Proposta e Análise de Desempenho de Dois Métodos de Seleção de Características para *Random Forests*

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

{denise.bastos,marcelolauretto}@usp.br, poulain.patricia@gmail.com

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)*. Comparative tests indicate that* DF *is a robust criterion, outperforming the* Error-Based Importance (EI) *and performing similarly to the* Gini Importance (GI)*, the two main criteria for* Random Forests *currently in use.*

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 relevantes para a classificação é fundamental, principalmente nos problemas em que os custos de coleta de dados são proporcionais à quantidade de atributos. Neste artigo, propomos dois critérios de seleção de atributos voltados para* Random Forests*, denominados* Fator de Incidência (FI) *e* Fator de Profundidade (FP)*. Testes comparativos indicam que o* FP *é um critério robusto, com desempenho superior ao da* Importância Baseada no Erro (IE) *e equivalente ao da* Importância de Gini (IG) – *os dois principais critérios para* Random Forests *atualmente em uso.*

1. Introdução

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 podem 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].

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.

[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.

Neste artigo, propomos duas novas medidas de importância de atributos, computadas sobre *Random Forests*, e comparamos empiricamente seu desempenho com as medidas IE e IG no contexto de seleção de características.

Na construção das árvores em uma RF, 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 proposta, aqui denominada *Fator de Incidência* (FI), busca medir 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), busca medir 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 descrevemos os experimentos numéricos e os resultados obtidos, e na Seção 5 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 1996a]. 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 $\{\mathcal{L}^{(k)}\}, k = 1, 2, \ldots, K$, cada um consistindo de N observações independentes provenientes da mesma distribuição subjacente ao conjunto $\mathcal{L}$. A idéia central é usar $\{\mathcal{L}^{(k)}\}$ para obter um preditor melhor do que o preditor simples $\psi(\boldsymbol{x}, \mathcal{L})$, tendo como restrição utilizar apenas a seqüência de preditores $\psi(\boldsymbol{x}, \mathcal{L}^{(k)})$. Indexando-se as classes por $c = 1, 2, \ldots, C$, um método de agregar os preditores $\psi(\boldsymbol{x}, \mathcal{L}^{(k)})$ é através de votação, escolhendo para $\boldsymbol{x}$ a classe mais votada entre os preditores. Formalmente, denotando por $N_c = |\{k \in \{1 \ldots K\} : \psi(\boldsymbol{x}, \mathcal{L}^{(k)}) = c\}|$ 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 $\{\mathcal{L}^{(k)}\}, k = 1, 2, \ldots, K$ é feita tomando-se reamostras *bootstrap* de $\mathcal{L}$, via sorteio com repetição, cada qual de tamanho N.

Em cada reamostra de treinamento bootstrap, aproximadamente 37% das instâncias do conjunto original não são utilizadas para o treinamento [Breiman 1996b]. Essas instâncias são usadas como um conjunto de teste, para estimar o erro de cada classificador e, a partir deste, o erro do classificador agregado. O erro *out-of-bag* de cada classificador $\psi(\boldsymbol{x}, \mathcal{L}^{(k)})$ é definido como o percentual do conjunto de teste (constituído por $\mathcal{L} \setminus \mathcal{L}^{(k)}$) classificado erroneamente.

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 $k = 1, 2, \ldots, K$; os atributos avaliados são indexados por $m = 1, 2, \ldots, M$; as classes são indexadas por $c = 1, 2, \ldots, C$. I_k denota o número de nós da k-ésima árvore.

- O par (k, i) denota o i-ésimo nó da k-ésima árvore, $k = 1, 2, \ldots, K$, $i = 1, 2, \ldots, I_k$.

- $T_k = \{(k,1),(k,2),\ldots,(k,I_k)\}$ 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.
- $r(k,i)$ denota o atributo que rotula o i-ésimo nó da k-ésima árvore. Para os nós terminais, define-se $r(k,i) = 0$.
- $d(k,i)$ denota a profundidade do i-ésimo nó da k-ésima árvore, ou seja, o comprimento do caminho da raiz da árvore k até o nó (k,i). Por definição, a profundidade da raiz de uma árvore é 0.

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 :

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

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 a em uma Floresta Aleatória pode ser dada pela soma dos decréscimos nos índices de Gini de todos os nós rotulados por a:

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

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 que estão mais próximos à raiz, e portanto 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 $H_k = \sum_{i \in T_k} H_k(m)$.

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. Experimentos Numéricos

Os experimentos numéricos foram baseados em nove *datasets* públicos obtidos da UCI Machine Learning Repository [Frank and Asuncion 2010], todos com 15 ou mais atributos, mais de 100 exemplos e sem valores faltantes. Os *datasets* de testes selecionados foram *Dermatology, Image Segmentation, Ionosphere, Letter Recognition, Landsat Satellite, Sonar, Vehicle Silhouette, Wave and WDBC (Wisconsin Diagnostic Breast Cancer)*, cujas descrições detalhadas estão disponíveis em [Frank and Asuncion 2010].

O primeiro passo consistiu em construir uma RF para cada *dataset* completo (com todos os atributos), e calcular as importâncias dos atributos sob cada um dos quatro critérios estudados (IE, IG, FI, FP). Dessa forma, para cada critério se obteve um *ranking* dos atributos do *dataset*, em ordem decrescente de importância.

O segundo passo consistiu em selecionar os atributos mais relevantes sob cada critério e comparar os erros *out-of-bag* obtidos pelas RFs sobre os subconjuntos gerados. Mais especificamente, para cada *dataset* foram definidos de 8 a 10 valores distintos de M (sendo M a quantidade de atributos selecionados), dentro da faixa de 15% a 67% da quantidade original de atributos, espaçados 2 a 2. Por exemplo, para o dataset Dermatology, os valores de M definidos foram $M \in \{5, 7, 9, 11, 13, 15, 17, 19, 21, 23\}$. Para cada valor de M e para cada critério, foram geradas 500 sub-amostras aleatórias, cada uma por sorteio simples sem reposição, contendo 60% dos exemplos do dataset original e composta apenas pelos M atributos mais relevantes. Para cada sub-amostra gerada, foi construída uma *Random Forest* e calculado seu respectivo erro *out-of-bag*. O desempenho de cada critério de seleção foi então avaliada pela média dos erros obtidos nas 500 sub-amostras contendo os M atributos mais relevantes.

O ambiente de teste foi implementado na linguagem R [R Core Team 2012], e para a construção e aplicações das RFs utilizou-se o Pacote *randomForest* [Liaw and Wiener 2002].

Nas Figuras 1, 2 e 3 são apresentados os gráficos dos erros *out-of-bag* médios obtidos pelos quatro critérios de seleção, em função da quantidade M de atributos selecionados. Nas legendas dos gráficos também são apresentados os erros médios finais obtidos pelos critérios, calculados sobre todos os valores de M.

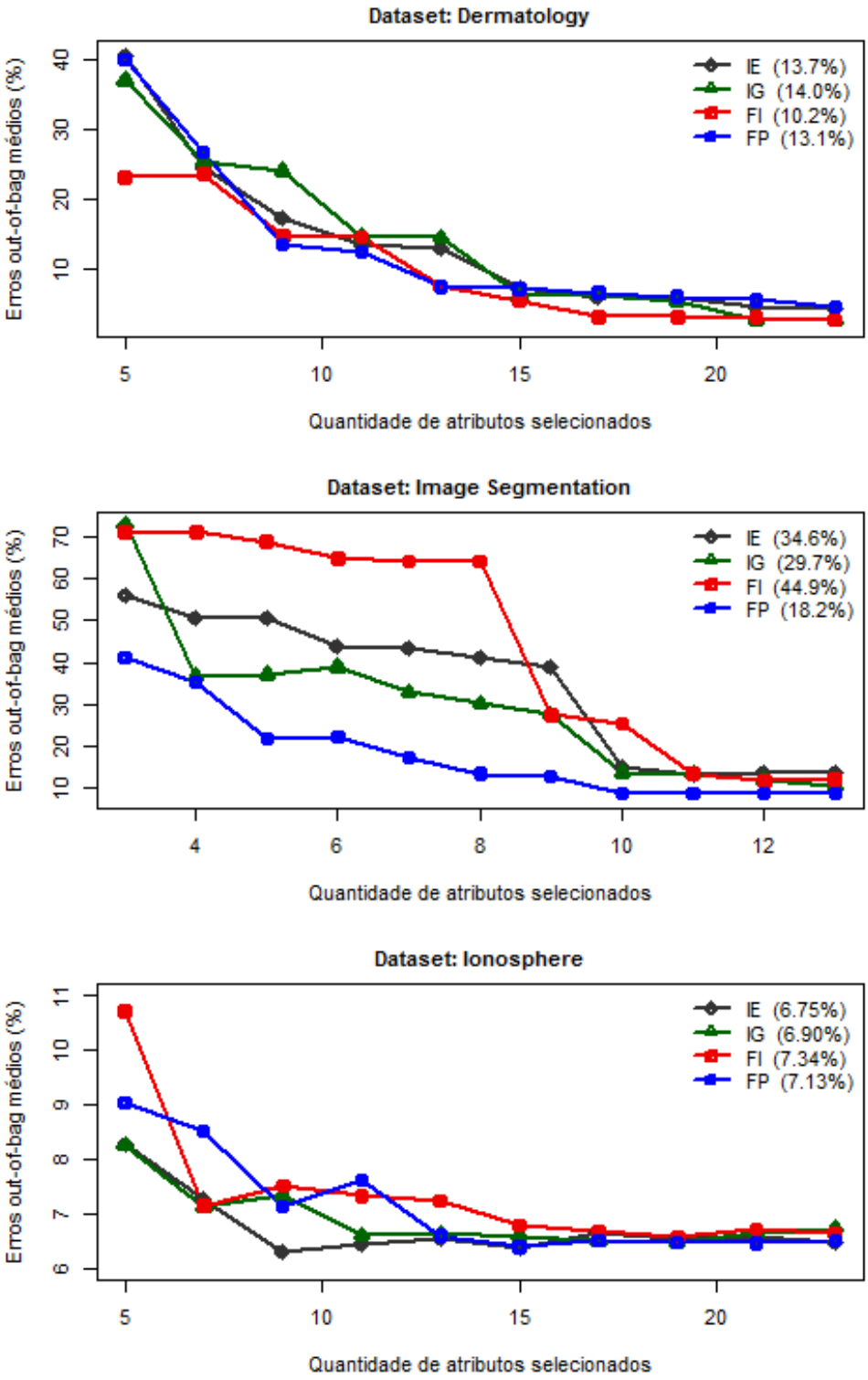

Figura 1. Taxas de erro *out-of-bag* em função da quantidade de atributos selecionados. Datasets: *Dermatology*, *Image Segmentation* e *Ionosphere*

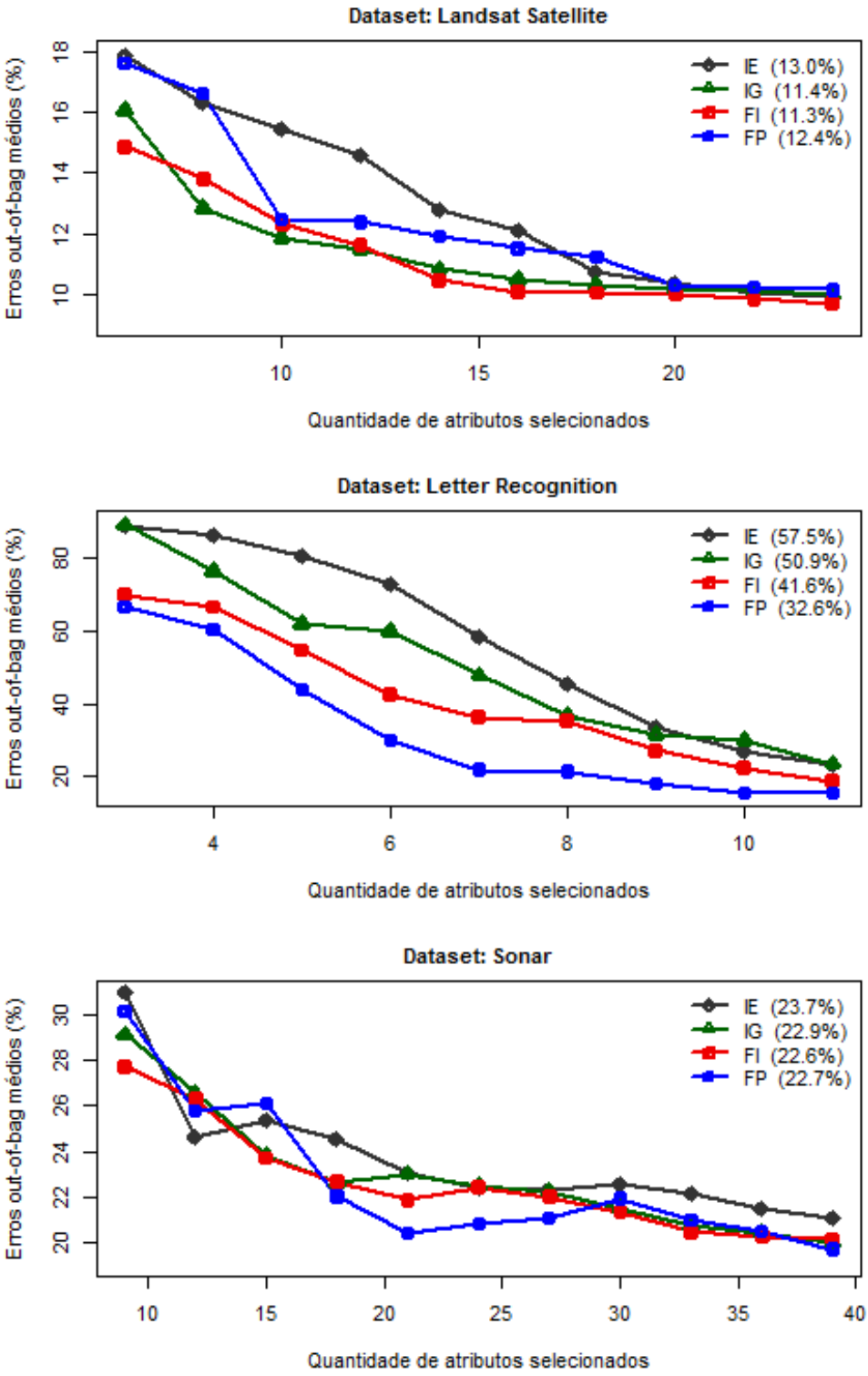

Figura 2. Taxas de erro *out-of-bag* em função da quantidade de atributos selecionados. Datasets: *Landsat Satellite*, *Letter Recognition* e *Sonar*

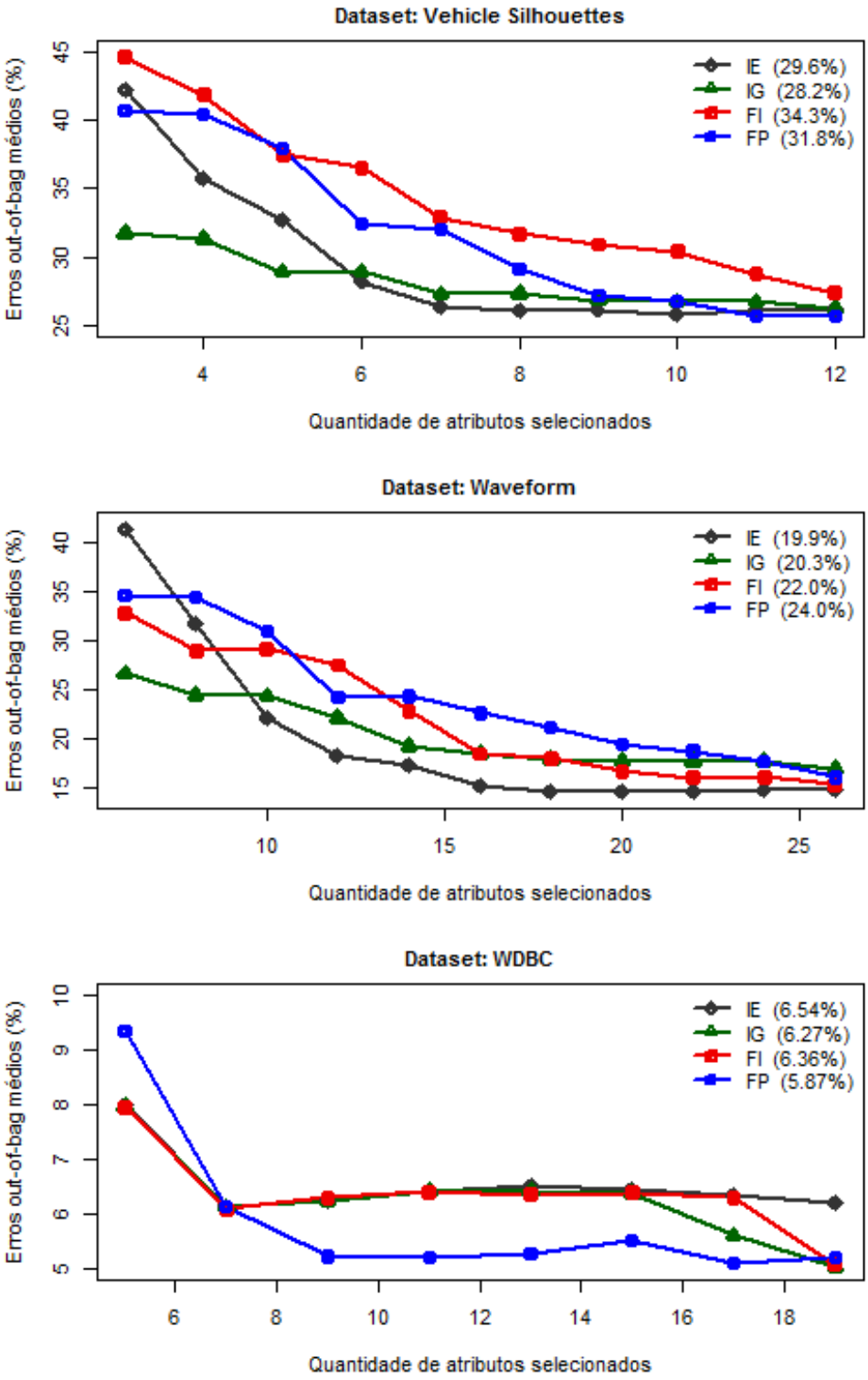

Figura 3. Taxas de erro *out-of-bag* em função da quantidade de atributos selecionados. Datasets: *Vehicle Silhouettes*, *Waveform* e *WDBC*

Nos *datasets* Dermatology, Ionosphere e Sonar, os desempenhos dos quatro critérios são bastante similares entre si, com ligeira vantagem do FI no Dermatology. No *dataset* Image Segmentation, foram observadas diferenças mais acentuadas entre os erros médios, na seguinte ordem de desempenho: FP, IG, IE, FI. No *dataset* Landsat Satellite, o FI e o IG obtiveram os melhores resultados, enquanto o IE obteve os maiores erros médios para alguns valores de M. No *dataset* Letter Recognition também se observaram significativas diferenças de erros entre os quatro critérios, sendo o ranking: FP, FI, IG, IE. Nos *datasets* Vehicle Silhouetes e Waveform, o IE e o IG obtiveram os menores erros médios. No *dataset* WDBC, o critério FP apresentou erros médios inferiores aos demais, para quase todos os valores de M..

A partir dos gráficos não é possível observar uma hegemonia clara de um critério sobre os demais, já que se verificam alternâncias dos desempenhos relativos dos critérios nos diferentes *datasets*. Para uma comparação mais ampla dos desempenhos, a Tabela 1 apresenta o erro médio final de cada critério sobre cada *dataset*. As células sombreadas em cinza escuro e cinza claro indicam, respectivamente, o primeiro e segundo menores erros médios para cada *dataset*.

A partir da Tabela 1, são computados:

- **MD**: Número de *datasets* em que cada critério obteve o melhor desempenho;
- **2MD**: Número de *datasets* em que cada critério ficou entre os dois de melhores desempenhos;
- **PD**: Número de *datasets* em que cada critério obteve o pior desempenho.

Esses indicadores são apresentados na Figura 4. Nota-se que o IE obteve resultados piores do que o FI e o FP, nos três quesitos. Além disso, embora o IE tenha pequena vantagem sobre o IG no quesito MD, apresenta resultados consideravelmente piores nos demais quesitos. Assim, consideramos o IE como o critério de pior resultado entre os quatro. O critério FP é superior ao FI, já que obteve resultados melhores nos quesitos 2MD e PD e apresentou um empate no quesito MD. Os dois critérios com melhores desempenhos parecem ser o IG e o FP. Ambos empatam no critério PD, e cada um apresenta ligeira vantagem sobre o outro nos critérios MD (3 para o FP, 1 para o IG) e 2MD (6 para o IG, 5 para o FP).

Assim, os resultados obtidos neste trabalho indicam que o FP é um bom competidor entre os critérios de seleção, com desempenho superior ao IE e equivalente ao IG. O critério FI, embora tendo apresentado desempenho inferior ao FP e ao IG, obteve resultados ainda superiores ao critério IE, nos três quesitos analisados.

5. Conclusões

Neste trabalho, foram propostas e avaliadas duas medidas de importâncias de atributos desenvolvidas para *Random Forests* (RFs), dentro do contexto de seleção de características em aprendizado supervisionado. Essas medidas, denominadas *Fator de Incidência (FI)* e *Fator de Profundidade (FP)*, são inspiradas 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.

Foram realizados experimentos numéricos baseados em nove problemas de classificação, para comparar o desempenho dessas duas medidas com os desempenhos

Tabela 1. Erros *out-of-bag* médios obtidos pelos critérios de seleção de atributos sobre cada *dataset*

Dataset	Erros médios (%)			
	IE	IG	FI	FP
Dermatology	13.70	14.00	**10.20**	13.10
Image Segmentation	34.60	29.70	44.90	**18.20**
Ionosphere	**6.75**	6.90	7.34	7.13
Landsat Satellite	13.00	11.40	**11.30**	12.40
Letter Recognition	57.50	50.90	41.60	**32.60**
Sonar	23.70	22.90	**22.60**	22.70
Vehicle Silhouettes	29.60	**28.20**	34.30	31.80
Waveform	**19.90**	20.30	22.00	24.00
WDBC	6.54	6.27	6.36	**5.87**

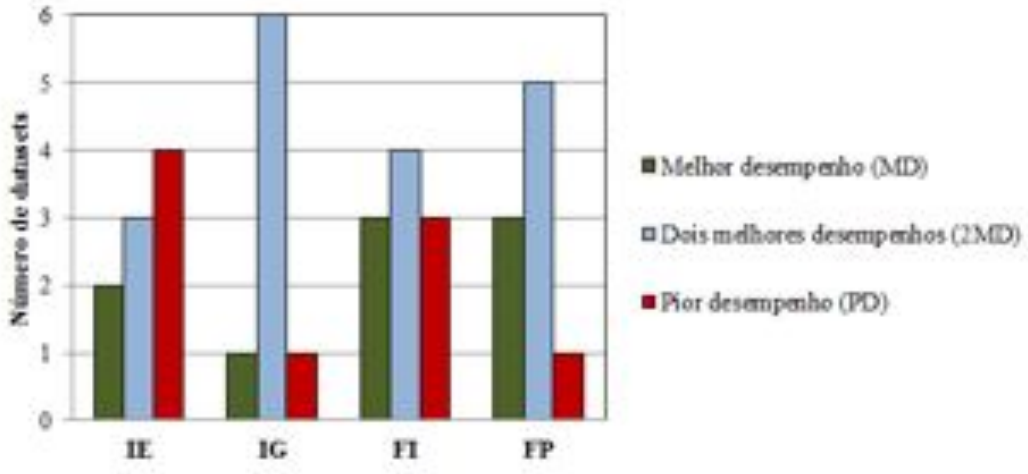

Figura 4. Número de *datasets* nos quais cada critério obteve o melhor desempenho (MD), um dos dois melhores desempenhos (2MD) e o pior desempenho (PD)

da *Importância Baseada no Erro* (IE) e da *Importância de Gini* (IG), ambas propostas por [Breiman 2001]. Os resultados obtidos sugerem que o critério FP é bastante robusto, tendo apresentado resultados superiores ao IE e comparáveis ao IG. O critério FI apresentou resultados inferiores ao FP e ao IG, porém superiores ao IE. Ou seja, o IE, embora intuitivo e um dos critérios mais utilizados atualmente, foi o que apresentou o pior desempenho entre os quatro.

Os critérios FI e FP são facilmente computáveis, 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 resultados obtidos motivam a realização de diversos estudos futuros, dentre os quais: comparações envolvendo um número maior de *datasets*; análises de desempenho em outros contextos além do aprendizado supervisionado, tais como problemas de regressão e aprendizado não supervisionado; comparação dos critérios propostos com ou-

tros critérios desenvolvidos para RFs [Altmann et al. 2010]; análise de associação entre o desempenho dos critérios e as características dos *datasets*.

Os autores são gratos pelo apoio e financiamento recebidos da EACH-USP, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Apoio à Pesquisa do Estado de São Paulo (FAPESP).

Referências

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

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

Breiman, L. (1996b). Out-of-bag estimation. Technical report, Technical report, Statistical Department, University of California Berkeley, Berkeley CA.

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.

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.

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.

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

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.