinte e seis bases foi dividida aleatoriamente em três conjuntos: treino, teste e validação. Neste processo, o conjunto de treino ficou com 50% do tamanho total da base, o de teste 25% e o conjunto de validação 25%. A divisão foi feita mantendo a proporção das classes do conjunto original. O primeiro conjunto é empregado no aprendizado do classificador. O segundo conjunto, Validação, é usado na definição de parâmetros do classificador. Já o conjunto de teste é utilizado na avaliação de acurácia do classificador.

O número total de instâncias de cada subconjunto é apresentado nas colunas "Treino", "Teste" e "Validação" da Tabela 3. Além disso, são especificados também o número de classes e atributos de cada uma das bases, bem como a taxa de desequilíbrio (na coluna "IR").

Na coluna "IR" é apresentado o valor referente ao *Imbalance Ratio* ou Taxa de Desequilíbrio que é um índice obtido pela divisão da quantidade de instâncias da classe mais representada pela menos representada (Silva et al., 2020). Esta proporção permite identificar se há desbalanceamento presente no conjunto ou se as classes têm a mesma representatividade. Valores próximos de 1 indicam que a proporcionalidade é respeitada. Por outro lado, valores maiores indicam que há classes com maior presença no conjunto de dados.

O processo de divisão dos dados de entrada, geração dos subconjuntos e demais etapas da validação do protocolo foram submetidos a 20 repetições, de forma a construir uma avaliação robusta dos métodos.

4.2 Geração de Subconjuntos

Com base nos conjuntos de treinos formados foram gerados subconjuntos de tamanhos variados, proporcionais ao conjunto base. Estes subconjuntos, que foram formados através dos algoritmos de *Bagging* e *Boosting*, possuem 10%, 33%, 50% e 66% do tamanho do conjunto de treino. Para cada proporção são gerados cem subconjuntos para cada base de dados, totalizando assim, dez mil e quatrocentos subconjuntos para cada um dos métodos de geração em cada uma das repetições.

Para a geração baseada em *Bagging* não foi necessária a especificação de parâmetros, uma vez que o método consiste basicamente em sortear as instâncias com reposição.

No caso do *Boosting*, durante a fase de geração, é preciso efetuar a classificação das instâncias que foram geradas em etapas anteriores. Inicialmente, todas as instâncias do

Tabela 3: Principais características das bases usadas nos experimentos

Base	Instâncias	Treino	Teste	Validação	Atributos	Classes	IR	Fonte
Adult	690	345	172	173	14	2	1,25	UCI
Banana	2000	1000	500	500	2	2	1,00	PRTools
Blood	748	374	187	187	4	2	3,20	UCI
CTG	2126	1063	531	532	21	3	9,40	UCI
Diabetes	766	383	192	191	8	2	1,86	UCI
Faults	1941	971	485	485	27	7	12,24	UCI
German	1000	500	250	250	24	2	2,33	STATLOG
Haberman	306	153	76	77	3	2	2,78	UCI
Heart	270	135	67	68	13	2	1,25	STATLOG
ILPD	583	292	145	146	10	2	2,51	UCI
Ionosphere	350	176	87	87	34	2	1,79	UCI
Laryngeal1	213	107	53	53	16	2	1,63	LKC
Laryngeal3	353	177	88	88	16	3	4,11	LKC
Lithuanian	2000	1000	500	500	2	2	1,00	PRTools
Liver	345	173	86	86	6	2	1,40	UCI
Mammo	830	415	207	208	5	2	1,06	KEEL
Monk	432	216	108	108	6	2	1,12	KEEL
Phoneme	5404	2702	1351	1351	5	2	2,41	ELENA
Segmentation	2310	1155	577	578	19	7	1,00	UCI
Sonar	208	104	52	52	60	2	1,14	UCI
Thyroid	692	346	173	173	16	2	5,00	LKC
Vehicle	847	423	212	212	18	4	1,10	STATLOG
Vertebral	300	150	75	75	6	2	2,10	UCI
WBC	569	285	142	142	30	2	1,68	UCI
WDVG	5000	2500	1250	1250	21	3	1,03	UCI
Weaning	302	151	75	76	17	2	1,00	LKC

conjunto possuem o mesmo peso. No entanto, a partir da segunda iteração, as instâncias tem seu peso aumentado com base nos erros de um classificador. Para tal processo utilizou-se o classificador KNN (K-Vizinhos Mais Próximos) com K igual a cinco.

4.3 Estimação de Complexidade

Para a estimação de complexidade, além das 12 medidas disponibilizadas na DCoL, duas foram implementadas: D2 e D3. Para a implementação destas medidas foi necessário a utilização do KNN, visto que tal medida verifica se um elemento é de uma região de fronteira, de acordo com as instâncias mais próximas. Para este processo, adotou-se um valor para K igual a sete.

Dentre as 12 medidas que estão presentes na DCoL cinco são parametrizáveis, sendo elas: L1, L2, L3, N3 e N4. Para as medidas L1, L2 e L3 é utilizado o classificador SVM com Kernel Linear. Nos dois descritores restantes, utiliza-se o KNN com K igual a um.

4.4 Análise da Correlação

Cada subconjunto gerado resultou em um vetor de descritores de complexidade. Assim, para cada uma das medidas levantadas presentes no vetor, analisou-se seu comportamento perante as dimensões dos subconjuntos gerados pelos métodos de *Bagging* e *Boosting*.

Visando ilustrar o processo de correlação o gráfico da Fig. 3 apresenta um exemplo do cálculo efetuado para uma medida de complexidade genérica. Na imagem, no eixo das abscissas são representadas as proporções, em termos de tamanho (10%, 33%, 50%, 66%), que cada subconjunto pode possuir. No eixo das ordenadas são representados os valores da medida de complexidade referente à cada subconjunto.

O cálculo do coeficiente de correlação foi realizado para cada uma das bases, seguindo a ideia representada na Fig. 3. Assim a correlação é estimada para um conjunto composto de quatrocentos elementos (cem para cada uma das proporções).

Visando representar de forma gráfica as análises realizadas, foram plotados gráficos de dispersão que exemplificam o comportamento dos dados em relação ao tamanho dos conjuntos.

Na Fig. 4 é representada uma correlação negativa observada ao longo dos experimentos. O gráfico foi gerado a partir da base Faults com relação à medida F4, na primeira das vinte repetições executadas. O valor da correlação observado neste caso foi de -0.9681, indicando uma forte correlação negativa (representada pela reta vermelha pontilhada). Analisando os valores é possível notar que, quanto maior o tamanho do conjunto, menor é o valor de F4, logo, menos discriminantes são os dados.

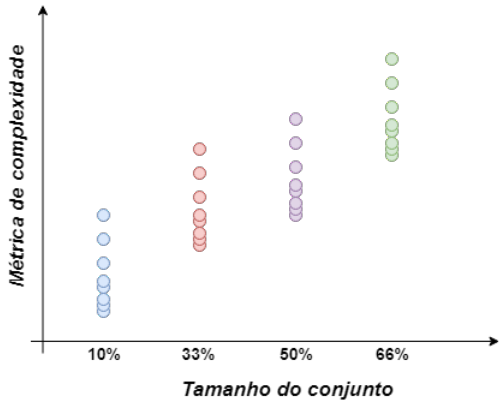

Figura 3: Gráfico representativo da análise de correlação entre medidas e proporção do subconjunto

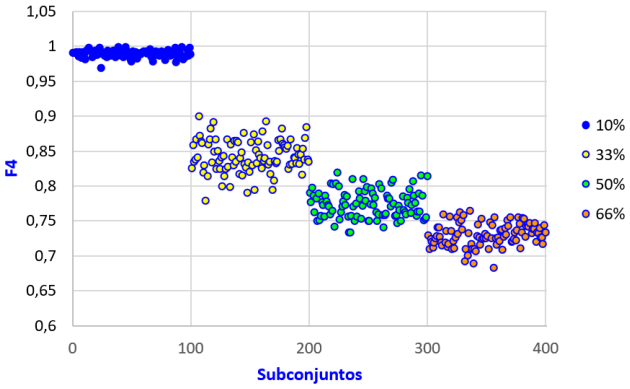

Figura 5: Dispersão dos dados entre a medida F4 e a variação dos subconjuntos

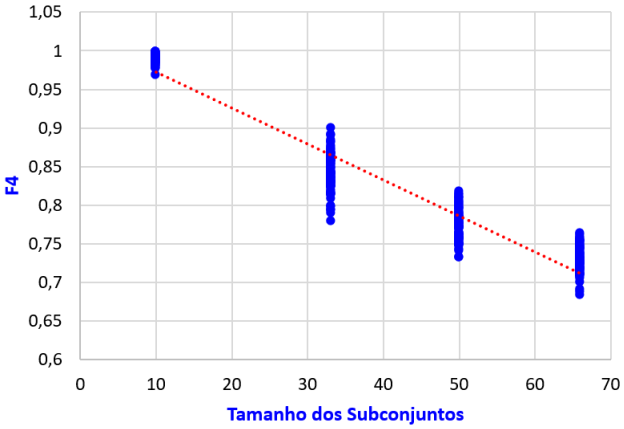

Figura 4: Correlação negativa para a medida F4 em relação a variação dos subconjuntos

Para melhor visualizar a variação do descritor F4 perante os quatrocentos subconjuntos gerados, a Fig. 5 detalha como estão distribuídos os valores do descritor para cada um dos conjuntos. Os primeiros cem elementos (representados em azul) correspondem aos subconjuntos com proporção de 10%, os elementos seguintes (em amarelo) referem-se aos subconjuntos de tamanho 33%. Na cor verde são apresentados os valores da medida obtidos a partir dos subconjuntos com 50% do tamanho do treino e, por fim, em laranja são ilustrados os elementos com 66% de proporção.

Na Fig. 6 é representada uma correlação positiva obtida durante as execuções dos experimentos. O gráfico ilustra um exemplo da base Faults com relação à medida L1, na primeira das vinte repetições executadas. O valor observado para a correlação neste caso foi de 0.9215, caracterizando uma forte correlação positiva (representada pela reta vermelha pontilhada). Analisando os valores é possível notar que, quanto maior o tamanho do conjunto, maior é o valor de L1, o que indica que o classificador linear está cometendo mais erros. Além de aumentar o percentual de erros este fato implica em um valor maior para a soma das instâncias classificadas incorretamente até a fronteira de separação.

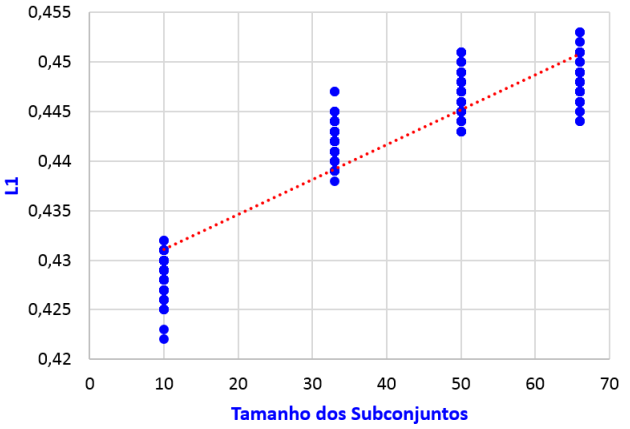

Figura 6: Correlação positiva de L1 em relação a variação dos subconjuntos

A Fig. 7 apresenta uma exposição mais detalhada dos valores de L1 ilustrados na Fig. 6. Nesta representação são exibidos os valores de complexidade para cada um dos subconjuntos. A representação visual segue o mesmo padrão de cores da Fig. 5. Nota-se o aumento nos valores de L1 conforme aumenta-se o tamanho dos subconjuntos.

Para exemplificar uma correlação nula, utilizou-se a 5ª repetição da base de dados Heart perante o comportamento da medida T1. Neste cenário, o coeficiente de correlação apresentou valor de 0,0035. Ilustrando tal comportamento, na Fig. 8 é representada a relação observada entre a variação no tamanho dos subconjuntos e o valor da medida. Notamos que a variação do tamanho do conjunto não possui influência no resultado da medida, ou seja, o número de esferas usadas na cobertura das classes não é influenciado por variações no tamanho dos subconjuntos.

A Fig. 9 apresenta uma distribuição mais detalhada dos

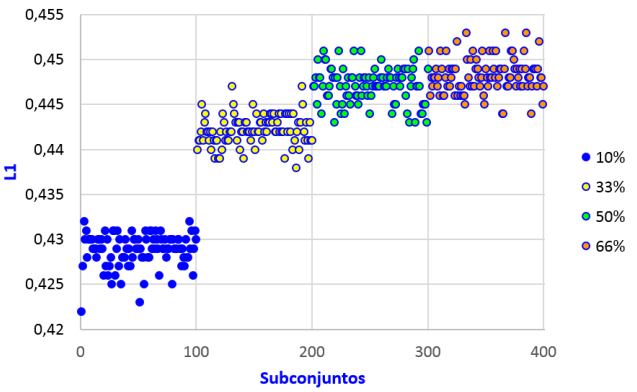

Figura 7: Dispersão dos dados obtidos entre a medida L1 e a variação dos subconjuntos

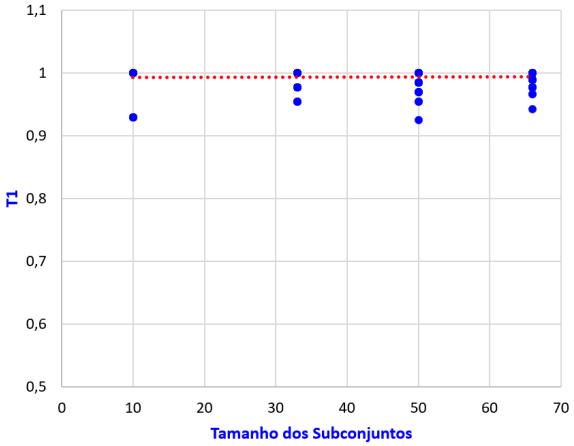

Figura 8: Correlação nula da medida T1 em relação a variação dos subconjuntos

valores de T1 ilustrados na figura anterior. Nesta representação são exibidos os valores de complexidade para cada um dos quatrocentos subconjuntos. A interpretação visual segue o mesmo padrão de cores da Fig. 5. Observando-se a variação dos valores de T1 no gráfico é possível perceber que a medida apresentou variação bastante pequena, onde os índices do descritor sempre estiveram próximos de 1.

Para cada repetição, obteve-se um valor de correlação entre cada medida de complexidade e cada uma das bases de teste. Assim, o coeficiente final foi obtido pela média de todas as 20 repetições. A partir destas médias foram geradas quatro tabelas, duas para o *Bagging* (Tabelas 4 e 5) e duas para o *Boosting* (Tabelas 6 e 7). Nas tabelas, cada linha corresponde a uma base de dados, enquanto as colunas remetem às medidas de complexidade.

Buscando um melhor entendimento das correlações obtidas entre cada uma das medidas e o tamanho de cada subconjunto, realizou-se uma análise individual para cada um dos descritores de complexidade em relação às médias apresentadas nas Tabelas 4 a 7. Tal análise teve por objetivo identificar medidas suscetíveis à variação no tamanho de um subconjunto.

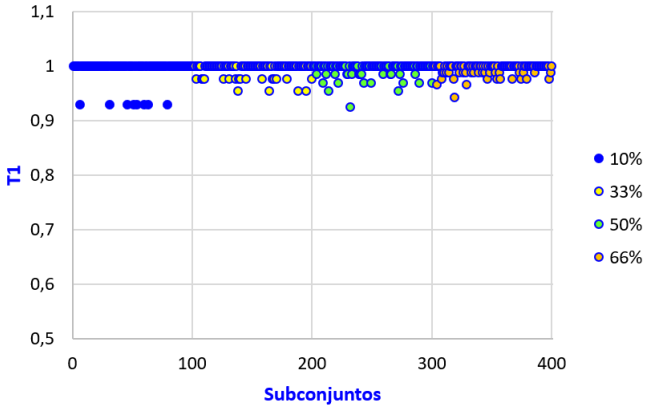

Figura 9: Dispersão dos dados entre a medida T1 e a variação dos subconjuntos

4.4.1 F1

A partir dos dados apresentados na Tabela 4, é perceptível o fato de que, para todas as bases, o valor final da média da correlação para F1 é sempre negativo, ou seja, utilizando o algoritmo de *Bagging*, o crescimento dos subconjuntos é inversamente proporcional ao valor de F1, de modo que, quanto maior o tamanho do subconjunto, menor é o valor da medida e, possivelmente, menos separáveis serão as classes do subconjunto.

O algoritmo de *Boosting* apresenta comportamento similar ao Bagging, ou seja, o aumento no tamanho dos subconjuntos de treino levam a uma diminuição no valor do índice F1.

4.4.2 F2

Apesar da variação do tamanho dos subconjuntos, a medida F2 teve uma correlação positiva tanto para o *Bagging* quanto para o *Bossting* em grande parte das bases, com algumas exceções (ILPD, Ionosphere, Laryngeal1, Segmentation, Sonar, Thyroid, Vehicle e WBC), onde os valores da correlação foram zero ou próximos a zero, ou seja, uma correlação pequena ou nula.

Para as outras bases a correlação foi positiva com média 0,281, caracterizando assim, uma correlação fraca. Como para F2, quanto maior seu valor maior sobreposição, infere-se que quanto maior o subconjunto mais alto é o valor da medida.

4.4.3 F3

Analisando os valores apresentados pelas médias do *Bagging* e *Boosting* para o descritor F3, percebe-se que o valor da correlação obtida é negativa para todas as bases. Para esta medida, a média da correlação foi de -0,658, que indica uma correlação moderada.

Como para F3 valores próximos a 0 remetem a um problema de alta complexidade, entende-se que, a medida que o tamanho do subconjunto aumenta mais complexo se torna o problema, já que o valor de F3 diminui.

Tabela 4: Coeficiente de Correlação médio entre os descritores de complexidade e o tamanho dos subconjuntos gerados por ***bagging*** ao longo das 20 repetições para cada uma das 26 bases - Parte 1

Base	F1	F2	F3	F4	L1	L2	L3
Adult	-0,1666	0,1778	-0,7217	-0,8567	-0,7158	-0,3463	-0,4155
Banana	-0,0959	0,5236	-0,5264	-0,7009	-0,845	-0,5647	-0,5281
Blood	-0,3054	0,2172	-0,7054	-0,7961	-0,0804	-0,0987	0
CTG	-0,268	0,1673	-0,6598	-0,9063	0,5016	0,0917	0
Diabetes	-0,2881	0,547	-0,7394	-0,8976	0,0799	-0,1049	-0,014
Faults	-0,4295	0,3462	-0,8442	-0,9634	0,8788	-0,0728	-0,1839
German	-0,5372	0,422	-0,7345	-0,8579	0,775	-0,3964	-0,3197
Haberman	-0,4225	0,6915	-0,7662	-0,8638	-0,4942	-0,5554	0
Heart	-0,1524	0,606	-0,7769	-0,3171	-0,7461	-0,7118	-0,7347
ILPD	-0,532	-0,0856	-0,7454	-0,897	-0,2104	-0,3017	0
Ionosphere	-0,4001	0	-0,7784	0,4359	-0,6867	-0,7401	-0,7299
Laryngeal1	-0,3957	0	-0,5351	0,6226	-0,2212	-0,3009	-0,2114
Laryngeal3	-0,5589	0,2689	-0,8478	0,4292	-0,1127	-0,095	0,0443
Lithuanian	-0,0998	0,5119	-0,5813	-0,745	-0,9146	-0,6667	-0,6845
Liver	-0,5571	0,4012	-0,7886	-0,8905	0,1128	0,0401	0,0484
Mammo	-0,3922	0,3452	-0,6544	-0,7519	-0,6031	-0,3418	-0,3982
Monk	-0,317	0,6805	-0,4383	0,5237	-0,8622	-0,7107	-0,7158
Phoneme	-0,3578	0,4672	-0,5051	-0,7851	0,9211	-0,3071	-0,2769
Segmentation	-0,2751	0	-0,7827	-0,5002	0,8714	-0,0157	0,0036
Sonar	-0,5421	0	-0,8539	0,5676	0,0271	-0,0811	-0,1683
Thyroid	-0,3893	0	-0,6264	0,5775	0,7666	-0,1015	0
Vehicle	-0,5696	-0,0356	-0,833	-0,8592	-0,3024	-0,3859	0
Vertebral	-0,391	0,3594	-0,5938	0,1212	0,22	0,0325	0,0106
WBC	-0,388	0	-0,5927	0,5466	0,1771	-0,5736	-0,5798
WDVG	-0,4866	0,7295	-0,8382	-0,903	0,3449	0,1062	0
Weaning	-0,3787	0,2091	-0,7109	0,4298	-0,5844	-0,1646	-0,0851

Tabela 5: Coeficiente de Correlação médio entre os descritores de complexidade e o tamanho dos subconjuntos gerados por ***bagging*** ao longo das 20 repetições para cada uma das 26 bases - Parte 2

Base	N1	N2	N3	N4	T1	D2	D3
Adult	-0,6035	-0,9038	-0,6033	0,5101	0,0999	-0,0003	-0,1806
Banana	-0,7825	-0,9413	-0,6209	0,219	-0,1636	-0,8921	-0,2971
Blood	-0,6621	-0,8348	-0,7072	0,4065	0,2283	-0,3657	-0,0371
CTG	-0,8329	-0,9587	-0,8164	0,2605	0,3105	-0,0003	-0,7028
Diabetes	-0,6807	-0,9108	-0,7183	0,6062	0,1267	-0,5359	-0,4024
Faults	-0,908	-0,9532	-0,8947	0,799	-0,0784	-0,1348	-0,862
German	-0,7331	-0,9446	-0,7838	0,7753	0,0207	0	-0,3006
Haberman	-0,54	-0,7788	-0,5807	0,5496	0,1915	-0,679	0,2597
Heart	-0,5325	-0,838	-0,48	0,5117	0,0963	-0,4222	-0,5224
ILPD	-0,6365	-0,8906	-0,6893	0,5983	0,159	-0,0063	-0,1006
Ionosphere	-0,7231	-0,843	-0,6749	0,2512	0,248	0	-0,7163
Laryngeal1	-0,5413	-0,8143	-0,4829	0,2528	0,1591	-0,0974	-0,7298
Laryngeal3	-0,6872	-0,8413	-0,6578	0,5161	0,304	-0,0926	-0,4508
Lithuanian	-0,8137	-0,9452	-0,7027	0,0111	-0,0321	-0,8963	-0,3104
Liver	-0,6031	-0,8501	-0,6612	0,5845	0,1	-0,1313	-0,5337
Mammo	-0,443	-0,8027	-0,5643	0,4138	-0,0353	-0,2779	-0,1427
Monk	-0,7373	-0,8861	-0,7162	0,2643	-0,1131	-0,8173	-0,5625
Phoneme	-0,8946	-0,9652	-0,9109	0,4092	0,1475	-0,8343	-0,8036
Segmentation	-0,9293	-0,8766	-0,9002	-0,3781	-0,2567	0	-0,8864
Sonar	-0,6893	-0,8622	-0,6853	0,47	0	0	-0,6064
Thyroid	-0,5391	-0,8914	-0,454	0,2931	0,2809	-0,1109	-0,5822
Vehicle	-0,7881	-0,9215	-0,815	0,5926	0,3799	-0,0181	-0,7963
Vertebral	-0,5812	-0,8147	-0,5483	0,3409	0,1796	-0,4436	-0,4909
WBC	-0,6285	-0,8886	-0,5151	0,0483	0,2334	-0,0847	-0,3858
WDVG	-0,8761	-0,9855	-0,9029	0,5969	0,0857	-0,7926	-0,503
Weaning	-0,6763	-0,8703	-0,6448	0,5149	0,0214	-0,4439	-0,6665

4.4.4 F4

Uma vez que F4 reflete o percentual de instâncias que podem ser separadas pelo conjunto de atributos, espera-se que o aumento do número de instâncias implique em um problema mais complexo e, consequentemente, este percentual decaia.

O valor médio observado da correlação indica que con-

Tabela 6: Coeficiente de Correlação médio entre os descritores de complexidade e o tamanho dos subconjuntos gerados por ***boosting*** ao longo das 20 repetições para cada uma das 26 bases - Parte 1

Base	F1	F2	F3	F4	L1	L2	L3
Adult	-0,2345	0,2596	-0,7433	-0,8894	-0,3567	-0,2983	-0,3855
Banana	-0,1369	0,2859	-0,3521	-0,5094	-0,7842	-0,5681	-0,5851
Blood	-0,4002	0,2284	-0,6323	-0,7604	-0,0499	0,0276	0,106
CTG	0,0254	0,2127	-0,569	-0,866	0,1646	0,0486	-0,1389
Diabetes	-0,5073	0,567	-0,7421	-0,8881	-0,0133	0,1337	0,2029
Faults	-0,5168	0,379	-0,7966	-0,9296	0,3466	0,3117	-0,0137
German	-0,4278	0,4932	-0,7398	-0,8155	0,0855	-0,3508	-0,4012
Haberman	-0,4849	0,6871	-0,7352	-0,8562	0,0666	0,3131	0,2895
Heart	-0,4276	0,638	-0,7903	-0,5819	-0,2709	-0,0646	-0,078
ILPD	-0,59	0,0264	-0,7413	-0,914	0,1253	0,0624	0,0132
Ionosphere	-0,5793	0	-0,6814	0,3996	-0,3315	-0,3519	-0,0892
Laryngeal1	-0,4471	-0,0005	-0,7249	0,4358	0,3318	0,2521	0,06
Laryngeal3	-0,5655	0,2965	-0,8401	0,4227	0,1718	0,2282	0,1569
Lithuanian	-0,1464	0,1593	-0,4579	-0,6085	-0,7948	-0,5352	-0,5476
Liver	-0,5607	0,4585	-0,7889	-0,8969	0,1874	0,2311	0,1824
Mammo	-0,4123	0,2898	-0,6007	-0,6962	-0,7424	-0,5044	0,2451
Monk	-0,4164	0,6826	-0,1174	0,2509	-0,6093	-0,099	-0,0363
Phoneme	-0,3405	0,3718	-0,4561	-0,6955	0,2872	0,2573	-0,1855
Segmentation	-0,3452	0	-0,4459	-0,7134	0,2256	0,1154	-0,0356
Sonar	-0,5369	0	-0,8533	0,5373	0,1193	0,1987	0,2083
Thyroid	-0,2775	0	-0,4926	0,3887	0,3158	0,1412	0
Vehicle	-0,6382	0,0227	-0,8551	-0,9022	0,5691	0,5681	0
Vertebral	-0,458	0,4307	-0,6854	-0,3815	0,2416	0,2123	0,0422
WBC	-0,4126	0	-0,6306	0,4629	0,0433	-0,3187	-0,3727
WDVG	-0,4107	0,6256	-0,7825	-0,8808	0,6545	-0,2341	0
Weaning	-0,435	0,1988	-0,7704	0,4127	0,1238	0,1354	0,0374

Tabela 7: Coeficiente de Correlação médio entre os descritores de complexidade e o tamanho dos subconjuntos gerados por ***boosting*** ao longo das 20 repetições para cada uma das 26 bases - Parte 2

Base	N1	N2	N3	N4	T1	D2	D3
Adult	-0,7472	-0,917	-0,814	0,5747	0,0408	-0,0518	-0,0638
Banana	-0,8267	-0,9334	-0,7959	0,0824	-0,1787	-0,888	-0,2758
Blood	-0,864	-0,8457	-0,8336	0,5328	-0,2587	-0,4212	0,1061
CTG	-0,8604	-0,8602	-0,8799	0,3358	0,4052	-0,0256	-0,1402
Diabetes	-0,7875	-0,9233	-0,8213	0,7012	0,0989	-0,5592	0,0163
Faults	-0,8161	-0,9528	-0,9068	0,7501	0,2408	-0,1313	-0,5549
German	-0,7806	-0,9401	-0,8531	0,8121	0,0109	0	0,0401
Haberman	-0,7102	-0,8209	-0,7143	0,6133	0,1379	-0,6969	-0,0756
Heart	-0,6265	-0,8661	-0,6761	0,663	0,056	-0,3872	-0,1986
ILPD	-0,7397	-0,8986	-0,7893	0,7009	0,1371	-0,0175	0,0666
Ionosphere	-0,8066	-0,871	-0,8046	0,3864	0,1883	0	-0,4344
Laryngeal1	-0,6214	-0,8326	-0,6814	0,4439	0,1079	-0,0575	-0,4517
Laryngeal3	-0,7082	-0,8533	-0,7301	0,682	0,2461	-0,1033	-0,3247
Lithuanian	-0,8381	-0,9329	-0,8256	0,0605	0,0448	-0,8815	-0,1735
Liver	-0,6792	-0,8603	-0,7217	0,6101	0,1278	-0,2242	-0,1783
Mammo	-0,8222	-0,8759	-0,8164	0,5355	-0,0233	-0,2715	-0,4376
Monk	-0,7425	-0,8858	-0,7698	0,416	-0,1393	-0,7958	-0,3059
Phoneme	-0,8906	-0,9589	-0,9384	0,4245	0,1905	-0,8364	-0,1134
Segmentation	-0,9098	-0,858	-0,8972	-0,0036	-0,0602	0	-0,8481
Sonar	-0,6805	-0,8668	-0,7149	0,4661	0,0234	0	-0,4352
Thyroid	-0,7059	-0,9022	-0,7301	0,3275	0,154	-0,0769	-0,2718
Vehicle	-0,8106	-0,9257	-0,8458	0,6951	0,3848	0,0212	-0,6232
Vertebral	-0,6828	-0,8542	-0,7057	0,4819	0,1404	-0,4648	-0,1674
WBC	-0,7074	-0,8631	-0,6515	0,1672	0,1507	-0,0764	-0,1888
WDVG	-0,742	-0,9377	-0,9229	0,4654	0,0689	-0,7579	0,2148
Weaning	-0,7044	-0,8755	-0,7291	0,6132	0,0006	-0,4076	-0,3427

forme o tamanho dos subconjuntos aumenta o valor de F4 diminui. No entanto, para algumas bases observou-se correlação moderada positiva, o que sugere um aumento no valor de F4 em relação ao crescimento dos subconjuntos. Analisando-se as bases em que o fato ocorreu notou-se que praticamente todas apresentam apenas duas classes e, geralmente, possuem os maiores números de atributos.

4.4.5 L1

O comportamento de L1 mostrou-se bastante variado em relação às dimensões dos subconjuntos. Em parte das bases, o aumento da dimensão dos subconjuntos implicou em diminuição do valor de complexidade. Este fato foi observado para as bases compostas por duas classes e, em sua maioria por poucos atributos. Por outro lado em bases onde o número de atributos era maior e o número de classes era superior a dois a correlação foi positiva, ou seja, a soma da distância dos erros aumentou. Acredita-se que tal correlação tenha se mostrado positiva pois um número maior de atributos e classes impactam em valores maiores para a distância dos erros até a fronteira de classificação.

4.4.6 L2

Para os subconjuntos gerados por *bagging* observou-se que a taxa de erro do classificador linear apresentou certa diminuição conforme o tamanho dos subconjuntos aumentou. Esta correlação negativa no entanto, mostrou-se fraca, com valores inferiores a 0,3.

A relação dos subconjuntos gerados por *boosting* perante L2, no entanto, mostrou-se muito próxima de zero, indicando que, neste caso, a variação do tamanho dos subconjuntos não causou variação na taxa de acerto do classificador linear.

4.4.7 L3

O comportamento apresentado pelo descritor L3 foi bastante similar ao observado para L2. Tal fato já era esperado uma vez que ambos baseiam-se na taxa de erro de um classificador linear.

4.4.8 N1

O aumento no tamanho dos subconjuntos teve influência direta no comportamento da medida N1. Os experimentos mostraram que, conforme o tamanho do subconjunto aumenta, o nível da complexidade estimado para a medida de complexidade diminui.

Como N1 reflete o número de vizinhos na região de sobreposição, acredita-se que conforme o número de instâncias foi aumentando, a maioria delas foi sendo distribuída nas regiões onde não há tanta sobreposição e, por isso, o valor da complexidade foi diminuindo, já que a proporção de instâncias na fronteira foi se tornando menor.

4.4.9 N2

Assim como ocorreu para a medida anterior, N2 apresentou uma alta correlação negativa perante o crescimento dos subconjuntos. O valor da correlação apresentou média aproximada a -0,89 para os grupos gerados por *Bagging* e por *Boosting*.

Tal comportamento indica que, conforme a dimensão dos subconjuntos aumentou, a distância entre os elementos de uma mesma classe tornaram-se menores, aumentando a coesão de cada classe. Essa diminuição implica em valores menores para F2 que mede a relação intra classes perante a inter classes.

4.4.10 N3

Considerando que N3 é a taxa de erro de um classificador KNN, era esperado que o valor da medida diminuísse conforme o tamanho do conjunto de treino (subconjuntos) aumentasse. Logo, os resultados obtidos foram os esperados. N3 apresentou média de 0,791 em seu coeficiente de correlação tanto para o *Bagging* quanto ao *Boosting*. Tal resultado representa uma correlação forte ou perfeita.

4.4.11 N4

Analisando-se o comportamento de N4 perante a variação do tamanho dos subconjuntos observou-se uma correlação positiva fraca (0,43). O que sugere que o classificador 1NN comete mais erros quando os subconjuntos possuem mais instâncias. Acredita-se que este fato ocorre pois conforme os subconjuntos vão sendo aumentados, a dificuldade do classificador também aumenta.

4.4.12 T1

O valor da correlação de T1 perante a variação do tamanho dos subconjuntos foi considerado pequeno ou nulo (aproximadamente 0,156), o que indica que a medida é pouco suscetível à oscilação na dimensão dos subconjuntos formados, tanto por *bagging* quanto por *boosting*.

4.4.13 D2

De acordo com a variação do tamanho dos subconjuntos o valor de D2 manteve-se negativo ou nulo. O valor médio da correlação de D2 calculado foi de -0,311 para o *Bagging* e -0,314 para o *Boosting*. Ambos os valores mostram que a medida tem um valor mais próximo a zero à medida que os subconjuntos crescem, resultando em uma correlação fraca.

4.4.14 D3

Para D3, quanto maior o número de elementos que pertencem a uma região de sobreposição, maior será seu valor. Levando em conta esta afirmação e analisando os valores coletados, nota-se uma correlação negativa fraca na utilização do algoritmo de *boosting*, pois o valor médio levantado para D3 foi de -0,2711 o que indica que a maior parte das instâncias dos conjuntos maiores são distribuídas em regiões de menor sobreposição.

Já para o *bagging* houve uma correlação pequena, devido ao fato de que o valor de D3 foi de -0,1558.

4.4.15 Bagging vs Boosting

Com o objetivo de comparar o comportamento das medidas de complexidade perante às variações do tamanho dos subconjuntos gerados por *bagging* e *boosting*, levantou-se o valor médio absoluto das correlações para as duas técnicas de geração, os quais são apresentados nas Tabela 8. O objetivo foi tentar mensurar o quanto o tamanho dos subconjuntos interfere nas medidas de complexidade, positiva ou negativamente.

Analisando-se os valores absolutos dos coeficientes de cada uma das medidas, nota-se que, em sua maioria, os algoritmos de geração não apresentaram discrepâncias consideráveis nas correlações apresentadas.

Por outro lado, observou-se que em quatro casos específicos (L1, N1, N3 e D3), a utilização do *bagging* e *boosting* teve influência direta nos resultados, ou seja, notou-se certa diferença entre as correlações.

Para L1 a diferença entre as abordagens foi de aproxima-

Tabela 8: Média dos valores absolutos para a correlação entre as medidas de complexidade e o tamanho dos subconjuntos

medidas	Bagging	Boosting
F1	0,373	0,413
F2	0,300	0,281
F3	0,699	0,655
F4	0,683	0,658
L1	0,502	0,308
L2	0,304	0,252
L3	0,237	0,170
N1	0,695	0,762
N2	0,885	0,889
N3	0,682	0,791
N4	0,430	0,483
T1	0,156	0,139
D2	0,311	0,314
D3	0,494	0,271

damente 20 pontos percentuais. Segundo o valor obtido pelo *bagging* a medida se enquadra na faixa de correlação moderada, já pelo *boosting*, o índice está na faixa de correlação fraca.

Com relação ao descritor N1, tem-se uma diferença menor, em torno de 10 pontos percentuais. No entanto, apesar da diferença ser menor, a correlação de cada estratégia de geração cai em uma faixa diferente de interpretação: para o *bagging* a relação entre tamanho e complexidade é apenas moderada, enquanto para o *boosting*, ela é considerada forte ou perfeita.

Como ocorreu para N1, o descritor N3, apresentou uma correlação moderada no uso do *bagging* e uma correlação forte ou perfeita no uso do *boosting*. Contudo, para D3, tanto para o *bagging* quanto para o *boosting* a correlação se mantém na faixa moderada, mas com uma diferença próxima aos 20 pontos percentuais entre as duas estratégias de geração.

Um dos objetivos do trabalho foi tentar identificar quais medidas de complexidade são mais suscetíveis à variação do tamanho dos conjuntos de dados. De forma a tentar responder esta pergunta, são apresentados na Tabela 9 o conjunto de descritores de complexidade analisados, divididos por faixa de correlação para os subconjuntos gerados pelo *Bagging* e *Boosting*.

A partir dos dados apresentados na tabela percebe-se que as medidas L3, T1, F1, F2, L2, N4, D2 e D3 são menos influenciadas pelo tamanho do conjunto em que é calculada. Entretanto, os descritores F3, F4, N1, N2 e N3 sofrem mais influência da quantidade de instâncias presentes no conjunto. A medida L1 apresentou um comportamento mais indefinido, variando entre correlação fraca e moderada.

Com o objetivo de analisar se os métodos de geração apresentam comportamento similar em termos de sentido da correlação entre a variação no tamanho dos subconjuntos perante os valores das medidas de complexidade são apresentados na Tabela 10 os valores médios das correlações obtidas.

Analisando-se os valores apresentados, observou-se que, tanto para o *Bagging* quanto para o *Boosting*, a correlação apresenta o mesmo sentido, ou seja, quando um índice de complexidade aumenta para o *Bagging* o mesmo aumenta para o *Boosting*. Este fato também é observado no sentido contrário, quando a medida diminui para os dois métodos de geração.

Além disso, percebeu-se que as medidas baseadas em classificador de vizinhança (N1, N2, N3 e N4) são mais suscetíveis ao *Boosting* enquanto as medidas de complexidade baseadas em um classificador linear (L1, L2 e L3) tem maior suscetibilidade aos conjuntos gerados pelo *Bagging*.

Na primeira coluna da Tabela 10 são apresentados sinais visuais. Logo, o sentido da seta aponta para qual valor da medida, se maior ou menor, indicará que o problema de classificação será mais simples. Sendo assim, quando a seta está voltada para cima, o índice aumenta, porém, a complexidade do conjunto diminui. Analogamente, as setas que apontam para baixo indicam que o valor medida decresce, assim como a complexidade do conjunto também diminui.

Os sinais das correlações apresentados na tabela indicam se a medida cresce ou diminui conforme o conjunto de dados aumenta. A correlação medida perante o sentido das setas exprime que, para os índices L1, L2, L3, N1, N2, N3 e D3 conforme o conjunto de treino aumenta, o problema torna-se menos complexo. Por outro lado, para as demais medidas (F1, F2, F3, F4, N4, T1 e D2) um aumento no tamanho dos conjuntos de treino indicam aumento na complexidade da classificação.

5 Conclusões

Neste trabalho, destaca-se o objetivo de estudar a relação entre as assinaturas de complexidade de um conjunto de dados perante o tamanho desse conjunto. Esperava-se, por meio desta pesquisa, identificar quais medidas de complexidade são mais suscetíveis à variação do tamanho do conjunto de dados. Diante desse desafio, a estratégia adotada foi a realização de experimentos com posterior análise dos resultados.

Para o início dos experimentos foram implementados dois algoritmos de geração de subconjuntos: *Bagging* e *Boosting*. Cada um gerou subconjuntos com tamanhos de 10%, 33%, 50% e 66% em relação ao tamanho da base de treinamento, que por sua vez correspondeu a 50% do total de observações registradas. Cada subconjunto teve a assinatura de complexidade estimada, possibilitando a avaliação da correlação entre o tamanho do subconjunto com o valor de cada descritor, usado para descrever a assinatura de complexidade. Para uma ampla avaliação, foi adotado um protocolo experimental robusto composto por 26 bases. Cada base foi submetida a 20 replicações, adotando o sorteio aleatório de amostras para aumentar a confiabilidade na análise por evitar situações de tendenciosas.

Diante dos resultados experimentais foi realizada a análise. Esta, sugere que os descritores se mostraram sensíveis a i) dimensionalidade das bases e ii) ao número de classes. As medidas de complexidade, apesar de mostrarem relação forte com o número de instâncias dos conjuntos de dados, apresentaram correlação positiva para algumas bases enquanto que para outras apresentou correlação negativa. Portanto, embora a relação se faça presente, não está totalmente elucidado o motivo de apresentar correlações positivas ou negativas quando analisadas todas as bases do estudo.

Assim, pode-se afirmar que as medidas que sofrem

Tabela 9: Comportamento das medidas de complexidade perante às faixas de classificação do coeficiente de correlação de Pearson

Faixa da Correlação	Bagging	Boosting
Pequena ou Nula	L3 e T1	L3 e T1
Fraca	F1, F2, L2, N4, D2 e D3	F1, F2, L1, L2, N4, D2 e D3
Moderada	F3, F4, L1, N1 e N3	F3 e F4
Forte ou Perfeita	N2	N1, N2 e N3

Tabela 10: Análise do sentido da correlação perante a complexidade do problema

Sentido	medidas	Bagging	Boosting
⇑	F1	-0,373	-0,411
⇓	F2	0,290	0,281
⇑	F3	-0,699	-0,655
⇑	F4	-0,355	-0,403
⇓	L1	-0,065	-0,283
⇓	L2	-0,283	-0,003
⇓	L3	-0,228	-0,051
⇓	N1	-0,695	-0,762
⇓	N2	-0,885	-0,889
⇓	N3	-0,682	-0,791
⇓	N4	0,401	0,482
⇓	T1	0,104	0,088
⇑	D2	-0,311	-0,312
⇓	D3	-0,474	-0,223

mais influência do tamanho do conjunto são: F3, F4, N1, N2 e N3, enquanto as menos suscetíveis foram: L3, T1, F1, F2, L2, N4, D2 e D3. Ou seja, estas últimas são mais indicadas em cenários em que pode haver variação do número de instâncias do conjunto.

Observou-se, através dos experimentos realizados, que as correlações existentes entre os algoritmos de *Bagging* e *Boosting* foram bastante similares, de forma que em apenas alguns descritores houve uma diferença significativa nos resultados obtidos da correlação.

Para uma análise mais detalhada entre os algoritmos de geração de subconjuntos, seria interessante a comparação entre *Bagging*, *Boosting* e *Random Subspace*. Dessa forma poderiam ser analisados fatores como a quantidade de atributos dos dados no comportamento das assinaturas de complexidade.

Poderia ser explorado também o comportamento das instâncias de um conjunto dentro do espaço de características e como estes atributos afetam a complexidade do conjunto.

Seria interessante também a proposição de novos descritores de complexidade, focando principalmente em uma medida que não seja suscetível ao tamanho do conjunto de dados e que também não varie de acordo com o número de atributos.

Referências

Alcalá-Fdez, J., Fernãndez, A., Luengo, J., Derrac, J., García, S., Sãnchez, L. and Herrera, F. (2011). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, *Journal of Multiple-Valued Logic and Soft Computing* **17**(2-3): 255–287. Disponível em https://tinyurl.com/yct8k3ba.

Augspurger, T., Bartak, C., Cloud, P. and Hayden, A. (2018). Pandas. Disponível em http://pandas.pydata.org/.

Breiman, L. (1996). Bagging predictors. machine learning, *Machine Learning* **24**(2): 123–140. Disponível em https://tinyurl.com/ycr4yagn.

Brun, A. L., Britto, A. S., Oliveira, L. S., Enembreck, F. and Sabourin, R. (2016). Contribution of data complexity features on dynamic classifier selection, *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 4396–4403. https://doi.org/10.1109/IJCNN.2016.7727774.

Brun, A. L., Britto, A. S., Oliveira, L. S., Enembreck, F. and Sabourin, R. (2018). A framework for dynamic classifier selection oriented by the classification problem difficulty, *Pattern Recognition* **76**: 175 – 190. https://doi.org/10.1016/j.patcog.2017.10.038.

Cavalcanti, G., Ren, T. and Vale, B. (2012). Data complexity measures and nearest neighbor classifiers: A practical analysis for meta-learning, *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, Vol. 1, pp. 1065–1069. https://doi.org/10.1109/ICTAI.2012.150.

Dua, D. and Graff, C. (2017). UCI machine learning repository. Disponível em http://archive.ics.uci.edu/ml.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm, *ICML'96 Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* **24**(2): 148–156. Disponível em https://tinyurl.com/yarxhfz6.

Ho, T. K. (1998). The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8): 832–844. https://doi.org/10.1109/34.709601.

Ho, T. K. and Basu, M. (2002). Complexity measures of supervised classification problems, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(3): 289–300. https://dl.acm.org/doi/10.1109/34.990132.

Ho, T. K., Basu, M. and Law, M. (2006). Measures of geometrical complexity in classification problems, *Data Complexity in Pattern Recognition, Advanced Information and Knowledge Processing* **16**: 1–23. https://doi.org/10.1007/978-1-84628-172-3_1.

Jain, A., Duin, R. P. W. and Mao, J. (2000). Statistical pattern recognition: a review, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(1): 4–37. http://dx.doi.org/10.1109/34.824819.

King, R., Feng, C. and Sutherl, A. (2000). Statlog: Comparison of classification algorithms on large real-world problems, *Applied Artificial Intelligence* **9**. https://doi.org/10.1080/08839519508945477.

Kittler, J., Hatef, M., Duin, R. P. W. and Matas, J. (1998). On combining classifiers, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **20**(3): 226–239. http://dx.doi.org/10.1109/34.667881.

Ko, A. H., Sabourin, R. and Britto Jr., A. S. (2008). From dynamic classifier selection to dynamic ensemble selection, *Pattern Recognition* **41**(5): 1718–1731. https://doi.org/10.1016/j.patcog.2007.10.015.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine Learning* **51**(2): 181–207. http://dx.doi.org/10.1023/A%3A1022859003006.

Kuncheva, L. L. (2004). *Combining Pattern Classifiers*, 1 edn, John Wiley & Sons, Inc, Hoboken, New Jersey.

Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P. and Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity, *ACM Comput. Surv.* **52**(5). https://doi.org/10.1145/3347711.

Luengo, J. and Herrera, F. (2010). Domains of competence of fuzzy rule based classification systems with data complexity measures: A case of study using a fuzzy hybrid genetic based machine learning method, *Fuzzy Sets Syst.* **161**(1): 3–19. http://dx.doi.org/10.1016/j.fss.2009.04.001.

Orriols-Puig, A., Macià, N. and Ho, T. K. (2010). Documentation for the data complexity library in c++, *Technical report*, Barcelona, Spain. Disponível em http://dcol.sourceforge.net/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**(12): 2825–2830. Disponível em http://www.techscience.com/CMES/v114n2/27377.

Ponti Jr., M. P. (2011). Combining classifiers: From the creation of ensembles to the decision fusion, *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2011 24th SIBGRAPI Conference on*, pp. 1–10. https://doi.org/10.1109/SIBGRAPI-T.2011.9.

Silva, R. A., Britto Jr, A. d. S., Enembreck, F., Sabourin, R. and de Oliveira, L. E. S. (2020). CSBF: A static ensemble fusion method based on the centrality score of complex networks, *Computational Intelligence* **36**(2): 522–556. https://doi.org/10.1111/coin.12249.

Skurichina, M. and Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers, *Pattern Analysis & Applications* **5**(2): 121–135. http://dx.doi.org/10.1007/s100440200011.

Sánchez, J. S., Mollineda, R. A. and Sotoca, J. M. (2007). An analysis of how training data complexity affects the nearest neighbor classifiers, *Pattern Anal. Appl.* **10**(3): 189–201. https://doi.org/10.1007/s10044-007-0061-2.

Sotoca, J. M., Sánchez, J. S. and Mollineda, R. A. (2005). A review of data complexity measures and their applicability to pattern classification problems, *Actas del III Taller Nacional de Minería de Dados y Aprendizaje*, TAMIDA,05, pp. 77–83. Disponível em https://tinyurl.com/ydhjub6t.

Tan, P.-N., Steinbach, M. and Kumar, V. (2009). *Introdução ao Data Mining. Mineração de Dados*, Ciência Moderna, Rio de Janeiro.

Vieira, S. (2015). *Introdução a Bioestatistica*, 5 edn, Elsevier Editora Ltda., Rio de Janeiro.