COMPARAÇÃO DE ALGORITMOS DE CLUSTERING HIERÁRQUICO EM DADOS REAIS: UM ESTUDO DE CASO NA AGRICULTURA[1]

TREVISAN, T. B. [1]; ZIGLIOLI, M. [1]; MALLMANN, A. A. [1]; METZ, J. [1]; PAULA FILHO, P. L. [1]

[1] Núcleo de Ciência da Computação.

Câmpus Medianeira

{thibtrevisan, marceloziglioli, a.a.mallmann, jmetzz, plpf2004}@gmail.com

Resumo

Um dos custos relacionados à atividade agrícola provém da aplicação de defensivos ou adubos na área de plantação. Usualmente a aplicação desses compostos é realizada em toda a área cultivada gerando gastos desnecessários. Com intuito de minimizar esses gastos, técnicas de agricultura de precisão são utilizadas para identificar as subáreas da lavoura nas quais se faz necessária a aplicação desses compostos. Neste trabalho apresentamos um estudo de caso considerando informações físico-químicas do solo e aspectos de coloração das folhas de soja para a análise exploratória desses dados por meio de aprendizado de máquina não-supervisionado. O objetivo é comparar diferentes algoritmos de agrupamento hierárquico para identificar os métodos mais adequados para esse domínio de aplicação e futuramente encontrar evidências para facilitar a identificação automática das necessidades específicas de cada subárea da lavoura. A partir dos resultados obtidos observa-se a presença de alguns padrões nos dados, os quais serão explorados e explicados em trabalhos futuros

Palavras-chave: Clustering hierárquico; Aprendizagem de máquina; Agricultura de precisão; Agrupamento de dados.

[1] Este trabalho faz parte do projeto de pesquisa "Análise inteligente de dados aplicados aos sistemas de produção agrícola", UTFPR - Medianeira.

Abstract

The use of pesticides and fertilizers in plantation areas represent a relevant cost in agriculture activity. The application of these composts is usually carried out over all the cultivated area causing unnecessary expenses. Aiming at minimizing these expenses, precision agriculture technics are used to identify the subareas where these composts should be applied. In this paper, we present a simple case study regarding soil physicochemical information and coloring aspects of soy leaves for exploratory data analysis by means of unsupervised clustering algorithms comparison. The aim is to compare different hierarchical clustering algorithms, identify appropriate methods for this application domain and find evidences to facilitate automatic identification of subareas specific needs. The initial results obtained show the presence of several patterns in the analyzed data, which will be further exploited in future work.

Keywords: Hierarchical clustering; Machine learning; Precision agriculture; Data clustering

INTRODUÇÃO

Com a evolução tecnológica cresceu também o número de alternativas de utilização da tecnologia em outras áreas. No campo, a cada dia se busca a adaptação de tecnologias para as necessidades rurais. Agricultores tentam, há muitos anos, maximizar a produção tanto no aspecto físico quanto econômico das culturas, variando a aplicação de defensivos e nutrientes de acordo com os tipos de solo e culturas [1].

A agricultura de precisão visa atender esta demanda de maneira eficiente e automatizada, com a criação de padrões e utilização de métodos inteligentes para se adaptarem as mudanças na lavoura. Nesse contexto, neste trabalho apresentamos um estudo de caso da aplicação de técnicas de agrupamento de dados provenientes da agricultura, mais especificamente considerando informações físico-químicas de solo e de coloração de folhas da soja coletadas em diversos pontos da plantação com auxílio de um sistema de posicionamento global (GPS). Como trabalho inicial, a motivação para a elaboração deste artigo é a comparação de diferentes métodos de clustering hierárquico sobre os dados brutos a fim de verificar a existência de padrões nos dados e a correlação entre as diferentes soluções de agrupamentos encontradas por esses métodos.

CONCEITOS E MÉTODOS

No método de agrupamento de dados o algoritmo de aprendizado analisa os exemplos fornecidos e os agrupa em clusters segundo algum critério de similaridade. Um desses métodos é chamado de clustering hierárquico. Existem duas abordagens para a construção de agrupamentos hierárquicos: (1) a abordagem aglomerativa, a qual começa com clusters unitários (compostos de apenas um exemplo) e segue agrupando os clusters

iterativamente, sempre considerando o par de cluster mais semelhantes com base na distância entre os mesmos [2]. Esse processo é repetido até que exista somente um único agrupamento contendo todos os exemplos; e (2) a abordagem divisiva, que realiza a mesma tarefa, porém, em ordem inversa [3].

Existem diversas maneiras para calcular a distância entre dois pontos, dentre as mais usadas estão a distância Euclidiana e a Manhattan. A medida Euclidiana é o comprimento do segmento de reta que une dois pontos. A medida Manhattan, por sua vez, representa a distância de dois pontos com base na soma das diferenças absolutas de dois objetos [4].

Para facilitar a compreensão do clustering hierárquico, os agrupamentos são graficamente representados por meio de um dendograma, o qual é uma estrutura especial do tipo árvore que facilita a visualização dos agrupamentos apresentado como os nós são combinados em cada etapa do processo [2].

Existem diversos métodos de agrupamento de dados hierárquico aglomerativo, como o single linkage, complete linkage, group-average, centroid e ward. O single linkage, também conhecido como nearest neighbor, é um dos métodos mais simples de agrupamento de dados. A principal característica desse método é que a distância entre grupos é definida pelo par de exemplos mais próximos pertencentes a clusters diferentes. O complete linkage, também conhecido furthest neighbor, é o oposto do single linkage, no sentido de que a distância entre grupos é agora definida pelos exemplos mais distantes de cada grupo. No caso do group-average, a distância entre dois clusters é definida pela média das distâncias entre todos os exemplos de cada grupo. O método baseado no centroide estima um ponto médio para cada cluster considerando todos os exemplos nele presentes e, a partir desse ponto médio (chamado centroide), é calculada a distância entre os clusters. O método ward considera a menor perda de informação possível para o agrupamento de dois clusters. O critério utilizado é o da soma dos quadrados dos erros [5]. Neste trabalho foram utilizados os métodos single linkage, complete linkage e group-average.

ESTUDO DE CASO

Na agricultura, após detectar a falta de nutrientes ou defensivos na plantação, o produtor frequentemente aplica produtos em toda a área de plantio, consequentemente, desperdiçando recursos e possivelmente tornando os alimentos menos saudáveis e, até mesmo, impróprios para consumo. Este projeto visa atender o produtor rural com a técnica de agricultura de precisão, na qual toda a plantação é dividida em setores, e cada setor é analisado individualmente para detectar as necessidades específicas de maneira independente dos demais facilitando a aplicação de nutrientes ou defensivos na plantação e principalmente minimizando os custos de produção.

Com as técnicas tradicionalmente utilizadas, a análise de informações setorizadas é altamente custosa e demorada além de necessitar de profissionais especializados, o que torna, por vezes, essa tarefa

inviável. Com métodos de Aprendizado de Máquina, essa tarefa pode ser automatizada ou pelo menos realizada de maneira semi-automática, simplificando a tarefa do especialista e agilizando o processo de análise das informações. Por meio do resultado obtido com a aplicação do clustering hierárquico, os agrupamentos são analisados em busca de padrões que possam ajudar o especialista a identificar as necessidades das áreas da plantação. Neste trabalho inicial a ideia é comparar as diferentes soluções de agrupamentos obtidos por meio de três métodos de clustering hierárquico e verificar se existe alguma estrutura inerente aos dados analisados. A comparação é feita por meio da correlação entre os agrupamentos encontrados pelos diferentes algoritmos. Casos com alta correlação significam que, embora a estratégias para identificar os agrupamentos seja diferente, ambas as soluções encontram grupos semelhantes, reforçando a presença dos padrões nos dados.

Conjunto de dados. Neste trabalho foram utilizados dois conjuntos de dados, cujas informações são sumarizadas na Tabela 1.

Tabela 1. Conjunto de dados.

NOME	NUM. ATRIBUTOS	NUM. EXEMPLOS
Folhas	105	3990
Solo	16	115

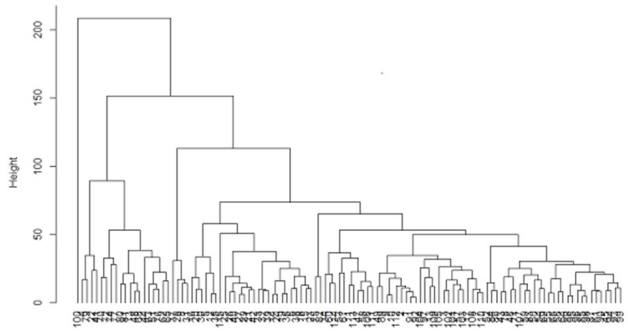
Os atributos do conjunto de dados Folhas são valores estatísticos como média, mediana, variância e curtose, com base em modelos de cor (RGB, HSV e outros) extraídos a partir de imagens de folhas da soja. Os exemplos correspondem a cada amostra em pontos geograficamente referenciados previamente escolhidos para abranger toda a área da plantação. O conjunto de dados Solo representa as análises físico-químicas do solo a partir de amostras coletadas nos mesmos pontos georeferenciados utilizados para coleta das folhas e foram obtidas através de análise laboratorial (Fósforo, Potássio, Cobre, Ferro, entre outros).

Resultados experimentais. Em virtude da limitação de espaço para elaboração deste trabalho, o dendograma (Figura 1a) e o mapa de intensidade de cores (Figura 1b) são apresentados apenas para o conjunto de dados Solo e considerando o método complete linkage. Por meio dessas figuras observa-se que há agrupamentos bem definidos nos dados, uma vez que as áreas com cores mais intensas no mapa de cores representam agrupamento de elementos semelhantes. Porém, com uma análise mais detalhada dos dendogramas resultantes dos outros métodos analisados, observamos que os mesmos são pouco correlacionados.

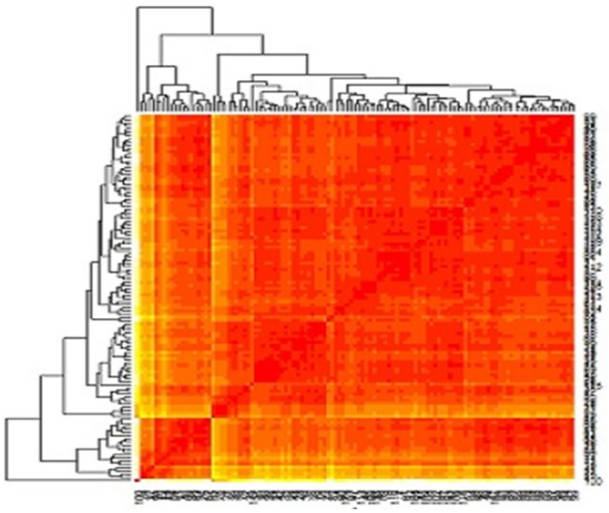
Se por um lado os métodos mostram que existem agrupamentos bem definidos nos dados, por outro lado, cada método identificou um tipo de padrão distinto dos demais métodos. Para validar essa interpretação dos dendogramas, foi calculada a correlação entre as soluções obtidas pelos diferentes algoritmos de clustering avaliados. O resultado é apresentado na Tabela 2. Quanto mais próximo de 1 o valor da correção, mas semelhantes são os agrupamentos identificados, ao passo que valores de correlação próximos de 0 indicam que os grupos são completamente distintos. Observa-se que para o conjunto de dados Solo os algoritmos single

e average linkage encontraram agrupamentos altamente correlacionados, o que indica que para esse conjunto de dados não há os dois algoritmos podem chegar a uma solução parecida.

(a)



(b)



linkage e (b) mapa de cores para o conjunto de dados Solo.

Tabela 2. Correlação dos diferentes métodos de linkage para cada base de dados.

SOLO			
	Single	Average	Complete
Single	1	0,86	0,48
Average	0,86	1	0,65
Complete	0,48	0,65	1

FOLHAS			
	Single	Average	Complete
Single	1	0,28	0,1
Average	0,28	1	0,59
Complete	0,1	0,59	1

CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado um estudo de caso inicial sobre a aplicação de agrupamento de dados hierárquico em bases de dados reais provenientes da área de agricultura. Como pretendido, foram identificados agrupamentos de dados nos conjuntos avaliados. Contudo, ainda há necessidade de explorar melhor esses resultados para interpretar os clusters construídos e investigar os padrões embutidos nesses dados. Essa análise mais criteriosa e detalhada fará parte de trabalhos futuros, nos quais a participação do especialista do domínio é de fundamental importância.

AGRADECIMENTOS

Agradecemos ao Professor Dr. Cláudio Leones Bazzi e seus alunos Douglas Castro Taube e Andressa Celant por terem disponibilizado o conjunto de dados do solo utilizados neste trabalhocoerente e preciso.

REFERÊNCIAS

[1] COELHO, A. M. Agricultura de precisão: manejo da variabilidade espacial e temporal dos solos e culturas. 1st. ed. Embrapa, 2005.

[2] JAIN, A. K.; DUBES, R. C. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, 1988.

[3] BERKHIN, P. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, 2002.

[4] HAIR, J. F.; BLACK, W. C.; BABIN, B.; ANDERSON, R. E. ; TATHAM, R. L. Análise multivariada de dados. 6th. ed. Bookman, 2009.

[5] EVERITT, B. S. Cluster Analysis. 3rd. ed. Arnold, 1993.