Novas Medidas de Relevância para Seleção *Lazy* de Atributos *

Douglas B. Pereira[1], **Alexandre Plastino**[1], **Rafael B. Pereira**[1]
Bianca Zadrozny[2], **Luiz Henrique de C. Merschmann**[3], **Alex A. Freitas**[4]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

[2]IBM Research Brasil
Rio de Janeiro – RJ – Brasil

[3]Instituto de Ciências Exatas e Biológicas – Universidade Federal de Ouro Preto (UFOP)
Ouro Preto – MG – Brasil

[4]Computing Laboratory – University of Kent
Canterbury – Kent – United Kingdom

douglasuff@vm.uff.br, {plastino,rbarros}@ic.uff.br, biancaz@br.ibm.com
luizhenrique@iceb.ufop.br, a.a.freitas@kent.ac.uk

Abstract. *Attribute selection is a data preprocessing step used to identify attributes relevant to the classification task. Recently, a lazy technique which postpones the choice of attributes to the moment an instance is submitted to classification was proposed. In the original lazy technique proposal, a measure based on the entropy concept was presented to evaluate the quality of the attributes. In this work, we propose four new measures, based on: the chi-square statistic test, the Cramer coefficient, the Gini index and the gain ratio concept. Experimental results show the relevance of this proposal since, for a large number of datasets, the best performance of the lazy selection strategy was achieved when the new measures were used.*

Resumo. *Seleção de atributos é uma etapa de pré-processamento que identifica atributos relevantes para a tarefa de classificação. Recentemente, foi proposta uma técnica lazy que adia a escolha dos atributos ao momento em que uma instância é submetida à classificação. Na proposta da técnica lazy original, uma medida baseada no conceito de entropia foi apresentada para avaliar a qualidade dos atributos. Neste trabalho, propõem-se outras quatro medidas, baseadas: no teste estatístico chi-quadrado, no coeficiente de Cramer, no índice Gini e no gain ratio. Experimentos evidenciam a importância dessa proposta uma vez que, para um número significativo de bases de dados, o melhor desempenho da seleção lazy foi obtido com a utilização das novas medidas.*

1. Introdução

Diversos estudos evidenciam que o desempenho de estratégias de classificação está diretamente relacionado, entre outros fatores, à qualidade dos dados da base de treinamento [Guyon et al. 2006, Liu and Motoda 2008]. Atributos redundantes e irrelevantes

*Trabalho resultante de pesquisa parcialmente financiada pelo CNPq e pela FAPERJ.

podem, não somente prejudicar a acurácia de um classificador, mas tornar o processo de construção do modelo ou a execução do algoritmo de classificação mais lento. Técnicas de seleção de atributos visam eliminar da base de treinamento atributos que não contribuem ou mesmo prejudicam o desempenho de um classificador. Desse modo, para bases que contêm atributos redundantes ou irrelevantes, técnicas de seleção de atributos são fundamentais para melhorar a qualidade dos dados que serão utilizados pelo classificador.

Tradicionalmente, técnicas de seleção de atributos são executadas na fase de pré-processamento dos dados e suas decisões são definitivas para a fase de construção do modelo ou classificação propriamente dita. Porém, em [Pereira et al. 2008, Pereira et al. 2011], foi proposta uma técnica de seleção de atributos cuja característica principal é adiar a seleção dos atributos relevantes – de forma *lazy* – ao momento em que uma instância é submetida ao processo de classificação, ao invés de se fazer a seleção previamente – de forma *eager*. Essa proposta tem como hipótese que o conhecimento dos valores dos atributos da instância a ser classificada pode contribuir para a identificação dos melhores atributos para aquela instância em particular. Dessa forma, para diferentes instâncias a serem classificadas, subconjuntos distintos de atributos, e customizados para cada instância, poderão ser selecionados. Nessa proposta, foi utilizado, como medida de qualidade dos valores dos atributos, o conceito de entropia da distribuição de classes.

Neste trabalho, serão propostas e avaliadas quatro novas medidas de qualidade de atributos para serem utilizadas com a abordagem *lazy* de seleção de atributos. Essas novas medidas são baseadas nos seguintes conceitos: teste estatístico chi-quadrado [Liu and Setiono 1995, Han and Kamber 2006], no coeficiente de Cramer [Spiegel 1993], no índice Gini [Breiman et al. 1984, Han and Kamber 2006] e no *gain ratio* [Quinlan 1986, Han and Kamber 2006]. Cabe informar que a proposta baseada no teste estatístico chi-quadrado evoluiu de uma versão preliminar apresentada em [Menezes et al. 2009].

O principal objetivo deste trabalho é mostrar a importância da utilização de outras medidas de relevância de atributos, junto à abordagem de seleção *lazy*, diferentes daquela baseada no conceito de entropia, definida em [Pereira et al. 2008, Pereira et al. 2011]. Os experimentos permitem verificar que, considerando-se diferentes bases de dados, com características variadas, cada uma das medidas avaliadas – a original e as quatro novas – pode levar, isoladamente, a técnica de seleção *lazy* a atingir seu melhor desempenho.

O restante deste trabalho está organizado da seguinte forma. Na Seção 2, realiza-se uma revisão sobre seleção de atributos e medidas de relevância de atributos comumente utilizadas. A seleção *lazy* de atributos e a adaptação do conceito de entropia para o contexto *lazy* são revisadas na Seção 3. As quatro novas medidas para seleção *lazy* são propostas na Seção 4. Na Seção 5, são analisados os experimentos computacionais e, finalmente, na Seção 6, são apresentadas as conclusões e direções para trabalhos futuros.

2. Seleção de atributos

Estratégias de seleção de atributos são categorizadas como: embutidas, *wrapper* ou do tipo filtro [Guyon et al. 2006, Liu and Motoda 2008]. Técnicas embutidas aparecem incorporadas ao algoritmo de classificação e realizam a seleção de atributos no seu processo de treinamento. Um exemplo típico são os algoritmos de indução de árvores de decisão, pois selecionam os atributos que farão parte da árvore. Técnicas *wrapper* e filtro pro-

curam pelo subconjunto de atributos mais adequado que será utilizado pelo algoritmo de classificação. No caso da seleção *wrapper*, o próprio algoritmo de classificação é utilizado para avaliar a qualidade dos atributos. Técnicas filtro são independentes do algoritmo de classificação e avaliam a qualidade dos atributos por meio de medidas específicas.

Uma maneira de se medir a qualidade de um atributo para classificação é avaliar o seu grau de associação com a classe. Para tanto, existem diversas medidas, que serão descritas a seguir. Considere que essas medidas sejam empregadas em uma base de dados $D(A_1, A_2, ..., A_n, C), n \geq 1$, com $n+1$ atributos, onde C é o atributo classe, com domínio $\{c_1, c_2, ..., c_m\}, m \geq 2$. Assume-se que os valores dos atributos e da classe são discretos.

2.1. Entropia

O conceito de entropia tem origem na área de teoria da informação e pode ser especificado da seguinte forma. A entropia da distribuição de classes em D, representada por $Ent(D)$ é dada por [Han and Kamber 2006]:

$$Ent(D) = -\sum_{i=1}^{m}[p_i \times log_2(p_i)], \tag{1}$$

onde p_i, $1 \leq i \leq m$, é a probabilidade de ocorrência de cada valor c_i da classe C, dada pela razão entre o número de instâncias em que c_i ocorre e o número total de instâncias.

Seja A um atributo de D, com domínio $\{a_1, a_2, ..., a_k\}, k \geq 1$, e seja D_{a_i} a partição de D composta por todas as instâncias cujo valor de A é igual a a_i. A entropia da distribuição de classes em D, condicionada aos valores do atributo A, representada por $Ent(D, A)$, é definida por:

$$Ent(D, A) = \sum_{i=1}^{k}\left[\left(\frac{|D_{a_i}|}{|D|}\right) \times Ent(D_{a_i})\right]. \tag{2}$$

Quanto mais informativo um atributo A for em relação à classe C, menor será a entropia condicional $Ent(D, A)$.

2.2. Chi-quadrado

A medida Chi-quadrado (ou χ^2) avalia a qualidade de um atributo de acordo com a sua correlação com a classe por meio de um teste estatístico χ^2. Para cada valor a_i do atributo A ($1 \leq i \leq k$) e para cada valor c_j da classe C ($1 \leq j \leq m$), existe uma frequência esperada quando ($A = a_i$) e ($C = c_j$), que pode ser calculada pela fórmula [Han and Kamber 2006]:

$$e_{ij} = \frac{count(A = a_i) \times count(C = c_j)}{n}, \tag{3}$$

onde n é o número total de instâncias, $count(A = a_i)$ é o número de instâncias em que ocorre o valor a_i do atributo A, e $count(C = c_j)$ é o número de instâncias que pertencem à classe c_j. A partir da frequência esperada de todas as combinações de valores i e j, pode-se calcular a medida χ^2 pela fórmula:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\left[\frac{(o_{ij} - e_{ij})}{e_{ij}}\right], \tag{4}$$

onde o_{ij} é a frequência observada da conjunção $(A = a_i) \wedge (C = c_j)$, ou seja, a razão entre o número de instâncias em que ocorre o valor a_i do atributo A simultaneamente com o valor c_j da classe C, e o número total de instâncias da base. O teste estatístico pode ser aplicado para determinar se o atributo A e a classe C são independentes, usando o número de graus de liberdade dado por $(k-1) \times (m-1)$, que é proporcional ao número de ocorrências distintas de valores de A multiplicado pela cardinalidade da classe C. Se essa hipótese puder ser rejeitada, de acordo com um nível de significância estatística pré-determinado e uma distribuição χ^2, significa que o atributo A é fortemente relacionado à classe C [Han and Kamber 2006].

2.3. Coeficiente de *Cramer*

O valor calculado pela medida Chi-quadrado está diretamente relacionado ao número de instâncias de uma base e à cardinalidade dos atributos. Já o coeficiente de *Cramer* não é afetado pelo tamanho da amostra, sendo muito útil quando se suspeita que um aumento significativo do Chi-quadrado é resultante do grande tamanho da amostra, em vez de uma relação entre os atributos. Esse coeficiente é interpretado como uma medida da correlação entre dois atributos e varia de 0 a 1. Quanto mais próximo de 0, menor a correlação e, quanto mais próximo de 1 maior a correlação. O valor do coeficiente de *Cramer* (V) de um atributo A é calculado pela fórmula [Spiegel 1993]:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}}, \tag{5}$$

onde χ^2 é o valor de Chi-quadrado de A, n o é total de instâncias da base e q é o mínimo entre o número de valores distintos do atributo e a cardinalidade da classe.

2.4. Índice *Gini*

O índice *Gini* é uma medida estatística de desigualdade (ou impureza), comumente utilizado para calcular a desigualdade de distribuição de renda. É utilizado na seleção de atributos no algoritmo CART [Breiman et al. 1984] de árvores de decisão, com o objetivo de encontrar o atributo que contém o melhor particionamento de valores. Esse índice é máximo (pior) quando as instâncias de uma determinada partição estão igualmente distribuídas entre todas as classes, e o índice é mínimo (melhor) quando todas as instâncias pertencem a uma única classe. O índice *Gini* de cada valor do atributo A é dado por:

$$Gini(D, A, a_i) = 1 - \sum_{j=1}^{m} (p_{j|i})^2, \tag{6}$$

onde m é o número de valores distintos da classe e $p_{j|i}$ é a probabilidade da instância pertencer à classe c_j $(1 \leq j \leq m)$ quando ocorre o valor a_i em A, dada pela razão entre o número de instâncias da classe c_j em que a_i ocorre e o número de ocorrências de a_i.

O índice *Gini* do atributo A é dado por [Breiman et al. 1984]:

$$Gini(D, A) = \sum_{i=1}^{k} p_i \times Gini(D, A, a_i), \tag{7}$$

onde k é o número de valores distintos do atributo A, p_i $(1 \leq i \leq k)$ a razão entre o número de instâncias da base em que ocorre o valor a_i para o atributo e o número total de instâncias e $Gini(D, A, a_i)$ é dado pela Equação 6.

2.5. *Gain ratio*

A medida *Gain ratio* [Quinlan 1986] é baseada no conceito de entropia, mais especificamente, no ganho de informação, ou seja, na diferença entre a entropia da classe e a entropia da classe condicionada a determinado atributo. Na tentativa de não priorizar atributos que têm muitos valores diferentes, a medida faz um tipo de normalização, que é a divisão do ganho de informação pelo $SplitInfo(D, A)$, definido por [Han and Kamber 2006]:

$$SplitInfo(D, A) = -\sum_{k=1}^{i} w_i \times log_2(w_i), \tag{8}$$

onde w_i pode ser descrito como o peso da partição que contém os valores a_i do atributo A, dado pela razão entre número de ocorrências de a_i e o total de instâncias da base.

O *Gain ratio* pode ser obtido através da fórmula [Han and Kamber 2006]:

$$GainRatio(D, A) = \frac{Ent(D) - Ent(D, A)}{SplitInfo(D, A)}, \tag{9}$$

onde $Ent(D)$ é a entropia da distribuição de classes em D (Equação 1) e $Ent(D, A)$ é a entropia de classe em D condicionada ao atributo A (Equação 2).

3. Seleção *lazy* de atributos

Estratégias convencionais de seleção de atributos – que neste trabalho serão referenciadas como estratégias de seleção *eager* – são executadas na etapa de pré-processamento e os atributos não selecionados são descartados da base de dados, não participando do processo de classificação. Proposta em [Pereira et al. 2008], a estratégia de seleção de atributos *lazy* baseia-se na hipótese de que o conhecimento dos valores dos atributos da instância a ser classificada pode contribuir para o processo de escolha dos atributos da base mais adequados para a classificação dessa instância em particular. A seleção de atributos, para cada instância, é, portanto, realizada apenas no momento da sua classificação. Para diferentes instâncias, subconjuntos distintos de atributos podem ser selecionados.

Essa proposta de seleção *lazy* de atributos pode ser aplicada em conjunto com qualquer classificador do tipo *lazy*, tais como o k-NN e o Naive Bayes, ou com qualquer versão *lazy* de classificadores tradicionalmente *eager*, como por exemplo as árvores de decisão *lazy* [Friedman et al. 1996].

O conceito definido na Equação 2 é usado pela estratégia *eager* de seleção de atributos conhecida como *Information Gain Attribute Ranking* [Yang and Pedersen 1997] para medir a capacidade de um atributo discriminar os valores da classe. Para a seleção *lazy* de atributos, proposta em [Pereira et al. 2008, Pereira et al. 2011], define-se, então, a entropia da distribuição de classes em D, restrita ao valor de a_i do atributo A, representada por $Ent(D, A, a_i)$, da seguinte forma:

$$Ent(D, A, a_i) = Ent(D_i), \tag{10}$$

onde a entropia $Ent(D_i)$ pode ser obtida através da Equação 1.

A Equação 10 é usada pela estratégia de seleção *lazy* para medir a capacidade de um valor especifico a_i, de um atributo particular A, de discriminar as classes. Quanto

mais próxima a entropia $Ent(D, A, a_i)$ é de zero, maior é a chance de o valor de a_i do atributo A determinar bem alguma classe.

Os parâmetros de entrada da estratégia *lazy* são: uma base de dados D com n atributos, uma instância $I = (v_1, v_2, ..., v_n)$ e um número $r, 1 \leq r \leq n$, que representa o número de atributos a serem selecionados. Para selecionar os r melhores atributos para classificar a instância I, os n atributos são avaliados baseados na medida *lazy* proposta em [Pereira et al. 2008, Pereira et al. 2011] (Equação 11) que estabelece que, para cada atributo A_j $(1 \leq j \leq n)$, se a capacidade discriminatória do valor específico v_j de A_j $(Ent(D, A_j, v_j))$ é melhor que (menor que) a capacidade de discriminação geral do atributo A_j $(Ent(D, A_j))$, então a primeira será considerada como medida *lazy* de A_j. Dessa forma, a medida proposta para avaliar a qualidade de cada atributo A_j é definida por:

$$Ent_{Lazy}(D, A_j, v_j) = Min(Ent(D, A_j, v_j), Ent(D, A_j)), \tag{11}$$

onde Min() retorna o menor entre seus parâmetros.

Após calcular o valor $Ent_{Lazy}(D, A_j, v_j)$ para cada atributo A_j, a estratégia *lazy* selecionará os r atributos que apresentem os r menores valores dessa medida.

O desempenho dessa proposta foi avaliado quando utilizada em conjunto com o algoritmo de classificação *k-NN* [Dasarathy 1991]. A partir de uma base de dados D, uma instância I e um valor de k, sendo $k \geq 1$, basicamente, o algoritmo *k-NN* atribui a I a classe majoritária entre os seus k elementos mais próximos. A distância entre I e as instâncias de D é calculada a partir de uma função definida sobre os valores dos atributos das respectivas instâncias. Dessa forma, a utilização da seleção *lazy* de atributos, em conjunto com o algoritmo *k-NN*, por exemplo, implica que o cálculo da distância entre instâncias poderá ser realizado utilizando-se diferentes subconjuntos de atributos, para diferentes instâncias de entrada.

A seleção de atributos *lazy* baseada no conceito de entropia apresentou resultados interessantes, demonstrando ser uma técnica bastante promissora. Para permitir uma avaliação mais abrangente, tentando dar evidências de que a técnica *lazy* pode ter um bom desempenho independentemente do tipo de medida utilizada, neste trabalho, propõem-se quatro novas medidas de relevância de atributos, que serão apresentadas a seguir.

4. Novas medidas para seleção lazy

As quatro novas medidas para seleção *lazy* de atributos, propostas neste trabalho, estão baseadas nos seguintes conceitos: teste estatístico chi-quadrado [Liu and Setiono 1995, Han and Kamber 2006], o coeficiente de Cramer [Spiegel 1993], o índice Gini [Breiman et al. 1984, Han and Kamber 2006] e a medida gain ratio [Quinlan 1986, Han and Kamber 2006].

4.1. Chi-quadrado *lazy*

No caso da seleção *lazy* de atributos, a medida chi-quadrado deve levar em consideração a capacidade de cada valor da instância a ser classificada determinar o atributo classe. Portanto faz-se necessária uma adaptação das Equações 3 e 4 para que o valor χ^2 possa ser utilizado também na abordagem *lazy* de seleção de atributos. Neste trabalho, propõe-se a Equação 12 que avalia o quanto o valor a_{ji} do atributo A_j está correlacionado com o

atributo classe C:

$$\chi^2(D, A_j, v_j) = \sum_{k=1}^{m} \frac{(o_k - e_k)}{e_k}, \tag{12}$$

onde m é o número de valores distintos do atributo classe C, o_k é a frequência observada do par $\{A_j = v_j, C = c_k\}$ e e_k é a frequência esperada do mesmo par. Quanto maior for o valor de χ^2, maior a chance de a_{ji} ser um bom determinante de uma classe.

A frequência esperada do par $\{A_j = v_j, C = c_k\}$ é definida na Equação

$$e_k = \frac{count(A_j = v_j) \times count(C = c_k)}{N}, \tag{13}$$

onde $count(A_j = v_j)$ é o número de instâncias em que ocorre o valor v_j do atributo A_j, $count(C = c_k)$ é o número de instâncias que pertencem à classe c_k e N é o número total de instâncias da base.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos pela medida χ^2_{Lazy} definida pela Equação 14, onde χ^2 é o valor de χ^2 calculado para o atributo A_j (Equação 4) e Max() é a função que retorna o maior valor entre seus parâmetros:

$$\chi^2_{Lazy}(D, A_j, v_j) = Max\left(\frac{\chi^2}{n_j}, \chi^2(D, A_j, v_j)\right), \tag{14}$$

onde n_j que é o número de valores distintos de A_j. Quanto maior o valor de χ^2_{Lazy}, maior a correlação entre o valor v_j da instância a ser classificada e o atributo classe C.

Como o valor de χ^2 é obtido a partir do acumulo dos n_j valores do domínio do atributo A_j e o valor de $\chi^2(D, A_j, v_j)$ é obtido por uma única parcela desta soma (referente ao valor v_j), existe a necessidade de torná-los comparáveis. Para tanto, foi introduzido o denominador n_j. Após calcular o valor χ^2_{Lazy} para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r maiores valores de χ^2_{Lazy}.

4.2. Coeficiente de *Cramer lazy*

A proposta de adaptação do coeficiente de Cramer para a abordagem *lazy* é representada pela Equação 15 que avalia o quanto o valor v_j do atributo A_j está correlacionado com o atributo classe C:

$$V(D, A_j, v_j) = \sqrt{\frac{\chi^2(D, A_j, v_j)}{N(q-1)}}, \tag{15}$$

onde $\chi^2(D, A_j, v_j)$ é o valor calculado de Chi-quadrado para o valor v_j do atributo A_j dado pela Equação 12, N o é total de instâncias e q é o mínimo entre o número de valores distintos do atributo A_j e a cardinalidade da classe C. Quanto maior for o valor de $V(D, A_j, v_j)$, maior a chance de v_j ser um bom determinante de uma classe.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos de acordo com a medida V_{Lazy} definida pela Equação 16, onde χ^2 é o valor de χ^2 calculado para o atributo A_j (Equação 4).

$$V_{Lazy}(D, A_j, v_j) = Max\left(\sqrt{\frac{\chi^2/n_j}{N(q-1)}}, V(D, A_j, v_j)\right) \tag{16}$$

Quanto maior o valor de V_{Lazy}, maior correlação entre o valor v_j e o atributo classe C.

Houve a necessidade de introduzir o denominador n_j na parcela referente ao valor de V obtido pela técnica *eager* pelo mesmo motivo apresentado no caso da medida Chi-quadrado *lazy*. Após calcular o valor V_{Lazy} para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r maiores valores de V_{Lazy}.

4.3. Índice *Gini lazy*

A adaptação *lazy* proposta neste trabalho para o índice Gini é definida por:

$$Gini(D, A_j, v_j) = 1 - \sum_{k=1}^{m} (p_k)^2, \tag{17}$$

onde m é o número de valores distintos do atributo classe, p_k é a razão entre o número de instâncias da base que pertencem à classe c_k em que ocorre o valor v_j, e o número de ocorrências do valor v_j do atributo A_j.

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos de acordo com a medida $Gini_{Lazy}$ definida pela Equação 18, onde $Gini(D, A_j)$ é o índice *Gini* calculado para o atributo A_j (Equação 7).

$$Gini_{Lazy}(D, A_j, v_j) = Min(Gini(D, A_j), Gini(D, A_j, v_j)) \tag{18}$$

Quanto menor o valor de $Gini_{Lazy}$ menor a desigualdade na relação dos valores do atributo A_j com os da classe C.

Após calcular o valor $Gini_{Lazy}$ para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r menores valores de $Gini_{Lazy}$.

4.4. *Gain ratio lazy*

No caso da estratégia *lazy*, a adaptação proposta para o cálculo do $SplitInfo(D, A_j, v_j)$ é definida por:

$$SplitInfo(D, A_j, v_j) = -log_2(w_j), \tag{19}$$

onde w_i é definido pela razão entre número de ocorrências de v_j no atributo A_j e o total de instâncias da base.

A partir das definições anteriores, pode-se calcular o valor de $GainRaito(D, A_j, v_j)$ a partir da Equação:

$$GainRatio(D, A_j, v_j) = \frac{Ent(D) - Ent(D, A, v_j)}{Split(D, A_j, v_j)}. \tag{20}$$

A seleção dos r melhores atributos para classificar a instância I é realizada a partir da avaliação dos n atributos pela medida $GainRatio_{Lazy}$ definida pela Equação 21.

$$GainRatio_{Lazy}(D, A_j, v_j) = Max(GainRatio(D, A_j), GainRatio(D, A_j, v_j)) \tag{21}$$

Um maior valor de $GainRatio_{Lazy}$ indica um aumento no ganho de informação da base obtido pela escolha do atributo A_j considerando o valor v_j.

Após calcular o valor $Gain_{Lazy}$ para os n atributos, a estratégia *lazy* selecionará os r atributos que apresentam os r maiores valores de $Gain_{Lazy}$.

5. Resultados Experimentais

Para realizar os experimentos computacionais, a estratégia de seleção *lazy* foi implementada utilizando cada uma das medidas propostas e integrada à ferramenta Weka (Waikato Environment for Knowledge Analysis) [Waikato 2009]. O efeito da seleção de atributos foi avaliado utilizando-se o classificador *k-NN*. A seleção ocorre quando o classificador recebe uma nova instância a ser classificada. Uma vez que, para cada nova instância, um subconjunto distinto de atributos pode ser considerado pelo classificador, os atributos não selecionados pela estratégia *lazy* para uma dada instância não são removidos do conjunto de dados, mas somente desconsiderados pelo classificador.

Os experimentos foram realizados com 40 bases de dados, que possuem uma ampla variedade de tamanho, complexidade e áreas de aplicação. Essas bases foram retiradas do repositório da UCI [Asuncion and Newman 2007]. O valor ausente de um atributo A em uma instância t foi substituído pela moda dos valores de A das instâncias da classe de t, no caso de A categórico, ou pela média dos valores de A das instâncias da classe de t, no caso de A contínuo. As medidas utilizadas para avaliar a qualidade dos atributos necessitam de valores de atributos discretos. Por esse motivo foi adotado um método supervisionado de discretização baseado em ganho de informação, proposto em [Fayyad and Irani 1993], para discretizar os atributos contínuos dessas bases. Sem atributos contínuos, não houve a necessidade do processo de normalização de atributos.

O algoritmo de classificação utilizado foi o *k-NN*, especificamente, a implementação da ferramenta Weka conhecida como *IBk*. Cabe ressaltar que o código do *k-NN* teve que ser adaptado para considerar, no cálculo da distância entre elementos, apenas os atributos selecionados de forma *lazy*.

Os resultados obtidos estão sumarizados na Tabela 1. A seleção *lazy* de atributos foi utilizada com o classificador *k-NN*, com k igual a 1, para as 40 bases de dados, em conjunto com cada uma das cinco medidas: ENT (entropia), CHI (chi-quadrado), CRV (coeficiente de Cramer), GIN (índice Gini) e GRT (*gain ratio*). Na primeira coluna, estão apresentados o nome da base de dados e, entre parênteses, respectivamente, o número de atributos, o número de classes e o número de instâncias da base. A última coluna representa a execução do classificador sem a execução da seleção de atributos, ou seja, com todos os atributos da base de dados. Para cada base de dados, o melhor comportamento está indicado por valores em negrito. Caso represente um melhor desempenho obtido isoladamente, o valor estará sublinhado.

Cada valor da tabela representa a melhor acurácia obtida pelo classificador, quando utilizada a seleção *lazy* e a medida em questão, ao se variar o percentual de atributos selecionados de 10% a 90% do total de atributos, com um incremento regular de 10%. Cada valor de acurácia é obtido a partir da média das acurácias resultantes do processo de validação cruzada com dez partições [Han and Kamber 2006], na qual cada partição foi obtida de maneira aleatória. A acurácia correspondente a cada execução do classificador é uma média dos valores obtidos em cada partição. As mesmas partições foram utilizadas nas execuções com as diferentes medidas.

Inicialmente, observa-se a importância da seleção de atributos: para apenas uma base, a *pendigits*, o desempenho do classificador sem seleção foi o melhor obtido sem que nenhuma outra estratégia tivesse levado ao mesmo valor. Para as outras 39 bases, a

Tabela 1. Comparativo entre medidas lazy usando 1-NN

Base	ENT	CHI	CRV	GIN	GRT	Sem sel.
anneal (38,5,898)	**99,55**	99,44	99,22	99,44	**99,55**	99,22
audiology (69,24,226)	76,99	78,32	77,88	77,88	**78,76**	76,11
autos (25,6,205)	87,32	**89,27**	86,83	87,80	88,78	85,85
breast-cancer (9,2,286)	74,13	**75,52**	**75,52**	73,08	73,43	69,93
breast-w (9,2,699)	**97,14**	**97,14**	**97,14**	97,00	95,99	**97,14**
chess-Kr-vs-Kp (36,2,3196)	**96,81**	96,50	96,50	96,75	96,21	96,56
credit-a (15,2,690)	85,51	**86,38**	**86,38**	85,22	86,23	82,32
diabetes (8,2,768)	**77,99**	74,61	74,61	68,49	73,05	76,43
flags (29,8,194)	60,82	60,82	61,34	61,34	**65,98**	59,79
glass (9,6,214)	**77,10**	**77,10**	**77,10**	**77,10**	**77,10**	**77,10**
heart-cleveland (13,2,303)	**82,84**	**82,84**	**82,84**	82,51	81,52	80,53
heart-hungarian (13,2,294)	81,63	82,65	82,65	81,63	**83,67**	80,27
hepatitis (19,2,155)	86,45	85,81	85,81	86,45	**87,10**	83,87
horse-colic (27,2,368)	83,70	**84,78**	**84,78**	84,24	**84,78**	78,53
hypo-thyroid (29,4,3772)	**96,98**	93,61	93,61	93,19	93,45	91,52
ionosphere (34,2,351)	**94,59**	93,45	93,45	94,30	93,16	92,59
labor (16,2,57)	**100,00**	98,25	98,25	**100,00**	**100,00**	96,49
letter-recog (16,26,20000)	91,93	**92,09**	90,06	91,95	91,86	91,87
lymph (18,4,148)	**85,14**	83,78	**85,14**	84,46	84,46	82,43
mol-bio-promot (57,2,106)	**89,62**	87,74	87,74	88,68	86,79	80,19
mol-bio-splice (60,3,3190)	90,72	88,21	88,21	87,49	**90,88**	73,32
mushroom (22,2,8124)	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
optdigits (64,10,5620)	94,84	94,73	94,73	**94,86**	94,52	94,25
pendigits (16,10,10992)	96,79	96,63	96,55	96,71	96,74	**97,05**
postoperative (8,3,90)	**71,11**	70,00	70,00	**71,11**	**71,11**	63,33
primary-tumor (17,21,339)	42,48	**43,07**	40,71	41,30	41,89	38,35
solar-flare1 (12,6,323)	71,52	70,90	70,28	**72,14**	70,59	65,94
solar-flare2 (12,6,1066)	**76,27**	75,89	74,95	**76,27**	75,05	73,45
sonar (60,2,208)	**86,54**	82,21	82,21	82,21	83,17	**86,54**
soybean-large (35,19,683)	**93,41**	91,80	92,09	93,27	91,51	92,24
spambase (57,2,4601)	93,68	93,61	93,61	**93,72**	**93,72**	92,98
statlog-heart (13,2,270)	**85,19**	84,07	84,07	**85,19**	84,07	84,07
statlog-segment (19,7,2310)	94,68	94,68	94,33	94,68	**94,76**	94,68
statlog-vehicle (18,4,846)	71,39	**71,87**	70,80	71,28	70,33	70,92
thyroid-sick (29,2,3772)	97,48	**97,83**	**97,83**	97,59	97,56	97,45
vote (16,2,465)	**96,09**	95,63	95,63	**96,09**	95,17	92,18
vowel (13,11,990)	89,80	89,80	88,18	**89,90**	89,80	89,80
waveform-5000 (40,3,5000)	75,52	75,54	75,54	**76,34**	75,86	73,82
wine (13,3,178)	**98,88**	**98,88**	97,75	**98,88**	**98,88**	98,31
zoo (17,7,101)	97,03	**98,02**	96,04	96,04	97,03	96,04
Num.Vitórias	19	14	9	13	14	5
Num.Vitórias.Isoladas	6	5	0	4	6	1

seleção de atributos com alguma das medidas obteve resultado melhor do que sem seleção em 35 bases, ou resultado igual ao do classificador sem seleção em quatro bases.

Na base da tabela, a linha com o rótulo *Num.Vitórias* representa o número de vezes em que a medida em questão levou o classificador a obter o melhor desempenho. A linha *Num.Vitórias.Isoladas* representa o número de vezes em que a medida levou o classificador ao melhor desempenho isoladamente, isto é, sem que nenhuma outra medida

obtivesse o mesmo desempenho.

O objetivo principal deste trabalho era evidenciar que a seleção de atributos *lazy* pode se beneficiar de outras medidas de relevância de atributo além da medida ENT, proposta originalmente. Observa-se então que a medida ENT, apesar de ter levado o classificador a obter o melhor desempenho em 19 bases, levou ao melhor desempenho isoladamente em apenas seis bases de dados. Para outras 15 bases, o melhor desempenho foi obtido isoladamente com a utilização de uma das três medidas: CHI (cinco bases), GIN (quatro bases) ou GRT (seis bases). Verifica-se ainda que em outras cinco bases (*breast-cancer*, *credit-a*, *horse-colic*, *spambase* e *thyroid-sick*) duas ou mais das novas medidas obtiveram resultado melhor do que a medida ENT. Observa-se que apenas a medida CRV não foi capaz de obter um melhor desempenho de forma isolada e contribuir para a confirmação da importância da existência de medidas alternativas à ENT.

Para algumas bases de dados, a acurácia obtida por uma das medidas propostas foi significativamente maior do que a obtida com a medida ENT. Por exemplo, as diferenças absolutas entre as acurácias obtidas pelas medidas GRT e ENT, para as bases *audiology*, *flags* e *heart-hungarian* foram, respectivamente: 1,77; 5,16 e 2,04.

Uma vez que se tenha uma base de treinamento sobre a qual será realizada a tarefa de classificação, a disponibilidade de um conjunto amplo de medidas permite que se faça uma avaliação, através de técnicas como validação cruzada ou *leave-one-out*, da medida mais adequada à base e, dessa forma, se obtenham melhores acurácias preditivas.

Vale informar que experimentos também foram realizados utilizando-se o classificador *k-NN* com os valores de k igual a 3 e 5. Porém, como o desempenho do *k-NN* com k igual a 1 foi, de maneira geral, superior ao das outras duas parametrizações e essas não permitiram conclusões adicionais, por economia de espaço, decidiu-se apresentar apenas a análise dos melhores resultados obtidos.

6. Conclusão

Neste trabalho, propôs-se um novo conjunto de medidas para seleção *lazy* de atributos baseadas no teste chi-quadrado, no coeficiente de *Cramer*, no índice *gini* e na medida *gain ratio*. O objetivo principal era disponibilizar diferentes medidas de relevância que pudessem ser adotadas pela abordagem de seleção *lazy*, além da originalmente definida em [Pereira et al. 2008, Pereira et al. 2011], baseada no conceito de entropia.

As avaliações das novas medidas foram realizadas a partir de 40 bases de domínio público adotadas frequentemente em experimentos de mineração de dados, obtidas do repositório da UCI [Asuncion and Newman 2007].

Os resultados experimentais mostraram que, para um número significativo de bases, as medidas propostas levaram o classificador a obter as melhores acurácias. Com isso, o paradigma de seleção *lazy* de atributos conta agora com um número maior de medidas de relevância de atributos, permitindo que, para diferentes bases de dados, medidas distintas e mais apropriadas sejam adotadas.

Pretende-se, mais adiante, adaptar, para o contexto de seleção *lazy*, medidas que avaliam subconjuntos de atributos, tais como o *Correlation-based Feature* [Hall 2000] e o *Consistency-based Feature* [Liu and Setiono 1996], que medem a capacidade de um determinado conjunto de atributos discriminar os valores do atributo classe.

Referências

Asuncion, A. and Newman, J. (2007). Uci machine learning repository. http://www.ics.uci.edu/ mlearn/mlrepository.html.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029.

Friedman, J., Kohavi, R., and Tun, Y. (1996). Lazy decision trees. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI'96)*, pages 717–724.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., editors (2006). *Feature Extraction, Foundations and Applications*. Springer.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Procs. of 17th Intl. Conf. on Machine Learning*, pages 359–366.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition.

Liu, H. and Motoda, H. (2008). *Computational Methods of Feature Selection*. Chapman & Hall/CRC.

Liu, H. and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *7th Intl. Conference on Tools with Artificial Intelligence*, pages 388–391.

Liu, H. and Setiono, R. (1996). A probabilistic approach to feature selection: A filter solution. In *Procs. of the 13th Intl. Conference on Machine Learning*, pages 319–327.

Menezes, R., Plastino, A., Zadrozny, B., Pereira, R., Merschmann, L. H., and Freitas, A. (2009). Avaliação de uma nova medida para seleção *lazy* de atributos baseada no teste chi-quadrado. In *Anais do V Workshop em Algoritmos e Aplicações de Mineração de Dados (WAAMD 2009/SBBD 2009)*, pages 58–65.

Pereira, R., Plastino, A., Zadrozny, B., Merschmann, L., and Freitas, A. (2008). Seleção lazy de atributos – uma nova perspectiva. In *Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados (WAAMD 2008/SBBD 2008)*, pages 1–9.

Pereira, R., Plastino, A., Zadrozny, B., Merschmann, L., and Freitas, A. (2011). Lazy attribute selection – choosing attributes at classification time. *Intelligent Data Analysis*, to appear.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

Spiegel, M. R. (1993). *Estatística*. Makron Books.

Waikato (2009). Weka (waikato environment for knowledge analysis) machine learning project http://www.cs.waikato.ac.nz/ml/weka/.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Procs. of the 14th Intl. Conf. on Machine Learning*, pages 412–420.