UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA

Wallace Soares dos Santos Júnior

Avaliação de Estratégias de Seleção de Instância e de Atributos em Série

Recife

2021

Wallace Soares dos Santos Júnior

Avaliação de Estratégias de Seleção de Instância e de Atributos em Série

Monografia apresentada ao Curso de Engenharia da Computação, como requisito parcial para a obtenção do Título de Bacharel em Engenharia de Computação.

Área de Concentração: Aprendizagem de Maquina

Orientador (a): Prof. George Darmiton Cavalcanti

Recife

2021

Folha de aprovação: Inserir a folha de aprovação enviada pela Secretaria do curso de Pós-Graduação. A folha deve conter a **data de aprovação**, estar **sem assinaturas** e em formato **PDF**.

Dedico este trabalho a todos os professores, pesquisadores e cientistas do Brasil. A ciência feita no Brasil é de enorme qualidade e deve ser valorizada. Chegou a hora de nos orgulharmos e estampar nos aviões da EMBRAER a frase: feito no Brasil.

AGRADECIMENTOS

Eu gostaria de agradecer a todos aqueles que me ajudaram nessa viagem que durou 10 anos. Passei por duas graduações e conheci bastante gente. Gostaria de agradecer em particular ao professor José Américo que me mostrou que ensinar é como contar uma história bonita e empolgante. Os aplausos à sua última aula mostraram que ensinar é uma arte que emociona. Gostaria de agradecer ao professor George Darmiton que mostrou o quão maravilhoso, apesar de cansativo, é ser um pesquisador. Sua calma é contagiante, suas aulas são inspiradoras. Credito muito minha crescente vontade de ser pesquisador a seus ensinamentos.

Também gostaria de agradecer a minha família que me apoiou durante a mudança de curso. Foi um dos momentos mais difíceis que enfrentei até hoje na minha vida. Sem o apoio deles eu, muito provavelmente, não conseguiria. Em particular gostaria de agradecer muito a minha mãe que, por muitas vezes me viu jogado no chão pensando no que fazer da vida, me chamava para conversar. Gostaria de agradecer a todos os meus amigos que me deram um enorme apoio nessa transição. Vocês não fazem ideia do quanto eu os agradeço por isso. Vocês são pilares que sustentam este trabalho.

Queria agradecer também a minha noiva que esteve comigo durante a escrita e pesquisa deste trabalho. Escrever este trabalho com você ao meu lado, me dando sua opinião, conversando sobre o tema com certeza fez tudo mais simples, principalmente durante essa pandemia. Te amo.

E finalmente hoje eu aprendo que não devo me arrepender das decisões que fiz e sim daquelas que não tive coragem de fazer.

RESUMO

Ao decorrer dos últimos anos a quantidade de dados produzidos pela humanidade vem atingindo níveis cada vez maiores. E ao mesmo tempo que estes dados trazem informações úteis para algoritmos de aprendizagem de máquina, também trazem ruídos que prejudicariam a formação de um modelo. Para estes tipos de cenários procedimentos de redução de informação e remoção de dados ruidosos, como seleção de instâncias e atributos, contribuiriam para formação de um modelo de classificador com maior otimização. Em face disso, o objetivo deste trabalho é apresentar os impactos em acurácia e em redução de dados ao se combinar um seletor de atributos à um seletor de instâncias nos dados apresentados a um algoritmo de aprendizagem baseado em instância. A partir dos resultados será possível então entender qual sequência de redução mais contribui para cada parâmetro estudado.

Palavras-chaves: Seleção de atributos, Seleção de instâncias, Classificadores, Aprendizagem de Máquina, Atributos, Algoritmos de aprendizagem

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KNN	*k-Nearest Neighbour Classifiers*
RBA	*Relief-based algorithms*
RIS	*Ranking-based Instance Selection*

SUMÁRIO

1 INTRODUÇÃO

Com o aumento exponencial da quantidade de dados produzidos pela humanidade, aumenta também a necessidade de possuir ferramentas de processamento cada vez mais robustas para pré-processar estes dados e melhorar a aprendizagem dos modelos. Dados ruidosos precisam ser retirados do aprendizado para que não gerem resultados inesperados. Alguns mecanismos já foram propostos por diversos pesquisadores da área: redução de dimensionalidade, redução de instâncias e redução de atributos. Estes mecanismos evitariam, no geral, a necessidade contínua de mais recursos computacionais para gerar um resultado, além de eliminarem pontos de dados que prejudicariam os algoritmos de aprendizagem de máquina. Miao e Niu (2016) relatam que a seleção de atributos geralmente pode levar a um melhor desempenho na aprendizagem, isto é, maiores acurácias, menores custos computacionais e uma melhor interpretação de modelos. O mesmo raciocínio se aplica à seleção de instâncias. Como demonstrado por Cavalcanti e Soares (2020), é possível selecionar um sub-conjunto de instâncias a partir de conjunto de treinamento obtendo um conjunto de dados que, oferecidos a um classificador, melhorariam seu desempenho. Estas duas técnicas, selecionar atributos e instâncias, teriam papel crucial desempenho de classificação Tsai et al. (2021).

Neste trabalho exploraremos os resultados da combinação entre o algoritmo de seleção de atributos, RelieF (KONONENKO, 1994), e o algoritmo de seleção de instâncias, *Ranking-based Instance Selection* (RIS) (CAVALCANTI; SOARES, 2020). Em particular faremos uma análise comparativa com os resultados encontrados por Cavalcanti e Soares (2020). Será observado que para alguns casos abordados por Cavalcanti e Soares (2020), a combinação do RelieF ao RIS se mostrou positiva, melhorando os resultados de classificação mesmo utilizando uma menor quantidade de dados.

2 OBJETIVO

O objetivo geral deste trabalho é avaliar o uso de algoritmos de seleção de instâncias em conjunto com algoritmos de seleção de atributos. Em particular utilizaremos o RelieF (KONONENKO, 1994) para seleção de atributos e o RIS (CAVALCANTI; SOARES, 2020) para seleção de instâncias. Deste objetivo geral podemos expandir e determinar que para o conclusão deste trabalho será necessário atingir os seguintes objetivos específicos:

- Apresentar a existência de benefícios à acurácia e redução de dados na combinação do redutor de atributos ao redutor de instâncias apresentado por Cavalcanti e Soares (2020).
- Demonstrar qual seria a melhor sequência de redução por base dados e realizar comparações através da acurácia, redução do número de instâncias e atributos. Este estudo será realizado permutando cada algoritmo em dois experimentos diferentes. No primeiro será realizado o RIS para depois executar o RelieF. E um segundo experimento realizando o inverso. Estes dois experimentos terão como base de comparação a execução do RIS isoladamente e utilizarão o *k-Nearest Neighbour Classifiers* (KNN)(CUNNINGHAM; DELANY, 2007) para gerar as métricas de acurácia.

A contribuição mais relevante deste trabalho será responder se há benefícios, para as bases de dados apresentadas, combinar redutores de instâncias e atributos aos dados apresentados a algoritmos de aprendizagem baseados em instâncias.

3 REVISÃO DA LITERATURA

Alguns trabalhos que combinam seleção de instâncias e a seleção de atributos já foram propostos. Tsai et al. (2021) apresentam esta combinação de seletores de dados adicionando ao procedimento a combinação de classificadores. Nesta abordagem, além dos seletores de dados, vários classificadores são agrupados e seus resultados combinados obtêm do resultado final. Entretanto, Tsai et al. (2021) focam na predição de dificuldades financeiras que podem atingir as instituições. Ou seja possuem como base um único tipo de problema, ao contrário da proposta deste trabalho que procura assim como Cavalcanti e Soares (2020) aplicar a técnica de seleção em múltiplos tipos de base de dados.

Uma outra proposta apresentada por Fragoudis, Meretakis e Likothanassis (2002) era combinar seleção de atributos e instâncias para classificação de textos. Entretanto, seguindo a mesma abordagem do artigo proposto por Tsai et al. (2021), o foco do artigo era avaliar o desempenho de classificação apenas em bases de dados textuais e não em múltiplos tipos de bases de dados.

Um outro trabalho proposto por Tang e Liu (2013) foca em diminuir a quantidade massiva de dados produzidos pelas redes sociais em matrizes de dados úteis. A técnica combinava a utilização de teorias de correlação social e seletores de instâncias e atributos para minerar os dados.

Como é possível observar até então já existem trabalhos propostos que se beneficiam do artifício de combinar seletores de atributos a seletores de instâncias. Entretanto, todos estes trabalhos otimizam os seletores de forma que o resultado obtido seja o melhor possível para um tipo pré-determinado de problema. A proposta aqui apresentada visa apresentar resultados em bases de dados dos mais diversos tamanhos e características e que demonstrará, através dos resultados, se sua aplicabilidade traz benefícios ou não.

4 CONCEITOS BÁSICOS

4.1 SELEÇÃO DE ATRIBUTOS

Existem vários *surveys* (MIAO; NIU, 2016)(LI et al., 2017)(CHANDRASHEKAR; SAHIN, 2014) que avaliam qual seletor de atributos é mais indicado para cada tipo de massa de dados. Alguns inclusive propõem formas selecionar o melhor seletor para cada tipo de bases de dados. Como demonstrado por Li et al. (2017) é possível separar os tipos de seletores de atributos por tipos de bases de dados. Os autores definem estes tipos de base como: convencionais, estruturados, heterogêneos e *streaming*.

Apesar de existirem diferentes métodos de seleção de atributos para diferentes tipos de dados, não é possível afirmar que exista um método que seria o melhor seletor e que teria o desempenho ótimo para todos os tipos de bases de dados (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2012). Esta afirmação seria, no mínimo, otimista. Outra forma de classificar os seletores de atributos seria sobre como os modelos são seleção são criados. Podemos dividir eles em: Filtradores, *Wrapper* e *Embedded* (CHANDRASHEKAR; SAHIN, 2014)(URBANOWICZ et al., 2018). Em particular, para o caso dos Filtradores, uma grande vantagem é sua portabilidade. Isto significa que os atributos escolhidos podem ser passados para qualquer algoritmo de modelagem. Por outro lado os métodos do tipo *Wrapper* necessitam de um nova fase de treinamento pós-seleção de atributos. Já os métodos do tipo *Embedded*, que realizam a fase de seleção durante o processo de modelagem, tendem a ser melhores que os do tipo *Wrapper*, mas ainda mais custosos em releção a custo computacional que os Filtradores. Entretanto, ao contrário da maioria dos métodos Filtradores, os tipos *Wrapper* e *Embedded* conseguem capturar interações e relações entre os atributos (URBANOWICZ et al., 2018).

Baseado neste entendimento e que seria necessário um método seletor que fosse veloz em entregar resultados e que também possuísse uma capacidade generalização, nos seletores analisados, ficou claro que o RelieF (KONONENKO, 1994)(URBANOWICZ et al., 2018) seria um método de seleção de atributos mais indicado. Outro fator que também levou a escolha do RefieF foi que ainda que preserve as características de métodos filtradores, o método RelieF de seleção de atributos é capaz de detectar dependências entre os atributos, sendo o único método filtrador capaz disso (URBANOWICZ et al., 2018).

4.2 RELIEF

O algoritmo RelieF é o *Relief-based algorithms* (RBA) mais utilizado entre todas as suas variações (URBANOWICZ et al., 2018). O objetivo desta seção não será detalhar o processo de seleção do RelieF, já que o mesmo foi bem exposto em outras publicações(ROBNIK-SIKONJA; KONONENKO, 2003)(KONONENKO; ROBNIK-SIKONJA; POMPE, 2000)(URBANOWICZ et al., 2018), mas será entregar uma visão geral do processo de seleção. Como definido, RelieF sendo uma RBA, é uma variação do algoritmo original do ReliefA. Em particular é a sexta (A-F) variação. Sua principal diferença com relação ao conceito base(e.g.: versão A) é a capacidade de lidar com problemas multi-classes e não somente com problemas binários. Entretanto existe outra capacidade incluída, ainda na quarta versão (D) do Relief, que foi lidar com dados nulos. Em observância a estas características e dado que algoritmos filtradores teriam, em princípio, menor custo computacional o RelieF seria o seletor mais indicado para uma proposta inicial que combina esforços da seleção de atributos à um seletor de instâncias.

De forma geral, e como detalhado por (URBANOWICZ et al., 2018), podemos demonstrar o Relief base como o seguinte pseudocódigo:

Algoritmo: **Pseudocódigo do Relief base**

Entrada: n: número de instâncias;

a: número de atributos;

m: número de instâncias aleatórias para treinamento

Retorno: W: vetor com os pesos estimando a relevância de cada atributo.

1 $W[A] \leftarrow$0.0

2 para cada $instancia$ em m faça

3 $R_i \leftarrow instanciaAleatoria()$

4 $H \leftarrow instanciaMaisProximaMesmaClasse()$

5 $M \leftarrow instanciaMaisProximaOutraClasse()$

6 para cada $atributo$ em a faça

7 $W[A] \leftarrow W[A] - diff(A, R_i, H)/m + diff(A, R_i, M)/m$

8 retorne W

Em que a função $diff$ é definida para valores discretos como:

$$diff(A, I_1, I_2) = \begin{cases} 0, & \text{se } valor(A, I_1) = valor(A, I_2) \\ 1, & \text{caso contrário} \end{cases} \tag{4.1}$$

E em casos de valores númericos a função $diff$ é definida por:

$$diff(A, I_1, I_2) = \frac{|valor(A, I_1) - valor(A, I_2)|}{max(A) - min(A)} \tag{4.2}$$

Onde, $valor(A, I_n)$ é o valor do atributo A na instância I_n. E max e min são os valores máximos e mínimos sobre todo o conjunto de instâncias.

Desta forma o algoritmo consegue capturar características geométricas entre as instâncias da base. O que é possível entender deste algoritmo é que ele busca maximizar o peso daqueles atributos que mais distinguem uma dada instância de outra. Quanto maior for o peso dado ao atributo maior será sua relevância para distinguir duas instâncias. Esse peso é chamado de *RelieF Score*. Vale salientar que para nossos testes o número de instâncias aleatórias de teste selecionadas foi o padrão considerado pela biblioteca, ou seja todas as instâncias da base apresentada. Destaco que este é um parâmetro que pode ser alterado.

Cada atributo receberá portanto uma pontuação a partir de sua relevância. Quanto maior o *RelieF Score* maior será a relevância do atributo na sua capacidade de diferenciar duas instâncias de classes distintas. Porém para selecionar os atributos é necessário haver um *threshold* que limite quantos atributos serão selecionados baseados em cada *RelieF Score* (KONONENKO, 1994). Entretanto, Urbanowicz et al. (2018) mostra que de forma prática é preferível definir um número fixo de atributos a serem selecionados baseado no poder computacional disponível. Como havia 22 bases de dados para serem utilizadas, neste trabalho, se fixou o número de atributos selecionados como sendo a metade do número de atributos presentes no conjunto de dados original como forma de simplificar o método, conforme demonstrado no pseudocódigo abaixo.

Algoritmo: **Selecionando atributos com RelieF**

Entrada: X: instâncias da base; y: classes de cada instância

Retorno: Índices dos atributos selecionados

1 $indicesSelecionados \leftarrow \{\}$

2 $relieFScores \leftarrow relieF(X, y)$

3 $numeroMaxAtributosReduzidos \leftarrow metadeDoTamanho(relieFScores)$

4 enquanto $tamanho(indicesSelecionados) < numeroMaxDeAtributosReduzidos$ faça

5 $\quad atributoMaisRelevante \leftarrow selecioneAtributoMaisRelevante(relieFScores)$

6 $\quad indicesSelecionados \leftarrow indicesSelecionados\ U\ atributoMaisRelevante$

7 $\quad removaAtributo(relieFScores, atributo)$

8 retorne $indicesSelecionados$

4.3 SELEÇÃO DE INSTÂNCIAS

Da mesma forma que mostrada na seção sobre seleção de atributos, a seleção de instâncias, ou redução de protótipos, tem como objetivo otimizar a base de dados sem comprometer a qualidade da informação do conjunto. Isto é, os métodos de redução de instâncias buscam representar uma base de dados completa a partir de um conjunto menor de protótipos. Existem algumas formas de diferenciar os tipos de seleção de instâncias. Uma delas é separar em 3 tipos de métodos: *Condensation*, *Edition* e *Hybrid*(KIM; OOMMEN, 2002).

Os algoritmos do tipo *Edition* buscam eliminar ruídos e *outliers* sem comprometer uma alta taxa de redução do conjunto de treinamento. Já os métodos do tipo *Condensation* focam em reduzir a base de dados, mantendo instâncias de "borda" e removendo instâncias que estão mais concentradas no centro da nuvem de dados. O tipo *Hybrid* busca combinar os dois métodos anteriores. Para tal, o método *Hybrid* seleciona instâncias que melhor representam a massa de dados aplicada, independente de sua localização na nuvem de dados(CAVALCANTI; SOARES, 2020). A importância deste tipo de redução é que a partir de uma quantidade de dados obter um conjunto de treinamento com menor ruído e com instâncias que consigam, durante a fase de treinamento, otimizar os modelos afim de obter resultados melhores para as abordagens propostas. Considerando a importância deste tipo de seleção de dados Cavalcanti e Soares (2020) propuseram mais uma variação: RIS.

4.4 RIS

O RIS é uma técnica de seleção de protótipos que pode ser definida em duas fases: ranqueamento e seleção. Um fato que reforçou a escolha do RelieF para se associado ao RIS, é de que o seletor de instâncias também é atribuído uma relevância a cada instância. E quanto maior for o *score* de relevância mais importante é a instância. Da mesma forma que a seção sobre o RelieF, o objetivo aqui é somente apresentar uma visão geral do RIS já que o mesmo já foi detalhado bem por Cavalcanti e Soares (2020).

Na fase de ranqueamento cada instância possuirá um *score* associado. Este *score* será maior se a instância de teste possuir outras instâncias de mesma classe ao seu redor na nuvem de dados. Por outro lado, instâncias de borda e *outliers* possuirão *scores* menores. O *score* é determinado da seguinte forma:

$$s_i = \sum_{j=1:j\neq i}^{m} \alpha(x_i, x_j) * sm(x_i, x_j, X) \tag{4.3}$$

Onde os parâmetros α e sm são calculados da seguinte forma:

$$\alpha(x_i, x_j) = \begin{cases} 1, & \text{se } classe(x_i) = classe(x_j) \\ -1, & \text{caso contrário} \end{cases} \tag{4.4}$$

$$sm(x_i, x_j, X) = \frac{exp(-d(x_i, x_j))}{\sum_{k=1}^{m} exp(-d(x_{xi}, x_k))} \tag{4.5}$$

A função α define o sinal do ajuste. Em poucas palavras, se x_i e x_j tem a mesma classe, o *score* de de x_i deve ser aumentado. Caso contrario, o *score* deve ser subtraído. Já a função sm calcula o valor absoluto do *score* de x_i dado um x_j a partir da distância entre os dois. Esta estratégia considera que instâncias que estiverem mais próximas da instância de teste tenham maiores contribuições na função sm do que instâncias que estão mais distantes. Após a avaliação de todo o conjunto de instâncias, os *scores* são normalizados entre 0 e 1.

$$scaling(s_i, S) = \frac{s_i - min(S)}{max(S) - min(S)} \tag{4.6}$$

Na fase de seleção o RIS busca remover instâncias que sejam redundantes e indesejadas. A remoção é feita com base em um *threshold* t que remove as instâncias após o processo

de $scaling$. Cavalcanti e Soares (2020) propuseram dois conceitos para selecionar as melhores instâncias para cada base de dados: *radius* e *relevant instance*. Estes dois conceitos são definidos da seguinte forma:

- ***Radius***: O *radius* de uma instância x, $radius(x)$, é dado pelo raio da maior hiperesfera centrada em x contendo somente instâncias da mesma classe de x.
- ***Relevant Instance***: Uma instância $x_i \in X$ é considerada relevante pelo RIS se não existir uma instância $x_r \in X$ de forma que:
  - $class(x_r) = class(x_i)$
  - $d(x_i, x_r) \leq radius(x_r); i \neq r$

De posse dos conceitos de *score, radius* e *Relevant Instance* o RIS agora possui informação suficiente para selecionar as instâncias que melhor representam o conjunto de dados. Este é o RIS1.

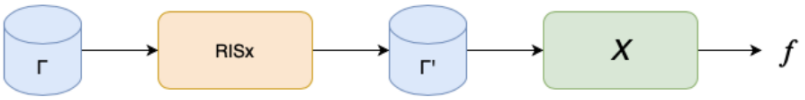
O processo de seleção se inicia escolhendo a instância (x_i) que possui o mais alto *score*. Esta instância é então inserida ao conjunto R que será entregue como resposta do algoritmo. Uma nova instância (x_j), a que possui o segundo maior *score*, é avaliada. Se a classe de x_i for diferente da classe de x_j, esta nova instância será inserida no conjunto de resposta R. Caso contrário x_j será selecionada se, e somente se estiver fora do raio de alcance da hiperesfera de x_i. Esta estrategia garante selecionar a melhor instância por "nuvem" por classe (CAVALCANTI; SOARES, 2020). Cavalcanti e Soares (2020) também apresentam em seu trabalho outras duas variações do RIS: RIS2 e RIS3.

O RIS2 tem como objetivo evitar, durante o $scaling$ apresentado na Equação 4.6, que os *scores* de uma classe influenciem no *score* de outra. A proposta para o RIS2 é de que o processo de $scaling$ seja feito por classe. Esta operação evitaria que classes de alto *score*, que seriam mais suscetíveis a permanecerem na seleção, influenciem nas classes com baixo *score* que tenderiam a ser removidas no processo.

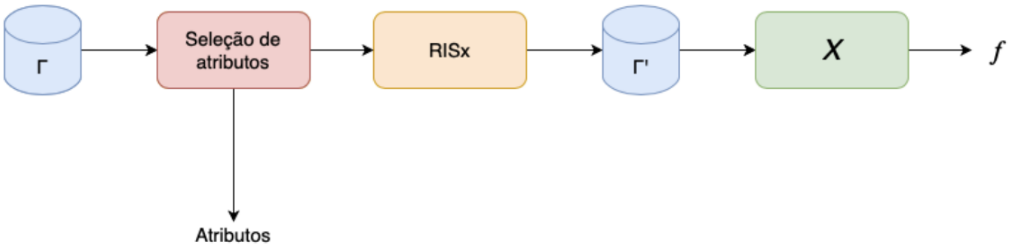
No RIS3, os autores propõem uma mudança no processo de geração das áreas de cobertura das hiperesferas. Durante esta geração é possível que duas instâncias de mesma classe tenha suas hiperesferas sobrepostas. O que Cavalcanti e Soares (2020) propõem é que as áreas de cobertura sejam geradas somente após a eliminação de instâncias que possuírem o *score* menor que um *threshold* t.

5 METODOLOGIA PROPOSTA

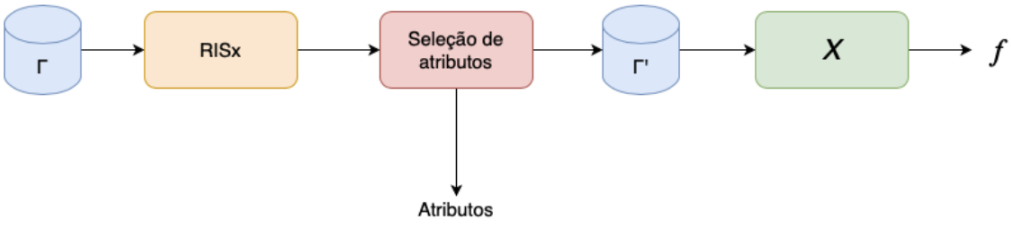
A metodologia proposta possui duas fases: treinamento e teste. Durante o treinamento, o conjunto de treino pode ser reduzido de 3 formas distintas. Inicialmente o conjunto é reduzido somente em instâncias seguindo a proposta de Cavalcanti e Soares (2020), como mostra a 1(a). Em uma segunda abordagem se reduziu primeiro os atributos, utilizando o RelieF, para depois reduzir as instâncias, utilizando o RIS, conforme a Figura 1(b). A terceira forma de redução seria aplicar o inverso, isto é, reduzir o conjunto de instâncias, utilizando o RIS, para depois reduzir o conjunto de atributos, utilizando o RelieF, como mostrado na Figura 1(c). Esta permutação entre seleção de atributos e instâncias é importante para realizar uma comparação dos resultados quando o RIS recebe o conjunto completo de treinamento (Γ) ou quando recebe os dados já reduzidos em atributos. Dessa forma é possível compararmos o resultados deste trabalho aos resultados produzidos por Cavalcanti e Soares (2020).



(a) – Procedimento de treinamento - Somente RIS



(b) – Procedimento de treinamento - FS $\rightarrow RIS$



(c) – Procedimento de treinamento - RIS $\rightarrow FS$

Figura 1 – Procedimentos de treinamento. Γ: Conjunto de treinamento, f: Classificador, Γ': Conjunto de treinamento reduzido, X: Fase de treinamento

Após a etapa final de redução, seja ela redução de instâncias ou redução de atributos, é

obtido um conjunto de dados que pode ser utilizado para treinamento de um classificador f. Este classificador deve ser capaz de, dado uma instância de teste X_q, realizar a predição sua classe.

Algoritmo: **Reduzindo os atributos do conjunto de teste**

Entrada: X: instâncias da base; y: classes de cada instância

Retorno: Conjunto de teste reduzido em atributos

1 $atributosSelecionados \leftarrow selecaoDeAtributos(X, y)$

2 $conjuntoTeste \leftarrow filtro(conjuntoTeste, atributosSelecionados)$

3 retorne $conjuntoTeste$

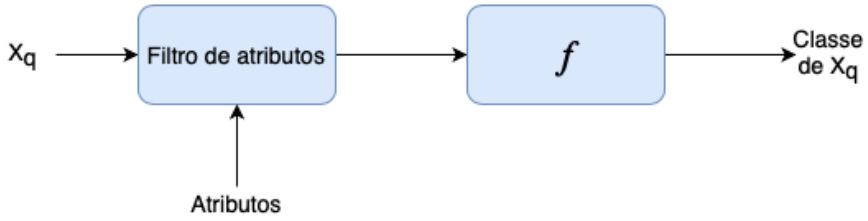


Figura 2 – Procedimento de classificação - RIS

Vale destacar que o conjunto de atributos selecionados durante o treinamento é extraído para ser posteriormente utilizado para filtrar os atributos no conjunto de teste durante o processo de classificação. Isso é feito passando os índices dos parâmetros selecionados durante a fase de treinamento para as instâncias de teste através de um filtro. Ao passar os índices, a instância de teste (X_q) resultante possuirá somente os atributos selecionados durante o treino conforme mostrado na Figura 2. É importante destacar que o conjunto durante a seleção de atributos pode ser todo conjunto Γ, como mostrado na Figura 1(b), ou pode ser somente as instâncias já selecionadas pelo RIS quando este é executado de forma prévia à seleção de atributos conforme a Figura 1(c).

Vale lembrar também que o módulo "Seleção de atributos" pode assumir qualquer algoritmo de redução de atributos. Como dito na seção 4.1, o uso do RelieF foi justificado por ser o método que melhor captura as relações entre os atributos e ao mesmo tempo possuir a rapidez e a modularidade de métodos do tipo Filtradores.

6 PROTOCOLO EXPERIMENTAL

6.1 BASES DE DADOS

Inicialmente, para cada uma das 22 bases de dados presentes na Tabela 1 é aplicado o procedimento de *10-fold cross-validation* em que sempre se garante, *a priori*, a probabilidade de aparecimento de cada classe. Um pré-processamento também foi aplicado normalizando os dados entre 0 e 1. Para cada um dos *folds* é realizado uma separação nos dados entre conjunto de teste e treino. Para o conjunto de validação são utilizados os dados do conjunto de treinamento (CAVALCANTI; SOARES, 2020).

Tabela 1 – Características das bases de dados utilizadas

Base de dados	**Número de instâncias**	**Número de atributos**	**Classes**
appendicitis	106	7	2
balance	625	4	3
bupa	345	6	2
coil2000	9822	85	2
contraceptive	1473	9	2
haberman	306	3	2
hayes	160	4	2
heart	270	13	2
Ionosphere	351	33	2
led7digit	500	7	10
Marketing	8993	13	9
monk-2	432	6	2
libras	360	90	15
pima	768	8	2
satimage	6435	36	7
segment	2310	19	7
titanic	2201	3	2
vowel	990	13	11
wine	178	13	3
red-wine	1599	11	11
white-wine	4898	11	11
yeast	1484	8	10

Para cada um dos experimentos executados neste trabalho foi verificado não somente

aquele possuía a melhor acurácia dentre todos, mas também aquele que, estatisticamente, também era melhor que os demais. Para realização deste teste estatístico foi utilizado o teste t das médias.

Visto que este trabalho tem como objetivo principal o entendimento do impacto da seleção de atributos adicionada ao estudo feito por Cavalcanti e Soares (2020), foram utilizadas, para melhor comparação, as mesmas bases de dados que foram utilizadas pelos autores conforme pode ser visto na Tabela 1. Como pode se observar as bases possuem as mais diversas características. Problemas binários à multi-classes. De centenas de instâncias à quase dezenas de milhares.

Um ponto a se destacar neste trabalho e de que devido a limitações técnicas que serão abordadas na seção 6.3, não foi possível realizar testes com as bases *adult* e *connect-4* como fez Cavalcanti e Soares (2020).

6.2 MÉTRICAS

Para cada um dos bancos de dados utilizados neste trabalho foram registrados métricas percentuais de avaliação para cada solução proposta. São elas: acurácia, redução do número de instâncias, redução no número de atributos e tamanho da matriz de dados. Para cada uma dessas métricas, com exceção do tamanho da matriz de dados, foram registradas suas respectivas médias e desvio-padrões com base nos melhores resultados do conjunto de validação por *fold*. Isto significa que, para cada *fold*, de cada base de dados, era escolhido o *threshold* *t* do RIS que obteve o melhor resultado de acurácia no conjunto de validação e para este resultado são recolhidas as métricas mencionadas anteriormente no conjunto de teste. Feito isso do *fold* 1 ao 10 se retira uma média entre todos os *folds*. Os valores estão presentes no anexo A. Os dados compilados seguem o algoritmo abaixo.

Algoritmo: **Compilando os resultados**
Entrada: $resultadosFold$: Resultados de todos os *folds* para todos os *thresholds* t
Retorno: $acuracia$: Média de acurácia dos resultados,
$reducaoInstancia$: Média de redução do número de instâncias,
$reducaoAtributos$: Média de redução do número de atributos,
$tamanhoMatriz$: Tamanho médio da matriz de dados pós-redução

1 $melhoresResultados \leftarrow \{\}$
2 para cada $fold$ em $resultadosFold$ faça
3 $\quad metricasMelhorAcuracia \leftarrow melhorMetricaValidacao(fold)$
4 $\quad metricasMelhorAcuracia\ U\ melhoresResultados$
5 $acuracia \leftarrow mediaAcuracias(melhoresResultados)$
6 $reducaoInstancia \leftarrow mediaReducaoInstancias(melhoresResultados)$
7 $reducaoAtributos \leftarrow mediaReducaoAtributos(melhoresResultados)$
8 $tamanhoMatriz \leftarrow mediaTamanhoMatriz(melhoresResultados)$
9 retorne $(accuracy, reducaoInstancia, reducaoAtributos, tamanhoMatriz)$

Em particular, o tamanho da matriz de dados merece uma atenção especial. O tamanho da matriz de dados (instâncias *versus* atributos) representa a quantidade de pontos de dados que sobraram após os métodos de redução. De forma prática, se ao início uma base de dados tinha 100 instâncias e 100 atributos, esta irá possuir 10000 pontos de dados. Se após os métodos de redução (RIS e RelieF) suas instâncias e atributos foram reduzidas a 40 e 60, respectivamente, seu tamanho agora será de 2400 pontos de dados, ou 24% do tamanho original.

6.3 PODER COMPUTACIONAL

É sabido que quanto mais pontos de dados for preciso processar, mais poder computacional é necessário. Duas bases de dados utilizadas por Cavalcanti e Soares (2020) possuíam uma quantidade de dados que ultrapassava o limite de processamento da máquina utilizada para realização dos testes: *adult* e *connect-4*. Esta circunstância levava o processo a abortar durante até mesmo a primeira execução, impedindo a obtenção dos resultados. Para este trabalho testes com bases maiores, que de fato trariam um resultado interessante, se tornou um fator limitante. As especificações técnicas da máquina utilizada para os testes estão descritas na Tabela 2. Há, entretanto duas abordagens que podem ser feitas para mitigar este fator limitante que já poderiam ser incluídas em melhorias futuras:

Tabela 2 – Especificações técnicas da máquina utilizada

CPU	2,3 GHz 8-Core Intel Core i9
GPU	AMD Radeon Pro 5500M 4 GB
Memória	16 GB 2667 MHz DDR4

- A mais direta seria aumentar o poder de processamento. Quanto mais recursos disponíveis, maior será *threshold* no qual o algoritmo conseguirá processar. Entretanto além de ser uma solução custosa, do ponto de vista monetário, com o aumento do tamanho das bases será necessário um novo investimento em processamento.

- Outra abordagem seria otimizar o algoritmo. Atualmente o código está escrito na linguagem *Python*. Como é sabido, apesar de ser uma linguagem de fácil leitura, possui uma menor velocidade de execução quando comparada a linguagens de mais baixo nível(e.g.: C, C++). Uma abordagem otimizando o código do RIS + FS em linguagem de mais baixo nível, utilizando bibliotecas como a Cython, sem dúvidas melhoraria seu desempenho.

7 RESULTADOS

Para a analise dos resultados é utilizado o teste t das médias. O teste valida, para cada melhor resultado, se o mesmo é estatisticamente melhor que os outros dois. A escolha é justificada já que o teste t valida dois conjuntos de amostras que não são dependentes entre si. Para este teste estatístico é considerado o nível de significância $\alpha = 5\%$. Os resultados que foram estatisticamente melhores estão representados nos gráficos das Figuras 3, 4 e 5 como a barra laranja. A barra azul é quantidade de vezes que o método avaliado foi melhor em valores absolutos. Nos experimentos é possível observar um padrão de melhor desempenho em acurácia para as bases de dados quando utilizamos o RelieF combinado com o RIS (barra à direita), nesta ordem. Com exceção do RIS3, tanto para o RIS1, assim como para o RIS2 é possível observar que o este método obteve os melhores resultados.

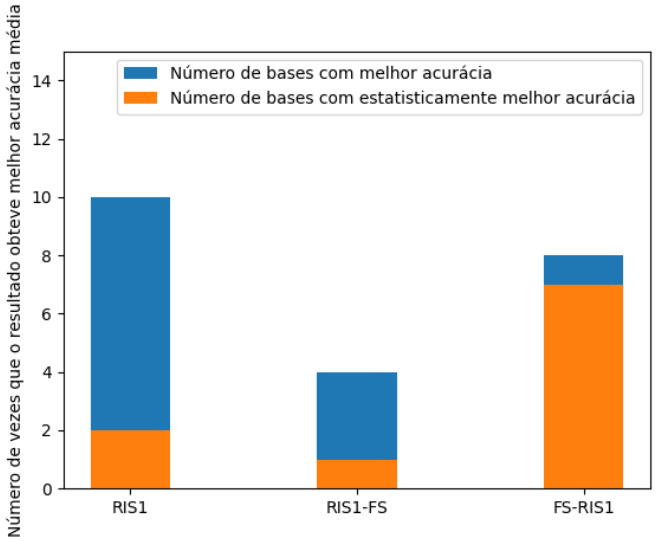


Figura 3 – Quantidade de bases de dados com melhor resultado X Método proposto

Como podemos ver na Figura 3, o método original do RIS1 é considerado melhor no geral do que se o utilizarmos combinado ao RelieF. Entretanto a quantidade de bases que obteve relevância estatística do resultado é baixa dado os resultados dos experimentos. Já os experimentos que envolveram seleção de atributos seguidos do RIS1 possui os melhores resultados confirmados estatisticamente.

O mesmo comportamento observado no RIS1, foi observado no RIS2. A combinação do RelieF com o RIS obteve a melhor acurácia em 8 bases de dados, dos quais todos obtiveram a validação estatística do teste t, como é possível ver na Figura 4.

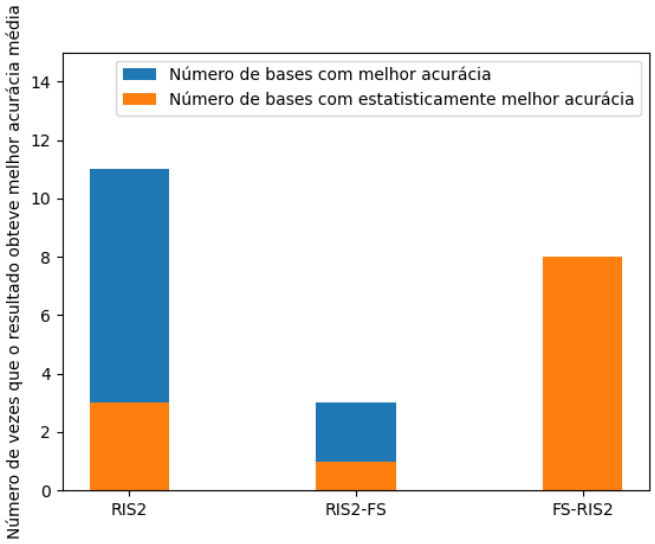


Figura 4 – Quantidade de bases de dados com melhor resultado X Método proposto

No caso do RIS3, seguindo o padrão observado no RIS1 e no RIS2 utilizar somente a seleção de instâncias obteve os melhores resultados, no geral. Porém, ao contrário do RIS1 e RIS2, o RIS3 obteve também os melhores resultados validados estatisticamente.

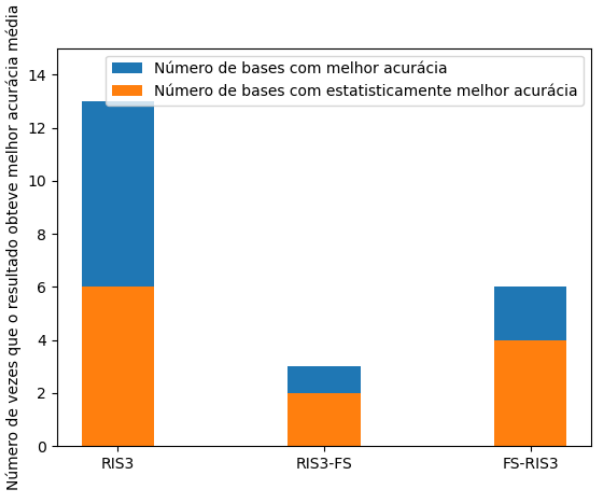


Figura 5 – Quantidade de bases de dados com melhor resultado X Método proposto

É interessante notar que o mesmo comportamento relacionado às instâncias após redução obtido por Cavalcanti e Soares (2020) foi também observado neste trabalho. Conforme podemos ver nos resultados do Anexo A, o RIS1 obteve a maiores médias de acurácia, enquanto obteve as menores médias de redução de instâncias. Enquanto isso, o RIS3 obteve a menores médias de acurácia, enquanto obteve as maiores médias de redução de instâncias. Como é possível observar na Tabela 3, mais da metade das bases de dados obtiveram os melhores resultados utilizando o RIS1 de alguma forma. Também, através da Tabela 3, é possível cons-

tatar que 13 das 22 bases testadas obtiveram os melhores resultados utilizando a seleção de atributos, seja como primeiro ou segundo passo de redução.

Tabela 3 – Métodos que obtiveram as melhores acurácias por base de dados.

Base de dados	**Método com maior acurácia**
appendicitis	RIS3 - FS
balance	RIS1*
bupa	RIS2 - FS
coil2000	FS-RIS1
contraceptive	FS-RIS1
haberman	FS-RIS3*
hayes	RIS1
heart	RIS2-FS*
Ionosphere	RIS3*
led7digit	RIS3*
marketing	RIS1-FS*
monk-2	RIS2-FS*
libras	RIS1*
pima	FS-RIS3
satimage	RIS1*
segment	FS-RIS1*
titanic	RIS1*
vowel	RIS1*
wine	FS-RIS1
red-wine	FS-RIS1
white-wine	FS-RIS2
yeast	RIS2*

* Não confirmado pelo teste estatístico.

Outro dado interessante de notar é a redução da matriz de dados representado pelas Tabelas 7, 11 e 15 do Anexo A. Utilizando a base *contraceptive* como exemplo, realizando somente a redução de instâncias com o RIS1 obteve-se uma massa de dados 73,86% do total. Quando combinado o RelieF ao RIS1, nesta ordem, obteve-se um conjunto de dados 38,52% do total. E mesmo com quase metade dos dados quando comparado a redução utilizando somente o RIS1, a combinação dos seletores de instâncias e atributos resultou em uma acurácia maior do que se utilizado somente o RIS1. Isso demonstra que a combinação da redução de atributos ao RIS trouxe melhorias ao modelo de classificador gerado.

8 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou os resultados da combinação de um seletor de atributos, o RelieF, a um seletor de instâncias, o RIS. Foi possível observar em nossos experimentos que não somente a redução do número de atributos teve o papel de aumentar a acurácia para alguns testes como, para a maioria das bases de dados testadas, a adição do RelieF ao pré-processamento obteve as maiores acurácias dentre todos os testes. Estes resultados demonstram que existem benefícios em se combinar estes dois algoritmos redutores de dados. Entretanto não foi possível afirmar qual seria a melhor combinação para ser utilizada durante o pré-processamento. Como cada base de dados apresentada possui características especificas e particulares que as diferenciam, é possível concluir que uma única abordagem de seleção seria incapaz de generalizar e obter modelos otimizados para todas as bases de dados apresentadas. Este resultado justifica a abordagem feita pelos trabalhos relacionados de definir previamente um seletor de dados para cada tipo específico de problema.

Trabalhos futuros seriam, a partir do tipo de base de dado apresentado como entrada do algoritmo de aprendizagem, adaptar e selecionar dinamicamente qual seletor de atributo e instância é mais indicado para ser aplicado.

REFERÊNCIAS

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, v. 34, 03 2012.

CAVALCANTI, G. D.; SOARES, R. J. Ranking-based instance selection for pattern classification. *Expert Systems with Applications*, v. 150, p. 113269, 2020. ISSN 0957-4174.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers Electrical Engineering*, v. 40, n. 1, p. 16–28, 2014. ISSN 0045-7906. 40th-year commemorative issue. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0045790613003066>.

CUNNINGHAM, P.; DELANY, S. k-nearest neighbour classifiers. *Mult Classif Syst*, v. 54, 04 2007.

FRAGOUDIS, D.; MERETAKIS, D.; LIKOTHANASSIS, S. Integrating feature and instance selection for text classification. In: . [S.l.: s.n.], 2002. p. 501–506.

KIM, S.-W.; OOMMEN, B. Creative prototype reduction schemes: A taxonomy and ranking. In: . [S.l.: s.n.], 2002. v. 7, p. 6 pp. vol.7–. ISBN 0-7803-7437-1.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: BERGADANO, F.; RAEDT, L. D. (Ed.). *Machine Learning: ECML-94*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994. p. 171–182. ISBN 978-3-540-48365-6.

KONONENKO, I.; ROBNIK-SIKONJA, M.; POMPE, S. Relieff for estimation and discretization of attributes in classification, regression, and ilp problems. 02 2000.

LI, J.; CHENG, K.; WANG, S.; MORSTATTER, F.; TREVINO, R. P.; TANG, J.; LIU, H. Feature selection: A data perspective. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 50, n. 6, dez. 2017. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3136625>.

MIAO, J.; NIU, L. A survey on feature selection. *Procedia Computer Science*, v. 91, p. 919–926, 2016. ISSN 1877-0509. Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016). Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050916313047>.

ROBNIK-SIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.*, Kluwer Academic Publishers, USA, v. 53, n. 1–2, p. 23–69, out. 2003. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1025667309714>.

TANG, J.; LIU, H. Coselect: Feature selection with instance selection for social media data. In: _____. [S.l.: s.n.], 2013. p. 695–703. ISBN 978-1-61197-262-7.

TSAI, C.-F.; SUE, K.-L.; HU, Y.-H.; CHIU, A. Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *Journal of Business Research*, v. 130, p. 200–209, 2021. ISSN 0148-2963. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0148296321001776>.

URBANOWICZ, R. J.; MEEKER, M.; La Cava, W.; OLSON, R. S.; MOORE, J. H. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, v. 85, p. 189–203, 2018. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046418301400>.

APÊNDICE A – TABELAS COM RESULTADOS

Tabela 4 – Media de acerto dos melhores thresholds por fold utilizando o RIS1 (%)

Base de dados	**RIS1**	**RIS1 - FS**	**FS - RIS1**
appendicitis	**80.79 ± 9.11**	78.21 ± 15.02	78.89 ± 8.96
balance	**86.72 ± 2.38**	69.91 ± 4.50	62.18 ± 3.98
bupa	**57.97 ± 5.63**	53.63 ± 8.25	56.54 ± 8.93
*coil*2000	93.78 ± 0.17	93.63 ± 0.44	**93.78 ± 0.24**
contraceptive	46.91 ± 4.41	47.45 ± 3.27	**48.21 ± 3.39**
haberman	64.98 ± 5.40	63.95 ± 10.96	**66.01 ± 4.90**
$hayes - roth$	**66.42 ± 12.21**	51.02 ± 10.58	48.69 ± 13.32
heart	74.44 ± 4.21	**79.26 ± 7.44**	73.70 ± 8.68
ionosphere	90.62 ± 4.35	**90.63 ± 5.63**	81.53 ± 6.87
*led*7*digit*	**72.37 ± 9.14**	56.09 ± 9.92	27.03 ± 6.03
marketing	27.63 ± 0.36	**27.87 ± 0.37**	18.37 ± 0.13
$monk - 2$	92.11 ± 3.81	**94.69 ± 3.59**	77.77 ± 3.17
$movement_libras$	**78.11 ± 11.69**	71.89 ± 11.79	65.44 ± 9.63
pima	63.40 ± 8.06	66.27 ± 6.55	**66.27 ± 6.80**
satimage	**26.36 ± 0.42**	25.92 ± 0.49	25.27 ± 0.59
segment	92.12 ± 2.77	87.84 ± 2.28	**92.86 ± 2.40**
titanic	**69.06 ± 0.76**	67.70 ± 0.09	67.70 ± 0.09
vowel	**87.88 ± 5.31**	77.27 ± 6.28	61.72 ± 6.44
wine	93.39 ± 5.40	89.93 ± 6.43	**93.85 ± 6.26**
$winequality - red$	45.83 ± 4.35	43.90 ± 3.58	**47.91 ± 4.68**
$winequality - white$	41.55 ± 3.01	39.08 ± 4.67	**41.81 ± 1.45**
yeast	**41.77 ± 4.39**	32.29 ± 3.65	25.39 ± 3.83
Mean	67.92	64.02	60.04

Tabela 5 – Media de redução de instâncias dos melhores thresholds por fold utilizando o RIS1 (%)

Base de dados	**RIS1**	**RIS1 - FS**	**FS - RIS1**
appendicitis	70.64 ± 2.18	74.72 ± 3.12	$\mathbf{84.38 \pm 7.24}$
balance	76.53 ± 1.25	$\mathbf{99.64 \pm 0.00}$	35.95 ± 2.62
bupa	$\mathbf{43.83 \pm 2.19}$	43.83 ± 2.19	31.65 ± 6.74
*coil*2000	10.85 ± 0.37	10.85 ± 0.37	$\mathbf{33.00 \pm 1.90}$
contraceptive	26.14 ± 0.63	26.14 ± 0.63	$\mathbf{30.66 \pm 1.74}$
haberman	$\mathbf{49.20 \pm 1.49}$	49.20 ± 1.49	28.43 ± 4.68
$hayes - roth$	44.79 ± 2.93	$\mathbf{49.74 \pm 3.62}$	11.23 ± 5.83
heart	50.99 ± 1.97	$\mathbf{57.45 \pm 4.90}$	16.58 ± 4.01
ionosphere	75.50 ± 0.64	$\mathbf{78.54 \pm 2.19}$	39.86 ± 2.51
*led*7*digit*	16.53 ± 0.89	$\mathbf{37.75 \pm 12.92}$	26.20 ± 12.36
marketing	0.16 ± 0.03	$\mathbf{6.87 \pm 10.24}$	0.09 ± 0.04
$monk - 2$	72.79 ± 0.98	$\mathbf{83.33 \pm 6.84}$	72.19 ± 13.34
$movement_libras$	62.28 ± 1.14	$\mathbf{62.75 \pm 1.50}$	52.32 ± 3.19
pima	53.78 ± 1.55	57.16 ± 4.89	$\mathbf{69.14 \pm 3.72}$
satimage	0.04 ± 0.06	0.04 ± 0.06	$\mathbf{0.08 \pm 0.08}$
segment	88.96 ± 0.26	$\mathbf{91.92 \pm 0.78}$	17.79 ± 2.00
titanic	17.61 ± 0.58	38.03 ± 0.13	$\mathbf{42.76 \pm 2.29}$
vowel	$\mathbf{76.90 \pm 0.36}$	76.90 ± 0.36	54.75 ± 1.27
wine	86.64 ± 0.82	$\mathbf{87.39 \pm 1.35}$	55.12 ± 9.70
$winequality - red$	$\mathbf{42.01 \pm 0.71}$	42.01 ± 0.71	25.54 ± 1.46
$winequality - white$	$\mathbf{41.83 \pm 0.36}$	41.83 ± 0.36	22.76 ± 0.70
yeast	$\mathbf{36.27 \pm 0.80}$	36.27 ± 0.80	0.01 ± 0.02
Mean	47.47	52.38	34.11

Tabela 6 – Media de redução de atributos dos melhores thresholds por fold para o teste do RIS1 (%)

Base de dados	**RIS1 - FS**	**FS - RIS1**
$appendicitis$	**42.86 ± 0.00**	42.86 ± 0.00
$balance$	**50.00 ± 0.00**	50.00 ± 0.00
$bupa$	**50.00 ± 0.00**	50.00 ± 0.00
$coil2000$	**49.41 ± 0.00**	49.41 ± 0.00
$contraceptive$	**44.44 ± 0.00**	44.44 ± 0.00
$haberman$	**33.33 ± 0.00**	33.33 ± 0.00
$hayes-roth$	**50.00 ± 0.00**	50.00 ± 0.00
$heart$	**46.15 ± 0.00**	46.15 ± 0.00
$ionosphere$	**48.48 ± 0.00**	48.48 ± 0.00
$led7digit$	**42.86 ± 0.00**	42.86 ± 0.00
$marketing$	**46.15 ± 0.00**	46.15 ± 0.00
$monk-2$	**50.00 ± 0.00**	50.00 ± 0.00
$movement_libras$	**50.00 ± 0.00**	50.00 ± 0.00
$pima$	**50.00 ± 0.00**	50.00 ± 0.00
$satimage$	**50.00 ± 0.00**	50.00 ± 0.00
$segment$	**47.37 ± 0.00**	47.37 ± 0.00
$titanic$	**33.33 ± 0.00**	33.33 ± 0.00
$vowel$	**46.15 ± 0.00**	46.15 ± 0.00
$wine$	**46.15 ± 0.00**	46.15 ± 0.00
$winequality-red$	**45.45 ± 0.00**	45.45 ± 0.00
$winequality-white$	**45.45 ± 0.00**	45.45 ± 0.00
$yeast$	**50.00 ± 0.00**	50.00 ± 0.00
$Mean$	46.26	46.26

Tabela 7 – Tamanho médio percentual do total da matriz de dados após as reduções de instâncias e atributos no RIS1

Base de dados	**Total**	**RIS1**	**RIS1 - FS**	**FS - RIS1**
appendicitis	100%	29.35%	14.44%	8.92%
balance	100%	23.47%	0.18%	32.03%
bupa	100%	56.17%	28.08%	34.17%
*coil*2000	100%	89.15%	45.10%	33.89%
contraceptive	100%	73.86%	41.03%	38.52%
haberman	100%	50.80%	33.87%	47.71%
$hayes - roth$	100%	55.21%	25.14%	44.38%
heart	100%	49.01%	22.91%	44.92%
ionosphere	100%	24.50%	11.06%	30.98%
*led*7*digit*	100%	83.47%	35.57%	42.18%
marketing	100%	99.84%	50.15%	53.80%
$monk - 2$	100%	27.21%	8.33%	13.91%
$movement_libras$	100%	37.72%	18.63%	23.86%
pima	100%	46.22%	21.42%	15.43%
satimage	100%	99.96%	49.98%	49.96%
segment	100%	11.04%	4.25%	43.27%
titanic	100%	82.39%	41.31%	38.16%
vowel	100%	23.10%	12.44%	24.37%
wine	100%	13.36%	6.79%	24.17%
$winequality - red$	100%	57.99%	31.63%	40.61%
$winequality - white$	100%	58.17%	31.73%	42.13%
yeast	100%	63.73%	31.87%	50.00%

Tabela 8 – Media de acerto dos melhores thresholds por fold utilizando o RIS2 (%)

Base de dados	**RIS2**	**RIS2 - FS**	**FS - RIS2**
appendicitis	79.14 ± 7.15	80.88 ± 10.07	$\mathbf{82.53 \pm 14.19}$
balance	$\mathbf{84.17 \pm 4.19}$	64.01 ± 12.20	46.71 ± 15.11
bupa	56.85 ± 3.51	$\mathbf{57.99 \pm 7.25}$	55.39 ± 7.75
*coil*2000	$\mathbf{70.69 \pm 20.11}$	66.23 ± 22.20	56.42 ± 22.87
contraceptive	47.05 ± 4.30	46.30 ± 4.17	$\mathbf{47.80 \pm 2.96}$
haberman	65.65 ± 7.40	61.40 ± 12.55	$\mathbf{67.93 \pm 6.74}$
$hayes - roth$	$\mathbf{63.49 \pm 12.95}$	52.64 ± 14.66	48.33 ± 14.32
heart	77.41 ± 5.35	$\mathbf{80.00 \pm 6.02}$	68.89 ± 10.89
ionosphere	$\mathbf{90.63 \pm 4.73}$	86.91 ± 6.55	80.05 ± 7.11
*led*7*digit*	$\mathbf{67.27 \pm 7.73}$	56.48 ± 8.81	40.79 ± 5.98
marketing	$\mathbf{13.24 \pm 1.30}$	12.14 ± 1.41	12.81 ± 1.42
$monk - 2$	92.60 ± 5.66	$\mathbf{97.23 \pm 1.98}$	97.23 ± 1.98
$movement_libras$	$\mathbf{60.56 \pm 10.50}$	55.89 ± 12.08	57.11 ± 9.90
pima	63.27 ± 7.40	66.53 ± 6.60	$\mathbf{66.92 \pm 6.34}$
satimage	13.33 ± 3.99	13.42 ± 4.03	$\mathbf{14.84 \pm 4.36}$
segment	$\mathbf{91.82 \pm 2.93}$	84.20 ± 3.22	91.56 ± 2.99
titanic	$\mathbf{56.44 \pm 15.21}$	52.98 ± 17.22	35.26 ± 8.85
vowel	$\mathbf{66.06 \pm 6.06}$	64.55 ± 5.83	52.83 ± 6.90
wine	91.62 ± 4.39	91.10 ± 4.22	$\mathbf{93.81 \pm 5.25}$
$winequality - red$	44.78 ± 5.61	43.91 ± 4.16	$\mathbf{46.77 \pm 5.70}$
$winequality - white$	41.81 ± 2.59	40.22 ± 3.48	$\mathbf{42.30 \pm 3.36}$
yeast	$\mathbf{42.39 \pm 3.39}$	31.15 ± 4.90	21.88 ± 9.97
$Mean$	62.74	59.37	55.83

Tabela 9 – Media de redução de instâncias dos melhores thresholds por fold utilizando o RIS2 (%)

Base de dados	**RIS2**	**RIS2 - FS**	**FS - RIS2**
appendicitis	72.42 ± 3.32	72.84 ± 3.72	$\mathbf{86.38 \pm 10.28}$
balance	84.76 ± 2.41	$\mathbf{97.63 \pm 1.51}$	36.93 ± 7.64
bupa	$\mathbf{50.18 \pm 5.31}$	50.18 ± 5.31	42.00 ± 6.24
*coil*2000	25.06 ± 23.94	16.55 ± 0.81	$\mathbf{44.30 \pm 13.59}$
contraceptive	38.36 ± 5.39	38.36 ± 5.39	$\mathbf{42.53 \pm 4.60}$
haberman	58.02 ± 2.39	$\mathbf{63.95 \pm 9.13}$	38.63 ± 6.16
$hayes - roth$	48.42 ± 9.92	$\mathbf{65.51 \pm 13.10}$	53.71 ± 17.99
heart	61.77 ± 4.35	$\mathbf{63.70 \pm 3.26}$	29.18 ± 8.30
ionosphere	74.64 ± 4.29	$\mathbf{75.05 \pm 4.18}$	47.04 ± 5.32
*led*7*digit*	57.76 ± 3.41	$\mathbf{90.02 \pm 11.83}$	67.17 ± 25.64
marketing	85.62 ± 1.61	85.62 ± 1.61	$\mathbf{87.18 \pm 6.65}$
$monk - 2$	83.04 ± 7.78	93.15 ± 3.15	$\mathbf{99.28 \pm 0.10}$
$movement_libras$	$\mathbf{70.73 \pm 1.76}$	70.73 ± 1.76	61.29 ± 2.51
pima	52.82 ± 1.82	56.77 ± 3.17	$\mathbf{74.18 \pm 4.80}$
satimage	$\mathbf{95.38 \pm 4.45}$	93.07 ± 5.80	89.40 ± 6.91
segment	90.51 ± 0.49	$\mathbf{93.39 \pm 0.97}$	46.62 ± 9.79
titanic	34.28 ± 2.05	40.63 ± 19.49	$\mathbf{98.92 \pm 0.11}$
vowel	$\mathbf{83.45 \pm 1.01}$	83.45 ± 1.01	74.41 ± 1.80
wine	86.15 ± 3.86	$\mathbf{92.76 \pm 2.31}$	62.83 ± 11.35
$winequality - red$	64.83 ± 6.25	$\mathbf{64.92 \pm 6.14}$	44.90 ± 6.08
$winequality - white$	$\mathbf{60.69 \pm 1.02}$	60.69 ± 1.02	39.54 ± 6.00
yeast	53.82 ± 5.43	54.77 ± 5.67	$\mathbf{60.81 \pm 11.35}$
$Mean$	65.12	69.26	60.33

Tabela 10 – Media de redução de atributos dos melhores thresholds por fold para o teste do RIS2 (%)

Base de dados	**RIS2 - FS**	**FS - RIS2**
$appendicitis$	**42.86 ± 0.00**	42.86 ± 0.00
$balance$	**50.00 ± 0.00**	50.00 ± 0.00
$bupa$	**50.00 ± 0.00**	50.00 ± 0.00
$coil2000$	**49.41 ± 0.00**	49.41 ± 0.00
$contraceptive$	**44.44 ± 0.00**	44.44 ± 0.00
$haberman$	**33.33 ± 0.00**	33.33 ± 0.00
$hayes-roth$	**50.00 ± 0.00**	50.00 ± 0.00
$heart$	**46.15 ± 0.00**	46.15 ± 0.00
$ionosphere$	**48.48 ± 0.00**	48.48 ± 0.00
$led7digit$	**42.86 ± 0.00**	42.86 ± 0.00
$marketing$	**46.15 ± 0.00**	46.15 ± 0.00
$monk-2$	**50.00 ± 0.00**	50.00 ± 0.00
$movement_libras$	**50.00 ± 0.00**	50.00 ± 0.00
$pima$	**50.00 ± 0.00**	50.00 ± 0.00
$satimage$	**50.00 ± 0.00**	50.00 ± 0.00
$segment$	**47.37 ± 0.00**	47.37 ± 0.00
$titanic$	**33.33 ± 0.00**	33.33 ± 0.00
$vowel$	**46.15 ± 0.00**	46.15 ± 0.00
$wine$	**46.15 ± 0.00**	46.15 ± 0.00
$winequality-red$	**45.45 ± 0.00**	45.45 ± 0.00
$winequality-white$	**45.45 ± 0.00**	45.45 ± 0.00
$yeast$	**50.00 ± 0.00**	50.00 ± 0.00
$Mean$	46.26	46.26

Tabela 11 – Tamanho médio percentual do total da matriz de dados após as reduções de instâncias e atributos no RIS2

Base de dados	**Total**	**RIS2**	**RIS2 - FS**	**FS - RIS2**
appendicitis	100%	27.57%	15.51%	7.79%
balance	100%	15.24%	1.18%	31.53%
bupa	100%	49.82%	24.91%	29.00%
*coil*2000	100%	74.94%	42.22%	28.18%
contraceptive	100%	61.64%	34.24%	31.93%
haberman	100%	41.98%	24.04%	40.91%
$hayes - roth$	100%	51.60%	17.26%	23.12%
heart	100%	38.23%	19.54%	38.14%
ionosphere	100%	25.36%	12.85%	27.28%
*led*7*digit*	100%	42.24%	5.70%	18.70%
marketing	100%	14.38%	7.74%	6.91%
$monk - 2$	100%	16.95%	3.42%	0.36%
$movement_libras$	100%	29.26%	14.63%	19.35%
pima	100%	47.18%	21.61%	12.91%
satimage	100%	4.62%	3.47%	5.30%
segment	100%	9.49%	3.48%	28.09%
titanic	100%	65.72%	39.58%	0.72%
vowel	100%	16.55%	8.91%	13.78%
wine	100%	13.86%	3.90%	20.03%
$winequality - red$	100%	35.17%	19.13%	30.06%
$winequality - white$	100%	39.31%	21.44%	32.98%
yeast	100%	46.17%	22.62%	19.59%

Tabela 12 – Media de acerto dos melhores thresholds por fold utilizando o RIS3 (%)

Base de dados	**RIS3**	**RIS3 - FS**	**FS - RIS3**
appendicitis	79.14 ± 7.15	$\mathbf{84.70 \pm 6.83}$	82.62 ± 12.76
balance	$\mathbf{81.48 \pm 5.93}$	62.24 ± 11.62	51.81 ± 17.04
bupa	57.14 ± 3.99	56.80 ± 7.81	$\mathbf{57.43 \pm 7.13}$
*coil*2000	$\mathbf{70.69 \pm 20.11}$	69.44 ± 21.84	56.41 ± 22.87
contraceptive	45.69 ± 5.01	44.26 ± 3.58	$\mathbf{47.87 \pm 2.62}$
haberman	66.34 ± 8.09	54.19 ± 9.48	$\mathbf{68.27 \pm 5.65}$
hayes − roth	$\mathbf{61.46 \pm 12.27}$	59.66 ± 10.78	48.52 ± 15.57
heart	$\mathbf{78.89 \pm 5.51}$	76.67 ± 5.51	70.74 ± 12.33
ionosphere	$\mathbf{90.92 \pm 4.84}$	87.50 ± 5.47	78.34 ± 8.53
*led*7*digit*	$\mathbf{74.30 \pm 8.05}$	56.16 ± 9.91	40.98 ± 6.06
marketing	$\mathbf{13.31 \pm 1.26}$	12.04 ± 1.48	12.87 ± 1.38
monk − 2	92.83 ± 4.58	$\mathbf{97.23 \pm 1.98}$	79.17 ± 4.27
movement$_l$*ibras*	$\mathbf{59.22 \pm 8.24}$	58.00 ± 8.77	57.67 ± 10.27
pima	63.27 ± 7.40	65.24 ± 4.07	$\mathbf{67.83 \pm 4.31}$
satimage	13.19 ± 4.06	13.44 ± 4.03	$\mathbf{15.03 \pm 4.61}$
segment	$\mathbf{91.65 \pm 2.58}$	81.00 ± 3.96	91.39 ± 3.11
titanic	$\mathbf{55.48 \pm 15.44}$	53.35 ± 16.81	35.26 ± 8.85
vowel	$\mathbf{65.56 \pm 6.27}$	59.70 ± 8.29	53.03 ± 7.30
wine	92.25 ± 5.04	$\mathbf{93.24 \pm 7.02}$	91.52 ± 3.87
winequality − red	44.72 ± 6.23	44.23 ± 3.98	$\mathbf{47.16 \pm 5.56}$
winequality − white	$\mathbf{42.16 \pm 2.36}$	40.22 ± 3.38	41.28 ± 5.25
yeast	$\mathbf{41.87 \pm 4.67}$	30.46 ± 4.14	22.87 ± 7.54
Mean	62.80	59.08	55.37

Tabela 13 – Media de redução de instâncias dos melhores thresholds por fold utilizando o RIS3 (%)

Base de dados	**RIS3**	**RIS3 - FS**	**FS - RIS3**
appendicitis	72.84 ± 3.59	76.28 ± 9.13	$\mathbf{89.00 \pm 12.70}$
balance	90.91 ± 6.78	$\mathbf{99.20 \pm 0.46}$	74.85 ± 13.77
bupa	$\mathbf{56.14 \pm 10.87}$	56.14 ± 10.87	48.73 ± 10.01
*coil*2000	49.07 ± 26.97	40.11 ± 19.39	$\mathbf{56.84 \pm 22.23}$
contraceptive	41.37 ± 6.37	41.37 ± 6.37	$\mathbf{46.09 \pm 5.90}$
haberman	64.78 ± 3.88	$\mathbf{85.57 \pm 11.08}$	50.33 ± 16.33
$hayes - roth$	51.47 ± 13.59	$\mathbf{66.62 \pm 13.58}$	56.28 ± 19.21
heart	$\mathbf{77.86 \pm 11.66}$	71.40 ± 14.24	36.38 ± 16.77
ionosphere	74.89 ± 4.94	$\mathbf{75.97 \pm 4.68}$	50.72 ± 8.63
*led*7*digit*	$\mathbf{92.53 \pm 12.10}$	83.26 ± 15.93	58.08 ± 22.94
marketing	85.69 ± 1.60	$\mathbf{87.76 \pm 5.07}$	87.29 ± 6.59
$monk - 2$	87.27 ± 6.22	94.13 ± 6.15	$\mathbf{99.02 \pm 1.39}$
$movement_libras$	$\mathbf{72.58 \pm 1.63}$	72.58 ± 1.63	62.55 ± 2.69
pima	53.27 ± 2.86	54.75 ± 4.17	$\mathbf{77.79 \pm 6.57}$
satimage	$\mathbf{95.43 \pm 4.49}$	93.08 ± 5.78	90.30 ± 7.04
segment	91.51 ± 0.61	$\mathbf{94.40 \pm 1.47}$	50.34 ± 11.20
titanic	68.34 ± 19.93	58.48 ± 12.88	$\mathbf{99.33 \pm 0.18}$
vowel	$\mathbf{84.39 \pm 1.25}$	84.39 ± 1.25	77.64 ± 2.15
wine	87.83 ± 4.27	$\mathbf{96.32 \pm 1.13}$	63.00 ± 9.58
$winequality - red$	69.35 ± 6.25	$\mathbf{69.86 \pm 6.38}$	48.57 ± 7.01
$winequality - white$	$\mathbf{62.91 \pm 1.27}$	62.91 ± 1.27	42.25 ± 6.75
yeast	58.24 ± 5.53	58.24 ± 5.53	$\mathbf{62.89 \pm 14.40}$
Mean	72.21	73.76	64.92

Tabela 14 – Media de redução de atributos dos melhores thresholds por fold utilizando o RIS3 (%)

Base de dados	**RIS3 - FS**	**FS - RIS3**
$appendicitis$	**42.86 ± 0.00**	42.86 ± 0.00
$balance$	**50.00 ± 0.00**	50.00 ± 0.00
$bupa$	**50.00 ± 0.00**	50.00 ± 0.00
$coil2000$	**49.41 ± 0.00**	49.41 ± 0.00
$contraceptive$	**44.44 ± 0.00**	44.44 ± 0.00
$haberman$	**33.33 ± 0.00**	33.33 ± 0.00
$hayes-roth$	**50.00 ± 0.00**	50.00 ± 0.00
$heart$	**46.15 ± 0.00**	46.15 ± 0.00
$ionosphere$	**48.48 ± 0.00**	48.48 ± 0.00
$led7digit$	**42.86 ± 0.00**	42.86 ± 0.00
$marketing$	**46.15 ± 0.00**	46.15 ± 0.00
$monk-2$	**50.00 ± 0.00**	50.00 ± 0.00
$movement_libras$	**50.00 ± 0.00**	50.00 ± 0.00
$pima$	**50.00 ± 0.00**	50.00 ± 0.00
$satimage$	**50.00 ± 0.00**	50.00 ± 0.00
$segment$	**47.37 ± 0.00**	47.37 ± 0.00
$titanic$	**33.33 ± 0.00**	33.33 ± 0.00
$vowel$	**46.15 ± 0.00**	46.15 ± 0.00
$wine$	**46.15 ± 0.00**	46.15 ± 0.00
$winequality-red$	**45.45 ± 0.00**	45.45 ± 0.00
$winequality-white$	**45.45 ± 0.00**	45.45 ± 0.00
$yeast$	**50.00 ± 0.00**	50.00 ± 0.00
$Mean$	46.26	46.26

Tabela 15 – Tamanho médio da matriz de dados do conjunto de treinamento para os melhores thresholds por fold utilizando o RIS3 (Instâncias X Atributos)

Base de dados	**Total**	**RIS3**	**RIS3 - FS**	**FS - RIS3**
$appendicitis$	100%	27.15%	13.54%	6.29%
$balance$	100%	9.08%	0.40%	12.58%
$bupa$	100%	43.86%	21.93%	25.64%
$coil2000$	100%	50.93%	30.30%	21.83%
$contraceptive$	100%	58.63%	32.57%	29.95%
$haberman$	100%	35.22%	9.61%	33.12%
$hayes-roth$	100%	48.54%	16.70%	21.84%
$heart$	100%	22.14%	15.40%	34.26%
$ionosphere$	100%	25.10%	12.38%	25.39%
$led7digit$	100%	7.51%	9.54%	23.90%
$marketing$	100%	14.31%	6.59%	6.85%
$monk-2$	100%	12.73%	2.93%	0.49%
$movement_libras$	100%	27.41%	13.70%	18.72%
$pima$	100%	46.73%	22.63%	11.10%
$satimage$	100%	4.57%	3.46%	4.85%
$segment$	100%	8.49%	2.95%	26.14%
$titanic$	100%	31.66%	27.68%	0.45%
$vowel$	100%	15.61%	8.41%	12.04%
$wine$	100%	12.17%	1.98%	19.93%
$winequality-red$	100%	30.65%	16.44%	28.05%
$winequality-white$	100%	37.09%	20.23%	31.50%
$yeast$	100%	41.76%	20.88%	18.55%