Aplicação do algoritmo SimpleKMeans em experimento de milho verde

Wesley Viana[1], **Prof. MSc. Marcos de Moraes Sousa** [1], **Prof. MSc. Júnio César de Lima**[1]
Prof. Dr. Milton Sérgio Dornelles [1]

[1]Instituto Federal Goiano – Campus Urutaí.
Fazenda Palmital, Km 2,5 Zona Rural 75790-000 – Urutai, GO – Brasil

viana.wesley@gmail.com, {marcos.moraes,junio.lima}@ifgoiano.edu.br,

msdornelles@yahoo.com.br

Abstract. *This article presents a data mining algorithm application in a corn experiment conducted at the Instituto Federal Goiano - Campus Urutaí. The aim was to apply clustering technique implemented by the WEKA software in productivity per hectometer variable. The algorithm used was SimpleKMean with 10 tests ranging from 2 to 11 clusters and then evaluate those with the lowest SSE (sum of squared errors). The results show that the cluster of higher productivity of the cultivar was number 2 with a linear density of 2.6667 grains per meter and 60 cm spacing.*

Resumo. *Este artigo apresenta uma aplicação de algoritmo de mineração de dados em um experimento de milho verde realizado no Instituto Federal Goiano - Campus Urutaí. O objetivo foi aplicar técnica de agrupamento implementado pelo software WEKA na variável produtividade por hectare. O algoritmo utilizado foi o SimpleKMean com 10 testes variando de 2 até 11 clusters para então avaliar os que apresentam o menor SSE (soma de quadrado de erros). Os resultados mostram que o cluster de maior produtividade foi do cultivar de número 2 com a densidade linear de 2,6667 grãos por metro linear e espaçamento de 60 cm.*

1 Introdução

Atualmente muitos dos experimentos científicos têm apresentado uma extensa quantidade de resultados, que exigem que o pesquisador tenha um conhecimento considerável sobre estatística e outras áreas da matemática, para que assim, este possa analisar de forma coerente os resultados coletados, para assim chegar a informações úteis.

Mineração de dados é o processo de extrair informações válidas a partir de grandes bases de dados (WESTPHAL e Braxton, 1998) e pode ser utilizada como uma ferramenta que objetiva analisar uma lista de resultados e gerar conhecimento a partir destes. A utilização desta ferramenta não reduz a importância da estatística e outras ferramentas, mais sim proporciona ao pesquisador um novo recurso. Assim os experimentos científicos se tornam um ótimo campo para mineração de dados.

Diversos usos de mineração de dados tem sido realizados na agricultura nos últimos anos. Criavelenti(2009) utilizou da mineração de dados para avaliar um banco de dados digitais para avaliar a relação solo-paisagem. Guimarães *et al* (2002) desenvolveram um algoritmo baseado em mineração de dados para determinar a produtividade de uma safra em função das propriedades físico-químicas do solo.

Ao longo do processo de mineração de dados são utilizadas de várias ferramentas estatísticas e matemáticas, mas de forma implícita, desta forma o pesquisador não precisa ter um real conhecimento de como esta a proceder.

Como a mineração de dados é rica em algoritmos, é necessário softwares que concentrem estes vários algoritmos em um só local, e os tornem acessíveis para o usuário. Assim um usuário leigo pode fazer uso da mineração de dados, não há assim necessidade conhecer computação de forma profunda. Por este motivo utilizou-se do software WEKA para este trabalho, que é um dos mais ricos softwares de mineração de dados da atualidade e é um software licenciado sobre a GPL.

Além disso, outro motivo da escolha do software WEKA, é que ele é multi-plataforma, desta forma pode ser usado em qualquer sistema operacional do mercado que disponha da máquina virtual Java.

Dos diversos tipos de algoritmos disponibilizados pelo WEKA, foi escolhido um algoritmo de agrupamento ou *clustering*. O agrupamento é uma técnica que é usada para particionar os registros de uma base de dados em subconjuntos ou *cluster*. Nesta tarefa os registros são agrupados segundo algum critério de semelhança(Dias,2001).

O objetivo deste artigo é mostrar a aplicação do algoritmo de agrupamento implementado pelo *WEKA* e baseado no *K-Mean*, que foi o *SimpleKMean*, em um resultado de experimento de cultivares de milho verde numero 002/2008 GPE do Instituto Federal Goiano - Campus Urutaí, para assim encontrar o *cluster* que apresenta a maior produtividade por hectare, para que em futuros plantios, possa ser utilizado dos dados coletados, a partir deste *cluster*, para que se ganhe maior produtividade e, consequentemente, maior lucratividade.

2 Metodologia

Uma base de dados é um aglomerado de informações que podem ser lidas e interpretadas. Tem-se como exemplos planilhas eletrônicas, bancos de dados, listas e qualquer forma de se representar diversas informações de forma organizada.

A base de dados utilizada foi uma planilha eletrônica que contem o resultado do projeto que é um experimento de plantio de milho verde, onde o pesquisador fez a variação do espaçamento, cultivar, densidade e bloco. O espaçamento é o espaço entre uma carreira de milho e outra, sendo este testado com os valores de 45, 60, 75, 90 centímetros. O cultivar é o tipo de semente usada, neste caso foi utilizado de dois cultivares diferentes. A densidade é a quantidade de semente em cada metro linear, assim utilizado de 4 diferentes densidades, que foram de 2, 3, 4, 5. Já o bloco é qual região em que foi plantado, assim utilizado 4 diferentes blocos.

Uma das respostas do experimento, entre outras, foi a produtividade/hectare. Esta também é a variável alvo da Mineração de Dados aplicada, gerando e analisando grupos paras as variáveis assim encontrando o grupo que apresenta a maior produtividade.

Foi utilizado o método de agrupamento para descoberta de grupos significativos, através do algoritmo *SimpleKMean*. Foram feitos 10 testes com 2 até 11 *clusters* para então avaliar os que apresentam o menor número de erro.

O *SimpleKMean* é um algoritmo que cria grupos fazendo uso da média aritmética, cria-se assim a quantidade de grupos solicitada pelo usuário, através da fórmula da figura

1, onde a média é a soma das observações dividida pelo número delas (Ferreira,2005).

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} (Xi)$$

Figure 1. Média Aritmética

Pela análise apresentada na figura 1, nota-se que n é o numero de observações, Xi, como o valor de determinada observação no índice i, chega-se assim que a média aritmética é a somatória de Xi, onde i varia de 1 a n, onde que n representa o número de observações.

Este algoritmo apresenta sua eficiência através de um dado chamado de *sum of squared errors* (SSE), que é a soma dos quadrados dos erros. O quadrado de erros (que pode ser obtido com a fórmula da figura 2) é uma medida para quantificar a diferença entre os valores da média aritmética, assim pode-se mostrar a precisão desta média.

$$MSE(\overline{X}) = E((\overline{X} - \mu)^2)) = (\frac{\sigma}{\sqrt{n}})^2$$

Figure 2. Quadrado de erros

Por fazer o uso de média aritmética, pode-se encontrar valores que não foram testados. Isso pode ter vantagens ou desvantagens, sendo que isso depende dos dados que estão sendo trabalhados.

A vantagem de encontrar dados que não foram testados pelo pesquisador é que esses trazem novas possibilidades e visões ao pesquisador, que agora tem uma sugestão para um futuro experimento e, assim, tentar melhorar a produtividade. A desvantagem é que não há como saber se estes dados encontrados realmente vão trazer o resultado esperado, necessitando assim fazer um novo experimento para confirmar estes dados.

Com o propósito de chegar ao melhor resultado foi aplicado o algoritmo *SimpleKMean* repetidas vezes, com a alteração somente da quantidade de *clusters* a serem gerados, tentando encontrar a quantidade que apresenta-se o menor SSE.

Para avaliar o algoritmo foram feitas 10 simulações para verificar em quantos *clusters* seria a quantidade eficiente para minimizar o SEE, visando não criar uma quantidade muito excessiva de *clusters*. Foi requisitado ao algoritmo que construísse primeiramente 2 *clusters*, logo em seguida 3, e assim sucessivamente até chegar-se a 11 *clusters*.

3 Resultados

A partir dos testes com diversos números de *clusters* no algoritmo *SimpleKMean* foi criada a tabela 1, que mostra a quantidade de *clusters* e o respectivo SSE. O gráfico da figura 3 foi gerado a partir dos dados da tabela 1. Desta forma verifica-se que com 11 *clusters* foi encontrado o menor SSE, assim 11 *clusters* pode ser considerado uma quantidade plausível de *clusters*.

Table 1. Quantidade de Cluster X SSE

Quantidade de Clusters	SSE
2 Clusters	39,56
3 Clusters	31,56
4 Clusters	28,05
5 Clusters	21,63
6 Clusters	17,95
7 Clusters	17,38
8 Clusters	12,64
9 Clusters	12,4
10 Clusters	8,92
11 Clusters	8,32

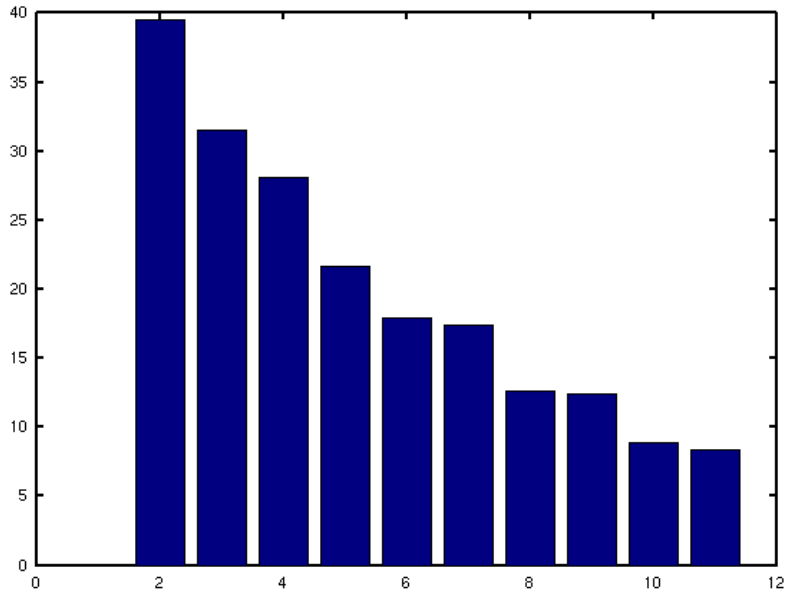

Figure 3. Quantidade de Cluster X SSE

Tem-se então que a quantidade de *cluster* a ser criada será de 11 *cluster*, pode-se ver que a tabela 2 apresenta os 11 *clusters* gerados pelo algoritmo *K-Mean*, com seus respectivos valores para cada um de seus atributos.

Table 2. clusters de gerados pelo SimpleKMean

CLUSTER	REGISTROS	ESPAÇAMENTO	CULTIVAR	DENSIDADE	PRODT.EC(kg/ha)
9	4	90	2	2	8472,2222
7	4	90	2	3	9687,5000
2	16	82,5	1	2,5	9798,6111
3	10	72	2	2,4	11106,9444
10	16	82,5	1	4,5	11604,1667
5	16	82,5	2	4,5	11977,4306
1	8	45	2	2,5	12199,0741
6	16	52,5	1	2,5	12630,2082
8	16	52,5	1	4,5	12977,4306
4	16	52,5	2	4,5	13906,2500
11	6	60	2	2,6667	15312,5000

De uma forma mais ampla de visualização tem-se o gráfico da figura 4 que mostra os *clusters* no eixo x e a produtividade no eixo y, onde os dados representados na cor laranja são os de cultivar de numero 2 e os azuis são do cultivar de numero 1.

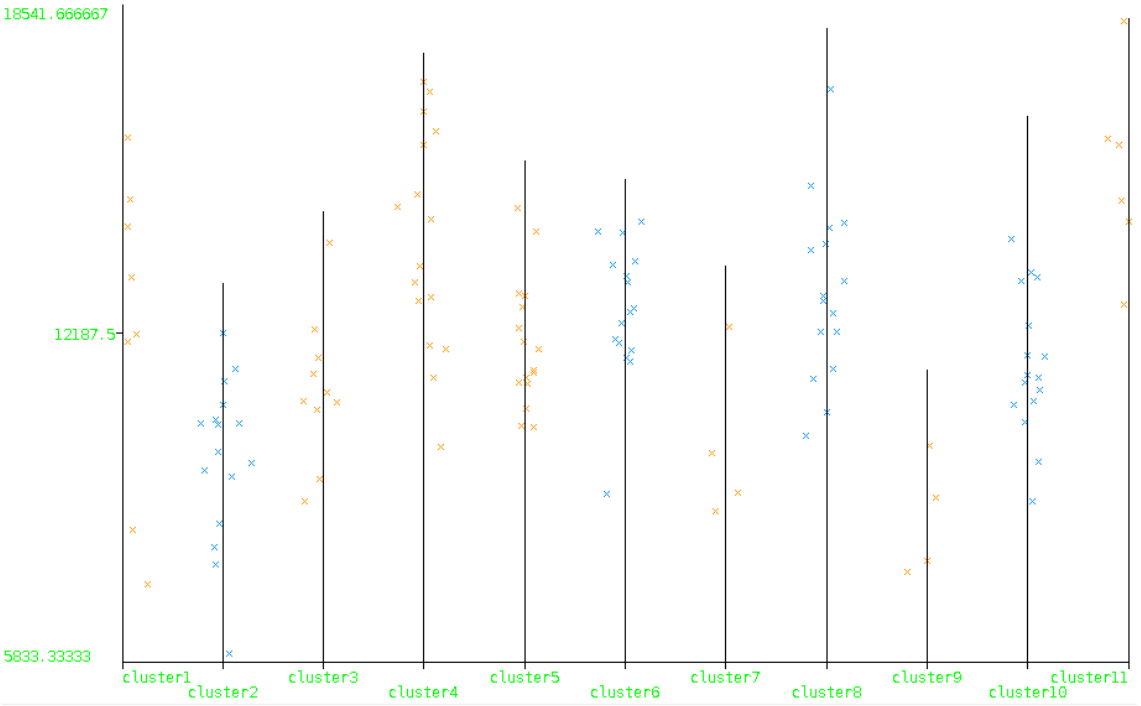

Figure 4. Representação gráfica dos clusters

Ao observar a tabela 2 nota-se que o *cluster* de número 11 apresentou a maior produtividade por hectare, sendo o objetivo da aplicação encontrar a maior produtividade possível. Pode-se notar que este *cluster* utilizou o cultivar de número 2, utilizou do valor de 2,6667 para sua densidade e espaçamento de 60 cm. Assim são essas as medidas medias para se alcançar o *cluster* de maior produtividade.

A partir da tabela 2 é possivel chegar a outras conclusões como: que o cultivar número 2 teve maior produtividade do que o cultivar número 1 e que a densidade do *cluster* que apresentou a maior produtividade possui um valor não existente no experimento, podendo assim enquadrar como 2 ou 3.

4 Considerações Finais

Os testes realizados neste trabalho foram baseados na aplicação do algoritmo de agrupamento implementado pelo *WEKA* e baseado no *K-Mean* em um resultado de experimento de cultivares de milho verde para encontrar o *cluster* que apresenta a maior produtividade por hectare, assim, não foram exploradas todas as informações apresentadas no relatório do algoritmo, mas sim encontrar um *cluster* que apresenta-se maior produtividade em relação aos outros.

Não foi estendido os testes para uma maior quantidade de *clusters* pois, a medida que aumenta-se a quantidade de *clusters*, diminui a quantidade de erro. Assim deve-se observar que chegaria um ponto onde a quantidade de *clusters* seria a mesma da quantidade de registros na base de dados, assim o algoritmo colocaria cada registro em um unico *cluster*, desta forma chega-se a erro zero, torna-se assim irrelevante as informações obtidas.

Os testes foram encerrados com 11 *clusters*, pois com essa quantidade de *clusters* foi encontrado um *cluster* com produtividade bem superior aos outros, pode-se assim enquadrar este *cluster* como o de maior produtividade.

Sugere-se para estudos posteriores a avaliação de outros algoritmos (como exemplo, o EM - *Expectation Maximization*) e inclusão de outras variáveis (como exemplo, a quantidade de espigas comerciais).

References

CRIVELENT. R. C. *Mineração de Dados para Inferência de Relação Solo-Paisagem em Mapeamentos Digitais de Solos.* Desertação de Mestrado. Campinas-SP: Instituto Agronômico, 2009.

DIAS. M. M. *Um modelo de formalização do processo de sistema de descoberta de Conhecimento em banco de dados.* Tese de Doutorado. Florianópolis-SC: Universidade Federal de Santa Catarina, 2001.

FERREIRA. D. F. *Estatística Básica.* 1 Ed. Lavras-MG: Universidade Federal de Lavras, 2005.

GUIMARÃES. A. M. ,VRIESMANN. L. M. ,CANTERI. M. G. ,CATANEO. A. ,MOLIN J. P. *Desenvolvimento de um Algoritmo Genético para Mineração de Dados na Agricultura de Precisão.* Viçosa-MG: Anais do 2o Simpósio Internacional de Agricultura de Precisão, 2002.

Machine Learning Group at University of Waikato. *WEKA* http://www.cs.waikato.ac.nz/ml/weka/

Projeto PIBIC/CNPQ N.002/GPE/2008. *Espaçamento entre linhas e densidade linear de cultivares de milho em Urutaí-GO* Cordenados Dr. Milton Sergio Dornelles.

WESTPHAL. C. e BRAXTON. T. *Data Mining Solutions: Methods and Tools for Solving Real-Word Problems.* New York, 1998.