ANÁLISE DE DADOS DE FISIOLOGIA DE PLANTAS APOIADA POR TÉCNICAS DE VISUALIZAÇÃO DE INFORMAÇÕES

Lidiane de Andrade Parisi[1], Almir Olivette Artero[2], Gustavo Maia Souza[3]

[1]FIPP - Faculdade de Informática – UNOESTE - Universidade do Oeste Paulista, Presidente Prudente-SP. [2]FCT - Faculdade de Ciências e Tecnologia – UNESP – Presidente Prudente-SP. [3]Laboratório de Ecofisiologia Vegetal – Universidade do Oeste Paulista,Presidente Prudente-SP. Email: gustavo@unoeste.br

RESUMO

Este artigo apresenta uma proposta alternativa para a análise de dados de fisiologia de plantas, usando como apoio, técnicas de visualização de informações. Estas técnicas conseguem gerar visualizações diretamente a partir de dados de alta dimensionalidade, que auxiliam a compreensão dos dados e servem para auxiliar a identificação dos atributos mais relevantes para a discriminação das espécies estudadas. A identificação dos parâmetros mais relevantes é fundamental para a construção de um modelo de classificação, seja usando um classificador Bayesiano ou mesmo uma Rede Neural Artificial.

Palavras-chave: Fisiologia de Plantas, Análise de Dados, Visualização de Informações.

DATA ANALYSIS OF PLANT PHYSIOLOGY SUPPORTED BY TECHNICAL INFORMATION VISUALIZATION

ABSTRACT

This paper presents an alternative proposal for the analysis of data from plant physiology, using as support, information visualization techniques. These techniques can generate visualizations directly from data of high dimensionality, which can help to understand the data and serve to support the identification of the most relevant attributes for the discrimination of species. The selection of the most important parameters is essential for the construction of a classification model, either using a Bayesian classifier or Artificial Neural Network.

Keywords: Plant Physiology, Data Analysis, Information Visualization.

INTRODUÇÃO

Os experimentos realizados na área da Botânica sempre geram uma grande quantidade de dados, com uma alta dimensionalidade, por causa da necessidade de se repetir os experimentos com diversos exemplares de cada espécie de planta e, também, pela necessidade de se colher medidas de diversos parâmetros. A determinação de quais fatores mais influenciam no desenvolvimento e funcionamento das plantas é um destes experimentos de interesse na área, em que se deseja analisar a resposta de uma ou mais espécies sob diferentes condições de cultivo. Por causa da grande quantidade de dados de alta dimensionalidade, a análise dos dados gerados tem sido uma tarefa muito dificil, pois não existe uma ferramenta específica para a sua realização. De fato, a grande quantidade de dados gerada (vários registros com muitos atributos), torna inviável o uso de técnicas de análise simples, usadas normalmente com dados de baixa dimensionalidade. Deste modo é muito difícil identificar modelos capazes de descrever o comportamento das plantas em função dos parâmetros medidos e, então, identificar a influência de cada um deles no desenvolvimento das plantas. Um possível objetivo desta análise é encontrar situações incomuns, também conhecidas com outliers, ou seja, situações em que o comportamento da planta é muito diferente de outras de sua espécie. Outro objetivo bastante frequente é a busca por agrupamentos, que servem para caracterizar um comportamento em comum para os registros dentro do mesmo agrupamento e, encontrar quais são os atributos que melhor contribuem para a construção destes agrupamentos.

Tradicionalmente a analise dos dados nesta área e feita sem a aplicação de técnicas específicas, sendo as conclusões baseadas em análise qualitativas ou, de forma puramente numérica, o que torna trabalho muito lento e, por vezes, inviável. O uso de técnicas estatísticas, também muito comum, tem sido uma importante linha para se conduzir análises de dados em qualquer área, porém, o uso de todos os parâmetros colhidos nos experimentos, muitas vezes conduz a resultados ruins. De fato, a alta dimensionalidade dos dados impossibilita uma exploração visual inicial. Com o objetivo de tratar a alta dimensionalidade dos dados, o uso da análise de componentes principais (PCA) [Pearson, 1901] tem se mostrado uma das abordagens mais interessantes, pois consegue gerar uma boa projeção de dados multidimensionais em um espaço 2D ou 3D, que pode ser representado graficamente. Entretanto, o uso de todos os parâmetros medidos nas plantas pode levar a resultados muito ruins, pois, entre os atributos sempre existem aqueles que não são interessantes para discriminar o comportamento das diferentes espécies de plantas.

Mais recentemente, foram propostas diversas técnicas de visualização de informações [Card et al., 1999], que conseguem gerar representações de dados multidimensionais e, deste modo, permitem que diversos parâmetros possam ser analisados visualmente, ao mesmo tempo. Deste modo, esta abordagem alia a capacidade de armazenamento e processamento dos computadores para gerar gráficos, com a capacidade das pessoas de identificarem padrões neles. Assim, esta área vem se tornando uma ferramenta muito valiosa em diferentes áreas de aplicação e, a Biologia não é uma exceção.

Este trabalho mostra como técnicas de visualização de informações podem ser usadas para ajudar na análise de dados na área da Biologia. As demais seções deste trabalho estão organizadas da seguinte maneira: Na Seção 2, é brevemente apresentada a área da Visualização

de Informações, com as técnicas usadas neste trabalho; Na Seção 3 é discutido como foi desenvolvido o processo de análise e a ferramenta o que o auxilia; Por fim, na Seção 4 é apresenta algumas conclusões e sugestões para trabalhos futuros.

TRABALHOS RELACIONADOS

Com o uso massivo dos computadores em todas as áreas observou-se um grande crescimento na quantidade de dados que são coletados nos mais diversos experimentos e, nas Ciências Biológicas isto não tem sido diferente. Para a análise destes dados, o mais comum é tentar organizá-los em tabelas e, de algum modo, tentar gerar gráficos como: Histogramas de Frequências, Gráficos de Setores, Gráficos de Dispersão 2D e 3D, etc. Porém, essas formas simples de representação funcionam satisfatoriamente apenas quando o volume e a dimensão dos dados analisados são relativamente pequenos. Entretanto, quando o volume e a dimensionalidade dos dados crescem muito, e este é o caso mais comum, são necessários outros tipos de representações para resolver este problema. Em seguida, é discutida a abordagem tradicional de análise de dados na área biológica e, também são apresentadas as técnicas de visualização de informação usadas neste trabalho.

Abordagem tradicional

Tradicionalmente os dados obtidos em estudos na área Biológica são analisados de maneira bastante elementar, normalmente buscando analisar os dados puramente numéricos o que por vezes se torna difícil ou até mesmo inviável quando o volume dos dados é muito grande. Um outro problema que ocorre com esta abordagem é a dificuldade de se identificar valores atípicos nos dados, por causa da dificuldade de se visualizar os valores em uma tabela. Uma alternativa que tem sido usada também é a análise individual dos atributos do conjunto de dado, entretanto, neste caso a análise perde uma faceta importante dos dados que é a capacidade de analisar a combinação entre os diferentes atributos e de que forma ela afeta o comportamento das plantas. Alguns pesquisadores utilizam a PCA [Pearson, 1901] para melhorar a capacidade de análise dos dados. Resumidamente, a PCA é uma técnica que busca reduzir a dimensionalidade dos dados a partir de combinações lineares dos atributos originais. Ela o eixo de miaor variabilidade nos dados e projeta os registros sobre esse eixo obtendo assim dados de menor dimensão, porém, a análise é feita sobre os dados numéricos, pouco conhecidos, ou seja, ainda não explorados pelo pesquisador, o que pode dificultar o processo. Assim, uma opção para facilitar o processo de análise é a utilização de técnicas de visualização da informação, que conseguem gerar representações gráficas dos dados, que são muito mais fáceis de analisar dos números em uma tabela. Esta visualização inicial é importante porque oferece ao pesquisador uma visão inicial de seus dados e, assim, pode se orientar melhor na condução das demais etapas do processo de análise, como: a identificação de valores atípicos, identificação dos atributos ruins para a discriminação das espécies, identificação de agrupamentos, etc.

Visualização de informações

A Visualização de Informações propõe várias técnicas que transformam uma série de dados abstratos em um gráfico, para que ele possa ser analisado por um usuário. A Visualização permite criar representações visuais diretamente dos dados ou, em casos mais complexos, permite representar o resultado de fórmulas ou algoritmos executados com o objetivo de extrair os dados mais importantes, tornando-os mais acessíveis,

compreensíveis e eficazes. O principal objetivo a Visualização de Informação é ampliar a cognição humana, gerando representações visuais de dados abstratos, de forma a permitir que estes sejam analisados e, deste modo, oferece recursos de exploração (visual) dos dados, capazes de aliar a flexibilidade, criatividade e o conhecimento humano à enorme capacidade de armazenamento e poder de processamento dos computadores [Keim, 2001]. Embora existam diversas técnicas de Visualização, que podem ser classificadas por diferentes critérios [Card et al., 1999; Oliveira e Levkowitz, 2003]. Este trabalho foi focado no estudo de duas Técnicas de Projeção Geométricas chamadas Coordenadas Paralelas e Viz3D.

Coordenadas Paralelas

A técnica chamada Coordenadas Paralelas [Inselberg, 1985; Inselberg e Dimsdale, 1990; Inselberg, 1997] mapeia um espaço de dimensão *n* em um espaço bidimensional, usando *n* eixos equidistantes e paralelos a um dos eixos principais. Deste modo, cada eixo representa um atributo e, normalmente, o intervalo de valores de cada atributo é mapeado linearmente sobre o eixo correspondente, ou seja, o intervalo representado pelo eixo é entre o menor e o maior valor existente no conjunto de dados. Nesta técnica, cada registro de dado é exibido como uma linha poligonal, que intercepta os eixos nos pontos que correspondem aos valores associados a eles. A interpretação é facilitada pela estimação imediata dos valores dos atributos ao longo dos eixos, o que não ocorre em muitas outras técnicas. Entre as vantagens desta técnica estão: a visualização da distribuição dos dados, dependências funcionais, os agrupamentos, etc. A Figura 1 apresenta em (a) um exemplo de conjunto de dados com quatro registros com cinco atributos e, em (b), a sua representação em coordenadas paralelas.

	a_1	a_2	a_3	a_4	a_5
Reg_1	50	10	0	38	19
Reg_2	40	11	2	60	21
Reg_3	10	78	39	21	42
Reg_4	12	68	31	69	39

(a)

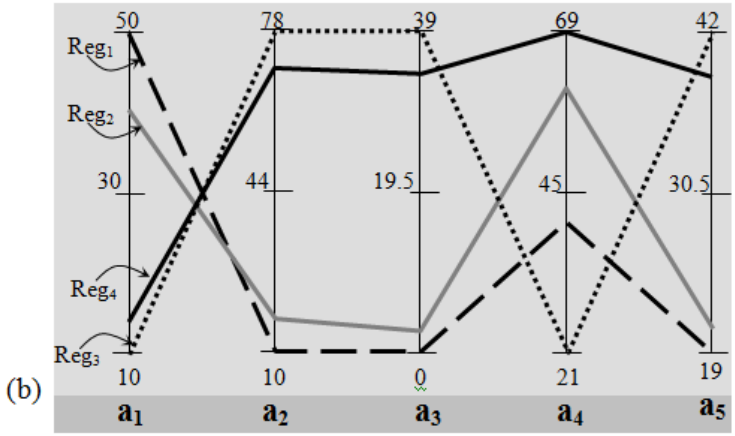


Figura 1. Visualização por Coordenadas Paralelas.

Viz3D

Esta técnica [Artero e Oliveira, 2004] projeta dados muiltidimensionais diretamente em um espaço tridimensional, obtendo resultados muito parecidos com os obtidos com PCA, entretanto, o Viz3D utiliza um mapeamento imediato nD para 3D e, consequentemente, muito rápido, o que a torna muito interessante para mapear grandes conjuntos de dados multidimensionais. O Viz3D projeta os dados na superfície e no interior de um cilindro 3D, cuja base consiste em sistema de eixos radiais que representam os atributos dos registros. Assim, dada uma matriz com os dados, D_{mxn}, o *Viz3D* mapeia as coordenadas *n*-

dimensionais dos *m* registros $d_{i,j}$ de *D* em coordenadas 3D (x_i,y_i,z_i) segundo as Equações 1

$$x_i = x_c + \frac{1}{n}\sum_{j=1}^{n}\frac{d_{i,j} - min_j}{max_j - min_j}\cos\left(\frac{2\pi j}{n}\right)$$

$$y_i = y_c + \frac{1}{n}\sum_{j=1}^{n}\frac{d_{i,j} - min_j}{max_j - min_j}\text{sen}\left(\frac{2\pi j}{n}\right) \qquad (1)$$

$$z_i = z_c + \frac{1}{n}\sum_{j=1}^{n}\frac{d_{i,j} - min_j}{max_j - min_j}$$

com: $i = 1,...,n$ and $j = 1,...,m$ e (x_c,y_c,z_c) sendo a origem do sistema radial 3D de eixos; $max_j = Max(d_{k,j})$; $min_j = Min(d_{k,j})$ para $k = 1,...,m$.

Na Figura 2 tem-se a aplicação dessa técnica no conjunto de dados apresentado na Figura 1(a).

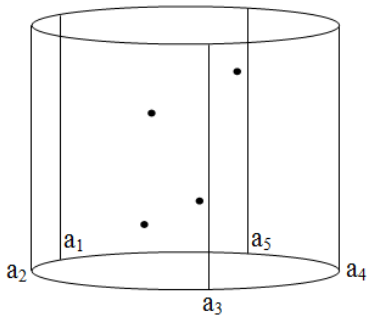


Figura 2. Visualização por Viz3D onde os registros são projetados de acordo com a classe a qual pertencem.

Entre as principais vantagens desta técnica [Artero e Oliveira, 2004] estão: o posicionamento de registros parecidos, próximos entre si no espaço 3D, facilitando a identificação de agrupamentos, identificação de outliers e estruturas geométricas. Além disso, esta técnica permite a interação com o usuário usando operações de rotação, translação e ampliação que permitem ao usuário verificar, em detalhes, os marcadores no espaço 3D.

Seleção de Atributos

Em muitos casos, como os dados que estão sendo explorados provêem de amostras diferentes (diferentes espécies de plantas), é esperado que os agrupamentos correspondentes sejam identificados, entretanto, nem todos os atributos contribuem para uma boa separação entre os grupos. De fato, alguns dos atributos apresentam valores tão idênticos entre os registros de classes diferentes, que é impossível distinguí-las usando eles. Assim, o uso de técnicas de seleção de atributos é bastante recomendado. Embora a técnica de visualização seja capaz de revelar os atributos menos relevantes, uma forma tradicional de determinar os atributos que são realmente relevantes na separação dos dados é a Analise de Variância (*ANova*). Esta técnica consiste em uma análise estatística muito difundida, que serve para comparar a variância das médias dentro dos grupos (Residual) com a variância das medias entre os grupos. A hipótese será rejeitada quando o valor calculado de *F* for maior que o valor crítico na distribuição de *Snedecor*, neste caso, o atributo analisado é considerado relevante e mantido na análise. A partir de uma amostra com *k* grupos (espécies diferentes), totalizando *n* registros, determina-se o valor crítico de *Snedecor* a partir dos parâmetros: *SQA* (soma dos quadrados das diferenças entre a média de cada amostra e a média global, multiplicadas pelos respectivos tamanhos das amostras); *SQR* (soma dos quadrados das diferenças entre cada observação e a média da amostra à qual ele pertence); *SQT* (soma dos quadrados das diferenças entre cada observação e a média global, ou seja, de todas as amostras reunidas); dados pelas Equações (2), (3) e (4).

$$SQA = n_j \sum \left(\bar{x}_j - \bar{x}\right)^2 \qquad (2)$$

$$SQR = \sum\sum \left(x_{ij} - \bar{x}_j\right)^2 \qquad (3)$$

$$SQT = \sum\sum \left(x_{ij} - \bar{x}\right)^2 \qquad (4)$$

Em seguida, a partir destes parâmetros são calculados QMA e QMR, dados por:

$$QMA = \frac{SQA}{k-1} \quad (5)$$

$$QMR = \frac{SQR}{n-k} \quad (6)$$

e, finalmente, o valor F de *Snedecor* é obtido com:

$$F = \frac{QMA}{QMR} \quad (7)$$

Valores grandes de *F* indicam uma diferença significativa entre as classes para o atributo.

ANÁLISE DE DADOS

Nesta seção é apresentado um exemplo de estudo de um conjunto de dados obtidos com um experimento onde duas espécies de plantas (soja e brachiaria) foram cultivadas sob três diferentes condições ambientes (temperatura de 20, 30 e 40ºC) e, para cada um dos diferentes cultivos foram feitas 8 repetições. Deste estudo foram coletadas os seguintes: *Fv/Fm, Fv'/Fm', PhiPS2, qP, NPQ, ETR, Trmmol, cond, EQA, Rd-luz, Amax-luz, Pcom-luz, Psat-luz, EFC, Amax-ci, Pcom-ci, Psat-ci, MSF, MSC, MSR, MST, PA/R, AF, MEF, ICC, EM, ls, DAE*. Dentre esses atributos existem os que foram medidos manualmente e aqueles que foram obtidos pela analise de uma máquina. A Tabela 1 mostra parte dos valores de alguns destes atributos. No caso, a tabela apresenta 19 dos 31 atributos e 16 dos 48 registros coletados, ou seja, a tabela apresenta apenas 20,43% dos dados contidos neste conjunto, chamado *soja-brachiaria.csv*, disponível para download em: *http://fipp.unoeste.br/~almir/ visualization*. Devido ao grande número de atributos e repetições que o estudo abrange, a análise dos dados puramente numéricos é de difícil execução e entendimento, assim é necessária a outra forma de análise desses dados. A análise através de representações visuais é mais simples e cognitiva.

Tabela 1. Tabela que contem dos dados do conjunto soja-brachiaria.csv que foi utilizado nas análises.

espécie	Temp(°C)	repetição	Fv/Fm	Fv'/Fm'	PhiPS2	qP	NPQ	ETR	Trmmol	cond	EQA	...	Psat-ci	MSF	MSC	MSR	MST	PA/R	AF	...
Soja	20	1	0.69	0.33	0.18	0.54	1.81	60.7	2.18	0.15	0.05	...	43.5	4.3	1.7	1.6	7.6	3.75	710.1	
Soja	20	2	0.73	0.36	0.2	0.56	1.99	69.1	2.66	0.18	0.06	...	43.4	2.8	1.2	0.7	4.7	5.71	512	
Soja	20	3	0.74	0.33	0.16	0.48	0.89	53.9	2.09	0.15	0.06	...	41.1	3.7	1.7	1.7	7.1	3.18	657	
Soja	20	4	0.75	0.36	0.19	0.53	2.55	64.8	2.24	0.15	0.05	...	47.5	3.5	1.4	0.9	5.8	5.44	622	
Soja	20	5	0.77	0.37	0.2	0.56	2.9	70.1	2.56	0.17	0.07	...	46	2.9	1.3	0.7	4.9	6	510.4	
Soja	20	6	0.72	0.33	0.16	0.49	2.6	55.4	2.23	0.15	0.05	...	46.1	3.5	1.7	1.2	6.4	4.33	587.2	
Soja	20	7	0.72	0.33	0.16	0.48	2.51	54.1	3.17	0.23	0.05	...	57.7	2.8	1.1	1	4.9	3.9	504.9	
Soja	20	8	0.76	0.38	0.21	0.56	2.53	73.3	2.19	0.14	0.06	...	52	3.5	1.5	1.8	5.8	6.25	627.5	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	....	...	...	...	...	
brachiaria	40	1	0.79	0.46	0.32	0.7	1.57	111	4.2	0.11	0.04	...	6.98	7.64	10.6	7.89	26.1	2.31	1500	
brachiaria	40	2	0.79	0.46	0.32	0.7	1.51	110	4.15	0.1	0.07	...	6.36	5.56	6.99	9.84	22.4	1.28	1169	
brachiaria	40	3	0.8	0.43	0.28	0.64	1.66	95.4	4.19	0.1	0.07	...	10	7.02	7.77	12	26.8	1.23	1298	
brachiaria	40	4	0.79	0.43	0.28	0.65	1.73	96.3	4.25	0.1	0.07	...	7.36	5.9	8.66	8.35	22.9	1.74	1191	
brachiaria	40	5	0.78	0.42	0.26	0.63	2	89.7	3.69	0.08	0.07	...	6.39	6.95	5.1	9.12	21.2	1.32	1469	
brachiaria	40	6	0.79	0.42	0.24	0.56	2.09	81	3.47	0.08	0.07	...	60.3	6.73	7.19	9.34	23.3	1.49	1425	
brachiaria	40	7	0.8	0.47	0.31	0.65	1.81	106	4.32	0.13	0.08	...	6.99	6.43	6.87	4.78	18.1	2.78	1258	
brachiaria	40	8	0.79	0.42	0.2	0.47	2.09	68	2.73	0.07	0.07	...	7.03	6.76	8.55	4.31	19.6	3.55	1266	

A sequência de análise proposta neste trabalho começa com uma visualização do conjunto de dados usando Coordenadas Paralelas. Através dela é possível visualizar a distribuição dos dados entre os diversos atributos e é possível identificar também os atributos que melhor classificam os grupos que existem dentro do conjunto de dados. A Figura 3 mostra a visualização obtida com todos os atributos coloridos de acordo com as duas espécies (soja e brachiaria), onde é possível ver que para alguns atributos, os marcadores escuros estão bem separados dos marcadores cinza, o que caracteriza atributos interessantes para fazer a discriminação entre os comportamentos das duas espécies.

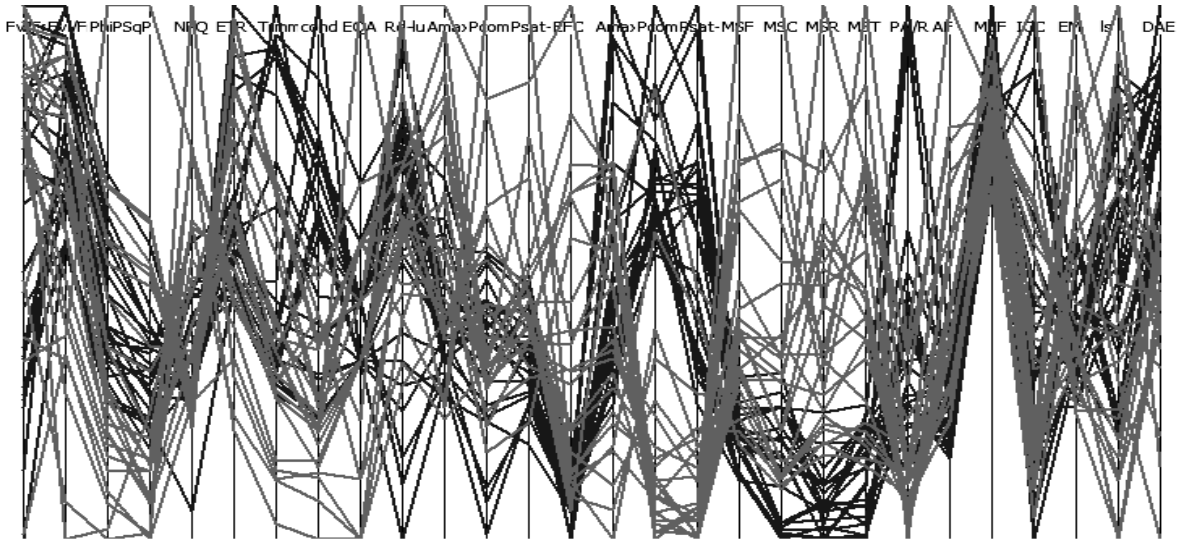


Figura 3. Visualização por *Coordenadas Paralelas* do conjunto de dados soja-brachiaria.csv com todos os atributos.

O mesmo conjunto de dados é visualizado usando o *Viz3D* na Figura 4, onde se nota que os marcadores, coloridos pelas espécies, ficam muito misturados. Isto ocorre porque nesta visualização foram usados todos os atributos do conjunto de dados.

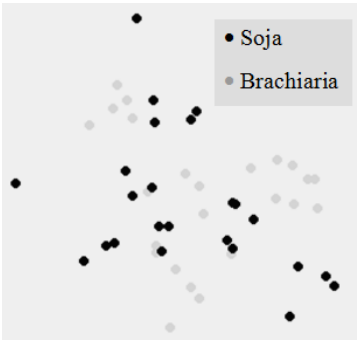


Figura 4. Visualização por *Viz3D* do conjunto de dados *soja-brachiaria.csv* com todos os atributos.

Embora os atributos possam ser selecionados visualmente, usando a visualização da Figura 3, a análise de variância é uma técnica bastante recomendada nesta etapa e, neste trabalho, é usada para rankear os atributos segundo o maior valor de *F* de Snedecor. Assim, o usuário pode ir selecionando quantos atributos deseja que sejam selecionados e, assim, apenas aqueles com os maiores valores de *F* são selecionados. A determinação da quantidade de atributos que deseja utilizar é uma tarefa interativa, em que o usuário define a quantidade, solicita a redução dos atributos para a quantidade desejada e, em seguida, analisa os resultados visualmente, tanto em coordenadas paralelas quanto no *Viz3D*. A Figura 5 mostra o resultado da projeção *Viz3D* depois de selecionados os 10 melhores atributos, onde é possível perceber um

resultado mais satisfatório na separação entre as amostras nas duas espécies (soja e brachiaria).

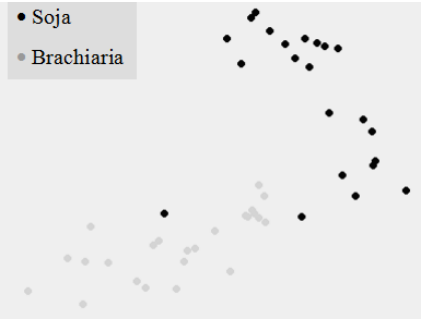
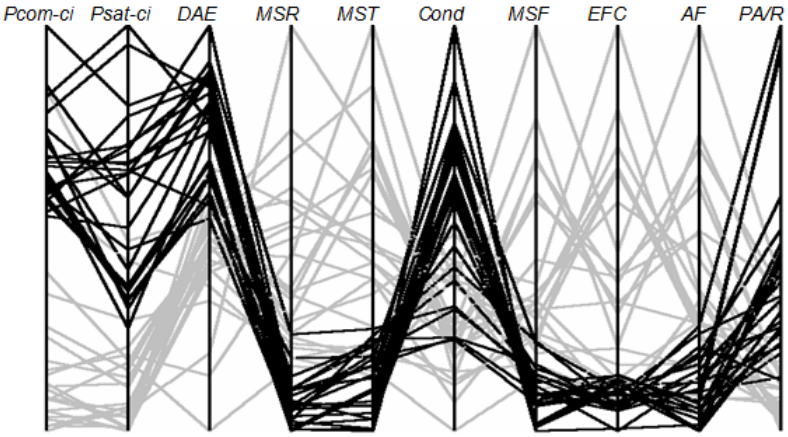


Figura 5. Visualização por *Viz3D* do conjunto de dados *soja-brachiaria.csv* selecionados os 10 melhores atributos segundo a análise de variância e separados por espécies.

A Figura 6 apresenta em (a) os dez atributos com os maiores valores de *F*, ou seja, os atributos que apresentam as maiores seprações de valores entre as classes soja e brachiaria. Em (b) tem-se a visualização destes atributos usando coordenadas paralelas, que evidencia claramente que estes atribu te são muito bons para discrimi duas espécies (soja/brachiaria).

Atributo	Resultado de F
Pcom-ci	114,9376
Psat-ci	82,14554
DAE	69,06803
MSR	58,51318
MST	43,84803
Cond	35,00708
MSF	31,46888
EFC	26,60279
AF	24,88176
PA/R	23,21417

(a)



(b)

Figura 6. a) Os 10 melhores atributos do conjunto soja-brachiaria.csv, segundo a análise de variância (*Anova*), com seus respectivos valores de *F*; b) Visualização em coordenadas paralelas usando os 10 atributos com os maiores valores de *F*.

Com o objetivo de melhorar ainda mais a separação entre os marcadores no espaço 3D e, consequentemente, a discriminação das duas espécies de plantas, foi identificada a melhor maneira de ordenar os atributos selecionados. A técnica usada para realizar esta tarefa é o *SBAA* (*Similarity- Based AttributeArrangement*) [Artero et al., 2006], disponível no software *MDV* (*Multidimensional Data Visualization*), disponível para Download em *http://fipp.unoeste.br/~almir/visualization*. A ordenação dos atributos usando o *SBAA* reduz a quantidade de cruzamentos de poligonais em coordenadas paralelas e, na projeção *Viz3D*, melhora a separação dos marcadores de classes diferentes. Usando os dez melhores atributos apresentados na Figura 6(a), ordenados com a técnica *SBAA*, foi obtida a visualização *Viz3D* ilustrada na Figura 7, que apresenta uma boa separação entre as duas espécies de de plantas.

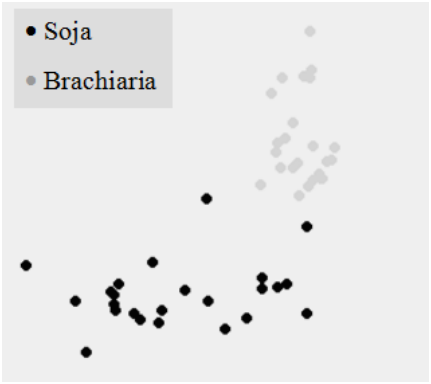
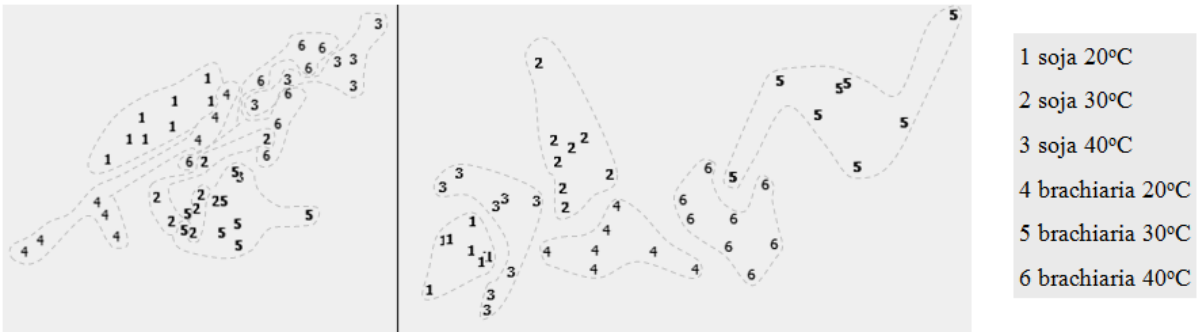


Figura 7. Visualização por *Viz3D* do conjunto de dados *soja-brachiaria.csv* selecionados e ordenados os 10 melhores atributos.

Nesta primeira análise, os marcadores foi considerada apenas a discriminação das duas espécies de plantas (soja e brachiaria), porém, como estas duas espécies foram submetidas a três condições de cultivo com variação de temperatura, isso também possibilita uma análise objetivando caracterizar estes tratamentos diferenciados, c total de seis classes diferentes, ou (20, 30 e 40ºC) e brachiaria (20, 30 e 40ºC). Assim, considerando o tipo de tratamento e também a espécie é possível refazer toda a análise anterior, levanto em consideração essa divisão em seis classes. A Figura 7 (a) mostra a projeção *Viz3D* com todos os atributos do conjunto de dados e, analogamente á análise anterior, pode-se observar uma divisão muito precária entre as seis classes (espécies/tratamento). Em seguida, utilizando apenas os dez melhores atributos da tabela na Figura 6(a), ordenados com a técnica *SBAA*, é obtida a visualização *Viz3D* da Figura 8 (b), que apresenta uma separação bem melhor entre os elementos destas classes.



(a) (b)

Figura 8. (a) Visualização *Viz3d* do conjunto de dados soja-brachiaria.csv com todos os atributos identificados pela classe ao qual pertencem; (b) visualização Viz3D do conjunto de dados soja-brachiaria.csv selecionados e ordenados os 10 melhores atributos

Após a determinação dos atributos mais relevantes para a discriminação das classes de espécies e espécies/tratamento, é possível realizar a classificação dos registros, usando algum algoritmo de agrupamento, ou alguma técnica de classificação supervisionada, como um Classificador Bayesiano ou uma Rede Neural. A etapa de classificação supervisionada usa os registros com classes conhecidas, para gerar um modelo de comportamento para as classes, de modo que, em seguida, é possível classificar registros de classes desconhecidas. A técnica de visualização *Viz3D* também pode ser usada na classificação, pois, se um registro com classe desconhecida é exibido em um mesmo gráfico, com registros de classes conhecidas, o registro de classe desconhecida irá pertencer à classe que ele estiver mais próximo no gráfico.

CONCLUSÕES

A modelagem do comportamento de plantas a partir de condições de cultivo é uma área de grande interesse e, os experimentos nesta área sempre geram uma grande quantidade de parâmetros fisiológicos das plantas sob estudo, entretanto, nem todos estes atributos contribuem para uma boa discriminação das espécies. Como a análise dos valores numéricos é uma tarefa muito difícil, o uso de técnicas de visualização se mostra uma importante ferramenta, pois gera representações visuais dos dados, que ajudam bastante a sua análise. A identificação dos parâmetros mais adequados para a caracterização das espécies é de grande importância para a sua classificação, usando ferramentas como o classificador Bayesiano ou mesmo as redes neurais, pois o uso de parâmetros que não contribuem para a discriminação das espécies não colaboram para a construção do modelo de classificação, além disso, aumentam o volume de processamento.

REFERENCIAS

ARTERO, A.O. Estratégias para apoiar a detecção de estruturas em visualizações multidimensionais perceptualmente sobrecarregadas. Tese Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, 2005.

ARTERO, A.O., OLIVEIRA, M.C.F., LEVKOWITZ, H., Uncovering Clusters in Crowded Parallel Coordinates Visualizations, *Proc. IEEE Symposium on Information Visualization (InfoVis2004)*, pp. 81-88, 2004.

ARTERO, A.O., OLIVEIRA, M.C.F., *Viz3D*: Effective Exploratory Visualization of Large Multidimensional Data Sets, Proc. *XVII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI2004)*, pp. 340-347, 2004.

ARTERO, A.O., Olivera, M.C.F., Levkowitz, H., Enhanced high dimensional data visualization through dimension reduction and attribute arrangement.In Proceedings of the conference on Infomation Visualization IV'06, pages707–712, IEEE, 2006.

BARTKE, KIM. 2D, 3D and High-Dimensional Data and Information Visualization. Seminar on Data and Information Management, University of Hannover Institut f¨ur Wirtschaftsinformatik (IWI), 2005.

CARD, S. K., MACKINLAY, J. D., SHNEIDERMAN, B., Readings in Information Visualization – Using Vision to Think, Morgan Kaufmann, p. 712, 1999.

HOFFMAN, P. E., Table Visualizations: A Formal Model and its Applications. *Doctoral Diss.*, Computer Science Department, University Of Massachusetts, Lowell, Ma,1999.

INSELBERG, A. The Plane with Parallel Coordinates, *The Visual Computer (Special Issue on Computational Geometry)*, vol. 1, n. 2, pp. 69-92, 1985. http://dx.doi.org/10.1007/BF01898350

INSELBERG, A., DIMSDALE, B., Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry. *Proc. IEEE Visualization'90,* pp. 361-375, 1990.

KEIM, D. A., Designing Pixel-Oriented Visualization Techniques:Theory And Applications, IEEE. http://dx.doi.org/10.1109/2945.841121

OLIVEIRA, M. C. F., LEVKOWITZ, H., From Visualization to Visual Data Mining: A Survey. IEEE Transactions on Visualization and Computer Graphics, vol. 9, n. 3, pp. 378-394, 2003. http://dx.doi.org/10.1109/TVCG.2003.1207445

PEARSON, K., On Lines and Planes of Closest Fit to System of Points in Space, Philosophy Magazine, vol. 6, pp. 559-572, 1901.

Transactions on Visualization and Computers Graphics, vol. 6, n.1, pp.59-78, 2000.