Mineração de dados georreferenciados aplicando diferentes funções de *fitness*

Thomas Jensen[1], Salua Sane Lima Azevedo[2], Sérgio Roberto Spitzner Júnior[3], Alaine Margarete Guimarães[4], Leila Maria Vriesmann[5]

[1]Acadêmico do Curso de Engenharia de Computação, Lab. InfoAgro, Universidade Estadual de Ponta Grossa

[2]Acadêmica do Curso de Engenharia de Computação, Lab. InfoAgro, Universidade Estadual de Ponta Grossa

[3]Acadêmico do Curso de Engenharia de Computação, Lab. InfoAgro, Universidade Estadual de Ponta Grossa

[4]Professora Doutora do Departamento de Informática da Universidade Estadual de Ponta Grossa, Pesquisadora Lab. InfoAgro

[5]Professora MSc. do Departamento de Informática da Universidade Estadual de Ponta Grossa, Pesquisadora Lab. InfoAgro

thomasjensen13@hotmail.com, salua_lim@yahoo.com.br, sergio_spitzner@hotmail.com, alainemg@uepg.br, leila@inf.ufpr.br

Resumo. *Na gestão do conhecimento é sabido que a visualização do mesmo e sua localização espacial podem contribuir para sua compreensão e para a adoção de novas estratégias. Esse trabalho teve como objetivos incorporar ao software MinAG, desenvolvido pelo laboratório InfoAgro – UEPG, um módulo de tratamento de diferentes funções de fitness e também contribuir para o desenvolvimento do projeto de visualização e interpretação de conhecimento.Foi utilizada uma base de dados georreferenciados, do setor agrícola. As regras geradas foram importadas no Sistema de Informações Geográficas SPRING e foram obtidos mapas apresentando a localização espacial das regras na propriedade agrícola.*

1 Introdução

A comunidade científica apresenta preocupação em agregar aos sistemas de informação novos conhecimentos, os quais possam ser apresentados em uma forma atrativa e de fácil compreensão dos mesmos. Alguns segmentos, em especial a agricultura, demanda de uma visualização em forma de mapas das informações e dos conhecimentos descobertos.

Nesse contexto, uma nova área de pesquisa, denominada mineração de dados georreferenciados, vem sendo explorada no sentido de gerar novos conhecimentos associados à localização espacial do mesmo.

Com o objetivo de implementar a associação de mineração de dados georreferenciados com recursos visuais, o grupo de pesquisa em informática aplicada a

agricultura (InfoAgro - UEPG) vem desenvolvendo pesquisas sobre mineração de dados e geocomputação (Gahegan, 1999), visando incorporar ao software de mineração de dados ambientais, denominado MinAG (Guimarães, 2005), um módulo que permita a visualização espacial das regras descobertas.

Esse trabalho teve dois principais objetivos: incorporar ao software MinAG, que realiza a tarefa de classificação de dados, um módulo de tratamento de diferentes funções de *fitness*; e também contribuir para o desenvolvimento do projeto de interface homem-máquina na construção e visualização de conhecimento.

2 Mineração de dados com diferentes funções de *fitness*

Uma das principais preocupações na mineração de dados está relacionada a como avaliar a qualidade de um conhecimento descoberto, já que este deve apresentar um bom grau de acerto. Um método utilizado é a função de avaliação, ou simplesmente, função de *fitness*.

Há várias formas de se calcular o *fitness* de uma regra. O software MinAG, que utiliza a técnica de Algoritmos Genéticos (Mitchell, 1997) , disponibiliza quatro delas, que são: precisão, sensibilidade, consistência e acurácia. O usuário pode então escolher, para cada mineração que for executar, qual a função de *fitness* a ser aplicada, sendo possível ainda utilizar uma combinação das mesmas. As funções de *fitness* (F) foram extraídas de Freitas (2002) e estão descritas no Quadro 1.

Quadro 1 Fórmulas das Funções de *Fitness*

Função	Fórmula
Precisão	$F = \frac{VP}{VP + FP}$
Sensibilidade	$F = \frac{VP}{VP + FN}$
Consistência	$F = \frac{VN}{VN + FP}$
Acurácia	$F = \frac{VP + VN}{VP + FP + FN + VN}$

Onde,

- VP (verdadeiro positivo): é o número de casos cobertos pela regra que tem a classe predita pela regra;

- FP (falso positivo): é o número de casos cobertos pela regra que tem uma classe diferente da classe predita pela regra;

- FN (falso negativo): é o número de casos que não estão cobertos pela regra, mas tem a classe predita pela regra;

- VN (verdadeiro negativo): é o número de casos que não estão cobertos pela regra e não tem a classe predita pela regra.

A avaliação da função de *fitness* é relativa visto que, por exemplo, uma regra pode apresentar *fitness* de sensibilidade próximo a 1, ou seja, ocorrer em grande parte da área de cultivo, mas, sendo assim, ela possivelmente não seria interessante, já que por ocorrer na maioria da área pode ser uma regra correta porém sem novidades.

Ou ainda, uma regra que obtenha *fitness* de precisão próximo a 1, não quer dizer que seja uma regra boa, pois pode acabar sendo uma exceção.

Uma forma de aperfeiçoar uma regra é considerar a combinação de funções de *fitness*, multiplicando uma função pela outra. Por exemplo, se combinarmos uma função de *fitness* de precisão com uma de sensibilidade, teremos uma regra que seja bem precisa e que ocorra em uma parte da área de cultivo que não a faça ser uma exceção.

3 Mineração de Dados Georreferenciados

Segundo Gahegan et al (2000), bases de dados geográficos continuam a se tornar mais complexas – várias áreas de análise geográfica agora têm acesso a vários tipos de dados digitais, e inversamente bases de dados convencionais agora contém também explicitamente ou implicitamente referências espaciais. Desta forma, os dados resultantes são vastos e previamente inexplorados, por serem ricos em termos de atributos e também por apresentarem um relevante número de objetos. Conseqüentemente, esses bancos de dados podem ajudar a estabelecer o conhecimento do mundo real ou a descobrir e entender o processo de construção do modelo da realidade humana, ou seja, representam desafios que possibilitam a extensão de diferentes áreas de pesquisa atuais.

A descoberta do conhecimento pode ser vista como um processo de análise para achar estruturas válidas e úteis em um conjunto de dados, reconhecendo-os e identificando meios de diferenciar padrões, induzindo a características que podem ser associadas a relacionamentos dentro do campo de interesse.

Descobrir dados ou estruturas de dados pode oferecer uma compreensão significativa para domínios complexos, levando a vantagens comerciais e aumento da produtividade. Como afirma Koperski (1999), as atividades de mineração de dados têm visto os dados geoespaciais como uma rica origem de estruturas e padrões, além de relacionar com as saídas do processo de construção de conhecimento.

Associar este conhecimento descoberto à visualização apresenta a vantagem complementar de aumentar o grau de interação com o usuário, fazendo com que ele explore mais os dados, elevando sua participação no processo.

4 Estudo de caso

Considerando-se que o grupo de pesquisa do Laboratório InfoAgro desenvolve soluções computacionais e sistemas de informação aplicados às ciências agrárias e ambientais, a base de dados utilizada nesse trabalho refere-se à cultura da soja, obtida por processos de agricultura de precisão.

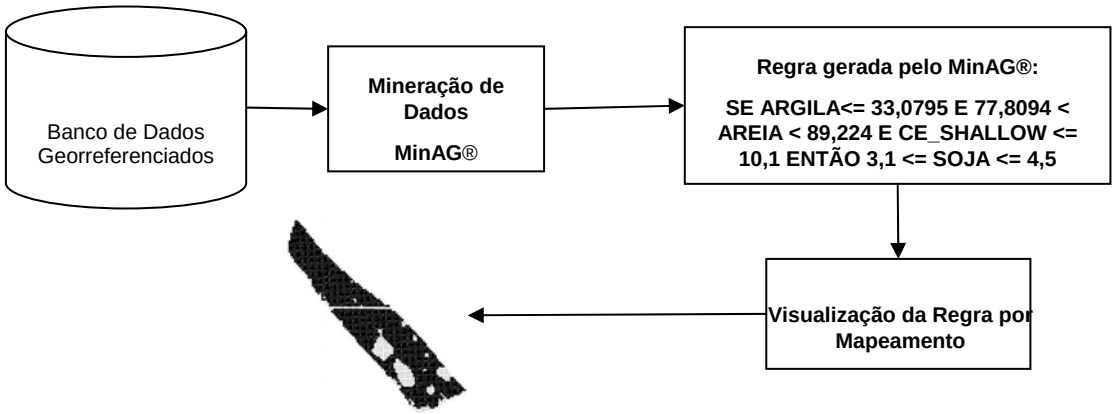
A variabilidade no índice de produtividade de uma cultura em diferentes pontos de uma determinada área de cultivo induz ao pensamento de que características ambientais e do solo exercem influência sobre os resultados obtidos. O mapeamento de fatores ambientais, de solo e de produtividade em um campo, segundo Stafford (2000), produz uma grande quantidade de dados que o produtor pode utilizar em um processo de decisão. Esses dados físico-químicos, quando combinados e bem manipulados, podem gerar regras que determinem alta produtividade. Mas essas regras podem ser avaliadas de diferentes formas.

Utilizou-se na aplicação uma base de dados agrícola, com referências da área de produção de soja em uma propriedade localizada em Campos Novos, SP. Constam na base todos os pontos georreferenciados na forma de coordenadas geográficas (X, Y), além de seus respectivos identificadores (chaves) e atributos físicos colhidos da análise do solo (como por exemplo, a quantidade de matéria orgânica, de areia e outros existentes em cada ponto). Dentre estes, está o índice de produtividade de soja, que na fase de visualização será constatado como categoria objeto.

Os dados foram submetidos ao software MinAG. Após 'minerados', os dados são apresentados sob a forma de regras (SE... ENTÃO) que são obtidas de acordo com o cálculo de *fitness* requisitado pelo usuário (acurácia, consistência, precisão e sensibilidade).

Com a descoberta e o reconhecimento de padrões na base é possível identificá-los na área de cobertura da propriedade. Serviu como recurso de conexão visual, o software SPRING versão 4.3Beta4, um Sistema de Informações Geográficas (SIG), para geração dos mapas. A Figura 1 apresenta o processo de construção de conhecimento associado a recursos visuais definido para esse trabalho.

Figura 1 Processo de Construção de Conhecimento com a Associação de Recursos Visuais



Construiu-se, a partir da base de dados com os pontos georreferenciados e atributos da área produtora, dois arquivos para o gerenciamento da unicidade do banco pelo SPRING. Um destes contém cada localização (em graus), identificação (rótulo), além da categoria objeto que irá ligar os arquivos, determinando os pontos 2D no mapa. O segundo arquivo consiste na descrição dos atributos e seus valores relativos para cada

ponto novamente identificado por um número único da base. Esses arquivos foram então convertidos para serem importados para o SPRING.

Já no ambiente do software para a visualização, ocorre a ativação de um banco de dados para concentrar todos os dados, figuras e um projeto que também deve ser definido para envolver todos os pontos. É também necessária a criação de duas categorias: a cadastral, para associar as representações gráficas (pontos 2D) e a temática, bem como as suas respectivas classes temáticas, para que cada regra seja associada a um plano de informação.

5 Resultados

Os parâmetros utilizados para o algoritmo genético foram: 0,544 para semente, 50% da base original para treinamento e 50% para teste. A classe procurada foi aquela com produtividade entre 3,1 e 4,5t/ha. Foram utilizadas 750 gerações. A probabilidade para cruzamento foi de 0,95 e para mutação 0,3. Usou-se torneio de seleção igual a 3 e porcentagem dos genes mutados igual a 0,5. A margem de mutação do peso de cada atributo foi 20%, enquanto que a probabilidade de mutar o operador foi de 66%. Permaneceram nas gerações apenas os atributos com peso maior que 0,6. Para a mutação do valor, adotou-se a probabilidade de 5%.

Para cada execução, usou-se uma combinação de função de *fitness*. As regras geradas são mostradas na Tabela 1 com a respectiva função de *fitness* e seus valores na base de teste.

Tabela 1 Exemplos de regras geradas pelo software MinAG® para as diferentes funções de *fitness* e seu respectivo valor na base de teste

Função de *Fitness*	**Regra Gerada**	***Fitness***
Precisão	SE 60,5>=MO>=14,7829 E 3,8792<Silte<10,94 E 86,6235<Areia<91,58 E CE_Shallow>9,0758	0,3793
Sensibilidade	SE Argila<=33,0795 E 77,8094<Areia<89,224 E CE_Shallow<=10,1	0,8272
Consistência	SE 46,3355>=MO>=27,14 E 91,58>=Areia>=81,4607 E 3,6144>=IC35_40mp>=1,8951 E CE_Shallow>10,1	0,9977
Acurácia	SE 27,14<MO<36,601 E Silte<=6,4657 E 11>=CE_Deep>=5,3299 E CE_Shallow>10,1	0,7301
Precisão* Sensibilidade	SE 14,5246<MO<60,332 E Silte<=10,1632 E 34,13>=Argila>=18,9802 E 85,97<Areia<91,58 E IC35_40mp<=1,8951 E CE_Shallow>7,6479	0,004

O mapa da área total da propriedade foi gerado a partir da classe cadastral, demonstrando cada ponto nela georreferenciado, agora sob a forma de objeto. Com o cursor de informações, ao clicar sobre qualquer objeto, obtém-se a identificação do

objeto, sua correta localização e os valores dos atributos pertinentes ao ponto. Usando o modelo de dados da classe temática, as regras foram submetidas através do recurso de geração de coleções, oferecido pelo software.

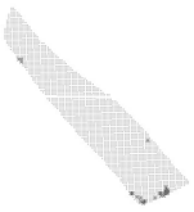
Cada coleção (satisfação da regra nos pontos) obteve sua marca visual possibilitando assim, observá-las simultânea ou individualmente sobre o mapa. Desenhando todas as regras desejadas, e clicando novamente sobre um ponto pode-se verificar além dos atributos, quais regras estão sendo satisfeitas em determinado local. A Figura 2 apresenta a visualização de quatro regras diferentes, sendo que cada uma foi obtida com cálculo de *fitness* específico.

Figura 2 Mapas de Visualização de Regras

Mapa 1 – Visualização da Regra Gerada por Acurácia: SE 24,14 < MO < 36,601 E SILTE <= 6,4657 E 11 >= CE_DEEP >= 5,3299 E CE_SHALLOW > 10,1

Mapa 2 – Visualização da Regra Gerada por Consistência: SE 46,3355 >= MO >= 27,14 E 91,58 >= Areia >= 81,46607 E 3,6144 >= IC >= 1,8951 E CE_SHALLOW > 10,1

Mapa 3 – Visualização da Regra Gerada por Precisão: SE 60,5 >= MO >= 14,7829 E 3,8792 < SILTE < 10,94 E 86,6235 < AREIA < 91,58 E CE_SHALLOW > 9,0758



Mapa 4 – Visualização da Regra Gerada por Sensibilidade: SE ARGILA <= 33,0795 E AREIA < 89,224 E CE_SHALLOW <= 10,1

6 Considerações finais

Baseando-se nos resultados parciais da pesquisa, confirmou-se a possibilidade da implementação de uma ferramenta para inspecionar, verificar e editar o conhecimento extraído. A requisição da apresentação de localidade dos dados georreferenciados e a representação do conhecimento geográfico mostram a maior problemática por conter linguagens universais do foco da geografia.

O ambiente de visualização deve facilitar um duplo caminho de trocas de conhecimento. O usuário pode ser capaz de realçar exemplos nos dados, selecionar interesses, refinando os resultados do aprendizado de máquina e talvez selecionar os métodos de mineração de dados apropriados ou objetivos. Reciprocamente, os resultados da mineração e construção do conhecimento são trazidos para o usuário na pretensão de promover a compreensão, para definir a funcionalidade requerida.

Considerando a evolução dos estudos realizados até o momento, observa-se a relevância da visualização de regras para o aferimento por parte do usuário, ampliando as possibilidades de construção de conhecimento.

Referências

FREITAS, A. (2002) **Data mining and knowledge discovery with evolutionary algorithms.** Berlin; New York: Springer.

GAHEGAN, M. (1999), **What is Geocomputation? Transactions in GIS.** Vol.3, No. 3, pp. 203-206.

GAHEGAN, M. et al., (2000) **The Integration of Geographic Visualization with knowledge Discovery in Databases and Geocomputation.** ICA Commission on Visualization: Working Group on Database-Visualization Links, Univerity Park, PA, USA.

GUIMARÃES, A. M. (2005) **Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo.** Tese de doutorado. Botucatu : [s.n.].

KOPERSKI, K. H, J. and ADHIKARY, J. (1999) **Mining knowledge in geographic data**. Comm. ACM (to appear). URL: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html.

MITCHELL, M. (1997) **An introduction to genetic algorithms**. Cambridge: Mit Press.

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