ALGORITMOS GENÉTICOS MULTI-OBJETIVOS NA MINERAÇÃO DE REGRAS INTERESSANTES E PRECISAS

GINA MAIRA BARBOSA DE OLIVEIRA[†], MARIA CAZUHO SAITO TAKIGUTI[£], LUIZ GUSTAVO ALMEIDA MARTINS[†]

[†] *Faculdade de Ciência da Computação,Universidade Federal de Uberlândia - UFU*
Av. João Naves de Ávila, 2121- Campus Santa Mônica, Bloco B, sala 1B60 CEP: 38400-902 Uberlândia, MG
E-mails: gustavo@facom.ufu.br, gina@facom.ufu.br

[£]*Programa de Pós-Graduação em Engenharia Elétrica, Universidade Presbiteriana Mackenzie*

Abstract— This work evaluates the use of multi-objective genetic algorithms (MOGA) in the mining of accurate and interesting rules. For this, a program was implemented based on the MOGA technique called nondominated sorting genetic algorithm (NSGA), which was applied in the database *Zoo* of public domain. The results of our experiments had been compared with those generated by a standard genetic algorithm in order to identify the benefits related to the multi-objective approach.

Keyword— *Multi-objective genetic algorithms, rules mining, data mining*.

Resumo— Este trabalho avalia o uso de algoritmos genéticos multi-objetivos (AGMO) no processo de mineração de regras precisas e interessantes. Implementou-se um programa baseado na técnica de AGMO conhecida como *nondominated sorting genetic algorithm*, o qual foi aplicado na base de dados de domínio público *Zoo*. Os resultados foram comparados com aqueles gerados por um algoritmo genético padrão, a fim de identificar as melhorias obtidas pela adoção de uma abordagem multi-objetivos.

Palavras-chave— Algoritmos genéticos multi-objetivos, extração de regras, mineração de dados.

1 Introdução

Com a evolução da tecnologia da informação, houve um crescimento acelerado na quantidade de dados produzida e armazenada. Esses dados possuem um enorme potencial de informações estratégicas não caracterizadas explicitamente, que necessitam ser extraídas. Quando produzidos em larga escala, sua leitura e análise por métodos manuais são inviáveis. *Data Mining* (DM) é a principal etapa do processo de descoberta de conhecimento em bancos de dados. Diferentes técnicas para DM baseadas em métodos evolutivos foram investigadas [Freitas, 2002]. Freitas afirma que o conhecimento descoberto pelo DM deve ser *exato*, *compreensível* e *interessante*, sendo essa terceira propriedade de difícil quantificação. A maioria dos trabalhos se propõe a avaliar métricas mais objetivas, como a acurácia e a compreensibilidade. Isso motivou a inclusão do objetivo "grau de interesse" na mineração de regras [Carvalho *et al*., 2003].

Neste trabalho, foi implementado um algoritmo genético (AG) multi-objetivos baseado na família de métodos *nondominated sorting genetic algorithm* (NSGA e NSGAII), para mineração de regras precisas e interessantes. Essas propriedades são avaliadas separadamente por duas métricas: a acurácia da predição e o grau de interesse das regras descobertas. O AG multi-objetivos é comparado com outro ambiente que utiliza um AG padrão, baseado no experimento de Noda e colaboradores (1999). O algoritmo genético padrão efetua a mineração das regras utilizando-se as mesmas métricas. Entretanto, elas são fundidas na função de avaliação em um único objetivo.

2 Algoritmos genéticos em DM

Data Mining, ou mineração de dados, é o processo de identificação e extração das informações úteis em bases de dados [Freitas, 2002], cujo objetivo é automatizar a descoberta dos conhecimentos que não são perceptíveis. O uso do DM é considerado o passo central de um processo maior chamado Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases-KDD*). O KDD inclui vários processos, que podem ser divididos em pré-processamento, mineração e pós-processamento [Freitas, 2002]. Os principais tipos de tarefas realizadas pelos algoritmos de DM são: classificação, análise de associação, clusterização, regressão, análise de seqüências, sumarização e modelos de dependência. A tarefa de classificação é a mais investigada, na qual se define um atributo objetivo e busca-se o conhecimento capaz de predizer o valor desse atributo, baseando-se em outros atributos de predição. Esse conhecimento é muitas vezes representado por regras do tipo SE-ENTÃO. A tarefa estudada neste trabalho é o modelo de dependência, que pode ser considerada uma generalização da tarefa de classificação. A meta também é descobrir regras de predição do valor do atributo objetivo (conseqüente), dado os valores dos atributos de predição (antecedente). Mas, no modelo de dependência, há mais de um atributo objetivo. Inicialmente, é especificado um conjunto pequeno de atributos objetivos que se está interessado em predizer. Esses atributos podem ocorrer no conseqüente de uma regra ou no antecedente de regras com outros atributos objetivos. Os outros atributos podem ocorrer somente no antecedente da regra.

Várias técnicas podem realizar a extração de informações das bases de dados. Elas estão diretamente ligadas ao tipo de tarefa onde se quer aplicar o DM. Algumas das principais técnicas usadas são: as redes neurais, as árvores de decisão e os algoritmos genéticos. Métodos evolutivos, especialmente os algoritmos genéticos [Goldberg, 1989], vêm sendo investigados em diversas tarefas [Freitas, 2002]. Os aspec-

tos mais relevantes na especificação dos AGs são a representação do indivíduo, que deve representar a informação minerada; e a função de avaliação, que deve avaliar a qualidade dessa informação.

Para representar regras do tipo SE-ENTÃO, têm-se duas abordagens: *Michigan* e *Pittsburgh* [Freitas, 2002]. Na abordagem *Michigan*, adotada nesse trabalho, cada indivíduo representa uma única regra e a população do AG representa um conjunto de regras. Na abordagem *Pittsburgh*, cada indivíduo representa um conjunto de regras. A avaliação de um indivíduo é baseada em métricas usadas para medir a qualidade da regra que ele codifica. Diversas métricas podem ser utilizadas, mas freqüentemente elas são relacionadas à precisão das regras: grau de confiança, cobertura, sensitividade e especificidade. Freitas (2002) destacou a importância de se utilizar também métricas de outra natureza na avaliação das regras, tais como, a compreensibilidade e o grau de interesse.

3 Otimização multi-objetivos baseada AGs

Muitos problemas do mundo real envolvem uma otimização simultânea de múltiplos objetivos [Coello, 1996]. Na otimização de um único objetivo, tenta-se obter o melhor resultado, usualmente o mínimo ou o máximo global. No caso de múltiplos objetivos, pode não haver uma melhor solução com respeito a todos os objetivos. Em um problema de otimização multi-objetivos, existe um conjunto de soluções que são superiores às demais dentro do espaço de busca [Srinivas e Deb, 1994]. Na figura 1, as funções devem ser maximizadas simultaneamente. Pode-se afirmar que a solução *A* é melhor que as soluções *C* e *D*, isto é, *C* e *D* são dominadas por *A*. Porém, no caso das soluções *A* e *B*, não é possível afirmar qual delas é a melhor. Assim, podemos dizer que as soluções *A* e *B* são não dominadas e ambas dominam as soluções *C* e *D*. O *Ótimo de Pareto* é o conjunto de soluções não dominadas considerando-se todo o espaço de busca.

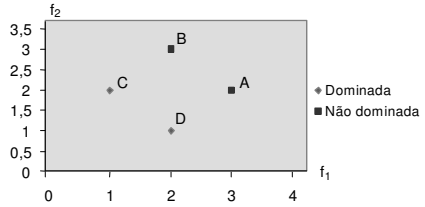

Figura 1. Dominância na maximização simultânea de 2 objetivos.

Os AGs podem ser adaptados facilmente para a manipulação simultânea das soluções não-dominadas, pois trabalham sobre uma população de soluções. Métodos multi-objetivos baseados em AGs foram propostos na literatura [Coello, 1996], sendo os algoritmos NSGA e NSGAII dentre os mais conhecidos. O AGMO conhecido por NSGA (*Nondominated Sorting Genetic Algorithms)* [Srinivas e Déb, 1994] é baseado em vários níveis de classificação de dominância dos indivíduos. Antes que a seleção seja executada, a população é agrupada em fronteiras hierárquicas onde todos os indivíduos não dominados entre si são classificados em uma mesma fronteira. A meta do algoritmo é obter, no final da execução, uma seqüência de fronteiras, onde a primeira corresponde à fronteira de Pareto. Inicialmente, para manter a diversidade da população, os indivíduos de uma mesma fronteira compartilham um mesmo valor de avaliação fictícia. Depois, dentro de uma mesma fronteira, é feita uma pequena diferenciação nessa avaliação, dependendo do quão isolado um indivíduo se encontra no espaço de busca. Essa diferenciação é chamada de operação de compartilhamento. Uma vez que todos os indivíduos da população tenham sido classificados em suas respectivas fronteiras e a avaliação compartilhada tenha sido calculada, a seleção para o *crossover* é realizada. Os indivíduos das primeiras fronteiras têm os maiores valores de avaliação e, portanto, têm mais cópias do que o resto da população. Isso permite procurar por regiões não dominadas e resulta em uma rápida convergência das populações em torno dessas regiões. Esse algoritmo é similar ao AG padrão, exceto pela classificação em fronteiras hierárquicas de dominância e pela operação de compartilhamento. O valor da função de compartilhamento entre indivíduos na mesma fronteira é dado por:

$$Sh(d_{ij}) = \begin{cases} 1-\left(\dfrac{d_{ij}}{\sigma_{share}}\right)^{\alpha}, & se \;\; d_{ij} < \sigma_{share} \\ 0, \;\; caso \;\; contrário \end{cases} \qquad (1)$$

sendo, d_{ij} a distância entre os fenótipos dos indivíduos *i* e *j* da mesma fronteira; σ_{share} a distância máxima permitida entre os fenótipos dos indivíduos para que eles sejam considerados membros do mesmo nicho; e α a potência que define a taxa na qual o grau de compartilhamento decresce com a proximidade.O *contador de nicho* é calculado pela somatória da função de compartilhamento para todos os indivíduos da fronteira. Os valores da avaliação compartilhados de cada indivíduo são calculados pela divisão de sua avaliação fictícia pelo seu *contador de nicho.*

O algoritmo NSGAII proposto em [Deb e Goel, 2001] é uma variação do NSGA original. A principal inovação em relação ao método anterior refere-se ao uso de uma estratégia elitista, através do uso de uma população auxiliar, que preserva os indivíduos das fronteiras de maior dominância. Além disso, o NSGAII utiliza um algoritmo de classificação das fronteiras de dominância mais eficiente.

Os AGMOs têm sido investigados em tarefas de DM. Um AGMO foi utilizado na busca de modelos não-lineares de marketing direto em [Bhattacharyya, 2000] e para a seleção de características para aprendizagem não supervisionada em [Kim *et al.,* 2000]. Em [Iglesia *et al.,* 2003], uma abordagem baseada no NSGAII é avaliada na tarefa de classificação (um único objetivo), utilizando-se apenas métricas relacionadas à precisão das regras: acurácia e cobertura. Um AG baseado em Pareto foi implementado para minerar regras na tarefa de associação em bases de dados do tipo *market-basket*, utilizando-se métricas para a acurácia, a compreensibilidade e o grau de interesse [Ghosh e Nath, 2004]. Uma abordagem

baseada no NSGAII também foi aplicada na tarefa de associação, para a maximização simultânea das métricas cobertura e confiança [Ishibuchi *et al.,* 2006]. Freitas (2004) apresentou uma análise crítica sobre as principais abordagens multi-objetivas em tarefas de DM: (i) composição das métricas através de uma soma ponderada, como o AG mono-objetivo descrito na próxima seção; (ii) maximização simultânea e independente das métricas com o uso de AGMOs, como a abordagem investigada nesse trabalho; (iii) abordagens lexográficas, onde as métricas são classificadas de acordo com uma ordem de prioridade.

4 Ambientes evolutivos

Utilizamos dois ambientes para minerar regras interessantes e precisas na tarefa de modelo de dependência: um AG mono-objetivo que avalia o grau de interesse e a precisão das regras através de uma média ponderada e um AG multi-objetivos, no qual essas métricas são avaliadas de forma independente.

O ambiente que utiliza um AG padrão foi elaborado fortemente baseado no descrito em [Noda *et al.*, 1999]. Para tal, coletamos dados adicionais com os autores. O AG realiza o processo de mineração de regras precisas e interessantes, utilizando-se uma métrica que mede a qualidade da predição realizada pela regra e outra que mede o seu grau de interesse, de forma objetiva. Cada indivíduo da população representa uma regra. No modelo de representação, o antecedente da regra é formado pela conjunção de atributos na forma $A_i = V_{ij}$, sendo A_i o i-ésimo atributo e V_{ij} o j-ésimo valor dentro do seu domínio. A parte conseqüente é uma condição simples da forma $G_k = V_{kl}$, sendo G_k o k-ésimo atributo objetivo e V_{kl} o l-ésimo valor de seu domínio. Cada cromossomo no AG representa a parte antecedente. A codificação do indivíduo é feita por uma string de tamanho fixo com n genes representando os valores correntes de cada um dos atributos, conforme ilustrado na Figura 2. Essa codificação representa uma regra de tamanho variável: se um atributo não ocorre na parte antecedente da regra, o valor do seu gene é "-1". Para cada cromossomo, o AG escolhe o melhor conseqüente, ou seja, os valores de k e l adequados à obtenção da melhor regra, visando maximizar sua avaliação.

A_1	A_2	A_3	...	...	...	A_n
V_{1k}	V_{2k}	V_{3k}	...	...	...	V_{nk}

Figura 2. Codificacão do cromossomo.

O grau de interesse da regra é usado na função de avaliação. Uma mensagem é informativa se seu conteúdo é pouco esperado ou desconhecido. Se for possível medir a probabilidade de ocorrência de uma mensagem, pode-se medir o seu conteúdo de informação de forma similar. A Teoria da Informação é baseada na freqüência relativa com que cada mensagem é emitida por uma fonte [Noda *et al*., 1999] e a métrica para avaliar o grau de interesse utiliza essa idéia. O cálculo avalia separadamente o antecedente e o conseqüente. Apesar dos atributos com alto ganho de informação serem bons preditores de classe, normalmente não apresentam informações interessantes, por serem provavelmente conhecidos pelo usuário. Regras cujos antecedentes contêm atributos com baixo ganho de informação podem apresentar informações mais interessantes. O grau de interesse da parte antecedente (*AntInt*) é baseado na entropia, dado por:

$$AntInt = 1 - \frac{\left(\sum_{i=1}^{n} InfoGain(A_i) \Big/ n \right)}{\log_2 \left(\left| dom(G_k) \right| \right)} \quad (2)$$

sendo, n o número de atributos da parte antecedente; $dom(G_k)$ o número de valores possíveis do atributo objetivo G_k no conseqüente e $InfoGain(A_i)$ o ganho de informação do atributo A_i, obtido por:

$$InfoGain(A_i) = Info(G_k) - Info(G_k | A_i) \quad (3)$$

$$Info(G_k) = -\sum_{l=1}^{n} (\Pr(V_{kl}) \log_2 (\Pr(V_{kl}))) \quad (4)$$

$$Info(G_k | A_i) = \sum_{j=1}^{n} (\Pr(V_{ij}) (-\sum_{l=1}^{mk} (\Pr(V_{kl} | V_{ij}) \log_2 (\Pr(V_{kl} | V_{ij})))) \quad (5)$$

sendo, mk o número de valores possíveis do atributo objetivo G_k; n o número de valores possíveis do atributo de predição A_i; $Pr(X)$ a probabilidade de X; e $Pr(X|Y)$ a probabilidade condicional de X dado Y. O grau de interesse da parte conseqüente da regra (*ConsInt*) é baseado na freqüência relativa do valor predito pelo conseqüente [Noda *et al*., 1999]. Quanto mais raro o valor do atributo objetivo, mais interessante é a predição. O grau de interesse do conseqüente é:

$$ConsInt = \sqrt{1 - \Pr(G_{kl})} \quad (6)$$

sendo, $Pr(G_{kl})$ a freqüência relativa do valor de G_{kl}.

A segunda parte da função de avaliação mede a acurácia da predição (*PredAcc*) da regra e é uma adaptação do indicador de sensibilidade preditiva:

$$PredAcc = \left(\left| A \,\&\, C \right| - 0{,}5 \right) \Big/ \left| A \right| \quad (7)$$

sendo, $|A\&C|$ o número de casos que satisfazem as partes antecedente e conseqüente da regra (verdadeiros positivos); e $|A|$ o número de casos que satisfazem o antecedente, independente do conseqüente (verdadeiros positivos e falsos positivos). A subtração de 0,5 do $|A\&C|$ foi acrescentada ao indicador original para penalizar regras que cobrem poucos exemplos de treinamento. Esse fator não é utilizado na medição do *PredAcc* na base de teste. A função de *fitness* usada no AG mono_objetivo é obtida pela média ponderada entre as medidas normalizadas do grau de interesse e da precisão da regra:

$$Fitness = w_1 \times \frac{AntInt + ConsInt}{2} + w_2 \times PredAcc \quad (8)$$

sendo, w_1 e w_2 os pesos das métricas. Adotou-se $w_1 = 1/3$ e $w_2 = 2/3$, dando maior peso à precisão da regra.

O AG usa a seleção por torneio simples, onde *Tour* indivíduos são aleatoriamente sorteados e aquele com aptidão mais alta é escolhido para reprodução. Foi utilizado operador de *crossove*r uniforme [Goldberg, 1989], com uma probabilidade de 70% para aplicação de *crossover* em um par de indivíduos e

outra probabilidade de 50% para trocar cada valor do atributo no antecedente da regra. Após o *crossover*, o AG verifica a existência de indivíduos inválidos e, se encontrado, emprega um operador de reparo. O operador de mutação transforma aleatoriamente o valor de um atributo em outro valor pertencente ao domínio daquele atributo. A taxa de mutação usada foi 5%. Além do *crossover* e da mutação, existem operadores de inserção e remoção, que tentam controlar diretamente o tamanho das regras. Esses operadores inserem ou removem aleatoriamente uma condição na regra antecedente. Suas probabilidades dependem do número de atributos no antecedente. O operador de inserção tem uma probabilidade nula de aplicação quando a regra tem o número máximo de atributos (especificado pelo usuário). O operador de remoção trabalha em caminho oposto. As taxas de inserção e remoção usadas são 50% e 70%, respectivamente. O elitismo [Goldberg, 1989] empregado mantém, de uma geração para outra, as M melhores regras de predição de cada par <*atributo objetivo, valor*>, dentre K pares existentes, correspondendo a um fator de elitismo de $K \times M$. Utilizou-se tamanho da população igual a 100 e número de gerações igual a 50. Após avaliações experimentais, algumas pequenas mudanças foram feitas em relação a [Noda *et al*., 1999]: (i) *Tour = 3*; (ii) critério para definir o vencedor do torneio: um par <*atributo objetivo, valor*> é sorteado, as regras participantes são comparadas em relação a esse par e aquela com maior avaliação é escolhida; (iii) número de regras mantidas por elistismo: 3 regras para cada conseqüente

A principal alteração efetuada no AG multi-objetivos refere-se à função de avaliação, que passou a ter dois objetivos separados: *AntInt* e *PredAcc*. A métrica *ConsInt* dada pela equação (6) não foi empregada, pois ela retorna um valor constante que depende unicamente do par <*atributo objetivo, valor*> e do número de suas ocorrências na base. Após a geração da população inicial, cada regra é avaliada em relação às duas métricas (equações (2) e (7)). Posteriormente, os objetivos são utilizados na classificação dos indivíduos nas fronteiras hierárquicas de dominância. Para cada fronteira, um valor de avaliação fictícia é atribuído inicialmente aos indivíduos. Depois esses valores são ajustados para cada regra da fronteira através do cálculo da função de compartilhamento, dada pela equação (1), e do contador de nicho. Cada indivíduo recebe uma avaliação escalar e a geração de filhos é processada como no caso mono-objetivo: (*i*) seleção para o *crossover* pelo torneio simples (*Tour* = 3), (*ii) crossover* uniforme com reparo eventual de inválidos, (*iii*) mutação aleatória, (*iv*) inserção e remoção. Nosso AGMO também usa elitismo na reinserção da população, como no NSGA-II. Entretanto, esse elitismo é similar ao realizado no ambiente mono-objetivo: para cada conseqüente, as três regras com maior avaliação compartilhada são mantidas. O restante a nova população é formada pelos filhos. Os demais parâmetros são os mesmos do AG mono-objetivo. Embora o objetivo do AGMO seja encontrar uma fronteira com diversas soluções não-dominadas, fez-se necessário estabelecer um critério para a seleção da melhor regra ao final da execução do AG. A equação (8) foi então aplicada para decidir qual é a melhor regra na fronteira não-dominada. Dessa forma, além de definirmos objetivamente a saída final do AG, selecionamos as regras de acordo com o mesmo critério utilizado para evoluir as regras no ambiente mono-objetivo.

5 Experimentos

A mineração de regras foi aplicada à base *Zoo*, também utilizada em [Noda *et al.,* 1999]. Essa base foi obtida no repositório de banco de dados da *UCI* e contém 101 registros e 18 atributos categóricos: *hair, feathers, eggs, milk, backbone, fins, legs, tail, catsize, airbone, aquatic, breathes, venomous, toothed, predator, domestic, type;* os três últimos são usados como objetivos. *Predator* e *domestic* são binários e *type* varia de 1 a 7, resultando em regras com onze conseqüentes diferentes: *predator=0*, *predator=1*, *domestic=0*, *domestic=1*, *type=1*, *type=2*, *type=3*, *type=4*, *type=5*, *type=6*, *type=7*, chamados, respectivamente, de *conseqüente 1,*., *conseqüente 11*.

Foram executados dois tipos de experimento. O primeiro, utilizou todo o conjunto de dados na evolução do AG obtendo as regras que melhor prediziam cada um dos onze pares <*atributo_objetivo,valor*>. No segundo, uma validação cruzada foi usada para avaliar a qualidade das regras descobertas, onde o conjunto de dados foi dividido em 5 partições mutuamente exclusivas e completas. Esse experimento é composto por cinco tipos de evolução do AG, sendo que em cada tipo, uma partição diferente foi usada como conjunto de teste. A partição de teste foi deixada à parte durante a evolução das regras e as restantes foram usadas como base de treino, correspondendo aos dados efetivamente utilizados pelo AG na busca das regras. Ao final, as regras são avaliadas quanto à sua capacidade de predição na partição de teste. O resultado desse experimento foi gerado através da média dos resultados das cinco partições de teste.

Os AGs mono-objetivo e multi-objetivos foram aplicados em 100% dos registros da base *Zoo*. Em cada experimento, foram realizadas 10 execuções diferentes do AG. A tabela 1 apresenta os valores das métricas *AntInt* e *PredAcc* para as melhores regras obtidas nessas execuções. Os dois ambientes conseguiram minerar regras com um alto grau de interesse e uma boa precisão para a maioria dos conseqüentes. O AG mono-objetivo encontrou regras com *AntInt* abaixo de 0,9 nos conseqüentes de 5 a 11 e, o AG multi-objetivos, apenas para os conseqüentes 7, 8 e 9. Nos demais, o *AntInt* foi acima de 0,95. Com relação à acurácia, tanto o AG padrão quanto o MO encontraram regras ruins (< 0,8) apenas para o conseqüente 4. Esses resultados são compatíveis com [Noda *et al*., 1999]. A diferença mais significativa ocorre em relação ao *AntInt*, sendo que o *PredAcc* foi praticamente

equivalente nos dois modelos. A figura 3 mostra os gráficos radiais da métrica (a) *AntInt* e (b) *PredAcc*, para os 11 conseqüentes. Com relação ao *AntInt*, o experimento multi-objetivos foi superior ao mono-objetivo em oito conseqüentes, sendo que em dois deles essa diferença foi superior a 10%. Com relação ao *PredAcc*, o multi-objetivos foi um pouco superior em três conseqüentes. Portanto, a abordagem multi-objetivos encontrou regras com maior grau de interesse, sem prejuízo da acurácia. A tabela 2 apresenta as melhores regras descobertas pelo AGMO. Em relação ao mono-objetivo, obtiveram-se antecedentes com um menor número de condições em seis regras e maior em apenas um. Portanto, em geral, a compreensibilidade das regras foi superior. Embora a predição dessas regras tenha sido avaliada na Tabela 1, sua qualidade é melhor estimada pela predição média obtida na validação cruzada.

Tabela 1. Comparação das métricas *AntInt* e *PredAcc* obtidas nos experimentos mono-objetivo e multi-objetivos.

Conseqüente	Mono-Objetivo		Multi_Objetivos	
	AntInt	***PredAcc***	***AntInt***	***PredAcc***
1	0,956	0,929	0,999	0,929
2	0,983	0,958	0,999	0,958
3	0,990	0,980	0,950	0,986
4	0,996	0,500	0,996	0,500
5	0,880	0,984	0,982	0,988
6	0,893	0,971	0,982	0,975
7	0,849	0,875	0,881	0,833
8	0,887	0,958	0,840	0,962
9	0,883	0,875	0,863	0,875
10	0,848	0,937	0,982	0,937
11	0,887	0,937	0,982	0,937

O método da validação cruzada foi usado para verificar a capacidade de generalização das regras mineradas pelos AGs. A base *Zoo* foi dividida em 5 partições e para cada partição-teste, foram feitas 10 execuções diferentes do AG. As melhores regras obtidas para cada conseqüente foram aplicadas na partição de teste para avaliar sua precisão em exemplos não vistos durante o treino. A tabela 3 apresenta o valor médio (nas 5 partições) do *PredAcc* das melhores regras obtidas em cada partição, tanto em treino quanto em teste. A Figura 4 mostra os gráficos radiais da métrica *PredAcc* em (a) treino e em (b) teste. Nos dois ambientes, a média do *PredAcc* na base de treino é próxima aos valores obtidos com a base completa. Porém, o resultado médio do *PredAcc* na base de teste decai razoavelmente para cinco ou seis conseqüentes, indicando que as regras mineradas no treino não possuem uma boa capacidade de generalização nesses conseqüentes. Comparando os resultados obtidos, nota-se que embora a performance tenha sido similar no treinamento, o AG multi-objetivos foi superior na base de teste. Em teste, os valores médios de *PredAcc* do AG multi-objetivos foram superiores em seis conseqüentes (significativamente em quatro). Pelos gráficos da Figura 4, vê-se que o ambiente multi-objetivo foi ligeiramente superior no treino, mas claramente superior no teste. Assim, conclui-se que as regras obtidas pelo AG mono-objetivo apresentam uma maior especialização ao conjunto de treinamento (*overfitting*) e as mineradas pelo AG multi-objetivos apresentam um melhor grau de generalização.

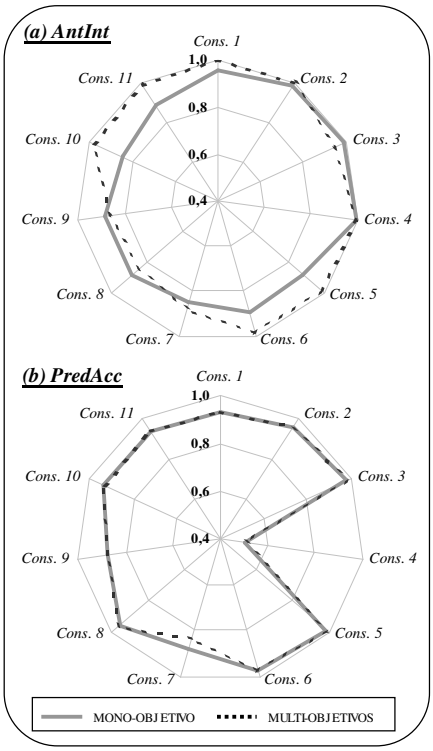

Figura 3. Comparação dos resultados mono e multi-objetivos na base completa: (a) métrica *AntInt* (b) métrica *PredAcc*.

Tabela 2. Regras obtidas pelo AG multi-objetivos na base completa.

Melhores Regras
SE (*domestic*=1) E (*catsize*=0) ENTÃO (*predator*=0)
SE (*airbone*=0) E (*aquatic*=1) E (*catsize*=1) ENTÃO (*predator*=1)
SE (*hair*=0) E (*predator*=1) ENTÃO (*domestic*=0)
SE (*eggs*=0) E (*venomous*=0) E (*tail*=0) E (*catsize*=0) ENTÃO (*domestic*=1)
SE (*milk*=1) E (*venomous*=0) ENTÃO (*type*=1)
SE (*feathers*=1) E (*venomous*=0) ENTÃO (*type*=2)
SE (*hair*=0) E (*aquatic*=0) E (*predator*=1) E (*toothed*=1) E (*domestic*=0) ENTÃO (*type*=3)
SE (*eggs*=1) E (*fins*=1) E (*domestic*=0) ENTÃO (*type*=4)
SE (*feathers*=0) E (*airbone*=0) E (*aquatic*=1) E (*breathes*=1) E (*catsize*=0) ENTÃO (*type*=5)
SE(*aquatic*=0)E(*fins*=0)E(*legs*=6)E(*catsize*=0) ENTÃO(*type*=6)
SE (*airbone*=0) E (*predator*=1) E (*backbone*=0) E (*domestic*=0) ENTÃO (*type*=7)

6 Conclusões

Investigamos a aplicação de um AG multi-objetivos, inspirado na família NSGA e NSGAII, para a mineração de regras precisas e interessantes na base de dados *Zoo*. Comparamos os resultados obtidos com aqueles gerados por um AG mono-objetivo, baseado no ambiente implementado em [Noda *et al*., 1999]. Observou-se que os experimentos que utilizaram 100% dos dados obtiveram resultados muito próximos para o grau de precisão das regras. Entretanto, o ambiente multi_objetivos retornou regras mais interesses. Acreditamos que essa diferença ocorreu devido à avaliação isolada das métricas e à dinâmica do

AGMO, que direcionou a busca para a região dos maiores valores em ambas as métricas. Já no mono-objetivo, o balanço entre as métricas foi definido *a priori*, dando um peso maior à previsão em relação ao grau de interesse. Na validação cruzada, nos dois AGs houve uma queda na precisão das regras avaliadas nas bases de teste. Apesar disso, constatou-se que os resultados no ambiente multi-objetivo foram superiores, resultando em um menor decaimento entre os valores de treinamento e teste. Isso indica que as regras obtidas nesse ambiente têm maior generalização que as do ambiente mono-objetivo.

Em [Noda *et al*., 1999] foram utilizas duas bases de dados públicas, *Zoo* e *Nursery*, com 101 e 12960 registros, respectivamente. Nos experimentos com a base *Nursery* chegamos a conclusões similares, que serão objeto de um artigo futuro. Muitos caminhos podem ser exploradas na área de AGMOs e DM como continuidade deste estudo, tais como: inclusão de outras métricas como a compreensibilidade; uso de outras técnicas de AGMO; utilização de outras funções de avaliação para as métricas, entre outras.

Agradecimentos

GMBO agradece ao CNPq e FAPEMIG pelo suporte.

Referências Bibliográficas

Bhattacharyya. S. (2000) Evolutionary algorithms in data mining: multi-objective performance modelling for direct marketing, *Proc. ACM SIGKDD Int. Conf. KDD-2000:*465-473.

Carvalho, D. R., Freitas, A. A., Ebecken, N. F. F. (2003) A critical review of rule surprisingness measures, *Proc. Data Mining IV - Int. Conf. on Data Mining*, 545-556.

Coello, C. A. C. (1996) An empirical study of evolutionary techniques for multiobjective optimization in engineering design, *Phd Thesis*, Tulane University.

Deb. K., Goel, T. (2001) Controlled Elitist Non-dominated Sorting Genetic Algorithms for Better Convergence, *Evolutionary Multi-Criterion Optimization*, 67-81.

Freitas, A. (2002) A survey of evolutionary algorithms for data mining and knowledge discovery, In: Ghosh and Tsutsui (ed). *Advances in Evolutionary Computation*. Springer-Verlag.

Freitas, A. (2004) A critical review of multi-objective optimization in data mining: a position paper, *ACM SIGKDD Explorations Newsletter*, 6:2.

Ghosh, A. and Nath, B. (2004) Multi-objective rule mining using genetic algorithms. *Information Sciences*, 163(1-3): 123-133.

Goldberg, D. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*, A-Wesley.

Iglesia, B., Philpott, M., Bagnall, A. and Smith, V. (2003) Data mining rules using multi-objective evolutionary algorithms, *Proc. of Congress on Evolutionary Computation*: 1552-1559.

Kim, Y., Street, W. N., Menczer, F. (2000) Feature selection in unsupervised learning via evolutionary search. *Proc. ACM SIGKDD Int. Conf. KDD-2000*: 365-369.

Noda, E., Freitas, A. A., Lopes, H. S. (1999) Discovering Interesting Prediction Rules with a Genetic Algorithm, *Proc .of Congress on Evolutionary Computation:* 1322-1329.

Srinivas, N., Deb, K. (1994) Multiobjective Optimization using nondominated sorting in genetic algorithms, *Evol. Computation*, 2(3): 221.

Takiguti, M. C. S. (2003) Utilização de Algoritmos Genéticos Multi-Objetivos na Mineração de Regras Precisas e Interessantes, *Dissertação de Mestrado*, Universidade Presb. Mackenzie.

Tabela 3. Comparação da métrica *PredAcc* de treino e teste nos experimentos mono e multi-objetivos.

Conseqüente	Mono-Objetivo		Multi-Objetivos	
	Treino	Teste	Treino	Teste
1	0,919	0,200	0,922	0,433
2	0,947	0,900	0,948	0,933
3	0,979	0,975	0,983	0,978
4	0,471	0,050	0,500	0,080
5	0,98	1,000	0,985	1,000
6	0,964	1,000	0,970	1,000
7	0,842	0,400	0,842	0,430
8	0,948	0,800	0,950	0,800
9	0,833	0,600	0,858	0,600
10	0,915	0,500	0,922	0,600
11	0,917	1,000	0,920	1,000

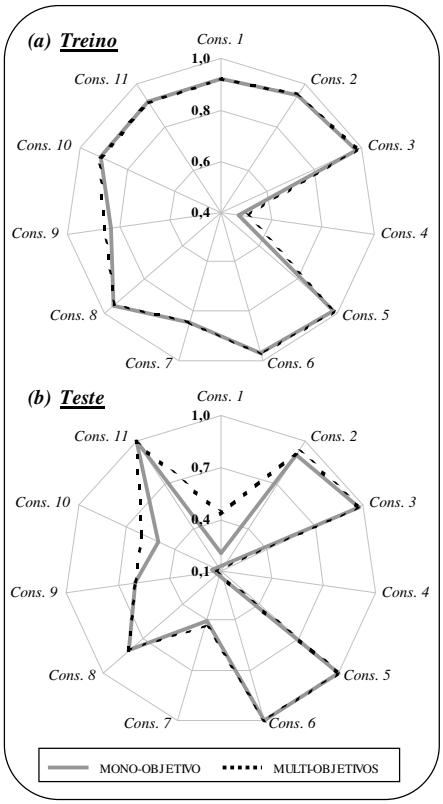

Figura 4. Comparação da métrica *PredAcc* dos ambientes mono e multi-objetivos na validação cruzada: (a) treino (b) teste.