Algoritmos Genéticos Multi-objetivo para a Seleção de Atributos

Newton Spolaôr[1,2,3], **Ana Carolina Lorena**[1], **Huei Diana Lee**[2,4]

[1]Grupo Interdisciplinar de Mineração de Dados e Aplicações (GIMDA)
Universidade Federal do ABC (UFABC) – Santo André, Brasil

[2]Laboratório de Bioinformática (LABI)
Universidade Estadual do Oeste do Paraná (UNIOESTE) – Foz do Iguaçu, Brasil

[3]Laboratório de Inteligência Computacional (LABIC)
Universidade de São Paulo (USP) – São Carlos, Brasil

[4]Serviço de Coloproctologia, Faculdade de Ciências Médicas (FCM)
Universidade Estadual de Campinas (UNICAMP) – Campinas, Brasil

{newtonspolaor,aclorena,hueidianalee}@gmail.com

Abstract. *The occurrence of irrelevant and/or redundant features in Databases can degrade the performance of computational processes for knowledge extraction, motivating the application of a Feature Selection process. Multi-objective Genetic Algorithms can help identifying subsets of features which optimize combinations of possibly conflicting feature importance measures. This paper presents the use of Multi-objective Genetic Algorithms in Feature Selection, investigating the use of different combinations of feature importance criteria in both labeled and unlabeled datasets.*

Resumo. *A ocorrência de atributos irrelevantes e/ou redundantes em Bases de Dados pode prejudicar o desempenho de processos computacionais de extração de conhecimento, o que motiva a aplicação da tarefa de Seleção de Atributos. Os Algoritmos Genéticos Multi-objetivo podem contribuir para identificar subconjuntos de atributos que otimizam combinações entre diferentes medidas ou critérios de importância de atributos, eventualmente conflitantes. Este trabalho apresenta o uso de Algoritmos Genéticos Multi-objetivo para a Seleção de Atributos, investigando o uso de distintas combinações de critérios de importância de atributos em dados rotulados e não-rotulados.*

1. Introdução

O progresso tecnológico tem possibilitado a construção de conjuntos de dados cada vez maiores, em distintas áreas, sob a forma de Bases de Dados (BD). Processos como a Mineração de Dados (MD) podem ser aplicados nas BD para gerar modelos computacionais que representem padrões presentes nos dados e auxiliem no processo de tomada de decisão [Han and Kamber 2006]. Esses modelos podem ser gerados pelo uso de técnicas de Aprendizado de Máquina (AM) [Mitchell 1997], em que um indutor, como um algoritmo de classificação ou agrupamento (*clustering*), constrói inferências (hipóteses) sobre os padrões presentes nos dados.

A Seleção de Atributos (SA) é uma tarefa de pré-processamento dentro do processo de MD. Ela abrange algoritmos para a identificação de subconjuntos de atributos que melhor representem os padrões inerentes aos dados. Esses algoritmos podem ser genericamente categorizados conforme a interação que realizam com o algoritmo de indução usado na extração de padrões. A abordagem filtro considera propriedades inerentes aos dados, como medidas estatísticas. Algoritmos da abordagem *wrapper* avaliam a importância de atributos com o auxílio do próprio algoritmo de indução de modelos.

De modo geral, a SA pode ser vista como um processo de busca combinatorial por subconjuntos de atributos importantes conforme um ou mais critérios, o que motiva a aplicação de métodos heurísticos como os Algoritmos Genéticos mono-objetivo (AG) e os Algoritmos Genéticos Multi-objetivo (AGM) neste processo. Os AGM possibilitam definir compromissos entre distintas medidas de importância e investigar se a combinação entre medidas, destacadas isoladamente, permite identificar atributos melhores.

Estudos na literatura recente envolvem a aplicação dos AGM ao problema de SA [Bruzzone and Persello 2009, Santana et al. 2009, Wang and Huang 2009]. A partir dos avanços do estado da arte correlato, apresenta-se neste trabalho um método baseado na aplicação de AGM ao problema de SA, combinando diferentes critérios que medem a importância dos atributos considerando a abordagem filtro. O método implementado nesse trabalho apresenta flexibilidade para incorporar com facilidade distintas medidas de avaliação individual ou de subconjuntos de atributos e o suporte para otimizar combinações entre essas medidas. Três medidas são adaptadas para tratar também atributos QL. Esses aspectos possibilitam realizar neste trabalho a SA em conjuntos de dados com propriedades variadas, como a presença ou ausência do atributo-classe e a ocorrência de atributos Quantitativos (QT) ou Qualitativos (QL), diferentemente de [Spolaôr et al. 2011, Wang and Huang 2009, Zaharie et al. 2007].

Este trabalho integra o projeto Análise Inteligente de Dados [Spolaôr et al. 2011, Lee et al. 2006] (AID), o qual é desenvolvido por meio de uma parceria entre a UFABC, o LABI/UNIOESTE, o LABIC/USP e o Serviço de Coloproctologia/UNICAMP. O trabalho é organizado do seguinte modo: na Seção 2 o AGM para SA proposto é apresentado. Os materiais e métodos usados na avaliação experimental desse algoritmo são apresentados na Seção 3. As avaliações experimentais realizadas, bem como os resultados alcançados, são abordados na Seção 4. A conclusão é descrita na Seção 5.

2. Algoritmos Genéticos Multi-objetivo na Seleção de Atributos

A aplicação da SA parte usualmente da representação do conjunto de dados em uma Tabela Atributo-Valor (TAV). Cada linha da TAV refere-se a um exemplo j, enquanto cada coluna corresponde a um atributo i de tipo QT (com valores exclusivamente numéricos) ou QL (não numéricos) para $j = \{1, \cdots, n\}$ e $i = \{1, \cdots, m\}$, em que n é a quantidade de exemplos e m é a quantidade de atributos. O atributo-classe $m + 1$ contém a classe dos exemplos se o conjunto de dados for rotulado. A identificação de um subconjunto S com m' atributos ($1 \leq m' \leq m$) pela SA permite definir uma projeção da TAV original em que os exemplos são descritos apenas pelos atributos selecionados em S.

A tarefa de SA pode ser formulada como um processo de busca por um subconjunto de atributos importantes, em termos de relevância e/ou não-redundância, que auxilie na identificação de padrões inerentes aos dados. A importância de subconjuntos de atri-

butos pode ser definida a partir de diferentes critérios [Liu and Motoda 1998], possivelmente conflitantes. Neste trabalho propõe-se a aplicação de AGM para SA, otimizando simultaneamente diferentes medidas de importância de atributos.

2.1. Medidas de Importância de Atributos

As medidas de importância de atributos existentes podem ser divididas em categorias conforme o tipo de propriedade que investigam [Liu and Motoda 2008]. Medidas de consistência podem reduzir a ocorrência de exemplos com valores similares nos atributos e distintos no atributo-classe. Critérios de dependência enfatizam a identificação de atributos correlacionados. Medidas de distância valorizam atributos que destacam propriedades espaciais dos dados. Critérios de informação buscam amenizar a incerteza presente nos dados. Por fim, a categoria de precisão considera alguma medida de desempenho do algoritmo de indução para estimar a importância dos atributos, sendo em geral associadas à abordagem *wrapper*. Neste trabalho foca-se no estudo e na aplicação das medidas descritas a seguir conforme a abordagem filtro.

Medida de Consistência: a medida Pares de exemplos Inconsistentes (PI) possibilita estimar o grau de consistência inerente a uma projeção correspondente ao subconjunto de atributos investigado [Arauzo-Azofra et al. 2008]. Esse grau é dado pela razão da quantidade de PI para a quantidade total de pares de exemplos da TAV. Essa medida pode ser utilizada para atributos QT e QL por meio de adaptações simples.

Medidas de Dependência: a Correlação Atributo-classe (CC) tem por intuito a investigação das mudanças nos valores dos atributos e das diferenças nos rótulos de classe [Zaharie et al. 2007]. A medida original é apresentada na Equação 1, sendo $x_j\,(i)$ o valor do atributo i no exemplo j, $\|.\|$ a função módulo e $C\,(i)$ uma formulação que destaca atributos com valores divergentes em classes distintas. Para dados QL, essa formulação utiliza $overlap(x_{j1}\,(i)\,,x_{j2}\,(i))$ no lugar da diferença $x_{j1}\,(i) - x_{j2}\,(i)$. A distância *overlap* retorna diferença 1 ou 0, respectivamente, entre atributos com valores distintos ou iguais [Wilson and Martinez 1997]. O peso w_i assume o valor 1 ou 0, respectivamente, se i é ou não é selecionado e $\phi\,(.,.) = 1$ se j_1 e j_2 pertencem a classes diferentes ou $\phi\,(.,.)$ =-0,05 caso contrário.

$$CC = \left(\sum w_i C\,(i)\right) / \left(\sum w_i\right) \qquad (1)$$
$$em\ que\ C\,(i) = \frac{\sum_{j_1 \neq j_2} \|x_{j1}\,(i) - x_{j2}\,(i)\|\,\phi\,(x_{j1},x_{j2})}{n\,(n-1)\,/2}$$

O critério Intra-Correlação (IC) avalia a correlação de Pearson c_p existente entre os atributos de um subconjunto S [Wang and Huang 2009]. Na Equação 2 a correlação global do subconjunto S é normalizada considerando $C\,(m',2)$ como a quantidade total de combinações entre os m' atributos selecionados em S tomados dois a dois. Essa medida pode ser aplicada tanto a conjuntos de dados rotulados quanto não-rotulados.

$$IC = \frac{1}{C\,(m',2)} \sum_{i_l=1}^{m'} \sum_{i_2=i_1+1}^{m'} |c_p\,(x\,(i_1)\,,x\,(i_2))| \qquad (2)$$

Medidas de distância: a Distância Inter-Classe (IE) mensura, por meio da Equação 3, a separabilidade existente entre b classes por meio da distância média entre: (1) o centróide de cada classe c_l, isto é, o exemplo obtido pela média aritmética dos exemplos de c_l e (2) o centróide dos dados $\vec{c}$ [Zaharie et al. 2007]. Nessa equação, $d(.,.)$ denota uma medida de distância e $\vec{c_{c_l}}$ e n_{c_l} representam, respectivamente, o centróide e a quantidade de exemplos em c_l. Para dados com atributos QT e QL empregou-se, respectivamente, a distância Euclidiana e a *overlap*.

$$IE = \frac{1}{n} \sum_{c_l=1}^{b} n_{c_l} d\left(\vec{c_{c_l}}, \vec{c}\right) \tag{3}$$

O *Laplacian Score* (LS) inspira-se no fato de que os exemplos da mesma classe em geral situam-se relativamente próximos uns aos outros. O poder de preservação de localidade pode ser explorado nesses casos para a SA [He et al. 2005]. A medida LS propõe identificar atributos que melhor atendem a estrutura geométrica local dos dados, modelada em um grafo de vizinhos mais próximos conforme a Equação 4. Nessa equação, $\vec{x}\left(i\right) = \left[x_1\left(i\right), x_2\left(i\right), \cdots, x_n\left(i\right)\right]^T$, $\mathbf{1} = \left[1, \cdots, 1\right]^T$ e a matriz D e o grafo Laplaciano L são definidos como $D = diag\left(G\mathbf{1}\right)$ e $L = D - G$. Nessas formulações, a função $diag(.)$ extrai a diagonal de uma matriz e G corresponde à matriz de pesos das arestas do grafo.

$$LS\left(i\right) = \frac{\tilde{x}\left(i\right)^T L \tilde{x}\left(i\right)}{\tilde{x}\left(i\right)^T D \tilde{x}\left(i\right)} \tag{4}$$
$$em\ que\ \tilde{x}\left(i\right) = \vec{x}\left(i\right) - \frac{\vec{x}\left(i\right)^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1}$$

A definição de vizinhança de cada exemplo pode ser realizada entre k vizinhos mais próximos ou entre todos os exemplos de sua classe. O primeiro tipo de vizinhança é aplicável tanto em dados rotulados quanto não-rotulados pois não requer a informação fornecida pelo atributo-classe.

Medida de Informação: a medida *Representation Entropy* (RE), apresentada na Equação 5, mensura a redundância de atributos de um subconjunto S com auxílio de formulações envolvendo os auto-valores extraídos de uma matriz de covariância [Yan 2007]. Esta matriz é construída a partir da projeção do conjunto de dados investigada sem utilizar o atributo-classe, o que agrega a esta medida a aplicabilidade a dados não-rotulados. A informação é distribuída uniformemente na projeção correspondente a S e há pouca redundância nos dados quando todos os auto-valores obtidos são iguais. O atributo i pode representar toda a informação dessa projeção se apenas o auto-valor λ_i for diferente de 0.

$$RE = -\sum \tilde{\lambda}_i \log \tilde{\lambda}_i \tag{5}$$
$$em\ que\ \tilde{\lambda}_i = \frac{\lambda_i}{\sum \lambda_i}$$

2.2. Algoritmos Genético Multi-objetivo

Os AGM vêm sendo considerados como uma alternativa para tratar o problema de SA. Neste processo, busca-se subconjuntos de atributos que otimizem um ou mais critérios

de importância. Uma grande parte dos trabalhos encontrados na literatura usam alguma medida de precisão dos modelos obtidos com o uso de atributos selecionados, configurando uma abordagem *wrapper*. Este trabalho apresenta como principal diferencial a investigação de múltiplas combinações multi-objetivo de medidas de importância de atributos conforme a abordagem filtro. Também diferentemente de trabalhos correlatos [Spolaôr et al. 2011, Wang and Huang 2009, Zaharie et al. 2007], foram analisados conjuntos de dados, rotulados e não-rotulados, com presença de atributos QT e QL. A adaptação das medidas PI, CC e IE para tratar atributos QL também contribui na flexibilidade do método. Esses fatores evidenciam a generalidade do método apresentado frente a diferentes tipos de dados e análises.

Utilizou-se para a otimização multi-objetivo o AGM *Non-Dominated Sorting Genetic Algorithm* (*NSGA-II*) [Deb et al. 2000], o qual se baseia na teoria de Pareto. Os indivíduos das populações do AGM são iniciados randomicamente e representam distintos subconjuntos de atributos, por meio da codificação binária. Cada subconjunto corresponde portanto a um cromossomo binário com m genes, de modo que um valor 1 no gene i indica a seleção do atributo i e um valor 0 no mesmo gene implica que este atributo não é selecionado. Os operadores genéticos utilizados foram o cruzamento de um ponto e a mutação por *bit flip*. O procedimento de seleção realizado foi o torneio binário.

O *NSGA-II* identifica um conjunto de soluções equivalentemente ótimas ao final do processo evolutivo, ou seja, diferentes subconjuntos de atributos representando distintos compromissos entre as medidas de importância combinadas. Para selecionar um único subconjunto de atributos entre eles, empregou-se a técnica relativamente simples *Compromise Programming* [Zeleny 1973], que propõe a seleção do indivíduo que possui a menor distância para uma solução ideal.

Conforme apresentado na seção anterior, foram selecionadas seis medidas de importância de atributos para serem investigadas como objetivos a serem otimizados. A escolha dessas medidas foi realizada de maneira a se ter ao menos um representante de cada categoria de importância de atributos, com exceção da de precisão. Todas são diretamente aplicáveis a dados rotulados com atributos QT. Três delas (PI, CC e IE) foram adaptadas para lidar também com atributos QL. E as medidas IC, LS e RE são aplicáveis a conjuntos de dados não-rotulados. Nota-se também que as medidas RE e IC oferecem suporte para a análise de redundância de atributos durante a SA, enquanto medidas como PI e IE permitem investigar a relevância dos atributos em relação à classe.

Todas as medidas, exceto a LS, medem a importância de subconjuntos de atributos. A medida LS avalia cada atributo individualmente. Optou-se então em aplicar esta medida para cada atributo e utilizar o valor médio obtido como o valor final para um determinado subconjunto de atributos.

Convém ressaltar que a otimização das medidas PI, IC e LS envolve a minimização dos valores objetivo correspondentes, enquanto que os critérios CC, IE e RE devem ser maximizados por definição. Transformou-se os três problemas de maximização em minimizações equivalentes, de maneira a obter uma uniformidade do método.

O método utilizado suporta atualmente a otimização de combinações multi-objetivo tomadas dois a dois e três a três entre os critérios mencionados na Seção 2.1. A sua flexibilidade também permite adicionar novas medidas e gerar combinações consi-

derando mais objetivos com simplicidade. Desse modo, torna-se possível a investigação experimental de complementaridades variadas entre esses critérios de importância.

3. Materiais e Métodos

Um conjunto de experimentos foi realizado para avaliar o AGM implementado na tarefa de SA. A seguir são apresentados os materiais e métodos empregados neste processo.

3.1. Conjuntos de Dados

Selecionou-se conjuntos de dados investigados em trabalhos correlatos [Spolaôr 2010] do repositório UCI [Asuncion and Newman 2007]: *Australian* (A), *Crx* (C), *Dermatology* (D), *German* (G), *Ionosphere* (I), *Lung cancer* (L), *Promoter* (P), *Sonar* (S), *Soybean small* (Y), *Vehicle*[1] (V), *Wisconsin Breast cancer* (B) e *Wine* (W). As informações sobre cada um desses conjuntos, apresentadas na Tabela 1, incluem a quantidade de atributos QT e QL e de classes (b) e o Erro da Classe Majoritária aproximado (ECM).

Tabela 1. Informações dos conjuntos de dados utilizados nos experimentos.

	A	**C**	**D**	**G**	**I**	**L**	**P**	**S**	**Y**	**V**	**B**	**W**
n	690	653	358	1000	351	32	106	208	47	846	569	178
m	14	15	34	20	34	56	57	60	35	18	30	13
QT	14	6	34	7	34	56	0	60	35	18	30	13
QL	0	9	0	13	0	0	57	0	0	0	0	0
b	2	2	6	2	2	3	2	2	4	4	2	3
ECM (%)	45	45	69	30	36	59	50	47	64	74	37	60

Os conjuntos de dados foram divididos conforme a Validação Cruzada Estratificada (VCE) [Han and Kamber 2006], com 10 partições.

3.2. Método Experimental

Para avaliação dos subconjuntos de atributos selecionados pelos AGM, foram induzidos modelos preditivos (de classificação) e descritivos (de agrupamento) com técnicas de AM nas projeções obtidas. Verificou-se em seguida possíveis ganhos de desempenho dos modelos derivados dos AGM frente àqueles gerados com o uso de todos os atributos. A SA foi aplicada apenas às partições de treinamento dos conjuntos de dados, reservando as partições de teste para uso na construção e avaliação dos modelos. Sendo os AGM estocásticos, eles foram aplicados cinco vezes em cada partição de treinamento, resultando em cinco subconjuntos de atributos por configuração multi-objetivo.

Os modelos de classificação foram induzidos com os métodos J48[2], *Support Vector Machines* (SVM), *k-Nearest Neighbour* (NN) com $k = 1$ e *Naïve Bayes* (NB). Foram utilizados vários algoritmos com o intuito de minimizar a influência que um deles poderia ter nos resultados [Witten and Frank 2005]. A construção de modelos descritivos foi realizada com o popular método de agrupamento *K-Means* (KM) [Jain and Dubes 1988].

Cada algoritmo para SA é avaliado por: (1) desempenho dos modelos derivados e (2) Porcentagens médias de Redução na quantidade original de atributos (PR). Para os

[1]Com apoio do Instituto Turing de Glasgow, Escócia.

[2]A SA embutida do J48, uma implementação do método C4.5, não foi investigada neste trabalho.

modelos de classificação, a medida de desempenho verificada é a taxa de acerto média no procedimento de VCE, em conjunto com seu desvio-padrão. Os modelos descritivos gerados foram avaliados por meio dos métodos de comparação Validação Cruzada Não-supervisionada (VCN) [Filho 2003] e Análise de Replicação (AR) [Morey et al. 1983] nas partições definidas pela VCE. Obteve-se em ambos os casos os valores médios do índice de validação *Rand* Corrigido (RC) [Jain and Dubes 1988]. O RC indica uma perfeita concordância entre agrupamentos pelo valor 1 e uma concordância gerada pelo acaso por valores negativos. Salienta-se que os rótulos não são utilizados para a SA e para a construção dos agrupamentos, apenas para a VCN.

A AR busca aferir a replicabilidade de um algoritmo de agrupamento por meio da técnica *nearest centroid*, a qual utiliza os centróides dos grupos identificados pelo algoritmo em cada partição de treinamento para agrupar os exemplos da partição de teste correspondente. Os grupos assim construídos são comparados aos que são gerados convencionalmente pelo algoritmo na partição de teste. A VCN utiliza o *nearest centroid* inicialmente e então compara os grupos obtidos em uma partição de teste com as classes dos exemplos nele.

3.3. Ferramentas e Configurações Adicionais

Os AGM são aplicados com os parâmetros: (1) *tamanho da população = 50 indivíduos*; (2) *probabilidade de cruzamento = 0,8*; (3) *probabilidade de mutação = 0,01*; (4) *critério de parada = 50 gerações*; (5) *quantidade de pais e de filhos após alteração = 50 indivíduos*. Esses valores foram selecionados a partir da literatura correlata [Spolaôr et al. 2011, Spolaôr 2010, Wang and Huang 2009]. A implementação do método baseou-se em módulos da ferramenta *PISA* [Bleuler et al. 2003]. Um desses módulos, originalmente proposto para um problema específico de otimização, foi adaptado neste trabalho para abstrair as medidas de importância de atributos como os objetivos a serem otimizados pelo AGM. A medida RE foi implementada com o apoio da *GSL*[3].

Os modelos preditivos foram gerados no trabalho com o auxílio da ferramenta *Weka* [Witten and Frank 2005], com valores de parâmetros como *default*. O KM utilizado foi o fornecido na ferramenta *R*[4] com valores dos parâmetros mantidos como *default*, com exceção da quantidade de grupos K. Esse parâmetro foi definido como $K = b$, isto é, K assume valor equivalente à quantidade de classes b no emprego do KM nos conjuntos de dados A, I, S, B, W, D e V.

Em cada conjunto de dados, o desempenho dos modelos gerados por um determinado classificador, obtidos antes e após a realização da SA, foi comparado pelo teste não-paramétrico de Kruskal-Wallis (KW) [Kruskal and Wallis 1952] a 95% de confiança. Ao todo, esse teste foi aplicado aos 12 conjuntos de dados rotulados da Tabela 1 e aos sete não-rotulados mencionados anteriormente. Foi utilizado um único controle, correspondente aos modelos construídos com todos os atributos (c_a), em cada avaliação para amenizar o efeito da multiplicidade [Salzberg 1997].

[3]http://www.gnu.org/software/gsl

[4]http://www.r-project.org

4. Resultados Experimentais

Os experimentos descritos a seguir foram organizados de acordo com o tipo de análise efetuada, ou seja, se sobre dados rotulados ou não-rotulados.

4.1. Seleção de Atributos em Dados Rotulados

Inicialmente, realizou-se um estudo acerca das medidas individuais e combinações de objetivos em duplas e triplas, envolvendo conjuntos de dados com atributos Quantitativos [Spolaôr 2010]. As combinações que se destacaram nesses experimentos foram então investigadas em um número maior de conjuntos de dados. Neste trabalho são apresentados os resultados das combinações IE+CC e IE+PI em conjuntos de dados com atributos QT e QL. Especificamente para dados com presença de atributos QL utilizam-se as medidas adaptadas PI, CC e IE.

A quantidade de modelos derivados de AGM com desempenho estatisticamente não inferior, isto é, superior (entre parênteses) ou equivalente, segundo o teste de KW, aos modelos construídos com todos os atributos é exibida na Tabela 2. Na Tabela 3 são apresentados, para cada modelo baseado no NN derivado de cada algoritmo de SA, as médias e os desvios-padrão, entre parênteses, das taxas de acerto (linha superior) e da PR (linha inferior), respectivamente. O valor médio e desvio-padrão das taxas de acerto e PR para todos os conjuntos de dados também é considerado no fim da tabela. A escolha por esse classificador justifica-se pela sensibilidade que apresenta à presença de muitos atributos (maldição da dimensionalidade) [Han and Kamber 2006] e pela sua relativa simplicidade. Os resultados correspondentes a outros algoritmos de classificação estão descritos em [Spolaôr 2010].

Tabela 2. Quantidade de modelos derivados de AGM para SA em dados rotulados com desempenho estatisticamente equivalente ou superior (entre parênteses) aos modelos c_a.

	IE+CC	IE+PI	**Total**
J48	10 (1)	12 (0)	22 (1)
SVM	8 (0)	12 (0)	20 (0)
NB	9 (1)	12 (0)	21 (1)
NN	9 (0)	11 (0)	20 (0)
Total	36 (2)	47 (0)	83 (1)

4.2. Discussão

Na Tabela 2, $87{,}5\%$ dos 96 modelos (12 conjuntos de dados $\times$ 2 AGM $\times$ 4 indutores) derivados dos AGM apresentam qualidade estatisticamente não inferior à atingida pelos modelos gerados com todos os atributos. Desse modo, nota-se a manutenção do desempenho com uma redução na quantidade de atributos e, consequentemente, no custo computacional envolvido na construção dos modelos.

A competitividade dos modelos derivados dos dois AGM em termos de desempenho preditivo ocorre não apenas em relação a c_a, como indicado anteriormente, mas também para modelos derivados de algoritmos filtro como *Correlation-based Feature Subset Selection*, *Consistency Subset Eval* e AG otimizando cada medida de importância descrita na Seção 2.1 [Spolaôr et al. 2011, Spolaôr 2010].

Tabela 3. Taxas médias de acerto e desvio-padrão (entre parênteses) dos modelos NN derivados e PR média e desvio-padrão dos algoritmos de SA, em cada conjunto e entre todos os conjuntos de dados.

	IE+CC	IE+PI	c_a
A	81,01 (3,73)	80,00 (4,08)	80,00 (4,26)
	57,14 (0,45)	0,00 (0,00)	0,00 (0,00)
C	70,45 (6,50)	81,94 (4,50)	81,32 (3,40)
	60,00 (0,00)	9,07 (0,69)	0,00 (0,00)
D	93,26 (3,50)	94,42 (2,16)	94,42 (2,26)
	22,94 (1,46)	0,00 (0,00)	0,00 (0,00)
G	61,16 (5,19)	70,72 (4,46)	71,90 (4,38)
	89,90 (0,14)	8,10 (1,21)	0,00 (0,00)
I	87,69 (3,84)	87,75 (4,49)	87,75 (4,68)
	32,59 (1,40)	1,41 (0,50)	0,00 (0,00)
L	52,17 (26,82)	46,67 (25,86)	46,67 (26,99)
	35,36 (1,96)	1,04 (0,73)	0,00 (0,00)
P	71,64 (10,91)	73,75 (9,08)	79,27 (7,30)
	24,60 (2,63)	11,68 (2,22)	0,00 (0,00)
S	85,05 (8,69)	85,61 (8,92)	86,98 (7,96)
	56,47 (2,46)	5,73 (1,70)	0,00 (0,00)
Y	100,00 (0,00)	100,00 (0,00)	100,00 (0,00)
	59,60 (0,50)	20,4 (1,65)	0,00 (0,00)
V	68,95 (3,04)	69,74 (2,99)	69,74 (3,12)
	7,44 (0,72)	0,00 (0,00)	0,00 (0,00)
B	93,46 (3,54)	95,25 (3,18)	95,25 (3,32)
	69,40 (0,44)	0,00 (0,00)	0,00 (0,00)
W	95,56 (4,89)	94,97 (4,67)	94,97 (4,87)
	37,69 (0,95)	0,00 (0,00)	0,00 (0,00)
Média	80,03 (15,07)	79,54 (16,83)	84,81 (9,05)
	46,09 (23,21)	4,16 (6,40)	0,00 (0,00)

Em termos de PR, nota-se na Tabela 3 que o AGM IE+CC obteve redução média entre todos os conjuntos de dados superior a 45%. Assim, esse algoritmo permitiu obter vários modelos com desempenho comparável a c_a e um custo computacional menor para a geração de modelos preditivos, pois menos atributos são utilizados durante a indução. O destaque geral da combinação IE+CC é compatível com o observado em relação ao uso de outros AGM em dados QT e QL [Spolaôr 2010].

Contudo, o AGM relacionado a IE+PI apresentou PR média próxima a 4%. Esse aspecto motiva estudos futuros sobre mecanismos que estimulem a identificação de subconjuntos com menos atributos nessa combinação, como a minimização explícita da quantidade de atributos em um objetivo adicional.

Em geral, o uso de algoritmos de classificação distintos pouco influencia na quantidade total de modelos estatisticamente não inferiores, obtida em cada classificador. A não interferência direta do classificador durante a SA, característica da abordagem filtro, pode auxiliar a justificar esse comportamento.

4.3. Seleção de Atributos em Dados Não-rotulados

Na Tabela 4 são quantificados os modelos derivados dos AGM que obtiveram índice RC estatisticamente não inferior, segundo o teste de KW, ao atingido por c_a nos métodos VCN e AR. Na Tabela 5 são apresentados, para cada modelo baseado no KM derivado de

cada algoritmo de SA e avaliado pelo método VCN, as médias e os desvios-padrão, entre parênteses, dos índices RC (linha superior) e da PR (linha inferior), respectivamente. Os resultados correspondentes para o método AR estão descritos em [Spolaôr 2010].

Tabela 4. Quantidade de modelos derivados de AGM para SA em dados QT não-rotulados com desempenho estatisticamente equivalente ou superior (entre parênteses) ao modelo c_a.

	IC+LS	RE+IC	RE+LS	**Total**
VCN	2 (1)	1 (2)	1 (2)	4 (5)
AR	2 (1)	4 (0)	4 (0)	10 (1)
Total	4 (2)	5 (2)	5 (2)	14 (6)

Tabela 5. Taxas médias de RC e desvio-padrão (entre parênteses) dos modelos KM derivados e PR média e desvio-padrão dos algoritmos de SA, em cada conjunto e entre todos os conjuntos de dados, avaliados pelo método VCN.

	IC+LS	RE+IC	RE+LS	c_a
A	0,00 (0,00)	0,19 (0,21)	0,20 (0,17)	0,00 (0,00)
	85,71 (0,00)	58,86 (0,82)	64,29 (0,00)	0,00 (0,00)
D	0,03 (0,04)	0,58 (0,12)	0,59 (0,13)	0,04 (0,04)
	94,12 (0,00)	66,35 (0,54)	58,88 (0,59)	0,00 (0,00)
I	0,07 (0,10)	0,14 (0,20)	0,05 (0,13)	0,17 (0,11)
	94,06 (0,14)	36,94 (0,64)	31,12 (0,54)	0,00 (0,00)
S	0,05 (0,15)	0,00 (0,07)	0,01 (0,07)	-0,01 (0,06)
	92,97 (1,66)	42,40 (1,63)	34,97 (2,13)	0,00 (0,00)
V	0,02 (0,02)	0,02 (0,03)	0,02 (0,02)	0,12 (0,05)
	88,89 (0,00)	88,89 (0,00)	88,89 (0,00)	0,00 (0,00)
B	0,20 (0,11)	0,00 (0,03)	0,00 (0,03)	0,48 (0,10)
	93,33 (0,00)	73,33 (0,29)	56,73 (0,25)	0,00 (0,00)
W	0,22 (0,15)	0,23 (0,15)	0,17 (0,12)	0,36 (0,15)
	84,62 (0,00)	76,92 (0,00)	62,31 (0,30)	0,00 (0,00)
Média	0,10 (0,10)	0,17 (0,19)	0,15 (0,19)	0,19 (0,19)
	89,79 (4,32)	65,08 (17,98)	57,44 (18,06)	0,00 (0,00)

4.4. Discussão

Na Tabela 4, $47{,}62\%$ dos 42 agrupamentos ([7 conjuntos de dados $\times$ 3 AGM para VCN] $+$ [7 conjuntos de dados $\times$ 3 AGM para AR]) derivados de AGM apresentam nível de concordância estatisticamente não inferior a c_a. Esses agrupamentos estão bem distribuídos entre os algoritmos de SA, mas o método VCN concentra os modelos com RC superior.

Observa-se na Tabela 5 que todos os algoritmos obtiveram uma PR média, entre todos os conjuntos de dados, superior a 55%. A agressividade na PR e o desempenho razoável dos modelos derivados, relacionados aos AGM RE+IC e RE+LS, também haviam sido observados anteriormente na SA supervisionada [Spolaôr 2010]. Experimentos complementares envolvendo mais algoritmos de SA e de agrupamento possibilitariam reforçar a competitividade desses AGM em relação a mais *baselines*.

Parte dos agrupamentos obtidos após a aplicação do AGM IC+LS atingiram taxas de concordância aproximadamente equivalentes às obtidas por c_a para ambos os métodos de comparação, apesar da PR ser superior a 85%. Assim, subconjuntos com poucos atributos podem contribuir para a geração de bons modelos, o que não ocorreu em relação a SA supervisionada [Spolaôr et al. 2011, Spolaôr 2010].

5. Conclusão

Neste trabalho apresentou-se um estudo da influência de combinações de medidas de importância de atributos em AGM conforme a abordagem filtro em dados QT e QL. O desempenho de modelos preditivos e descritivos derivados de AGM é comparado à qualidade dos modelos gerados com todos os atributos. A Porcentagem de Redução na quantidade de atributos obtida pela SA também é considerada.

Em geral, uma grande parte dos modelos construídos após a SA foram considerados competitivos em termos de desempenho, enquanto a PR variou conforme o AGM utilizado. Assim, a adaptação dos critérios para tratar dados com atributos QL e a aplicação de medidas para dados não-rotulados foram relativamente bem-sucedidas. O método implementado neste trabalho demonstrou flexibilidade para incorporar distintas medidas para SA em dados com propriedades distintas e suporte para estudar combinações entre elas.

O uso dos AGM para SA em dados artificiais e reais com mais atributos está em andamento. Para dados artificiais, pretende-se realizar análise de relevância e de redundância dos atributos. Trabalhos futuros incluem uma análise ampla da complementaridade entre medidas de importância envolvendo, por exemplo, a intensidade de conflito entre os critérios, e a inclusão de algoritmos *wrapper* e outras heurísticas como *baselines*.

Agradecimentos

A UFABC, CAPES, FAPESP e CNPq pelo apoio financeiro recebido para este trabalho e a Ronaldo C. Prati, Ivan G. C. Filho e colaboradores do projeto AID pela contribuição.

Referências

Arauzo-Azofra, A., Benitez, J. M., and Castro, J. L. (2008). Consistency measures for feature selection. *Journal of Intelligent Information Systems*, 30(3):273–292.

Asuncion, A. and Newman, D. (2007). UCI machine learning repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Bleuler, S., Laumanns, M., Thiele, L., and Zitzler, E. (2003). PISA — a platform and programming language independent interface for search algorithms. In *Evolutionary Multi-Criterion Optimization*, pages 494–508.

Bruzzone, L. and Persello, C. (2009). A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE transactions on geoscience and remote sensing*, 47:3180–3191.

Deb, K., Agrawal, S., Pratap, A., and Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. Technical report, Indian Institute of Technology Kanpur - India.

Filho, I. G. C. (2003). Comparative analysis of clustering methods for gene expression data. Dissertação de mestrado, Universidade Federal de Pernambuco.

Han, J. and Kamber, M. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

He, X., Cai, D., and Niyogi, P. (2005). Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, pages 507–514.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., New Jersey, Estados Unidos.

Kruskal, W. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *American Statistical Association*, 47:583–621.

Lee, H. D., Monard, M. C., and Wu, F. C. (2006). A fractal dimension based filter algorithm to select features for supervised learning. In *Ibero-American Conference on Artificial Intelligence - Brazilian Symposium on Artificial Intelligence*, pages 278–288.

Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

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

Mitchell, T. M. (1997). *Machine Learning*. Hardcover.

Morey, L. C., Blashfield, R. K., and Skinner, H. A. (1983). A comparison of cluster analysis techniques within a sequential validation frame work. *Multivariate Behavioral Research*, 18:309–329.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–328.

Santana, L. E. A., Silva, L., and Canuto, A. M. P. (2009). Feature selection in heterogeneous structure of ensembles: a genetic algorithm approach. In *International Joint Conference on Neural Networks*, pages 1491–1498.

Spolaôr, N. (2010). Aplicação de algoritmos genéticos multiobjetivo ao problema de seleção de atributos. Dissertação de mestrado, Universidade Federal do ABC.

Spolaôr, N., Lorena, A. C., and Lee, H. D. (2011). Multiobjective genetic algorithm evaluation in feature selection. In Takahashi, R. H. C., Deb, K., Wanner, E. F., and Greco, S., editors, *Lecture Notes in Computer Science (Evolutionary Multi-criterion Optimization Proceedings)*, pages 462–476. Springer-Verlag.

Wang, C.-M. and Huang, Y.-F. (2009). Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications*, 36(3):5900–5908.

Wilson, D. R. and Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Yan, W. (2007). Fusion in multi-criterion feature ranking. In *International Conference on Information Fusion*, pages 01–06.

Zaharie, D., Holban, S., Lungeanu, D., and Navolan, D. (2007). A computational intelligence approach for ranking risk factors in preterm birth. In *International Symposium on Applied Computational Intelligence and Informatics*, pages 135–140.

Zeleny, M. (1973). An introduction to multiobjetive optimization. In Cochrane, J. L. and Zeleny, M., editors, *Multiple criteria decision making*, pages 262–301. University of South Carolina Press.