Análise Empírica de Desempenho de Quatro Métodos de Seleção de Características para *Random Forests*

Denise G. D. Bastos, Patricia S. Nascimento, Marcelo S. Lauretto[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH-USP)
Rua Arlindo Bettio, 1000 – 03828-000 – São Paulo – SP – Brazil

marcelolauretto@usp.br

Abstract. *In supervised learning, it is very usual the ocurrence of datasets containing irrelevant attributes. Under such circumstances, it is crucial to apply some feature selection criterion, mainly in learning problems where data acquisition costs are proportional to the number of attributes. In this paper, we introduce two attribute selection criteria designed for* Random Forests*, named* Incidence Factor (IF) *and* Depth Factor (DF)*. We also describe a detailed empirical analysis comparing the performance of these criteria with the* Error-Based Importance (EI) *and* Gini Importance (GI)*, the two main criteria for* Random Forests *currently in use. Results indicate tht* DF *is a robust criterion, outperforming both the* EI *and* GI *criteria.*

Resumo. *Em aprendizado supervisionado, é comum a ocorrência de bases de dados contendo atributos irrelevantes. Sob tais circunstâncias, a adoção de critérios de seleção de características é fundamental, principalmente nos problemas em que os custos de coleta de dados são proporcionais à quantidade de atributos. Neste artigo, são apresentados dois critérios de seleção de atributos voltados para* Random Forests*, denominados* Fator de Incidência (FI) *e* Fator de Profundidade (FP)*. Uma análise empírica detalhada é apresentada, comparando o desempenho desses critérios com a* Importância Baseada no Erro (IE) *e a* Importância de Gini (IG) *– os dois principais critérios para* Random Forest *atualmente em uso. Os resultados indicam que o* FP *é um critério robusto, com desempenho superior aos critérios* IE *e* IG*.*

1. Introdução

As técnicas de mineração de dados, e mais especificamente de aprendizado de máquina, têm se tornado bastante populares, passando a ser incorporadas nos Sistemas de Informação para Apoio à Decisão, Previsão de Eventos e Análise de Dados. Por exemplo, sistemas de apoio à decisão na área médica e ambientes de *Business Intelligence* fazem uso intensivo dessas técnicas, envolvendo particularmente árvores de decisão [Morais et al. 2012, Microsoft 2006]. A mineração de informação e conhecimento a partir de grandes bases de dados tem sido reconhecida como tema chave de pesquisa em sistemas de banco de dados e aprendizado de máquina [Chen et al. 1996].

Em aprendizado supervisionado, é bastante frequente a ocorrência de bases de dados contendo grande número de atributos, muitos dos quais irrelevantes ou com alta correlação entre si. O primeiro impacto imediato do treinamento de algoritmos de aprendizado com essas características é o fenômeno de *overfitting* – um ajustamento excessivo

dos modelos ao conjunto de treinamento, que compromete a acurácia na classificação de novos casos. O segundo aspecto a ser considerado é que, em diversos domínios, tais como diagnósticos médicos baseados em exames clínicos/genéticos ou problemas de decisão baseados em entrevistas, existem custos associados à obtenção dos atributos, muitas vezes variáveis [Mitchell 1997, He et al. 2012].

Sob tais circunstâncias, a adoção de critérios de seleção de características relevantes para a classificação é fundamental no processo de aprendizado computacional. Assim, a adoção de procedimentos de seleção de atributos pode trazer diversas vantagens a um sistema de classificação supervisionada, tais como o aumento da acurácia do sistema, a diminuição dos custos de aquisição, o aumento da simplicidade e entendimento do modelo de classificação e uma maior compreensão dos processos que originam os dados [Inza et al. 2010]. Tal identificação é usualmente o ponto de partida para investigações mais aprofundadas de possíveis associações causais entre os atributos e as classes. Análises de risco de crédito e pesquisas médicas são exemplos de aplicações fortemente baseadas nesse paradigma.

Neste artigo, nosso interesse se concentra nos métodos de seleção de características baseadas no algoritmo *Random Forests*, proposto por [Breiman 2001]. Uma *Random Forest* (RF) é um classificador formado por uma coleção de árvores de classificação, cada qual construída a partir de uma reamostra aleatória do conjunto de treinamento original. A classificação de um vetor de características $\boldsymbol{x}$ é feita por votação, submetendo-se o vetor às árvores da floresta e atribuindo-se a $\boldsymbol{x}$ a classe mais votada. RFs suprem uma lacuna importante em relação às árvores isoladas. É sabido que algoritmos de construção de árvores de classificação são instáveis em relação ao conjunto de treinamento, no sentido de que perturbações nos atributos de entrada ou a inclusão de novos exemplos podem resultar em árvores consideravelmente diferentes, com diferentes erros classificação [Briand et al. 2009].

Por outro lado, florestas aleatórias não possuem uma interpretação direta como ocorre com as árvores de classificação individuais, pela dificuldade prática de se analisar e comparar as centenas de árvores constituintes da floresta. Outro aspecto é que, para garantir uma baixa correlação entre as classificações das árvores individuais (condição necessária para a obtenção de florestas com baixo erro de generalização), é adotada uma seleção aleatória dos atributos candidatos para partição de cada nó. Isso pode implicar na eventual seleção de atributos com baixo poder preditivo.

[Breiman 2001] propôs duas medidas de importância de atributos para utilização com florestas aleatórias. O primeiro, aqui denominado *Importância Baseada no Erro* (IE), mede o aumento do erro quando se permutam os valores do atributo de interesse. O segundo, denominado *Importância de Gini* (IG), é baseado na soma dos decréscimos do índice de Gini em todos os nós rotulados pelo atributo. Cada uma dessas medidas pode ser utilizada como um critério de seleção de características, através do qual são selecionados os atributos com maior importância [Guyon and Elisseeff 2003]. Por essa razão, adotaremos nesse trabalho os termos *medida de importância* e *critério de seleção* indistintamente.

Em trabalho anterior [Bastos et al. 2013], foram apresentadas duas novas medidas de importância de atributos, denominadas *Fator de Incidência* (FI) e *Fator de Profundi-*

dade (FP). Uma análise empírica preliminar foi apresentada, comparando o desempenho dessas medidas com os critérios tradicionais IE e IG. Todavia, aquele estudo se limitava a uma coleção restrita de *datasets*, e não incluía análises de significância para comparação entre os resultados.

O objetivo deste artigo é reintroduzir as duas medidas propostas, fixando algumas notações em relação ao trabalho anterior, bem como apresentar uma análise empírica mais completa das quatro medidas (IE, IG, FI, FP). Essa análise empírica contempla uma coleção mais extensa de *datasets*, bem como análises de significância que nos permitem aferir de forma mais completa os desempenhos relativos dos quatro critérios.

As novas medidas propostas são inspiradas em uma propriedade intrínseca ao processo de construção das árvores em uma *Random Forest*: atributos com maior relevância global tendem a ser escolhidos antes dos atributos com relevância local. Logo, tendem a aparecer nos nós mais próximos à raiz, sobre os quais incidem as maiores quantidades de exemplos. Com base nessas premissas, a primeira medida, denominada *Fator de Incidência* (FI), avalia a quantidade relativa de exemplos do conjunto de treinamento que incidem sobre nós rotulados por cada atributo; a segunda medida, denominada *Fator de Profundidade* (FP), avalia as profundidades relativas dos nós rotulados pelo atributo, ou seja, suas distâncias em relação à raiz.

O artigo está organizado da seguinte maneira. Na Seção 2 apresentamos brevemente as definições de *Random Forests* e seu método básico de construção. A Seção 3 descreve as medidas de importância de atributos, sendo que as duas primeiras subseções descrevem as medidas definidas por [Breiman 2001], e as duas últimas apresentam as novas medidas propostas. Na Seção 4 detalhamos a metodologia empregada na avaliação empírica, e na Seção 5 são apresentados os resultados obtidos. Finalmente, na Seção 6 apresentamos nossas conclusões.

2. *Random Forests*

As *Random Forests* (RFs) são obtidas através de ***b**ootstrapping **agg**regat**ing*** (ou simplesmente *bagging*), um método utilizado para gerar múltiplas versões de um preditor [Breiman 1996]. Tais versões são construídas a partir de reamostras do conjunto original, obtidas via sorteio simples com reposição.

Apresentamos a seguir a notação sugerida por [Breiman 2001]. Um conjunto de treinamento é denotado por $\mathcal{L} = \{(\boldsymbol{x}_n, y_n), n = 1, 2, \ldots, N\}$, onde N é a quantidade de exemplos, $\boldsymbol{x}_n$ é o vetor de atributos e $y_n \in \{1, 2, \ldots, C\}$ é a classe verdadeira do n-ésimo exemplo. Os atributos são indexados por $m = 1, 2, \ldots, M$, e assim o vetor de atributos do n-ésimo exemplo é denotado por $\boldsymbol{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,M})$.

Denote por $\psi(\boldsymbol{x}, \mathcal{L})$ um preditor para a classe de $\boldsymbol{x}$ construído a partir do conjunto de treinamento $\mathcal{L}$. Suponha que exista uma seqüência finita de conjuntos de treinamento $\{\mathcal{L}^{(s)}\}, s = 1, 2, \ldots, S$, cada um consistindo de N observações independentes provenientes da mesma distribuição subjacente ao conjunto $\mathcal{L}$. A idéia central é usar $\{\mathcal{L}^{(s)}\}$ para obter um preditor melhor do que o preditor simples $\psi(\boldsymbol{x}, \mathcal{L})$, tendo como restrição utilizar apenas a seqüência de preditores $\psi(\boldsymbol{x}, \mathcal{L}^{(s)})$. Indexando-se as classes por $c = 1, 2, \ldots, C$, um método de agregar os preditores $\psi(\boldsymbol{x}, \mathcal{L}^{(s)})$ é através de votação, escolhendo para $\boldsymbol{x}$ a classe mais votada entre os preditores. Formalmente, denotando por

$N_c = |\{s \in \{1 \dots S\} : \psi(\boldsymbol{x}, \mathcal{L}^{(s)}) = c\}|$ o número de "votos" na classe c, o classificador agregado pode ser definido por $\psi_A(\boldsymbol{x}) = \arg\max_c N_c$. O subscrito A em ψ_A denota agregação.

A obtenção de $\{\mathcal{L}^{(s)}\}, s = 1, 2, \dots, S$ é feita tomando-se reamostras *bootstrap* de $\mathcal{L}$, via sorteio com repetição, cada qual de tamanho N.

Na formulação das RFs propostas por [Breiman 2001], o algoritmo básico de construção das árvores é o CART – Classification and Regression Trees [Breiman et al. 1984]. As árvores são expandidas ao máximo, sem poda. Para a divisão de cada nó, um subconjunto de tamanho fixo dos atributos de entrada é selecionado aleatoriamente, escolhendo-se a divisão ótima dentro desse subconjunto.

3. Índices de Importância de Atributos

Nos algoritmos de construção de árvores de classificação tradicionais, os atributos mais relevantes para classificação são selecionados graças aos procedimentos de pré e pós poda [Breiman et al. 1984]. Nas RFs, por sua vez, a identificação dos atributos relevantes não é imediata, devido ao grande número de árvores geradas e devido à ausência de procedimentos de poda na construção das árvores.

Assim, são adotadas algumas métricas de avaliação da importância de cada atributo. [Breiman 2001] sugere duas medidas de importância, descritas nas próximas subseções.

Neste artigo, apresentamos a notação a seguir. Denotamos por K o número de árvores da floresta, M o número de atributos e C o número de classes. As árvores são indexadas por $k = 1, 2, \dots, K$; os atributos avaliados são indexados por $m = 1, 2, \dots, M$; as classes são indexadas por $c = 1, 2, \dots, C$. I_k denota o número de nós da k-ésima árvore.

- O par (k, i) denota o i-ésimo nó da k-ésima árvore, $k = 1, 2, \dots, K$, $i = 1, 2, \dots, I_k$.
- $r(k, i)$ denota o atributo que rotula o i-ésimo nó da k-ésima árvore. Para os nós terminais, define-se $r(k, i) = 0$.
- $T_k = \{(k, 1), (k, 2), \dots, (k, I_k)\}$ denota o conjunto dos nós da árvore k.
- $T_k(m) \subseteq T_k$ denota o subconjunto dos nós de T_k rotulados pelo atributo m:

$$T_k(m) = \{i \in T_k | r(k, i) = m\} \tag{1}$$

- $n(k, i)$ é o número de exemplos do conjunto de treinamento que incidem sobre o nó i da árvore k.
- $n(k, i, c)$ é o número de exemplos de classe c do conjunto de treinamento que incidem sobre o nó i da árvore k.

3.1. Importância Baseada no Erro (IE)

Essa técnica consiste em, uma vez construída a floresta aleatória, permutar aleatoriamente os valores do atributo m entre os exemplos do conjunto de teste. Aplicam-se os exemplos com o m-ésimo atributo permutado sobre as árvores, analisando-se os erros resultantes. O aumento do erro de classificação sobre os exemplos permutados em relação aos exemplos originais fornece a medida de importância do atributo.

Formalmente, denotemos por err_k e por err_k^m o percentual de exemplos *out-of-bag* classificados incorretamente pela árvore k, respectivamente antes e após a permutação dos valores do atributo m. O índice de importância do atributo m baseado no erro (IE) é dado por :

$$\text{IE}(m) = \frac{1}{K} \sum_{k=1}^{K} \frac{err_k^m - err_k}{err_k} \tag{2}$$

3.2. Importância de Gini (IG)

Na metodologia CART para construção de árvores de classificação, a escolha das partições ótimas dos nós utiliza como critério de pureza o Índice de Gini [Breiman et al. 1984]. Esse índice é utilizado para avaliar a distribuição das classes em cada nó. A divisão de cada nó é feita de maneira a resultar em nós filhos mais "puros" do que o pai original, ou seja, com maiores concentrações de exemplos em certas classes.

Dado um nó i de uma árvore k, denotemos por $p_c = n(k,i,c)/n(k,i)$ as proporções de exemplos de i pertencentes à classe c. O índice de diversidade Gini é definido como

$$G(k,i) = \sum_{c_1 \neq c_2} p_{c_1} p_{c_2}. \tag{3}$$

Note que esse índice tem seu valor máximo quando todas as classes são equiprováveis, ou seja, quando $p_c = \frac{1}{C}$, $c = 1 \ldots C$; e é igual a zero quando uma das classes tem proporção 1 (e conseqüentemente as demais têm proporção 0).

Para escolher a divisão de um nó i de uma árvore k, o índice de Gini é utilizado como segue. Seja (m, s) uma divisão candidata representando uma restrição $x_m \leq s$, onde s é um número real. Suponha que (m, s) divide o nó em dois nós filhos, i_v (correspondente às instâncias que obedecem à restrição) e i_f (correspondente às demais instâncias). A qualidade da divisão de (m, s) é medida pelo decréscimo no índice de Gini:

$$\Delta G(k,i,m,s) = G(k,i) - \frac{n(k,i_v)}{n(k,i)} G(k,i_v) - \frac{n(k,i_f)}{n(k,i)} G(k,i_f). \tag{4}$$

Para expandir o nó i, escolhe-se a divisão (m^*, s^*) que maximiza $\Delta G(k,i,m,s)$.

A medida de importância de cada atributo m em uma Floresta Aleatória pode ser dada pela soma dos decréscimos nos índices de Gini de todos os nós rotulados por m:

$$\text{IG}(m) = \frac{1}{K} \sum_{k \in K} \sum_{i \in T_k(m)} \Delta G(k,i,m,s^*) \tag{5}$$

3.3. Fator de Incidência (FI)

A primeira medida de importância proposta nesse artigo leva em consideração o número relativo de exemplos que são afetados pela presença de cada atributo, ou mais especificamente, o número relativo de exemplos incidentes sobre os nós rotulados pelo atributo. Como essa medida é, em média, proporcional à freqüência do atributo nos nós das árvores geradas e inversamente proporcional à profundidade do atributo nas árvores, essa é uma medida baseada (indiretamente) na topologia das árvores geradas.

A soma das quantidades de exemplos incidentes sobre os nós da k-ésima árvore rotulados pelo atributo m é denotado por $N_k(m)$: $N_k(m) = \sum_{i \in T_k(m)} n(k,i)$. Note que, na soma acima, um exemplo pode ser computado mais de uma vez.

Definimos o *Fator de Incidência Local* (FIL) do atributo m na k-ésima árvore por

$$\text{FIL}_k(m) = N_k(m)/N_k, \tag{6}$$

onde $N_k = \sum_{i \in T_k} n(k,i)$ denota a soma das quantidades de exemplos incidentes sobre todos os nós da árvore k.

O Fator de Incidência (FI) do atributo m é definido como a média de seus fatores de incidência locais sobre todas as árvores:

$$\text{FI}(m) = \frac{1}{K} \sum_{k=1}^{K} \text{FIL}_k(m). \tag{7}$$

3.4. Fator de Profundidade (FP)

A segunda medida de importância proposta parte do princípio de que os atributos mais relevantes tendem a rotular os nós mais próximos à raiz, e portanto nós de menor profundidade. Assim, definimos uma função de importância inversamente proporcional às profundidades dos nós rotulados pelo atributo na *Random Forest*.

Denotamos por $d(k,i)$ a profundidade do i-ésimo nó da k-ésima árvore da floresta. Dada uma árvore k, $H_k(m)$ representa a soma das inversas das profundidades dos nós da k-ésima árvore rotulados pelo atributo m:

$$H_k(m) = \sum_{i \in T_k(m)} \frac{1}{d(k,i)+1}. \tag{8}$$

(A adição $d(k,i)+1$ no denominador é utilizada para tratar a raiz, que tem profundidade zero.)

Definimos o *Fator de Profundidade Local* (FPL) do atributo m na k-ésima árvore por

$$\text{FPL}_k(m) = \frac{H_k(m)}{H_k}, \tag{9}$$

onde $H_k = \sum_{m=1}^{M} H_k(m)$.

O Fator de Profundidade (FP) do atributo m é definido como a média de seus fatores de profundidade locais sobre todas as árvores:

$$\text{FP}(m) = \frac{1}{K} \sum_{k=1}^{K} \text{FPL}_k(m). \tag{10}$$

4. Análise de Desempenho

Foi realizada uma análise comparativa do desempenho dos quatro critérios de seleção de atributos estudados (IE, IG, FI, FP). Os experimentos foram baseados em uma coleção de 36 *datasets* obtidos da UCI Machine Learning Repository [Frank and Asuncion 2010],

todos contendo entre 10 e 800 atributos, e no mínimo 100 exemplos. A Tabela 1 contém um sumário de cada *dataset*: nome e abreviação, número de exemplos, número de atributos, número de classes, percentual médio de dados faltantes (sobre todos os atributos) e percentual máximo de dados faltantes em um único atributo.

Os experimentos foram realizados no ambiente R [R Core Team 2013], e para a construção e aplicações das RFs utilizou-se o Pacote *randomForest* [Liaw and Wiener 2002].

Uma etapa inicial consistiu em eliminar, de cada *dataset*, atributos categóricos com mais de 32 categorias, uma vez que a implementação das *Random Forests* adotada somente suporta atributos categóricos até este limite. Também, em virtude dos experimentos computacionalmente intensivos (vide descrição abaixo), para cada *dataset* com mais de 40 mil exemplos (Bankmark, Connect4 e Covertype), foi selecionado aleatoriamente um subconjunto de 40 mil exemplos, via sorteio simples sem reposição.

A avaliação dos critérios sobre cada *dataset* foi feita por reamostragem, em 100 iterações. Cada iteração $s \in \{1, 2, \ldots, 100\}$ consistiu nos seguintes passos:

a) Sorteou-se uma sub-amostra com 50% dos exemplos do *dataset* original, via sorteio simples sem reposição;
b) Sobre a sub-amostra gerada, foi construída uma RF, a partir da qual foram calculadas as importâncias dos atributos sob cada um dos quatro critérios estudados. Para cada critério de seleção, foram selecionados os M' atributos de maiores importâncias, para três valores de M': $M' \in \{M/4, M/3, M/2\}$, onde M denota o número de atributos do conjunto de treinamento original.
c) Foi sorteada uma nova sub-amostra com 50% do *dataset* original, também por sorteio simples sem reposição. Construiu-se uma RF para cada critério de seleção $c \in \{\text{IE}, \text{IG}, \text{FI}, \text{FP}\}$ e para cada quantidade de atributos M'. Denotamos por $\psi_{c,M',s}$ a RF induzida a partir da sub-amostra da s-ésima iteração, restrita aos M' atributos selecionados pelo critério c.
d) Para avaliação de cada RF $\psi_{c,M',s}$, foi utilizado o conjunto complementar da sub-amostra gerada no passo anterior, constituído pelos exemplos não sorteados. As medidas de desempenho adotadas são:
    - A acurácia, que corresponde ao percentual de exemplos de teste classificados corretamente.
    - A área sob a Curva ROC [Fawcett 2006], aqui abreviada por AUC. Neste trabalho, utilizamos a extensão da AUC para classes múltiplas, proposta por [Hand and Till 2001] e dada por

$$AUC = \frac{2}{C(C-1)} \sum_{i,j \in \{1,\ldots,C\}, i<j} \text{AUC}(i,j), \tag{11}$$

      onde C denota o número de classes e $\text{AUC}(i,j)$ denota a área sob a curva ROC computada exclusivamente sobre as classes i e j. Essa extensão está implementada no Pacote *pROC* [Robin et al. 2011].

O sorteio de duas amostras separadas – uma para seleção dos atributos e outra para avaliação das RFs – foi feito para evitar medidas superestimadas das acurácias e das AUCs. Também tomou-se o cuidado de, em cada iteração, aplicar os quatro critérios

Tabela 1. Sumários dos *datasets* utilizados

Nome / Abreviação *Dataset*	#Exemp	#Atr	#Class	% *Missing Values*	
				Média	Máx
Arrhythmia (Arrhythmia)	452	279	13	0.3	83.2
Audiology–Stand (Audiolstand)	226	69	24	2.0	98.2
Bank Marketing (Bankmark)	45211	16	2	0.0	0.0
Breast Cancer (Breastcancer)	569	30	2	0.0	0.0
Cardiotocography (Cardiotoc)	2126	35	10	0.0	0.0
Connect Bench–Sonar (Cbsonar)	208	60	2	0.0	0.0
Chess-KR vs. KP (Chesskr)	3196	36	2	0.0	0.0
Climate Model Crashes (Climatefail)	540	20	2	0.0	0.0
Congressional Voting (Congrvoting)	435	16	2	5.6	23.9
Connect-4 (Connect4)	67557	42	3	0.0	0.0
Covertype (Covertype)	581012	54	7	0.0	0.0
Credit Approval (Creditapp)	690	15	2	0.6	1.9
Cylinder Bands (Cylbands)	540	37	2	5.0	28.9
Dermatology (Dermatology)	358	34	6	0.0	0.0
Hepatitis (Hepatitis)	155	19	2	5.7	43.2
Horse Colic (Horsecolic)	368	27	3	19.4	81.3
Indian Liver Patient Dataset (ILPD)	583	10	2	0.1	0.7
Ionosphere (Ionosphere)	351	34	2	0.0	0.0
Letter Recognition (Letterrec)	7742	16	26	0.0	0.0
Multiple Features (Multiplfeat)	2000	649	10	0.0	0.0
Mushroom (Mushroom)	8124	22	2	1.4	30.5
Page Blocks Classif (Pageblocks)	5473	10	5	0.0	0.0
Parkinsons (Parkinsons)	195	22	2	0.0	0.0
Pittsburgh Bridges (Pittbridges)	105	11	6	5.3	22.9
Planning Relax (Plannrelax)	182	12	2	0.0	0.0
Semi-conductor Manuf (Semicond)	1567	590	2	4.5	91.2
Soybean (Soybean)	307	35	19	6.6	13.4
Spambase (Spambase)	4601	57	2	0.0	0.0
Spectf Heart (Spectf)	267	44	2	0.0	0.0
Statlog–Austr Credit (Stlogaustcred)	690	14	2	0.0	0.0
Statlog–German Credit (Stloggercred)	1000	24	2	0.0	0.0
Statlog–Image Segm (Stlogimageseg)	2310	19	7	0.0	0.0
Statlog–Vehicle (Stlogvehicle)	846	18	4	0.0	0.0
Steel Plates Faults (Stplatefaults)	1941	33	2	0.0	0.0
Waveform DB Generator (Waveform)	5000	40	3	0.0	0.0
Wine (Wine)	178	13	3	0.0	0.0

de seleção sobre as mesmas sub-amostras aleatórias, para evitar que eventuais diferenças entre as acurácias observadas pudessem ser atribuídas às flutuações decorrentes da amostragem [Mitchell 1997].

Para a comparação entre os desempenhos dos quatro critérios de seleção, para cada *dataset* e para cada valor de M' foram calculadas, sobre as 100 iterações dos testes, as médias e desvios-padrão das acurácias e das AUCs.

Realizou-se uma análise de significância das diferenças entre as acurácias dos quatro critérios. Foi adotado um procedimento de teste de sinal, comparando as acurácias do melhor método de seleção (ou seja, aquele com maior acurácia média) e as acurácias de cada um dos demais critérios. Os métodos cujo nível descritivo do teste (p-valor) era maior do que um nível crítico p foram considerados estatisticamente equivalentes ao melhor método em termos de desempenho. A fim de avaliar a sensibilidade dos desempenhos comparativos dos critérios de seleção em relação à escolha do nível crítico, neste estudo foram adotados os valores mais usuais: $p \in \{0.1, 0.05, 0.01\}$.

Um dos procedimentos de teste mais usuais é o teste *t* pareado [Mitchell 1997], que assume que as diferenças entre as observações sigam distribuição normal. A fim de confirmar se a premissa de normalidade se aplicava às acurácias estimadas, foi aplicado o teste de normalidade de Shapiro-Wilk [Royston 1982] sobre as acurácias obtidas nas 100 iterações, para cada um dos quatro métodos de seleção. Em virtude de se observar desvios significantes da distribuição normal em diversos casos (p-valor $<$ 0,1), buscou-se um procedimento de teste que dispensasse a premissa de normalidade dos dados e que tratasse adequadamente os empates entre as observações.

[Coakley and Heise 1996] descreveram e compararam diversos métodos de testes pareados, e indicaram o teste de sinal proposto por [Putter 1955] (nomeado *ANU* – Asymptotic uniformly most powerful nonrandomized test), por ser conceitualmente simples e por apresentar bons resultados empíricos. Dessa forma, esse teste foi implementado no ambiente R [R Core Team 2013] e utilizado para a comparação entre as acurácias.

Nas análises comparativas consolidadas, para cada critério c, para cada *dataset* D e para cada configuração M' de atributos ($25\%, 33\%$ e 50% dos atributos originais), os seguintes quesitos binários foram definidos:

- *MA*(c, D, M'): Indica se o critério c obteve a melhor acurácia média sobre o *dataset* D na configuração M';
- *AE*(c, D, M', p) Indica se o critério c obteve a melhor acurácia ou acurácia estatisticamente equivalente ao melhor critério sobre o *dataset* D na configuração M', sob o nível crítico p;
- *PA*(c, D, M'): Indica se o critério c obteve a pior acurácia média sobre o *dataset* D na configuração M'.

As frequências das ocorrências de cada um desses quesitos entre os critérios de seleção também foram comparadas através de testes de significância. Para cada quesito e para cada par de critérios (c_1, c_2), construiu-se uma tabela X de dimensão 2×2 contendo as frequências das ocorrências cruzadas do quesito no par (c_1, c_2). Por exemplo, para o quesito *MA*, a tabela X possui as seguintes componentes:

- X_{00}: Número de *datasets* e configurações em que *MA*$(c_1, D, M') = 0$ e *MA*$(c_2, D, M') = 0$, ou seja, nenhum dos dois critérios de seleção obteve a melhor

acurácia;

- X_{01}: Número de *datasets* e configurações em que $MA(c_1, D, M') = 0$ e $MA(c_2, D, M') = 1$, ou seja, somente o critério c_2 obteve a melhor acurácia;
- X_{10}: Número de *datasets* e configurações em que $MA(c_1, D, M') = 1$ e $MA(c_2, D, M') = 0$, ou seja, somente o critério c_1 obteve a melhor acurácia;
- X_{11}: Número de *datasets* e configurações em que $MA(c_1, D, M') = 1$ e $MA(c_2, D, M') = 1$, ou seja, os dois critérios de seleção obtiveram simultaneamente a melhor acurácia.

A partir desta tabela, o problema consiste em analisar a significância estatística da diferença entre X_{01} e X_{10} [McNemar 1947], e para isso foi adotada a abordagem de cálculo exato do Teste de McNemar implementada no Pacote *exact2x2* [Fay 2010].

Procedimento similar ao descrito acima foi realizado também para a análise das áreas sob a curva ROC (AUCs) obtidas pelos critérios de seleção.

5. Resultados e Discussões

5.1. Acurácias

As Tabelas 4, 5 e 6, apresentadas no Apêndice, contêm as médias e desvios-padrão das acurácias em cada *dataset*, obtidas pelos quatro critérios de seleção sobre sub-amostras geradas respectivamente com $25\%, 33\%$ e 50% dos atributos originais. Em cada linha, a célula sombreada indica a maior acurácia média obtida, e os sobrescritos a, b e c indicam, respectivamente, as acurácias médias estatisticamente equivalentes à maior acurácia sob cada nível crítico (p-valor>0.1, p-valor>0.05 ou p-valor>0.01).

A Figura 2, obtida a partir dessas tabelas, apresenta o número de *datasets* em que cada critério obteve a maior acurácia média (*MA*), o número de *datasets* em que cada critério obteve a maior acurácia ou acurácia estatisticamente equivalente ao melhor critério sob cada nível crítico (*AE*(0.1), *AE*(0.05) e *AE*(0.01)), e o número de *datasets* em que o critério obteve a pior acurácia média (*PA*). A figura é subdividida em três gráficos, correspondentes às três configurações consideradas: (a) 25%, (b) 33%, e (c) 50% dos atributos originais.

Através dos gráficos nota-se que o FI apresentou desempenho bastante inferior aos demais critérios, em todos os quesitos analisados, não se mostrando portanto um bom critério de seleção de atributos. Dessa forma, as análises comparativas posteriores concentram-se exclusivamente nos critérios IE, IG e FP.

O critério FP apresentou, de forma geral, desempenho superior aos demais critérios. Obteve as maiores frequências nos quesitos *MA* (exceto na configuração (c), em que o IE foi ligeiramente superior), *AE*(0.1), *AE*(0.05) e *AE*(0.01) (exceto na configuração (a), em que os critérios IE e IG mostraram-se superiores). Obteve ainda frequências de *PA* consideravelmente menores do que os demais critérios, nas três configurações. O critério FP também apresentou variações menores do que os outros dois critérios no quesito *AE* sob os três níveis críticos (0.1, 0.05 e 0.01), o que sugere menor sensibilidade do FP quanto à escolha do nível crítico para o quesito *AE*.

Entre os critérios IE e IG, não foram observadas evidências de superioridade de um método em relação ao outro. As maiores diferenças observadas foram no indicador

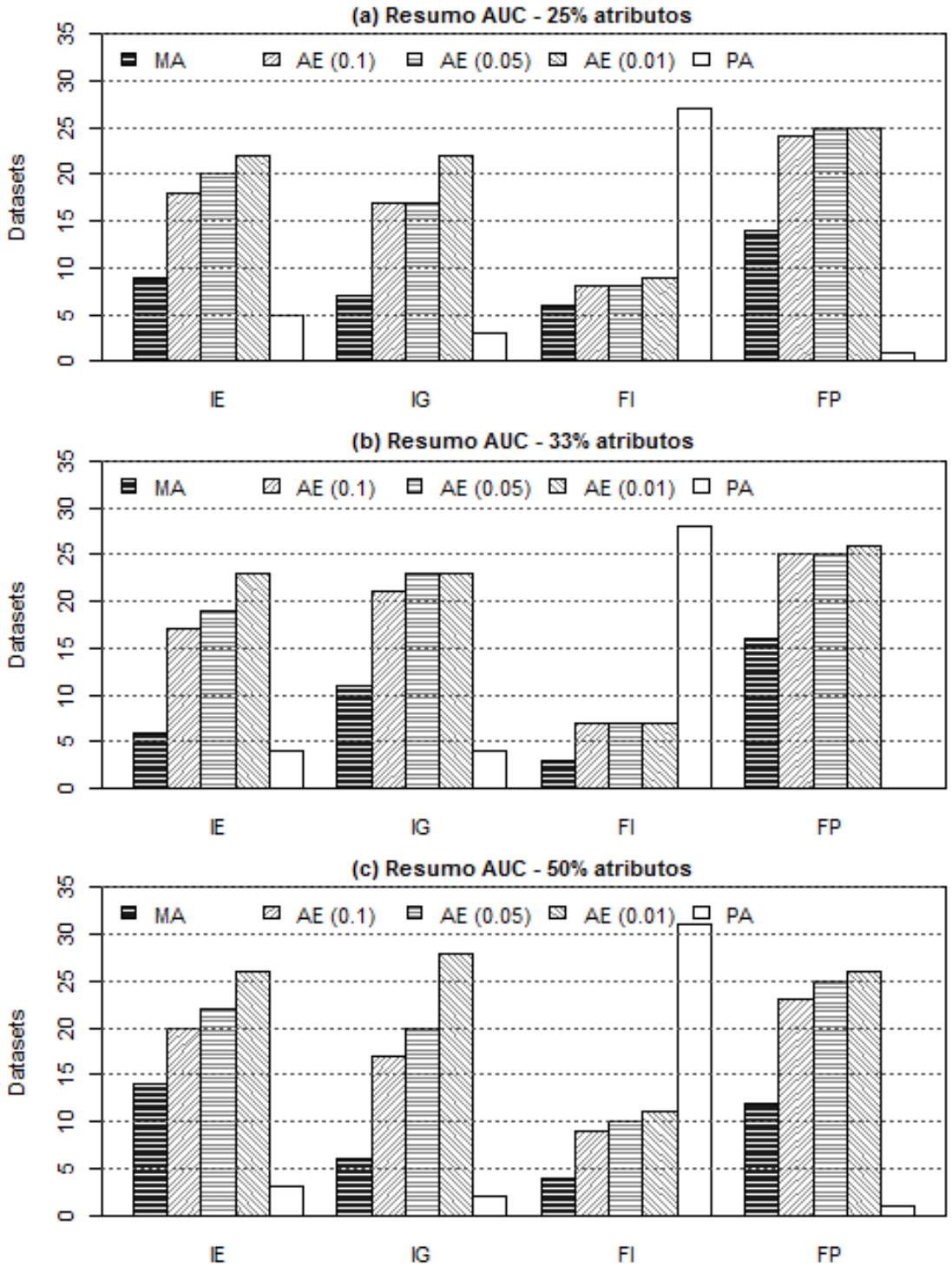

Figura 1. Número de *datasets* nos quais cada critério obteve a maior AUC média (MA), AUC estatisticamente equivalente à do melhor método (AE) e a pior AUC média (PA)

MA, em que o critério IE foi superior nos cenários (a) e (c), enquanto o critério IG foi superior no cenário (b).

A Tabela 2 apresenta as frequências totais de *datasets* e de configurações em que

Tabela 2. Frequências totais (sobre todos os *datasets* e configurações) dos quesitos *MA*, *AE* e *PA* para as acurácias obtidas pelos três melhores métodos (IE, IG e FP), e respectivos níveis descritivos (p-valores).

Quesito	Frequências			p-valores		
	IE	IG	FP	IE×IG	IE×FP	IG×FP
MA	29	25	43	0.68	0.12	0.04**
AE (0.1)	53	53	67	1.00	0.10	0.07*
AE (0.05)	62	57	72	0.56	0.23	0.04**
AE (0.01)	67	68	75	1.00	0.34	0.40
PA	12	9	2	0.66	0.01**	0.07*

*p-valor$<$0.1; **p-valor$<$0.05

cada critério atingiu os quesitos *MA*, *AE* e *PA*, bem como os níveis descritivos (p-valores) do teste de McNemar entre cada par de critérios. Uma vez que o critério FI apresentou desempenho estatisticamente inferior aos demais critérios em todos os quesitos (p-valor$<$0.01), preferiu-se manter na tabela apenas os resultados dos três melhores critérios (IE, IG e FP), para maior facilidade de análise.

Na tabela observa-se uma considerável superioridade numérica dos resultados do FP em relação aos critérios IE e IG, já que o FP apresentou maiores ocorrências nos quesitos positivos (*MA* e *AE*) e menores ocorrências no quesito negativo (*PA*). Observa-se também que o FP teve desempenho estatisticamente superior ao IG em quase todos os quesitos (exceto em *AE*(0.01)) e superior ao IE no quesito *PA*.

Comparando-se o IE e o IG, não foram identificadas diferenças estatisticamente significantes de desempenho.

5.2. Áreas sob a curva ROC (AUC)

As Tabelas 7, 8 e 9, apresentadas no Apêndice, contêm as médias e desvios-padrão das AUCs em cada *dataset*, obtidas pelos quatro critérios de seleção sobre sub-amostras geradas respectivamente com $25\%, 33\%$ e 50% dos atributos originais.

A Figura 1, obtida a partir dessas tabelas, apresenta o número de *datasets* em que cada critério obteve a maior AUC média (*MA*), o número de *datasets* em que cada critério obteve a maior AUC ou AUC estatisticamente equivalente ao melhor critério sob cada nível crítico (*AE*(0.1), *AE*(0.05) e *AE*(0.01)), e o número de *datasets* em que o critério obteve a pior AUC média (*PA*).

Como foi observado na análise das acurácias, o FI também apresentou desempenho bastante inferior aos demais critérios, em todos os quesitos analisados, e portanto as análises comparativas posteriores concentram-se exclusivamente nos critérios IE, IG e FP.

Com relação às AUCs, a superioridade numérica de desempenho do critério FP sobre os critérios IE e IG foi ainda mais expressiva do que aquela observada nas acurácias. O FP apresentou os maiores valores de *MA*e *AE*, e os menores valores de *PA*. A única exceção é no cenário (c), em que o IG apresenta um valor ligeiramente superior no quesito *AE*(0.01). O critério FP também apresenta menor sensibilidade do FP quanto à escolha do nível crítico para o quesito *AE*.

Entre os critérios IE e IG, não foram observadas evidências de superioridade de

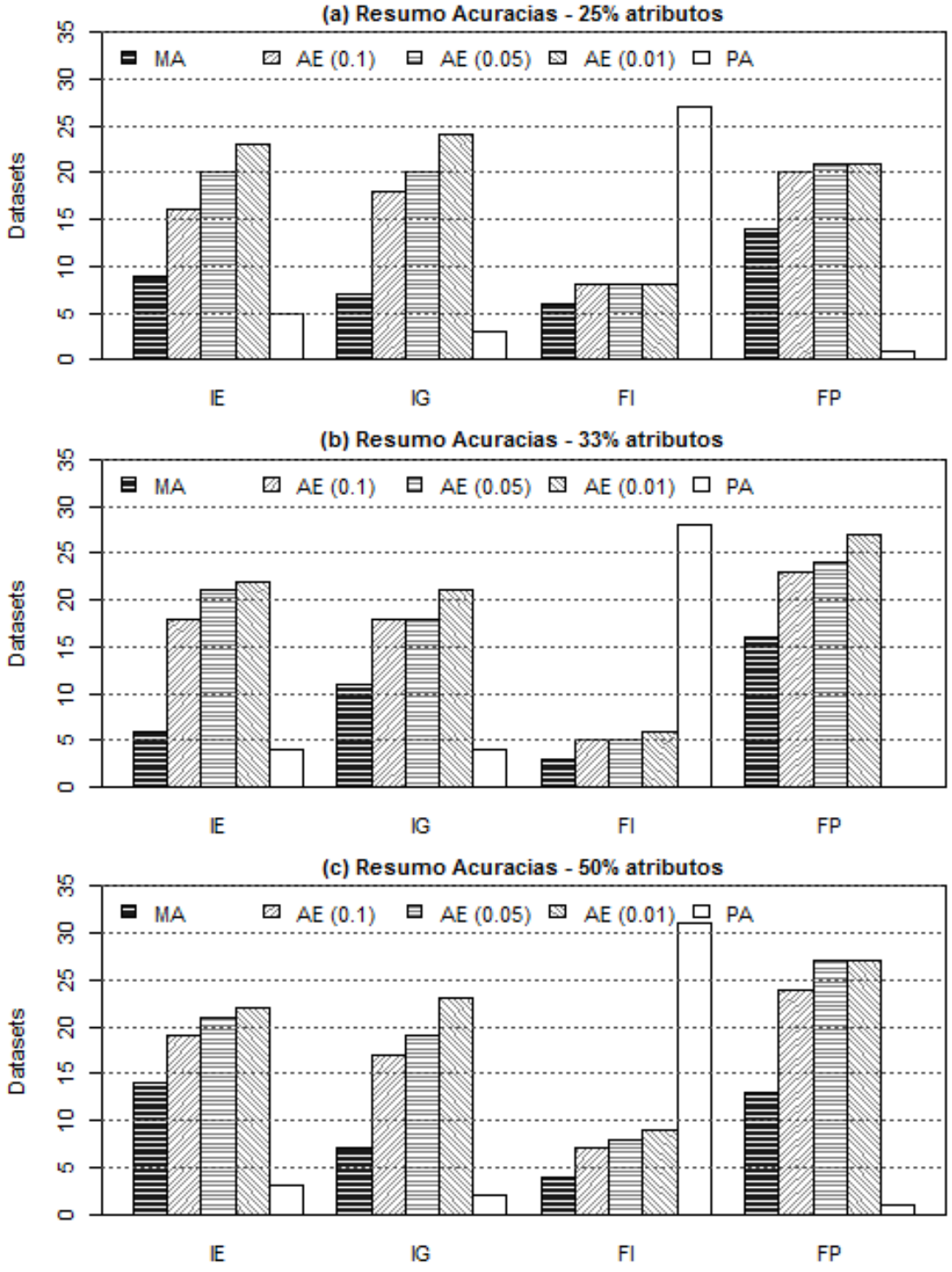

Figura 2. Número de *datasets* nos quais cada critério obteve a maior acurácia média (MA), acurácia média estatisticamente equivalente à do melhor método (AE) e a pior acurácia média (PA)

um método sobre o outro. Observam-se pequenas diferenças pontuais entre os quesitos, porém sem padrão definido.

A Tabela 3 apresenta as frequências totais de *datasets* e de configurações em que

Tabela 3. Frequências totais (sobre todos os *datasets* e configurações) dos quesitos *MA*, *AE* e *PA* para as AUCs obtidas pelos três melhores métodos (IE, IG e FP), e respectivos níveis descritivos (p-valores).

Quesito	Frequências			p-valores		
	IE	IG	FP	IE×IG	IE×FP	IG×FP
MA	29	24	42	0.58	0.15	0.04**
AE (0.1)	55	55	72	1.00	0.04**	0.02**
AE (0.05)	61	60	75	1.00	0.08*	0.04**
AE (0.01)	71	73	77	0.86	0.47	0.64
PA	12	9	2	0.66	0.01**	0.07*

*p-valor$<$0.1; **p-valor$<$0.05

cada critério atingiu os quesitos *MA*, *AE* e *PA*, bem como os níveis descritivos (p-valores) do teste de McNemar entre cada par de critérios. Uma vez que o critério FI apresentou desempenho estatisticamente inferior aos demais critérios em todos os quesitos (p-valor$<$0.01), preferiu-se manter na tabela apenas os resultados dos três melhores critérios (IE, IG e FP), para maior facilidade de análise.

Assim como observado para as acurácias, o FP teve resultados superiores ao IE e IG, já que apresentou maiores ocorrências nos quesitos positivos (*MA* e *AE*) e menores ocorrências no quesito negativo (*PA*). Observa-se também que o FP teve desempenho estatisticamente superior ao IG em quase todos os quesitos (exceto em *AE*(0.01)), e superior ao IE nos quesitos *AE*(0.1), *AE*(0.05) e *PA*.

De forma análoga às acurácias, comparando-se o IE e o IG, não foram identificadas diferenças estatisticamente significantes de desempenho.

5.3. Discussões

Os experimentos numéricos realizados sugerem que, de maneira geral, o critério FP tende a apresentar desempenho superior aos critérios IE e IG, tanto sob o indicador de acurácia como de área sob a curva ROC (AUC). Essa afirmação decorre do fato do FP ter apresentado maiores ocorrências do que os demais critérios nos quesitos positivos (*MA* e *AE*), e menores ocorrências no quesito negativo (*PA*). Várias dessas diferenças foram confirmadas por análise de significância, com destaque para o quesito *PA*, no qual o FP apresentou frequências estatisticamente menores do que o IE e o IG no quesito *PA*. Em outras palavras, o FP se mostrou pelo menos tão eficiente quanto os demais critérios nos quesitos positivos, e estatisticamente mais eficiente do que os demais no quesito negativo.

O quesito *AE*(0.01) foi o único no qual os critérios IE e IG tiveram desempenho similar ao critério FP. Todavia, embora o nível crítico 0.01 seja um dos níveis tradicionalmente considerados nos testes de significância, recomendamos certa cautela com seu uso no presente contexto, pela razão exposta a seguir.

Sob o quesito *AE*(0.01), um critério de seleção de atributos só não é considerado tão bom quanto o melhor critério se houver uma forte evidência do contrário (ou seja, se p-valor$<$0.01). É importante notar que, se por um lado esse baixo nível crítico pode propiciar uma menor taxa de erro do Tipo I (probabilidade de dois métodos serem considerados não-equivalentes em termos de desempenho quando na verdade o são), por outro

lado tende a aumentar consideravelmente a taxa de erro do Tipo II (probabilidade de dois métodos serem considerados equivalentes em termos de desempenho quando na verdade não o são). Assim, o quesito $AE(0.01)$ pode ser visto como pouco rigoroso na comparação entre critérios de seleção distintos.

Tendo em vista a importância de se adotar bons métodos de seleção de características, nossa recomendação é pelo uso de quesitos mais rigorosos (ou seja, nos quais dois métodos de seleção de atributos sejam considerados como não equivalentes se a evidência contra a hipótese de equivalência for apenas moderada). Por essa razão, recomendamos o uso do quesito $AE(0.1)$ como mais adequado para a comparação entre critérios de seleção.

Uma versão preliminar do presente estudo havia sido realizada por [Bastos et al. 2013], que compararam os desempenhos dos quatro critérios aqui analisados, sobre nove *datasets*. Naquele trabalho, o critério FP havia apresentado resultados superiores ao IE, porém comparáveis ao IG. Todavia, não é possível comparar os resultados daquele trabalho com os resultados aqui obtidos, pois este estudo envolve um número consideravelmente maior de *datasets*.

6. Conclusões

Neste trabalho, apresentamos uma análise empírica detalhada de quatro critérios de seleção de características para *Random Forests* (RFs), sendo dois critérios tradicionais – *Importância Baseada no Erro* (IE) e *Importância de Gini* (IG) [Breiman 2001] – e dois critérios introduzidos – *Fator de Incidência (FI)* e *Fator de Profundidade (FP)* [Bastos et al. 2013].

Os critérios FI e FP são inspirados em uma propriedade fundamental do processo de construção das árvores de decisão: atributos mais relevantes tendem a rotular nós com mais exemplos incidentes e mais próximos à raiz.

O critério FP apresentou, de maneira geral, desempenho consideravelmente superior aos demais critérios, tanto em termos de acurácia como em termos da área sob a curva ROC (AUC). Um aspecto importante é que o critério FP foi pelo menos tão eficiente quanto os demais critérios nos quesitos positivos (frequências de casos em que o critério apresentou o melhor desempenho ou foi estatisticamente equivalente ao melhor), e foi estatisticamente melhor do que os demais no quesito negativo (frequências de casos em que o critério apresentou o pior desempenho).

O critério FP tem ainda a vantagem de ser computacionalmente simples de ser calculado, com custo linear no número total de nós das árvores (custo equivalente ao do IG), não trazendo nenhum impacto significativo no custo computacional de treinamento.

Os critérios IE e IG apresentaram desempenhos bastante similares entre si.

O critério FI, por sua vez, apresentou desempenho bastante inferior aos demais sob os quesitos analisados, não se mostrando portanto um bom critério de seleção.

Os resultados obtidos motivam a realização de diversos estudos futuros, dentre os quais: análises de desempenho em outros contextos além do aprendizado supervisionado (tais como problemas de regressão e aprendizado não supervisionado), comparação com outros critérios desenvolvidos para RFs [Altmann et al. 2010] e análise de desempenho

dos critérios para determinação automática do número de atributos.

Os autores são gratos pelo apoio e financiamento recebidos da EACH-USP, da Pró-Reitoria de Pesquisa da Universidade de São Paulo (Programa Novos Docentes e NAP eScience) e da Fundação de Apoio à Pesquisa do Estado de São Paulo (FAPESP 2011/50761-2 e 2012/04788-9). Parte dos experimentos numéricos foi realizada na Rede *Vision* do IME-USP.

Referências

Altmann, A., Tolosi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Bastos, D. G. D., Nascimento, P. S., and Lauretto, M. S. (2013). Proposta e análise de desempenho de dois métodos de seleção de características para random forests. In *IX Simpósio Brasileiro de Sistemas de Informação*, pages 49–60, João Pessoa.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Freadman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International, CA.

Briand, B., Ducharme, G. R., Parache, V., and Mercat-Rommens, C. (2009). A similarity measure to assess the stability of classification trees. *Comput. Stat. Data Anal.*, pages 1208–1217.

Chen, M., Han, J., and Yu, P. S. (1996). Data mining: An overview from database perspective. *IEEE Xplore Digital Library*, 15(6):866–883.

Coakley, C. W. and Heise, M. A. (1996). Versions of the sign test in the presence of ties. *Biometrics*, 52(4):1242–1251.

Fawcett, T. (2006). An introduction to {ROC} analysis. *Pattern Recognition Letters*, 27(8):861 – 874. {ROC} Analysis in Pattern Recognition.

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *R Journal*, 2(1):53–58.

Frank, A. and Asuncion, A. (2010). Uci machine learning repository.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hand, D. and Till, R. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186.

He, H., III, H. D., and Eisner, J. (2012). Cost-sensitive dynamic feature selection. In *International Conference on Machine Learning (ICML) workshop on Inferning: Interactions be- tween Inference and Learning*, Edinburgh, Scotland.

Inza, I., Calvo, B., nanzas, R. A., Bengoetxea, E., naga, P. L., and Lozano, J. A. (2010). Machine learning: An indispensable tool in bioinformatics. In Matthiesen, R., editor, *Bioinformatics Methods in Clinical Research*, volume 593 of *Methods in Molecular Biology*, chapter 2, pages 25–48. Humana Press.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Microsoft (2006). *SQL Server 2005 Analysis Services Tutorial*.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Redmond, WA.

Morais, D. C. S., Morais, B. C. S., J.V., J. V. M. J., and Gusmão, C. M. G. (2012). Sistema móvel de apoio a decisão médica aplicado ao diagnóstico de asma - intelimed. In *VIII Simpósio Brasileiro de Sistemas de Informação*, São Paulo.

Putter, J. (1955). The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics*, 26:368–386.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Royston, J. P. (1982). An extension of shapiro and wilk's w test for normality to large samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2):115–124.

Apêndice: Tabelas de médias e desvios-padrão de acurácias e áreas sob a curva ROC

Tabela 4. Médias e desvios-padrão das acurácias obtidas pelos quatro critérios de seleção de variáveis, sobre sub-amostras geradas com 25% dos atributos

Dataset	Acurácias (%)							
	IE		IG		FI		FP	
Arrhythmia	64.6	(3.7)	65.1	(4.1)	63.2	(3.0)	71.8	(2.7)
Audiolstand	70.7^{c}	(4.0)	70.6^{b}	(4.7)	71.1^{a}	(4.1)	71.5	(3.8)
Bankmark	90.1	(0.2)	89.0	(0.1)	88.2	(0.2)	88.6	(0.2)
Breastcancer	94.4^{b}	(1.0)	94.3	(1.0)	93.3	(1.3)	94.8	(1.2)
Cardiotoc	97.0^{a}	(0.9)	95.2	(1.0)	77.4	(1.1)	97.1	(1.4)
Cbsonar	79.5	(4.1)	78.0	(4.2)	74.8	(3.8)	78.0	(4.3)
Chesskr	95.7^{a}	(0.5)	95.8	(0.6)	65.2	(0.9)	91.8	(0.7)
Climatefail	93.7	(1.4)	93.7^{a}	(1.3)	91.8	(1.1)	93.5^{a}	(1.3)
Congrvoting	95.5^{a}	(1.2)	95.4^{c}	(1.2)	95.3	(1.1)	95.6	(1.2)
Connect4	73.9	(0.4)	74.4	(0.3)	67.7	(0.2)	72.2	(1.2)
Covertype	79.3	(0.8)	79.6	(1.8)	82.2	(0.3)	82.9	(0.3)
Creditapp	84.8	(1.4)	84.7^{a}	(1.3)	60.3	(1.9)	67.9	(2.4)
Cylbands	78.6	(2.5)	77.6^{c}	(2.2)	75.0	(2.2)	77.2	(2.2)
Dermatology	60.2	(8.3)	86.1	(4.1)	75.8	(3.1)	82.3	(4.5)
Hepatitis	82.3^{a}	(3.5)	82.6	(3.5)	76.6	(3.3)	81.9	(3.5)
Horsecolic	97.4	(1.1)	97.6	(1.2)	97.3	(1.0)	97.5^{a}	(1.2)
Ilpd	68.3	(2.4)	67.9	(2.0)	71.3	(1.8)	68.0	(2.0)
Ionosphere	92.7^{c}	(1.9)	93.0^{a}	(1.9)	93.1	(1.6)	93.1^{a}	(1.9)
Letterrec	66.7	(0.9)	66.3	(1.1)	14.5	(0.5)	69.5	(0.7)
Multiplfeat	98.6^{b}	(0.3)	98.7	(0.3)	97.7	(0.5)	98.7^{a}	(0.3)
Mushroom	99.6	(0.2)	99.8	(0.2)	99.3	(0.1)	99.9	(0.1)
Pageblocks	95.5^{b}	(0.4)	94.9	(0.3)	95.6	(0.3)	95.5	(0.3)
Parkinsons	88.0^{a}	(3.2)	87.9^{a}	(3.1)	87.5^{a}	(2.9)	88.2	(3.2)
Pittbridges	55.6^{b}	(6.2)	57.5^{a}	(6.3)	45.4	(5.9)	57.5	(6.9)
Plannrelax	64.1^{a}	(4.7)	64.7^{a}	(4.8)	64.8	(4.1)	64.8^{a}	(4.6)
Semicond	93.2^{c}	(0.6)	93.3^{a}	(0.6)	93.3	(0.6)	93.2	(0.6)
Soybean	77.5	(4.3)	75.2	(5.7)	52.7	(3.4)	80.8	(3.2)
Spambase	93.6	(0.5)	93.6^{a}	(0.5)	86.0	(0.6)	93.4	(0.5)
Spectf	81.1	(2.6)	81.0^{a}	(2.7)	76.2	(2.6)	80.8^{a}	(2.7)
Stlogaustcred	84.8	(1.6)	84.6^{b}	(1.7)	60.6	(2.3)	68.4	(2.9)
Stloggercred	74.3	(1.4)	74.4^{c}	(1.4)	75.0	(1.4)	74.4	(1.4)
Stlogimageseg	96.5^{a}	(1.5)	96.4^{c}	(1.2)	57.4	(1.5)	96.8	(0.6)
Stlogvehicle	68.3	(2.6)	64.3	(2.2)	63.6	(1.9)	67.7^{b}	(2.3)
Stplatefaults	96.8	(0.6)	97.0	(0.6)	72.6	(1.2)	87.9	(2.9)
Waveform	81.2	(0.8)	83.7^{a}	(0.5)	78.0	(0.7)	83.8	(0.6)
Wine	89.1	(3.7)	94.4^{a}	(2.9)	79.1	(3.7)	94.7	(2.7)

^{a} p-valor > 0.1; ^{b} p-valor > 0.05; ^{c} p-valor > 0.01;

Tabela 5. Médias e desvios-padrão das acurácias obtidas pelos quatro critérios de seleção de variáveis, sobre sub-amostras geradas com 33% dos atributos

Dataset	Acurácias (%)							
	IE		IG		FI		FP	
Arrhythmia	65.6	(3.5)	65.7	(3.8)	65.7	(3.1)	71.2	(3.0)
Audiolstand	70.9	(3.9)	69.7	(4.9)	69.0	(4.6)	72.2	(3.8)
Bankmark	90.5	(0.2)	89.6	(0.1)	88.3	(0.2)	88.8	(0.2)
Breastcancer	94.4	(1.0)	94.4	(1.0)	93.1	(1.3)	95.1	(1.2)
Cardiotoc	98.8	(0.5)	98.6	(0.9)	86.3	(1.0)	99.2	(0.5)
Cbsonar	80.0	(4.3)	79.2	(4.5)	75.9	(3.5)	79.0	(4.3)
Chesskr	96.2^{b}	(0.9)	96.5	(0.8)	71.2	(0.9)	95.3	(0.4)
Climatefail	93.2^{a}	(1.5)	93.2	(1.5)	91.9	(1.3)	93.2^{a}	(1.5)
Congrvoting	95.4^{b}	(1.4)	95.4^{a}	(1.3)	95.2	(1.1)	95.6	(1.2)
Connect4	77.2	(0.3)	77.2^{a}	(0.3)	68.4	(0.2)	74.8	(0.3)
Covertype	81.9	(0.5)	81.2	(1.7)	82.1	(0.3)	83.7	(0.3)
Creditapp	84.7^{a}	(1.3)	84.9	(1.3)	62.2	(2.0)	70.9	(3.4)
Cylbands	79.2	(2.7)	78.5^{c}	(2.4)	74.0	(2.5)	77.9	(2.6)
Dermatology	80.7	(5.2)	93.8	(2.5)	85.8	(2.2)	92.8^{a}	(3.2)
Hepatitis	82.5	(3.5)	82.4^{a}	(3.4)	78.3	(3.2)	82.2^{a}	(3.5)
Horsecolic	97.4^{a}	(1.1)	97.4^{a}	(1.2)	97.2	(1.1)	97.5	(1.2)
Ilpd	70.2^{c}	(2.1)	69.5	(2.1)	70.9	(1.7)	69.5	(1.9)
Ionosphere	93.0^{a}	(1.7)	93.2	(1.7)	93.1^{a}	(1.5)	93.2^{a}	(1.6)
Letterrec	77.4^{a}	(1.6)	77.2	(1.0)	20.3	(0.6)	77.9	(0.7)
Multiplfeat	98.6^{b}	(0.3)	98.7	(0.3)	98.2	(0.4)	98.6^{a}	(0.3)
Mushroom	99.9	(0.1)	99.9	(0.1)	99.3	(0.1)	100.0	(0.1)
Pageblocks	96.7	(0.3)	96.1	(0.5)	95.4	(0.3)	96.9	(0.3)
Parkinsons	88.3^{a}	(3.0)	88.2^{a}	(3.2)	87.4^{c}	(2.9)	88.4	(3.1)
Pittbridges	56.8	(5.8)	60.1^{a}	(6.8)	46.9	(5.3)	60.7	(6.4)
Plannrelax	63.2	(4.1)	64.8^{c}	(4.2)	65.7	(4.2)	64.7^{c}	(4.4)
Semicond	93.3^{a}	(0.6)	93.3	(0.6)	93.3	(0.6)	93.3^{c}	(0.6)
Soybean	82.2^{a}	(3.7)	79.8	(4.5)	60.8	(3.4)	82.6	(3.1)
Spambase	94.1^{a}	(0.4)	94.1^{c}	(0.5)	89.3	(0.5)	94.2	(0.5)
Spectf	81.2^{a}	(2.8)	81.6	(2.6)	78.6	(2.4)	81.3^{b}	(2.7)
Stlogaustcred	85.2^{a}	(1.7)	85.4	(1.6)	67.4	(2.0)	72.3	(4.0)
Stloggercred	75.4	(1.4)	75.6	(1.4)	74.8	(1.4)	75.6^{a}	(1.4)
Stlogimageseg	96.8^{a}	(0.8)	96.6	(0.8)	68.0	(1.4)	96.8	(0.5)
Stlogvehicle	71.6	(2.1)	67.3	(2.0)	71.2^{a}	(1.9)	70.9^{c}	(2.2)
Stplatefaults	97.6	(0.5)	97.9	(0.6)	77.3	(1.2)	91.0	(3.1)
Waveform	83.1	(0.6)	85.2	(0.6)	82.4	(0.6)	85.1^{a}	(0.6)
Wine	91.6	(3.0)	95.3^{a}	(2.6)	82.7	(3.3)	95.7	(2.7)

^{a} p-valor > 0.1; ^{b} p-valor > 0.05; ^{c} p-valor > 0.01;

Tabela 6. Médias e desvios-padrão das acurácias obtidas pelos quatro critérios de seleção de variáveis, sobre sub-amostras geradas com 50% dos atributos

Dataset	Acurácias (%)							
	IE		IG		FI		FP	
Arrhythmia	68.0	(3.5)	68.1	(3.4)	67.5	(3.1)	70.5	(3.1)
Audiolstand	70.2	(4.0)	68.9	(4.8)	67.6	(4.8)	72.1	(4.1)
Bankmark	90.6	(0.2)	90.5	(0.1)	88.4	(0.2)	90.2	(0.1)
Breastcancer	95.2	(1.1)	95.2	(1.0)	93.1	(1.3)	95.7	(1.0)
Cardiotoc	99.7	(0.3)	99.7^{c}	(0.2)	87.1	(1.0)	99.6	(0.3)
Cbsonar	80.8	(4.3)	80.4^{a}	(4.1)	77.3	(3.6)	80.2^{a}	(4.1)
Chesskr	97.5	(0.4)	97.5^{a}	(0.4)	77.3	(0.9)	96.9	(0.5)
Climatefail	92.8	(1.5)	92.8^{a}	(1.6)	91.8	(1.3)	92.8^{a}	(1.5)
Congrvoting	95.5^{b}	(1.3)	95.5^{a}	(1.2)	95.4^{c}	(1.1)	95.7	(1.2)
Connect4	79.4	(0.3)	79.7	(0.3)	75.0	(0.3)	79.3	(0.3)
Covertype	81.8	(0.3)	81.0	(1.3)	79.3	(0.3)	82.5	(0.3)
Creditapp	85.9	(1.3)	86.1^{c}	(1.3)	70.2	(1.6)	86.3	(1.3)
Cylbands	79.7	(2.4)	79.3^{c}	(2.3)	78.5	(2.1)	78.9	(2.3)
Dermatology	90.9	(2.3)	95.8	(1.4)	93.8	(1.4)	96.5	(1.3)
Hepatitis	83.4	(3.8)	82.9^{c}	(3.5)	77.9	(3.7)	83.2^{a}	(3.3)
Horsecolic	97.2^{a}	(1.5)	97.2^{a}	(1.3)	97.2	(1.3)	97.2^{a}	(1.4)
Ilpd	70.7	(2.2)	70.6^{a}	(2.1)	70.3^{a}	(2.0)	70.4^{a}	(2.1)
Ionosphere	93.2	(1.7)	93.1^{a}	(1.6)	93.0^{a}	(1.5)	93.1^{a}	(1.6)
Letterrec	87.5	(0.5)	87.6	(0.5)	56.9	(0.6)	89.1	(0.5)
Multiplfeat	98.6^{c}	(0.3)	98.7	(0.3)	98.4	(0.3)	98.6^{b}	(0.3)
Mushroom	100.0	(0.0)	100.0	(0.0)	99.9	(0.1)	100.0	(0.0)
Pageblocks	97.1^{a}	(0.4)	97.0	(0.3)	96.7	(0.3)	97.3	(0.3)
Parkinsons	88.3^{a}	(3.3)	88.2^{a}	(3.0)	86.6	(3.6)	88.4	(3.0)
Pittbridges	60.2	(6.4)	62.7	(6.2)	53.8	(5.9)	62.5^{a}	(6.3)
Plannrelax	65.2^{a}	(3.9)	65.7	(3.9)	65.2^{a}	(4.2)	65.7^{a}	(3.6)
Semicond	93.3	(0.6)	93.3	(0.6)	93.3	(0.6)	93.3	(0.6)
Soybean	85.2	(3.5)	83.7	(3.9)	73.4	(3.7)	84.8^{b}	(3.0)
Spambase	94.5	(0.5)	94.4	(0.5)	92.0	(0.5)	94.4	(0.5)
Spectf	81.5^{a}	(2.7)	81.6	(2.5)	78.0	(2.7)	81.6^{a}	(2.7)
Stlogaustcred	86.0^{b}	(1.5)	86.3^{a}	(1.5)	70.8	(1.6)	86.3	(1.4)
Stloggercred	75.7	(1.5)	75.5^{a}	(1.5)	75.4^{b}	(1.3)	75.4^{b}	(1.4)
Stlogimageseg	96.7	(0.6)	96.6	(0.6)	93.8	(0.7)	96.7^{a}	(0.6)
Stlogvehicle	73.0	(2.0)	70.3	(1.8)	73.6	(1.8)	72.7	(2.0)
Stplatefaults	98.0	(0.6)	98.3	(0.5)	80.2	(1.1)	95.6	(2.9)
Waveform	83.2	(0.7)	85.5^{b}	(0.6)	85.6	(0.5)	85.5^{a}	(0.5)
Wine	95.2	(2.4)	96.8^{b}	(1.9)	89.3	(3.1)	97.0	(1.7)

^{a} p-valor > 0.1; ^{b} p-valor > 0.05; ^{c} p-valor > 0.01;

Tabela 7. Médias e desvios-padrão das AUCs obtidas pelos quatro critérios de seleção de variáveis, sobre sub-amostras geradas com 25% dos atributos

Dataset	Acurácias (%)							
	IE		IG		FI		FP	
Arrhythmia	68.8	(4.2)	68.9	(4.1)	67.9	(3.9)	73.3	(4.0)
Audiolstand	87.7[a]	(2.5)	87.9[a]	(2.4)	87.4[c]	(2.5)	87.7	(2.7)
Bankmark	69.6	(0.7)	65.4	(0.5)	51.2	(0.2)	61.0	(0.4)
Breastcancer	93.9[c]	(1.2)	93.7	(1.2)	92.6	(1.6)	94.3	(1.4)
Cardiotoc	95.7[a]	(1.3)	90.8	(2.3)	79.4	(1.5)	93.7	(3.8)
Cbsonar	79.4	(4.1)	77.9	(4.2)	74.6	(3.7)	77.9	(4.3)
Chesskr	95.7[a]	(0.5)	95.8	(0.6)	64.4	(0.9)	91.6	(0.7)
Climatefail	66.7	(5.5)	66.8[a]	(5.4)	59.9	(4.5)	66.0[a]	(4.9)
Congrvoting	95.6	(1.3)	95.4[c]	(1.3)	95.3[a]	(1.3)	95.5	(1.2)
Connect4	59.6	(4.7)	62.8	(3.9)	51.7	(0.2)	54.7	(0.8)
Covertype	75.6	(0.8)	76.6	(1.6)	78.9	(0.6)	79.3	(0.6)
Creditapp	85.0	(1.6)	84.8[a]	(1.3)	59.4	(2.0)	67.7	(2.4)
Cylbands	77.2	(2.6)	75.7	(2.4)	72.1	(2.2)	75.2	(2.3)
Dermatology	75.4	(7.7)	88.8	(3.2)	81.2	(3.5)	89.0[a]	(3.5)
Hepatitis	67.9[a]	(5.6)	68.7	(6.3)	53.0	(4.1)	67.7[a]	(6.2)
Horsecolic	86.9	(6.0)	87.5	(6.4)	87.4	(6.1)	87.3[a]	(6.1)
Ilpd	55.7	(3.9)	57.3	(2.4)	50.7	(1.0)	57.1	(2.5)
Ionosphere	91.5[b]	(2.3)	91.8[a]	(2.2)	91.9	(1.8)	91.9[a]	(2.2)
Letterrec	78.1	(1.0)	78.7	(1.0)	56.1	(0.6)	82.4	(0.5)
Multiplfeat	99.2[a]	(0.3)	99.3	(0.2)	98.6	(0.4)	99.3[b]	(0.2)
Mushroom	99.6	(0.2)	99.8	(0.2)	99.3	(0.1)	99.9	(0.1)
Pageblocks	73.3	(4.6)	77.8	(2.4)	76.9	(2.6)	77.1[a]	(2.2)
Parkinsons	81.4[a]	(5.3)	81.1[a]	(5.1)	81.3[a]	(4.8)	81.4	(5.1)
Pittbridges	74.0[a]	(5.1)	72.7[c]	(5.7)	69.2	(4.8)	74.1	(5.5)
Plannrelax	49.7[a]	(4.5)	49.7[a]	(4.2)	49.9	(3.9)	49.8[a]	(3.8)
Semicond	50.1[b]	(0.3)	50.2[a]	(0.5)	50.0	(0.2)	50.2[a]	(0.5)
Soybean	89.9[c]	(2.9)	90.1	(3.1)	76.1	(2.6)	89.3	(2.2)
Spambase	92.9	(0.6)	92.8[a]	(0.6)	84.2	(0.7)	92.7	(0.5)
Spectf	62.9	(4.8)	63.1[a]	(5.0)	51.4	(2.9)	62.8[a]	(5.1)
Stlogaustcred	84.8	(1.6)	84.6[c]	(1.7)	59.6	(2.2)	67.8	(3.0)
Stloggercred	65.7	(2.0)	65.8	(1.9)	66.6	(1.9)	65.7	(1.9)
Stlogimageseg	98.3[a]	(0.6)	98.3[c]	(0.6)	83.3	(0.7)	98.5	(0.3)
Stlogvehicle	85.3	(1.8)	83.6	(1.9)	80.9	(1.5)	85.8[a]	(1.4)
Stplatefaults	97.4	(0.5)	97.6	(0.5)	67.4	(1.5)	87.5	(3.3)
Waveform	84.9	(0.6)	86.2[c]	(0.5)	82.1	(0.6)	86.3	(0.6)
Wine	95.2	(1.7)	97.2[a]	(1.8)	79.6	(4.0)	97.2	(2.0)

[a] p-valor > 0.1; [b] p-valor > 0.05; [c] p-valor > 0.01;

Tabela 8. Médias e desvios-padrão das AUCs obtidas pelos quatro critérios de seleção de variáveis, sobre sub-amostras geradas com 33% dos atributos

Dataset	Acurácias (%)							
	IE		IG		FI		FP	
Arrhythmia	69.0	(4.7)	68.2	(4.1)	68.6	(3.7)	72.5	(4.0)
Audiolstand	87.3[a]	(2.6)	87.0[a]	(2.9)	87.2[a]	(2.9)	87.1	(2.9)
Bankmark	70.2	(0.7)	67.4	(0.5)	50.6	(0.2)	61.9	(1.0)
Breastcancer	93.8	(1.2)	93.8	(1.1)	92.5	(1.6)	94.6	(1.4)
Cardiotoc	97.5	(1.0)	96.3	(2.8)	87.3	(1.5)	98.4	(1.2)
Cbsonar	80.0	(4.2)	79.1	(4.5)	75.8	(3.5)	78.9	(4.3)
Chesskr	96.2[b]	(0.9)	96.5	(0.8)	70.7	(0.8)	95.3	(0.4)
Climatefail	61.4[a]	(4.9)	61.5	(4.9)	55.6	(4.1)	61.4[a]	(4.9)
Congrvoting	95.4[a]	(1.5)	95.4[a]	(1.4)	95.2[a]	(1.3)	95.4	(1.2)
Connect4	67.7	(0.4)	67.7[a]	(0.4)	52.1	(0.2)	62.3	(4.6)
Covertype	78.2	(0.8)	78.1	(1.5)	78.7	(0.7)	80.6	(0.6)
Creditapp	84.7[a]	(1.4)	85.0	(1.3)	60.8	(2.0)	70.6	(3.4)
Cylbands	77.7	(2.9)	76.5	(2.6)	71.0	(2.6)	75.9	(2.7)
Dermatology	85.4	(2.6)	92.4	(3.0)	85.3	(4.0)	93.9	(2.8)
Hepatitis	68.0	(6.1)	68.2[a]	(5.9)	57.9	(5.9)	67.6[a]	(6.4)
Horsecolic	87.5[a]	(6.0)	87.7[a]	(6.0)	87.7	(6.0)	87.9	(5.9)
Ilpd	58.5	(2.7)	58.0	(2.5)	51.1	(1.3)	58.0	(2.5)
Ionosphere	91.8[a]	(1.9)	92.0	(1.9)	92.0[a]	(1.7)	92.0[a]	(1.8)
Letterrec	85.6	(1.5)	85.9	(0.7)	57.6	(0.6)	86.5	(0.7)
Multiplfeat	99.2[a]	(0.2)	99.3	(0.2)	99.0	(0.3)	99.3[a]	(0.2)
Mushroom	99.9	(0.1)	99.9	(0.1)	99.2	(0.2)	100.0	(0.1)
Pageblocks	76.3	(3.4)	76.7	(2.7)	77.0	(2.4)	80.1	(3.6)
Parkinsons	81.6[a]	(5.2)	81.5[a]	(5.2)	81.2[a]	(4.7)	81.5	(5.4)
Pittbridges	76.8[a]	(5.1)	76.1[a]	(5.1)	73.2	(5.2)	77.5	(4.2)
Plannrelax	48.5	(4.1)	49.6[b]	(3.7)	50.8	(3.8)	49.4[c]	(3.7)
Semicond	50.1[a]	(0.3)	50.1[a]	(0.3)	50.0	(0.2)	50.1[a]	(0.3)
Soybean	90.7[c]	(2.4)	91.2	(2.6)	78.2	(3.1)	90.1	(2.2)
Spambase	93.5[c]	(0.5)	93.5[b]	(0.5)	88.3	(0.6)	93.6	(0.5)
Spectf	61.8[c]	(4.7)	62.8	(5.1)	53.4	(2.8)	62.4[a]	(4.5)
Stlogaustcred	85.1[b]	(1.6)	85.3	(1.6)	66.8	(2.0)	71.8	(4.1)
Stloggercred	66.1[c]	(1.9)	66.5	(1.9)	65.4	(1.9)	66.4[a]	(1.8)
Stlogimageseg	98.4[a]	(0.4)	98.4[a]	(0.4)	86.4	(0.7)	98.5	(0.3)
Stlogvehicle	87.8	(1.3)	85.8	(1.5)	86.3	(1.2)	87.6[a]	(1.2)
Stplatefaults	97.8	(0.5)	98.0	(0.6)	72.4	(1.5)	90.6	(3.4)
Waveform	85.9	(0.6)	87.2	(0.6)	86.3	(0.6)	87.2[a]	(0.5)
Wine	96.3	(1.3)	97.7[a]	(1.4)	85.5	(3.4)	97.9	(1.5)

[a] p-valor > 0.1; [b] p-valor > 0.05; [c] p-valor > 0.01;

Tabela 9. Médias e desvios-padrão das AUCs obtidas pelos quatro critérios de seleção de variáveis, sobre sub-amostras geradas com 50% dos atributos

Dataset	Acurácias (%)							
	IE		IG		FI		FP	
Arrhythmia	72.0[a]	(4.5)	70.9[c]	(4.2)	72.2[a]	(4.0)	71.7	(4.0)
Audiolstand	86.8[b]	(2.5)	86.6	(3.1)	86.5	(2.5)	87.3	(2.7)
Bankmark	68.5	(0.7)	70.3	(0.5)	50.4	(0.1)	70.0	(0.6)
Breastcancer	94.7	(1.2)	94.8	(1.2)	92.5	(1.6)	95.2	(1.2)
Cardiotoc	99.4	(0.7)	99.4[a]	(0.5)	86.3	(1.6)	99.2	(0.6)
Cbsonar	80.7	(4.3)	80.3[a]	(4.1)	77.1	(3.6)	80.1[a]	(4.1)
Chesskr	97.5	(0.4)	97.5[a]	(0.4)	77.0	(0.9)	96.9	(0.5)
Climatefail	58.9	(4.6)	58.6[a]	(4.8)	54.4	(3.7)	58.9[a]	(4.8)
Congrvoting	95.4[a]	(1.4)	95.5[a]	(1.3)	95.3[c]	(1.1)	95.6	(1.2)
Connect4	69.9	(0.4)	70.1	(0.4)	58.2	(4.1)	69.6	(0.4)
Covertype	80.1[c]	(0.9)	79.9[a]	(1.2)	75.5	(0.5)	80.1	(0.6)
Creditapp	85.9	(1.3)	86.1[c]	(1.3)	69.9	(1.6)	86.3	(1.3)
Cylbands	78.1	(2.6)	77.4	(2.5)	77.0	(2.2)	76.9	(2.5)
Dermatology	89.8	(3.3)	92.8	(3.0)	92.4	(4.1)	96.0	(2.3)
Hepatitis	69.6	(6.5)	68.7[b]	(6.5)	58.8	(5.3)	69.1[b]	(6.1)
Horsecolic	84.3[b]	(7.7)	84.5[c]	(7.2)	85.4	(7.1)	84.5[c]	(7.4)
Ilpd	59.9	(2.8)	59.5[b]	(2.5)	60.1[a]	(2.7)	59.4[b]	(2.7)
Ionosphere	92.0	(1.9)	91.8[a]	(1.7)	91.7[b]	(1.7)	91.9[a]	(1.7)
Letterrec	92.3	(0.4)	92.5	(0.4)	74.5	(0.5)	93.0	(0.4)
Multiplfeat	99.2[c]	(0.2)	99.3	(0.2)	99.0	(0.3)	99.2[a]	(0.2)
Mushroom	100.0	(0.0)	100.0[a]	(0.0)	99.9	(0.1)	100.0[a]	(0.0)
Pageblocks	80.5	(3.8)	80.1[c]	(4.0)	82.5[a]	(2.7)	82.4	(3.2)
Parkinsons	81.5[a]	(5.6)	81.6[a]	(5.3)	79.2	(5.5)	81.7	(5.3)
Pittbridges	78.1[a]	(4.2)	77.8	(4.1)	74.6	(5.0)	78.3[a]	(4.3)
Plannrelax	49.0[a]	(3.2)	49.2	(3.4)	49.6[a]	(3.6)	49.4[a]	(3.3)
Semicond	50.0[c]	(0.3)	50.1[c]	(0.3)	50.0	(0.1)	50.0	(0.3)
Soybean	92.1	(2.0)	92.0[a]	(2.4)	85.6	(2.8)	91.4	(2.1)
Spambase	93.9	(0.5)	93.8	(0.5)	91.2	(0.6)	93.8	(0.5)
Spectf	61.3[a]	(5.0)	61.8	(4.6)	52.2	(2.8)	61.8[a]	(4.7)
Stlogaustcred	85.8[c]	(1.5)	86.2[a]	(1.5)	70.1	(1.6)	86.3	(1.4)
Stloggercred	66.3	(2.1)	66.0[c]	(2.0)	65.9[a]	(1.8)	65.9[a]	(1.9)
Stlogimageseg	98.4	(0.4)	98.4[c]	(0.3)	96.5	(0.4)	98.4[a]	(0.3)
Stlogvehicle	88.7	(1.1)	87.7	(1.0)	89.1	(1.0)	88.8	(1.1)
Stplatefaults	98.0	(0.6)	98.3	(0.5)	75.7	(1.4)	95.4	(3.2)
Waveform	86.0	(0.6)	87.6[c]	(0.6)	87.7	(0.5)	87.6[a]	(0.5)
Wine	97.9	(1.1)	98.6[b]	(0.9)	92.5	(2.5)	98.7	(0.8)

[a] p-valor > 0.1; [b] p-valor > 0.05; [c] p-valor > 0.01;