Revista Brasileira de Agrocomputação, v.2, n.1, p.11-18, Jun.2004
Ponta Grossa-PR, DEINFO/UEPG - http://www.agrocomputacao.uepg.br

ANÁLISE DE RESULTADOS OBTIDOS POR TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL NA MINERAÇÃO DE DADOS DE PRODUTIVIDADE DO SOLO

Leila Maria Vriesmann[1], Alaine Margarete Guimarães[2], Marcelo Giovanetti Canteri[3], José Paulo Molin[4], Angelo Cataneo[5], Danilo Kovalechyn[6]

Recebido para publicação em 29/03/2003
Aprovado para publicação em 20/04/2004

RESUMO: *O mapeamento de fatores ambientais, de solo e de produtividade de uma determinada cultura produz uma grande quantidade de dados que podem conter informações importantes a serem usadas em processos de tomada de decisão sobre ações no campo. A extração de informações potencialmente úteis e implícitas em bases de dados é o principal objetivo da área denominada Mineração de Dados, a qual utiliza técnicas de Inteligência Artificial, como Árvore de Decisão (AD) e Algoritmos Genéticos (AGs), na execução de suas tarefas. O objetivo desse trabalho foi comparar resultados obtidos por essas técnicas na mineração de dados de características físico-químicas do solo e de produtividade da soja, obtidos experimentalmente. Os códigos para a Árvore de Decisão e para o Algoritmo Genético foram implementados nas linguagens Prolog e Borland Delphi® Professional, respectivamente. Os resultados foram apresentados na forma de regras de produção, tendo sido a meta obter regras que predissessem índices de produtividade acima de 2t/ha, com 100% de confiança. As regras geradas pelo algoritmo de Árvore de Decisão utilizaram os operadores < ou >= para relacionar um determinado valor a cada um dos atributos analisados. O Algoritmo Genético, pela facilidade em manipular números contínuos, possibilitou o uso de mais operadores, aceitando também nas regras a adoção de intervalo de valores para um atributo. Quanto à composição das regras, a AD apresentou maior variedade de atributos, enquanto que no AG a variabilidade notada foi mais especificamente nos valores, uma vez que as regras concentraram-se em torno de certos atributos, considerados os mais importantes. Assim, conclui-se que o Algoritmo Genético possibilitou um melhor tratamento dos dados pela diversidade de operadores, pela busca global e por simplificar o pré-processamento dos atributos, manipulando valores contínuos.*

Palavras-chave: *Algoritmo Genético, Árvore de Decisão, características físico-químicas do solo*

[1] Bolsista Pibic/CNPq, Acadêmica do Curso de Bacharelado em Informática, Laboratório InfoAgro, Departamento de Informática, Universidade Estadual de Ponta Grossa – Ponta Grossa/PR – Brasil, *leilavriesmann@bol.com.br*.
[2] Doutoranda em Agronomia, FCA/UNESP – Botucatu/SP – Brasil , *alainemg@fca.unesp.br.* Professora no Departamento de Informática, Universidade Estadual de Ponta Grossa – Ponta Grossa/PR – Brasil.
[3] Doutor em Agronomia, Professor no Departamento de Informática, Universidade Estadual de Ponta Grossa – Ponta Grossa/PR – Brasil, *mgcanter@uepg.br*.
[4] Doutor em Agronomia, Professor na Escola Superior de Agricultura Luis de Queiroz, Universidade de São Paulo – Piracicaba/SP – Brasil, *jpmoli@carpa.ciagri.usp.br*.
[5] Doutor em Agronomia, Professor no Departamento de Gestão e Tecnologia Agroindustrial, FCA/UNESP - Botucatu/SP - Brasil, *angelo@fca.unesp.br*.
[6] Acadêmico do Curso de Bacharelado em Informática, Departamento de Informática, Universidade Estadual de Ponta Grossa, Ponta Grossa/PR, Brasil.

ANALYSIS OF RESULTS OBTAINED THROUGH ARTIFICIAL INTELLIGENCE TECHNIQUES IN SOIL YIELD DATA MINING

ABSTRACT: *The environmental factors, soil and productivity mapping of a certain culture produces a great amount of data, which may contain important information to be used in the decision making processes concerning actions in the field. The extraction of implicit and potentially useful information from databases is the main objective of the area denominated Data Mining. This area uses Artificial Intelligence techniques such as Decision Tree and Genetic Algorithms, in order to accomplish its tasks. The objective of this work was to compare results obtained through those techniques in data mining about soil physiochemical characteristics and soy productivity, obtained experimentally. The Decision Tree and Genetic Algorithm were implemented in Prolog and Borland Delphi® Professional languages, respectively. The results were presented in the form of production rules, being the goal to obtain rules to predict productivity indexes above 2t/ha, 100% reliable. The rules generated by the algorithm of the Decision Tree used the operators < or >= to relate a certain value to each of the analyzed attributes. The Genetic Algorithm, due to the easiness in handling continuous values, enabled the use of more operators, also accepting the adoption of values interval for an attribute in the rules. As for the composition of the rules, AD presented larger variety of attributes, while in AG the variability was noticed more specifically in the values, once the rules pondered around certain attributes, considered the most important. Therefore, the conclusion is that the Genetic Algorithm enabled better treatment of the data due to the diversity of operators, global search and pre-process simplification by manipulating continuous values.*

Keywords: *Genetic Algorithm, Decision Tree, soil physical-chemical properties*

1 INTRODUÇÃO

A variabilidade no índice de produtividade de uma cultura em diferentes pontos de uma determinada área de cultivo induz ao pensamento de que características ambientais e do solo exercem influência sobre os resultados obtidos. O mapeamento de fatores ambientais, de solo e de produtividade em um campo, segundo Stafford (2000), produz uma grande quantidade de dados que o produtor pode utilizar em um processo de decisão. Esses dados, quando bem manipulados, podem contribuir para o aumento da produtividade a custos reduzidos.

Segundo Molin et al. (2001), literaturas recentes apresentam muitos exemplos onde os fatores de produtividade têm sido listados baseados na correlação entre parâmetros de fertilidade do solo e produtividade. A descoberta de tais padrões constitui-se de uma tarefa que pode ser executada utilizando-se a técnica de Mineração de Dados.

A Mineração de Dados é uma das etapas do processo de descoberta de conhecimento em banco de dados (KDD – Knowledge Discovery in Databases) e tem por objetivo extrair informações implícitas e potencialmente úteis de dados (Fayyad et al., 1996). Suas técnicas envolvem fundamentos computacionais, ligados intimamente à área de Inteligência Artificial, que propiciam a construção de algoritmos que possibilitam a busca por padrões implícitos.

Os parâmetros de fertilidade do solo (características físico-químicas) e o índice de produtividade podem ser rotulados, respectivamente, como atributos preditivos (ou previsores) e atributo objetivo (ou meta). Segundo Lopes (1999), a descoberta de algum tipo de relacionamento entre os atributos preditivos e o atributo objetivo, de modo a se obter um conhecimento que possa ser utilizado para prever a classe de uma tupla (registro) desconhecida, ou seja, que ainda não possui uma classe definida, é princípio da tarefa de classificação. As técnicas de Inteligência Artificial que aplicam algoritmos de classificação incluem, de acordo com Langley (1996), Árvore de Decisão (AD) e Algoritmos Genéticos (AGs).

Uma Árvore de Decisão é utilizada para descobrir regras e relacionamentos partindo e subdividindo a informação contida nos dados (Chou, 1991). Um objeto é classificado seguindo o caminho da raiz da árvore até a folha de acordo com os valores de seus atributos.

Os Algoritmos Genéticos são algoritmos de busca e otimização baseados na analogia com os processos de seleção natural e genética evolucionária (Goldberg, 1989). A essência do método consiste em manter uma população de indivíduos (cromossomos), os quais representam possíveis soluções para um problema específico. Melhores soluções podem ser atingidas por meio de um processo de seleção competitiva, envolvendo cruzamentos e mutações (Herrera, 1996).

Segundo Carvalho & Freitas (2000), a maioria dos algoritmos de indução de regras (como Árvore de Decisão) trabalham uma única solução candidata ao mesmo tempo, e normalmente avaliam uma solução candidata parcial, baseada somente em informação local enquanto que os AGs trabalham com uma população de soluções candidatas, que são avaliadas totalmente pela função de *fitness*[1].

A performance de um algoritmo de classificação depende muito do domínio da aplicação (Freitas, 2000). Um algoritmo de mineração pode utilizar uma ou mais técnicas associadas. Recentes pesquisas têm mostrado que para alguns domínios é interessante o uso de mais de uma técnica associada, formando algoritmos híbridos, como em Kim & Ham (2000), onde se trabalha com Algoritmos Genéticos e Redes Neurais, e em Carvalho & Freitas (2000), onde são utilizados Árvores de Decisão e Algoritmos Genéticos. Dessa maneira, faz-se necessário avaliar as técnicas de forma individual para perceber as vantagens e as desvantagens de cada uma delas em um determinado domínio.

O presente trabalho teve por objetivo analisar resultados obtidos por técnicas de Inteligência Artificial na mineração de dados de características físico-químicas do solo associadas à produtividade da soja produtividade. As técnicas utilizadas foram Árvores de Decisão, que emprega o método de busca local em seu espaço de estados, e Algoritmos Genéticos, que executam uma busca global.

2 MATERIAL E MÉTODOS

2.1 Base de dados

Os dados foram coletados na região de Campos Novos, SP. Uma área foi dividida em 2388 células quadradas, onde cada célula continha informações sobre produtividade da soja (atributo meta) e características físico-químicas do solo (atributos previsores). As informações foram agrupadas e coletadas usando aparelhos de GPS (Global Position System).

O atributo meta considerado foi a produtividade (*Produt*) e os atributos predicados considerados foram: *pH* (acidez do solo), *Ctc* (capacidade de troca catiônica), *V* (saturação), *H+Al* (hidrogênio e alumínio), *Ca* (cálcio), *Mg* (magnésio), *Mn* (manganês), *P* (fósforo), *K* (potássio), *Bo* (boro), *Zn* (zinco), *Cu* (cobre), *Fe* (ferro), *M.O.* (matéria orgânica), *Areia*, *Silte* e *Argila*. Todos os valores dos atributos foram obtidos no ano 2000, exceto a *Areia*, o *Silte* e a *Argila* que foram coletados em 1999. Um exemplo de parte da base de dados pode ser visualizado na Tabela 01, que possui uma coluna nomeada *Id* representando o identificador de cada célula.

Tabela 01 - Exemplo de parte da base de dados.

Id	Produt	Zn	V	Ph	...	Silte	Argila
163	2,9760	3,1880	71,2500	6,0330	...	5,0780	22,1200
164	3,0040	3,0890	71,2500	6,0330	...	5,0050	22,1200
255	2,7950	3,6430	70,5000	6,0080	...	5,2470	24,3700
256	2,8060	3,4880	71,4200	6,0420	...	5,1270	23,0500
257	3,0250	3,3340	71,2500	6,0330	...	5,0940	22,1200
258	3,1430	3,1770	71,2500	6,0330	...	5,0050	22,1200
259	3,0280	3,0320	71,2500	6,0420	...	4,8270	22,2500
...	...	...	...	...	...	...	...
8843	2,5030	0,5727	68,9000	5,7500	...	3,6280	20,2400
8935	2,5710	0,5807	69,2500	5,7750	...	3,5560	20,2300
8936	2,5660	0,6063	68,9000	5,7500	...	3,6830	20,2400

[1] *Fitness* é uma função matemática que associa um valor numérico a uma solução candidata, representando uma medida de qualidade.

A base de dados foi adquirida junto a Fundação ABC – Castro-PR e sofreu alguns refinamentos na fase de pré-processamento, objetivando preparar os dados para a mineração. Os refinamentos constituíram-se da eliminação de ruídos, como dados imprecisos ou ausentes. Para a aplicação do algoritmo de Árvore de Decisão, os atributos foram discretizados, ou seja, transformados em 2 classes distintas (yes/no, indicando se o valor de cada atributo é ou não maior que um determinado limiar estabelecido). Para o uso do AG, a base permaneceu como mostra a Tabela 01, não havendo necessidade de discretização. Dessa maneira, obtiveram-se 2 arquivos, um para cada técnica.

Na tarefa de classificação é comum utilizar uma base de dados de treinamento e outra base de dados para teste. A base de treinamento é responsável pela geração de regras, enquanto que a base de teste visa comprovar os resultados obtidos. Dessa maneira, a base de dados utilizada foi particionada aleatoriamente em 70% para treinamento e 30% para teste. A divisão foi realizada de forma independente no arquivo com dados discretizados e no arquivo com dados sem discretização.

2.2 Representação do conhecimento

Antes da modelagem das técnicas utilizadas para a Mineração de Dados, analisou-se como o conhecimento descoberto deveria ser apresentado, uma vez que o formato da saída pode influenciar na maneira de representar os dados e suas transformações no algoritmo. No contexto da tarefa de classificação, o conhecimento descoberto muitas vezes é expresso como um conjunto de regras de classificação do tipo SE-ENTÃO, uma vez que esse tipo de representação do conhecimento é intuitivo para o usuário (Carvalho & Freitas, 2000).

Segundo Romão et al. (2000), regras do tipo SE-ENTÃO são chamadas regras de produção, constituindo uma forma de representação simbólica e possuindo a seguinte forma:

SE <antecedente> ENTÃO <conseqüente>

O antecedente é formado por expressões condicionais envolvendo atributos do domínio da aplicação existentes nos bancos de dados. O conseqüente é formado por expressões que indicam a previsão de algum valor para um atributo meta a partir dos valores dos atributos previsores. Para o domínio em estudo, o antecedente foi constituído pelas características físico-químicas do solo e seus valores e o conseqüente pela classe procurada no índice de produtividade.

2.3 Técnicas de Inteligência Artificial utilizadas

2.3.1 Árvores de Decisão

A Árvore de Decisão baseou-se no algoritmo C45 (Quinlan, 1993). Os nós da árvore representaram os atributos, as arestas os valores possíveis e as folhas as classes.

Para se construir a árvore, algumas decisões tiveram que ser tomadas. Primeiramente houve a necessidade de decidir quais atributos físico-químicos do solo seriam considerados mais importantes, os quais deveriam ser alocados em nós próximos ao topo da árvore, enquanto que os menos relevantes são considerados nos nós próximos às folhas. Depois, estudou-se o valor (limiar) a ser utilizado como teste em cada nó. A Tabela 02 apresenta os limiares para cada um dos atributos físico-químicos.

A Árvore de Decisão foi projetada para ser do tipo binária, ou seja, cada nó teria duas arestas: uma para quando o valor encontrado na base é menor do que o limiar e outra em caso contrário. Finalmente, estipulou-se o número de classes e os valores para cada uma delas, constituindo assim as folhas. Foram utilizadas duas classes: uma para indicar que a regra predizia um índice de produtividade maior ou igual a 2 t/ha e outra para índice menor que 2 t/ha.

Tabela 02 - Limiares para cada um dos atributos físico-químicos.

Atributo	Zn	V	Ph	M.O.	Mg	Fe	Ca	B	P
Valor	1,221	70,86	5,97	19,705	12,265	18,15	28,865	0,1596	38,265
Atributo	H+Al	Cu	K	Mn	Ctc	Areia	Silte	Argila	
Valor	17,46	0,7476	1,542	9,5045	62,94	80,725	4,354	19,819	

Cada objeto foi classificado seguindo um caminho da raiz da árvore até uma folha.

O algoritmo de Árvore de Decisão foi implementado em linguagem Prolog para Linux.

2.3.2 Algoritmos Genéticos

A menor unidade de um AG é chamada gene. Um gene representa uma unidade de informação do domínio do problema, ou no âmbito de Mineração de Dados, um valor de um atributo. Uma série de genes, ou seja, um cromossomo, representa uma possível solução completa para o problema, denominada regra candidata (Lopes, 1999).

A codificação dos genes baseou-se na consideração de que, tendo-se uma base de dados com vários atributos contínuos predicados (características físico-químicas do solo) e um atributo meta (produtividade da soja), um cromossomo corresponde a um registro e, um gene corresponde a um atributo predicado do registro. Os genes foram tratados de forma posicional, ou seja, o primeiro gene correspondeu ao primeiro atributo, o segundo gene correspondeu ao segundo atributo, e assim sucessivamente.

Embora os valores de cada atributo estivessem armazenados no gene, fez-se necessário desenvolver uma maneira de tratar esses valores no decorrer do processo evolutivo do Algoritmo Genético. À medida que um indivíduo evolui muitas vezes faz-se necessário eliminar de seu cromossomo genes não expressivos. A representação desse conhecimento no algoritmo foi feita por meio da adoção de um campo, denominado Peso, em cada gene. Esse campo corresponde a uma variável real pertencente ao intervalo de 0 a 1 que indica se aquele gene será ou não considerado na regra, de acordo com um limiar configurável. Quanto maior o limiar, menos chances um atributo tinha de pertencer à regra. O limite para o atributo Peso foi de 0,95.

Em relação aos operadores a serem considerados nas regras, optou-se por implementar o AG de forma a permitir o uso dos operadores =, <>, < e >=. Com objetivo de gerar regras mais eficientes foi projetada no algoritmo a possibilidade de que, em cada gene pudesse ser considerada uma faixa de valores para o atributo referenciado. Assim, um atributo poderia, por exemplo, ser maior ou igual a 62,1 e menor ou igual a 72,45. Em função disso incorporou-se à estrutura dos genes, campos para armazenar o operador e o(s) valor(es) do(s) atributo(s).

A Tabela 03 demonstra parte de um cromossomo (indivíduo) com seus campos preenchidos.

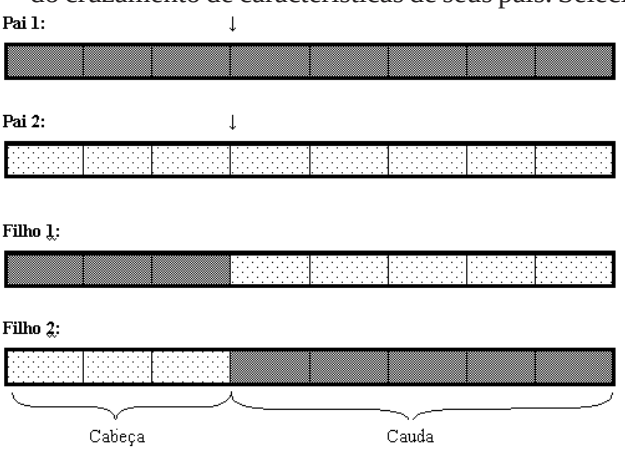
Tabela 03 - Parte de um cromossomo.

...	Gene[3]			...	Gene[12]				...
...	Peso	Operador	Valor	...	Peso	Operador	Valor		...
...	1	=	1,275	...	0,96	>=	3,2	1,21	...

Foram escolhidos 50 indivíduos para evoluírem por 50 gerações procurando por intervalo entre 2 e 5 t/ha no índice de produtividade por meio de mutação e cruzamento (crossover), que são os processos de reprodução do AG. O método utilizado para seleção dos indivíduos foi a roleta. Maiores detalhes desse método podem ser vistos em Goldberg (1989).

A mutação é responsável pela mudança aleatória de um gene. Essa mudança pode ocorrer nos campos Operador, Peso e Valor. A probabilidade de mutação adotada foi de 50%, almejando que cada indivíduo selecionado tivesse a mesma probabilidade de ser ou não modificado. A taxa de mutação, assim como os demais parâmetros utilizados, não foram baseados em critérios específicos, uma vez que não tinham sido realizados testes anteriores no algoritmo para justificação da aplicação de um determinado valor ao invés de outro.

O processo de crossover cria novos indivíduos por meio do cruzamento de características de seus pais. Seleciona-



mos
ato-
uda.
ndo
01).
enes
over

Figura 01 - Cruzamento entre dois indivíduos.

Baseado na lei da sobrevivência, onde somente os melhores permanecem, um novo indivíduo só é aceito na nova população se o seu valor de *fitness* é maior ou igual a um valor estipulado (0,65). Esse valor, pertencente ao intervalo entre 0 e 1, é fornecido por uma função de *fitness*, que indica a qualidade do indivíduo avaliado. A função utilizada foi:

$$fitness = \frac{tp}{(tp + fp)} \qquad (01)$$

onde:

tp = número de verdadeiros positivos encontrados na base de dados consultada, ou seja, número de registros onde os valores dos genes dos indivíduos foram encontrados predizendo o intervalo desejado no índice de produtividade (entre 2 e 5);

fp = número de falsos positivos encontrados na base de dados consultada, ou seja, número de registros onde os valores dos genes dos indivíduos foram encontrados não predizendo o intervalo desejado no índice de produtividade (entre 2 e 5).

Durante a criação das gerações, a função de *fitness* é avaliada na base de dados de treinamento. Na última geração, é também avaliada na base de teste, objetivando a confirmação ou não da qualidade da regra.

Para implementação do código do Algoritmo Genético foi utilizada a linguagem Borland Delphi® Professional, versão 5 em ambiente Windows 98.

3 RESULTADOS E DISCUSSÃO

As Tabelas 04 e 05 apresentam resultados, ou seja, regras na forma SE-ENTÃO obtidas pela Árvore de Decisão e pelo Algoritmo Genético, respectivamente. A parte SE constituiu-se dos valores dos atributos predicados (características físico-químicas do solo) e a parte ENTÃO demonstrou o valor do atributo meta (índice de produtividade).

Somente regras com 100% de confiança na base de treinamento e de teste foram selecionadas, ou seja, somente regras onde todos os registros considerados possuíam valores de produtividade dentro do intervalo procurado.

Tabela 04 - Regras para produtividade maior ou igual a 2 t/ha com 100% de confiança, obtidas por Árvore de Decisão.

Id	SE	E	E	E	E
1	Argila<19,815				
2	pH>=5,97	P>=38,265	Cu>=0,7476	Bo<0,1596	
3	pH>=5,97	P>=38,265	M.O.<19,705	Zn<1,221	
4	pH<5,97	V>=70,86	K<1,542	Zn>=1,221	Silte>=4,354
5	P>=38,265	Ctc<62,94	Zn<1,221	Bo>=0,1596	
6	P>=38,265	K>=1,542	Zn<1,221	Bo>=0,1596	
7	P>=38,265	K>=1,542	Zn>=1,221	Bo<0,1596	Areia>=80,725
8	Ctc>=62,94	Mn>=9,5045			

Tabela 05 - Regras obtidas por Algoritmos Genéticos, com 100% de confiabilidade, para produtividade entre 2 t/ha e 5 t/ha.

Id	SE	E	E	E	E
1	V<=67,45	Areia<=82,7561	M.O.<=23,55		
2	V<=67,45	Fe>22,4995			
3	V<=67,45	Fe>23,1457			
4	K=1,275	8,26>=Silte>=3,28			
5	V<=67,45	23,55>=M.O.>=19,99	Fe>22,4995	pH=5,79	Silte<=8,26
6	Mg=13,85	Mn=2,9			
7	V<=67,45	8,11>=Silte>=5,039	pH=5,79		
8	Mg=13,85	8,11>=Silte>=5,039	Fe=24,55	Mn=2,9	

As duas tabelas apresentam regras para o índice de produtividade superior a 2 t/ha, uma vez que não existiam registros na base de dados com valor superior a 5 t/ha de produtividade.

A primeira coluna das tabelas constitui-se de um número que identifica cada regra. As demais colunas demonstram os atributos predicados com seus respectivos valores. O atributo meta de todas as regras é "*Produtividade* >= 2". Assim, tomando-se como exemplo a linha da Tabela 04 cujo valor de "*Id*" é 1, tem-se a regra "Se *Argila* < 19,815 então *Produtividade* >= 2". Já na linha onde o valor de "*Id*" é 2, tem-se a regra "Se *pH* >= 5,97 e *P* >= 38,265 e *Cu* >= 0,7476 e *Bo* < 0,1596 então *Produtividade* >= 2".

A Tabela 05 tem o mesmo formato da Tabela 04. Assim, a regra obtida na primeira linha da Tabela 05 é "Se *V* <= 67,45 e *Areia* <= 82,7561 e *M.O.* <= 23,55 então *Produtividade* >= 2".

As regras obtidas por Árvore de Decisão demonstraram sempre um determinado valor para uma característica, e usaram apenas os operadores < e >= . Os Algoritmos Genéticos permitiram quatro diferentes operadores, além do uso de intervalos de valores, o que é importante quando se trata de dados contínuos.

O AG obteve dois resultados bastante semelhantes nas regras 2 e 3, sendo que, o que diferencia um do outro é o valor do atributo *Fe*, que na regra 2 apresenta um valor um pouco inferior ao da regra 3. Isso não seria possível com a Árvore de Decisão, uma vez que demonstra somente um determinado valor para o ponto de teste de cada atributo.

Dos atributos utilizados nas regras finais de um método, nem todos foram levados em consideração pelo outro método. O *Fe* e o *Mg* não aparecem nas regras da Tabela 04 e a *Argila*, *P*, *Cu*, *Bo*, *Zn* e *Ctc* não são observados na Tabela 05.

Observou-se que o Algoritmo Genético gerou regras concentrando-se em torno de certos atributos, variando valores, enquanto que a Árvore de Decisão combinou atributos. Essa última afirmação pode ser comprovada nas regras 2 e 3, assim como nas regras 6 e 7 da Tabela 04, onde os dois primeiros atributos são mantidos, mudando somente os demais. Isso ocorreu justamente pelo fato de que o AG evolui preservando os melhores elementos, ou seja, os atributos predicados que mais interferem no atributo meta.

4. CONCLUSÕES

Os resultados obtidos permitiram concluir que, para o domínio estudado, o uso de Algoritmos Genéticos apresentou algumas vantagens em relação à Árvore de Decisão:

- O AG possibilitou utilizar vários operadores relacionais, enquanto que a AD fez uso somente de dois operadores (>= e <).
- O AG desenvolvido permitiu a geração de intervalos de valores para os atributos nas regras, valorizando a característica dos dados serem contínuos. Já a AD empregou um único valor para cada atributo, onde a regra gerada revelava se o atributo estava abaixo ou acima do valor estipulado.
- Não foi preciso realizar uma discretização prévia dos atributos envolvidos no estudo de caso para o AG, preservando seus valores originais, enquanto que na AD foi necessária a realização da fase de discretização.

5 REFERÊNCIAS BIBLIOGRÁFICAS

CHOU, P.A. Optimal partitioning for classification and regression trees. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** 13, p.340-354, 1991.

CARVALHO, D.R., FREITAS, A.A. A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in data mining. **Proc.** 2000 Genetic and Evolutionary Computation Conf. (GECCO-2000), 1061-1068. Las Vegas, NV, USA. July 2000.

FAYYAD, U.M., PIATETSKI-SHAPIRO, G., SMYTH, P. From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining, 1996. Menlo Park: AAAI Press, 1996, p.11-34.

FREITAS, A.A. Undestanding the crucial differences between classification and discovery of association rules. A position paper. **SIGKDD Explorations**, v.2,n.1, p.65-69, 2000.

GOLDBERG, D.E. **Genetic algorithms in search, optimization, and machine learning**. New York: Addison-Wesley, 1989, 412p.

HERRERA, F. **Tackling real-coded genetic algorithms: operators and tools for behavioral analysis.** 1996.

KIM, K., HAN, I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. **Expert Systems with Applications,** 19, p.125-132, 2000.

LANGLEY, P. **Elements of machine learning.** San Francisco, CA: Morgan Kaufmann, 1996, 419p.

LOPES, C.H.P. **Classificação de registros em banco de dados por evolução de regras de associação utilizando algoritmos genéticos**. **R**io de Janeiro, 1999. 136f. Dissertação (Mestrado em Engenharia Elétrica/Sistemas de Computação) – Pontifícia Universidade Católica do Rio de Janeiro.

MOLIN, J.P., COUTO, H.T.Z., GIMENEZ, L.M., PAULETTI, V., MOLIN, R., VIEIRA, S.R. Regression and correlation analysis of grid soil data versus cell spatial data. In: THIRD EUROPEAN CONFERENCE ON PRECISION AGRICULTURE, 2001. **Anais...** Montpellier: Agro Montpellier, 2001, p.449-453.

QUINLAN, J.R. **C4.5: Programs for Machine Learning.** San Francisco, CA: Morgan Kaufmann, 1993.

ROMÃO, W., FREITAS, A.A.; PACHECO, R.S. Uma revisão de abordagens genético-difusas para descoberta de conhecimento em banco de dados. **Acta Scientiarum**, v.22, n.5, p.1347-1359, 2000.

STAFFORD, J.V. Implementing precision agriculture in the 21st century. **J. agric. Engng Res. 2000**, p.267-275, 2000.