XI Jornadas de Classificação e Análise de Dados
Associação Portuguesa de Classificação e Análise de Dados
Lisboa, 1 a 3 de Abril de 2004

Biplots PMD - Data Mining Centrada em Biplots. Apresentação de um Protótipo

Valter Martins Vairinhos[1], M. Purificación Galindo[2]
1 – Universidade Independente – Lisboa
2 – Universidade de Salamanca – Departamento de Estatística

Resumo: O objectivo deste trabalho é mostrar que se pode basear a interface gráfica para sistemas de *data mining* no conceito de biplot.
A justificação para esse facto assenta na interpretabilidade dos biplots e na observação de que a grande maioria dos resultados das análises de dados multivariados pode ser expressa por listas de indivíduos, de variáveis ou listas mistas de indivíduos e variáveis.
Esta observação empírica tem como consequência a de que todos os resultados expressáveis desta forma podem, naturalmente, ser representados por configurações de marcadores de variáveis e indivíduos em biplots.
Daqui resulta a possibilidade de comparar graficamente resultados obtidos por técnicas de análise distintas ou resultantes de parametrizações diferentes da mesma técnica.
Quando *n* e *p* são elevados – como sucede em problemas de *data mining* – a realização de operações de interpretação levanta problemas que resultam das limitações de memória de trabalho dos seres humanos e da dificuldade que estes têm em realizar sem erros lógicos ou omissões, as operações inerentes à interpretação.
Do reconhecimento destes factos resulta a necessidade de desenvolver sistemas capazes de sugerir possíveis interpretações ao analista e o auxiliem nas operações intermédias de interpretação de resultados.
A concepção de algoritmos com essa capacidade exige uma formulação matemática do problema de interpretação.
Neste trabalho apresentam-se resultados nesta direcção obtidos no Departamento de Estatística da Universidade de Salamanca, consubstanciados num sistema-protótipo que implementa, de forma experimental mas encorajante, algumas dessas ideias.

Abstract: The objective of this work is to show that biplots can be used as the kernel for visual data mining systems. This claim can be justified observing that biplots produce results eminently interpretable and that an important class of results generated by distinct multivariate data analysis methods can be expressed by lists of variables, individuals or both. This empirical finding has two consequences: 1-All results that can be expressed this way can be represented by configurations of individuals and variables in biplots. 2-Results obtained by distinct methods or distinct parameter values of the same method can be visually compared inspecting its configurations on the same biplot. When *n* and *p* are big – which is common in data mining systems- the human interpretation of results poses important problems resulting from limitations of the human work memory and other shortcomings such as logical limitation and data processing capabilities. This means that systems must collaborate with the analyst suggesting possible interpretations and supporting him with other interpretation oriented tasks. The development of algorithms with such capability means a mathematical formulation of the interpretation problem. In this work we present results in that direction obtained at the Departamento de Estadística da Universidade de Salamanca, resulting in an experimental system whose practical application to data sets from distinct sources is considered encouraging.

Palavras Chave: Biplots, Interpretação de Resultado , Análise Preliminar de Dados, *Data Mining*.

1. INTRODUÇÃO.

Neste artigo pretende-se mostrar que é possível basear no conceito de biplot a interface gráfica de um sistema de *data mining*.
No número quatro deste trabalho apresenta-se a funcionalidade básica, ilustrada com aplicações a dados reais, de um sistema protótipo – designado Biplots PMD (Biplots Para Mineria de Datos) – que realiza esta ideia.
Em geral, aceita-se que as actividades de *data mining* formam a fase crucial do chamado processo de extracção de conhecimentos a partir de bases de dados (KDD : *Knowledge Discovery in Data Bases*). Ver FAYYAD *et al* (1996) e BORGELT *et al* (2001).
Uma definição em geral aceite de KDD e também de *data mining* é, pois a seguinte: «*a descoberta de conhecimentos não triviais e a identificação de padrões válidos, novos, potencialmente úteis e interpretáveis nos dados*». FAYYAD *et al* (1996).
Excluída a referência às bases de dados, esta definição coincide com a definição de TUKEY (1966) para a análise de dados: «*The basic general interest of data analysis is simply stated: to seek throught a body of data for interesting relationship and information and to exhibit the result in such a way as to make them recognizable to the data analyst and recordable for posterity*».
Quando se pretende usar uma técnica específica como núcleo de sistemas de *data mining*, exige-se que essa técnica satisfaça pelo menos três princípios básicos: incrementabilidade, escalabilidade e interpretabilidade. Ver BORGELT (2001).

O modelo mais conhecido de biplots GABRIEL (1971) assenta na decomposição em valores e vectores singulares (SVD) de uma matriz de observações formada por números reais. Isso significa que os biplots que obedecem a este paradigma – ver abaixo número dois – herdam do SVD a incrementabilidade e escalabilidade. Ver GOLUB *et al* (1996), HALL *et al* (2000, 2002).
A interpretabilidade dos biplots explica em parte o seu uso crescente. Esta interpretabilidade assenta, por um lado, em resultados teóricos que relacionam os biplots com quase todas as técnicas de análise de dados multivariados -VICENTE VILLARDÓN (1992) e GOWER *et al* (1996) - e, por outro lado, no facto de sobre os biplots se poder visualizar e comparar todos os resultados obtidos por qualquer técnica desde que expressáveis por listas de marcadores de variáveis e de indivíduos. Ver número três deste artigo.
Todas as obras de análise de dados, incluindo as mais antigas, atribuem grande importância à interpretação dos resultados. Ver TUKEY (1966) e BENZÉCRI (1973). Apesar disso, a essa questão pouco esforço de teorização tem sido dedicado, o que se traduz pela quase ausência de publicações sobre o assunto.
, No nosso entender, a razão deste facto explica-se pela pequena dimensão (*n* e *p* pequenos) dos conjuntos de dados tratados pela análise multivariada clássica – o que

tornava possível que todas as tarefas de interpretação fossem realizáveis pelo analista humano.
O surgimento de bases de dados gigantes em que a mera listagem do significado das variáveis exige dicionários de dados com centenas ou milhares de entradas, torna inviável esta abordagem e exige que o sistema sugira possíveis interpretações desses resultados ou, no mínimo, colabore eficazmente com o analista no âmbito do processo de interpretação. Essa funcionalidade pressupõe uma base teórica sobre a qual desenvolver algoritmos que a permitam.
Um esforço nesse sentido foi desenvolvido no Departamento de Estatística da Universidade de Salamanca, tendo resultado dessa tentativa o sistema protótipo que se apresenta no número quatro deste artigo.

2. BIPLOTS

2.1. Conceitos de biplot.

Segundo GABRIEL (1971), «*Toda a matriz de característica 2 pode ser representada graficamente por um biplot que consiste num vector para cada linha e num vector para cada coluna, escolhidos de modo que cada elemento da matriz seja exactamente o produto interno desses dois vectores. Se a matriz de dados tem característica superior a dois, essa matriz pode ser representada, de modo aproximado, por um biplot de uma matriz de característica 2*».
O prefixo «***bi***» no termo biplot significa que, no gráfico que representa *X*, existem dois tipos de marcadores: os marcadores representativos dos indivíduos (a_i, i= 1…n) e os marcadores representativos das variáveis (b_j, j= 1…p). O gráfico pode ser *bi* ou ***tri***dimensional.
Segundo a definição anterior, dada a matriz de dados X ($n \times p$), então

$$X = \left[x_{ij}\right] = \left[a_i^T \; b_j\right] = \left[< a_i , b_j > \right] = \left[\left| a_i \right| \left| b_j \right| \cos\theta_{ij} \right].$$

No caso de r = ***carac***t (X) = ***min*** (n, p) > 3, os biplots são sempre aproximações dos dados. Quando $r = 2$ ou 3, podem construir-se biplots (de 2 ou 3 dimensões) que representem exactamente os dados.
GABRIEL (1971) usa a decomposição de X em valores e vectores próprios como técnica para obter os marcadores das linhas/indivíduos e das colunas/variáveis.
Isto é: $X = \left(U \Sigma^{\alpha}\right)\left(V \Sigma^{1-\alpha}\right)^T = AB^T$ com $A = U \Sigma^{\alpha}$ e $B = V \Sigma^{1-\alpha}$, sendo α um parâmetro que varia de modo contínuo no intervalo [0, 1].
, Se pretendermos visualizar os biplots quando $r > 3$, é necessário escolher uma dimensão $d \leq r$ para o espaço de representação aproximada de X.
Em particular, para representações planas ($d= 2$), $X \cong U_{(2)} \Sigma_{(2)}^{(\alpha)} \Sigma_{(2)}^{(1-\alpha)} V_{(2)}^T$, em que os índices (2) significam que se estão a considerar apenas 2 vectores singulares (esquerdos e direitos) e os 2 valores singulares correspondentes.

Neste caso, $x_{ij} \cong a_i^T b_j$ em que a_i é uma linha genérica, de $A = U_{(2)} \Sigma_{(2)}^{(\alpha)}$, representativa dos indivíduos e b_j uma linha genérica de $B = V_{(2)} \Sigma_{(2)}^{(1-\alpha)}$ representativa das variáveis.

Observe-se que de $X = U \Sigma^{\alpha} \left(V \Sigma^{1-\alpha} \right)^T$ resulta $X V = U \Sigma$, projecção das linhas de X sobre as direcções principais e $X^T U = V \Sigma$, projecção das variáveis sobre as direcções principais.

As propriedades deste tipo de biplot podem ver-se em GABRIEL (1971, 2002). Nessas referências pode verificar-se que, quando $\alpha \neq ½$. a qualidade da representação dos indivíduos e das variáveis é diferente. Quando $\alpha = ½$ atinge-se a mesma qualidade de representação para indivíduos e variáveis mas não a máxima que é possível separadamente para indivíduos e variáveis.

GALINDO (1985) propõe um novo tipo de biplot que generaliza a qualquer matriz de números reais, o conceito de representação simultânea de BENZÉCRI (1973).

A definição de GALINDO (1985) é a seguinte: dada a decomposição $X = (U \Sigma)(V \Sigma)^T$, usam-se para marcadores dos indivíduos as linha da matriz $A = U \Sigma$ e para marcadores das variáveis as linha da matriz $B = V \Sigma$.

O critério usado por GALINDO (1985) é o de obter um biplot em que tanto variáveis como indivíduos tenham a mesma e máxima qualidade de representação que é possível, sendo o biplot uma verdadeira representação conjunta de indivíduos e variáveis, referidos à mesma escala e não uma mera representação simultânea.

Neste caso, $X = U \Sigma V^T \neq A B^T = U \Sigma^2 V^T$, não se garantindo a decomposição dos valores observados como produtos internos de marcadores, mas garantindo-se, no entanto, para lá da já mencionada qualidade de representação uniforme para indivíduos e variáveis, as chamadas formulas de transição, características da Análise Factorial de Correspondências. Ver GALINDO (1985).

Quando $r = caract\,(X) = min\,(n, p) \geq 3$, se se pretende uma aproximação de ordem $d = 2$, os biplots são construídos usando apenas dois dos vectores singulares e valores singulares correspondentes.

VICENTE VILLARDÓN (1992), no seguimento dos trabalhos de GALINDO (1985) e usando a decomposição em valores e vectores próprios generalizada (SVDG) de GREENACRE (1984) apresenta o conceito de biplot generalizado, mostrando que os métodos mais usados de análise de dados multivariados- incluindo a Análise Canónica e o MDS- podem ser vistas como casos particulares de biplots generalizados.

Em GOWER (1995, 1996) é apresentado um novo tipo de biplot assente no conceito de MDS. Neste tipo de biplot, a localização dos marcadores dos indivíduos num plano ou num espaço de representação de 3 dimensões obtém-se por MDS a partir das dissemelhanças entre indivíduos (medidas eventualmente por distâncias). As variáveis são posicionadas *à posteriori* usando regressão múltipla.

Representando os indivíduos por pontos e as variáveis por linhas, as direcções destas linhas no novo biplot – designadas os eixos do biplot– são as direcções, no espaço do biplot, dos versores dos eixos coordenados dos dados originais, não existindo agora quaisquer referências aos eixos factoriais.
Uma outra característica deste conceito de biplot de GOWER é a distinção entre eixos de ***interpolação*** e eixos de ***predição***.
A ***interpolação*** consiste em localizar num biplot – por interpolação entre os marcadores presentes – a posição de uma observação adicional. A ***predição*** consiste em, dado um ponto no biplot, identificar – no referencial inicial dos dados – o conjunto de observações cuja imagem no biplot é esse ponto.
GOWER (1996) mostra que também este tipo de biplot serve de conceito unificador dos métodos habituais de análise de dados multivariados (MDS, Análise em Componentes Principais, Análise Crónica, Análise Factorial de Correspondência Simples e Múltipla).

2.2. Interpretação de biplots.
Referindo nos apenas ao conceito de biplot baseado na decomposição em valores e vectores próprios, GABRIEL (1971) mostra que os cosenos dos ângulos entre os vectores representativos das variáveis num biplot são os coeficientes de correlação entre as variáveis respectivas. As proximidades, sobre o biplot, entre os marcadores dos indivíduos representam semelhanças entre os indivíduos: dois pontos próximos correspondem a dois indivíduos com respostas semelhantes; dois pontos afastados correspondem a dois indivíduos com respostas tanto mais díspares quanto maior o afastamento, sobre o *biplot*, dos marcadores respectivos.

Isto significa que quando maior é o valor da projecção de um indivíduo sobre uma variável - medida a partir do centro - maior é o valor dessa variável sobre o indivíduo e maior é a preponderância da variável na explicação do comportamento ou resposta do indivíduo. Isto significa também que quanto menor o ângulo entre os vectores definidos pelo centro do biplot e os marcadores de um indivíduo e de uma variável, maior a afinidade entre esse individuo e essa variável, no sentido descrito. Quando mais próximo da direcção de uma variável está o ponto representativo de um indivíduo e maior for o afastamento do indivíduo em relação ao centro, maior a preponderância ou importância dessa variável na explicação dos resultados obtidos por um indivíduo.
Uma vez que nos biplots de GALINDO (1985) indivíduos e variáveis estão representados na mesma escala, faz sentido interpretar as distâncias entre indivíduos e variáveis como preponderância de uma variável para explicar um indivíduo ou como contribuição de um indivíduo para os valores de uma variável.
Em BRADU e GABRIEL (1978), DIAZ-LENO (1995), BLAZQUEZ ZABALLOS (1998), GABRIEL *et al* (1998) mostra-se que é possível usar os biplots em problemas de diagnóstico visual de modelos log-lineares usando os alinhamentos segundo linhas rectas dos marcadores das variáveis e dos indivíduos. Em SEPÚLVEDA (2004) esta mesma ideia é usada em problemas de inferência relativas a dependências locais em análise de classes latentes.

3. PROBLEMAS DE INTERPRETAÇÃO DOS RESULTADOS EM ANÁLISE DE DADOS MULTIVARIADOS.

As questões de interpretação dos resultados obtidos pelas diversas técnicas de análise de dados multivariados sempre foram consideradas questões importantes. Ver TUKEY (1966), BENZÉCRI (1973), JAMBU (1991), BORG et al (1997), KRZANOWSKI (2002).

Com o advento da *data mining*, face aos valores muito elevados de *n* e *p* torna-se necessário que os sistemas colaborem mais intensamente com os analistas na procura do significado dos resultados. Com efeito, face às conhecidas limitações da memória de trabalho dos seres humanos - ver ANDERSON (1990) - não se pode esperar que seja o analista a realizar sem erros todas as operações lógicas e de processamento de dados inerentes a esta fase, em que conhecer o significado de cada uma das variáveis e cada um dos seus valores é crucial.

Isto significa que se torna necessário criar um corpo teórico acerca dos problemas de interpretação, sobre o qual basear algoritmos que implementem essa nova funcionalidade. Não basta que o ambiente de trabalho seja amigável: deve ser amigável e colaborante.

Nesta perspectiva, as razões fundamentais que estiveram na base da escolha dos métodos biplot para núcleo de um sistema de *data mining* foram as seguintes:

1 - Os métodos de análise de dados multivariados mais usados (ANÁLISE CANÓNICA, ANÁLISE EM COMPONENTES PRINCIPAIS, ANÁLISE FACTORIAL DE CORRESPONDÊNCIAS, MDS, ANÁLISE DE CLUSTERS, ÁRVORES DE CLASSIFICAÇÃO e REGRESSÃO) geram como principais resultados listas de indivíduos/observações, listas de variáveis, listas mistas de indivíduos e variáveis e partições desses conjuntos. VAIRINHOS (2003).

2 - Do ponto de vista teórico, os métodos de análise de dados multivariados mais usados podem ser vistos como casos particulares de biplots. Isto é, assim quer se trate dos conceitos de Biplot de GABRIEL (1971) e GALINDO (1985) - ver, neste caso, VICENTE VILLARDÓN (1992) - quer se trate do conceito de biplot de GOWER (1996).

3 - Todos os resultados gerados por métodos de análise de dados multivariados que assumam a forma de listas de indivíduos, de variáveis e de listas mistas de indivíduos e variáveis podem ser representados em biplots por configurações de marcadores e, por isso, comparados e estudados visualmente. VAIRINHOS (2003).

Em síntese, os biplots constituem um instrumento natural, tanto do ponto de vista teórico como prático, para raciocinar visualmente acerca dos resultados obtidos por métodos distintos.

A realização de um sistema capaz de colaborar eficazmente com o analista na fase de interpretação, sugerindo possíveis interpretações, exige, contudo, uma formulação matemática do problema da interpretação em si, vista como uma fase específica, com necessidades teóricas específicas.

Uma formulação possível para este problema pode ver-se em VAIRINHOS (2003) e VAIRINHOS E GALINDO (2004).

A ideia básica é a seguinte: quando, através de uma técnica de análise de dados multivariados - por exemplo, análise de cluster - obtemos como resultado uma lista de indivíduos, o significado deste resultado fica, formalmente, «conhecido», visto que o significado dos indivíduos que integram essa lista, bem como o significado das variáveis observadas e seus valores, é conhecido.
Nesta perspectiva, o ***problema de interpretação*** não é, pois, o de obter o significado do resultado mas o de traduzir esse significado conhecido numa linguagem próxima da linguagem humana. É o problema de tornar inteligível para o ser humano o resultado que a máquina apresenta sob a forma de listas de identificadores.
Como se sabe, um conceito pode ser definido por ***intenção*** ou por ***extensão***.
Contudo, quando a extensão de um conceito (lista dos objectos que o integram) tem baixo cardinal, as pessoas não têm dificuldade em explicar o respectivo significado, em obter uma relação entre variáveis (a intenção) que expresse esse significado.
Quando o cardinal da extensão de um conceito é elevado, a elaboração de uma expressão que traduza essa lista não está, em geral, ao alcance da generalidade das pessoas. Ora é sempre possível representar um subconjunto de indivíduos ou uma partição desse conjunto de indivíduos usando a função característica do grupo ou uma variável qualitativa para a partição. Daí a justificação para seguinte definição preliminar do problema de interpretação. VAIRINHOS (2003), VAIRINHOS *et al* (2004).

Definição 1 (**Problema de interpretação)**
Dado um resultado representado por uma variável qualitativa *R*, interpretar este resultado é expressar a variável *R* em função das variáveis observadas e respectivos valores, usando expressões de uma linguagem próxima da linguagem humana.

Nesta definição assume-se que as únicas fontes de significado para o processo de interpretação são os significados dos objectos observados, das variáveis observadas e dos respectivos valores.
Em resumo: o problema de interpretação também pode ser visto - VAIRINHOS (2003) e VAIRINHOS *et al* (2004) - na forma seguinte:

Definição 2 (**Interpretação)**
Seja *R* uma variável qualitativa, definida no conjunto de indivíduos, que representa um resultado (grupo ou partição) obtido por um método de análise de dados multivariados.
Interpretar esse resultado é aproximar os significados das classes de equivalência definidas por *R* usando uma certa classe de funções das variáveis observadas e seus valores. A qualidade desta aproximação deve ser medida usando uma função de perda adequada.

A linguagem que se adoptou é formada por conjunções de expressões atómicas do tipo *(Variável Observada = valor)*, cujo significado, indecomponível, não oferece qualquer ambiguidade. Quando necessário, estas expressões conjuntivas são combinadas por disjunções.

Em resumo, o problema de interpretação pode ser formulado como um problema de aproximação do conjunto ($R = r$) por expressões do tipo $\underset{i \in I}{\wedge} \left(X_i = x_i \right)$ em que X_i são variáveis observadas, x_i são valores observados e $I \subseteq \{1...p\}$.
Como se mostra em VAIRINHOS (2003) e VAIRINHOS et al (2004), esta formulação permite a representação matemática do problema de interpretação sobre um grafo de intersecção cujos vértices são os átomos das variáveis observadas.

4. BIPLOTS PMD – UM SISTEMA PROTÓTIPO DE *DATA MINING* BASEADO EM BIPLOTS.

4.1. Síntese da funcionalidade do sistema.

A ideia condutora da concepção e desenvolvimento do sistema protótipo, para além da concretização das ideias teóricas referidas, foi a de construir a respectiva funcionalidade em torno das seguintes operações básicas de interpretação.

1 – Identificar grupos de objectos e variáveis em biplots.
Como se viu em 3, uma parte significativa dos resultados da análise de dados multivariados traduz-se em subconjuntos e/ou partições dos conjuntos de indivíduos e variáveis.

2 – Caracterizar grupos.
Uma vez identificado um grupo, importa caracterizá-lo em termos das variáveis observadas e dos seus valores, usando expressões adequadas.

3 – Comparar dois grupos.
Uma vez identificados e caracterizados dois grupos, a operação natural seguinte é tentar descobrir *«o que é que»* separa um do outro.

4 – Atribuir nomes a grupos.
Caracterizados e comparados os grupos, o analista tenta de seguida atribuir nomes ou etiquetas *significativas* a esses grupos. Esta atribuição traduz ou simboliza a ligação do grupo ao conhecimento, *à priori*, do analista.

5 – Reconhecer padrões em grupos.
Numa fase mais avançada procura-se detectar nos grupos identificados, a ocorrência de padrões lineares ou outras relações me temáticas e estatísticas entre as variáveis observadas.

Na figura 4.1 pode ver-se a janela principal do sistema experimental BiplotsPMD, desenvolvido em torno destas ideias para demonstrar a respectiva exequibilidade. Os botões principais (veja **figura 4.1.**) estão associados à realização das tarefas básicas ou elementares de interpretação. O sistema foi desenvolvido usando o sistema Delphi da BORLAND, ambiente de programação para o sistema WINDOWS e LINUX, assente sobre a linguagem PASCAL, que permite programação por objectos e integra, a funcionalidade de bases de dados relacionais.
O funcionamento do sistema experimental assenta na decomposição dos dados em átomos (de significado); ou seja, expressões com significado elementar definido por ($V= v$) em que V é uma variável observada e v um valor observado. Esta expressão representa por *intenção* um conjunto de indivíduos cujo significado não oferece ambiguidades. A extensão correspondente é o conjunto ou lista $\{i: X(i)= v\}$ em que $i \in I$, sendo I o conjunto de indivíduos observados. Esta representação permite usar

as operações básicas de união e intersecção de conjuntos para implementar –todas as operações de interpretação. VAIRINHOS (2003).
Da funcionalidade do sistema experimental destacam-se, ainda, os seguintes aspectos:

- Edição e limpeza dos dados brutos, preparando-os para análise.
- Agregação de átomos por união de vários átomos, no caso em que as variáveis são contínuas gerando átomos de cardinal muito pequeno.
- Definição e visualização sobre o biplot em estudo de expressões do tipo:
  $(X_1 = v_1) \wedge (X_2 = v_2) \wedge \ldots \wedge (X_k = v_k)$ *com* $1, 2, \ldots, k \subset \{1 \ldots p\}$
- Definição de hierarquias de partes do conjunto de indivíduos e sua visualização sobre o biplot através dos fechos convexos ou cores.
- Projecção dos indivíduos e das variáveis sobre uma direcção arbitrária, o que equivale a definir relações de ordem correspondentes a perspectivas definidas pelo analista.
- Geração de expressões de tipo conjuntivo para caracterização de grupos isolados de indivíduos.
- Geração de regras de classificação para partições do conjunto de indivíduos, usando expressões conjuntivas e disjuntivas dos átomos de variáveis observadas e seus valores.

4.2. Exemplo de aplicação a dados reais.

Seguidamente, apresenta-se um exemplo de aplicação do sistema protótipo a dados reais. O conjunto de dados tem 249 linhas – correspondentes à maior parte dos municípios do Continente – e 27 colunas, respeitantes a variáveis ecológicas (as 11primeiras) e às percentagens obtidas pelos partidos e coligações que se candidataram a eleições legislativas e autárquicas entre 1975 e 1979. Os dados foram obtidos por DANIEL NATAF, no âmbito de um estudo de ecologia política de Portugal.Ver NATAF (1995). Seguidamente apresenta-se o significado das variáveis:

Variáveis ecológicas:

SN	Sul/Norte 0- SUL; 1- NORTE
LS	Large/Small 0- Large 1-Small
UR	Urbano/Rural 0-Urbano 1- Rural
ISO3	%Trabalhadores Isolados do sector Terciário
TRAB3	% Trabalhadores Assalariados do Terciário
ISO1	% Trabalhadores Isolados na Agricultura
TRAB1	% Trabalhadores Assalariados na Agricultura
ISO2	% trabalhadores Isolados na Indústria
TRAB2	% Trabalhadores Assalariados na Indústria
RELIGIAO	% Com Confissão Religiosa
DIMEXPLAG	Dimensão da Exploração Agrícola

Variáveis com resultados obtidos pelos partidos:

PCP75	% Do PCP Eleições de 1975
PCP76	% Do PCP Eleições de 1976
PCP79	% Do PCP Eleições de 1979

PCP88	% Do PCP Eleições de 1988
PS75	% Do PS Eleições de 1975
PS76	% Do PS Eleições de 1976
PS79	% Do PS Eleições de 1979
PS88	% Do PS Eleições de 1988
PPD75	% Do PPD Eleições de 1975
PPD76	% Do PPD Eleições de 1976
CDS75	% Do CDS Eleições de 1979
CDS76	% Do CDS Eleições de 1988
AD75	% Da AD Eleições de 1975
AD76	% Da ADEleições de 1975
POUS	% Da POUS Eleições de 1975

No biplot que é apresentado na **figura 4.1.** as linhas correspondem às variáveis e os pontos correspondem aos municípios. Usou-se para representar cada município o valor da variável UR .Verifica-se que UR e LS formam um ângulo pequeno – o que corresponde a uma elevada correlação: os pequenos municípios têm, em geral, carácter rural e opõem-se no biplot aos municípios em que a percentagem de trabalhadores de serviços(TRAB3) e da indústria (TRAB2) é elevada (à esquerda). O eixo horizontal do biplot corresponde, pois, à oposição cidade-campo – o que está visível no facto de os 1's estarem à direita e os 0's à esquerda. Ao examinar o biplot, o utilizador marcou o Grupo nº 1 (primeiro quadrante) rodeando os pontos respectivos por um polígono. Seguidamente, o sistema apresenta a lista dos indivíduos e variáveis que integram esse grupo, atribui-lhe uma cor e um nome provisório, apresenta um resumo estatístico e uma síntese do grupo, guardando a respectiva definição na base de dados de grupos.

Na **figura 4.2.** observa-se o momento em que, tendo decidido realizar uma classificação hierárquica excedente (distância euclidiana, método de agregação de Wrad) o usuário *corta* a árvore de classificação a um dado nível do índice de dissemelhança (veja tracejado do vertical do lado esquerdo da **figura 4.2.**) definindo assim uma partição no conjunto de indivíduos.

Esta partição é em seguida, visualizada sobre o biplot – veja **figura 4.3.** – pelos fechos convexos dos grupos de indivíduos que a integram.

Examinando esses grupos (a que o sistema atribuiu, agora, os números 2, 3 e 4) verifica-se que o grupo 2 (à esquerda da figura) está associado a valores elevados das percentagens de trabalhadores dos serviços (TRAB3) e da indústria (TRAB2).

O grupo 3 está associado a valores elevados de praticantes da religião católica (RELIGIÃO), a trabalhadores isolados (ISOL1) e o grupo do 4º quadrante associado a assalariados rurais (TRAB1) e à grande dimensão da exploração agrícola (DIMENSÃO).

Constata-se ainda, que sobre esta **figura 4.3**, é visível o Grupo nº 1, definido pelo analista.

Este biplot está, pois, a permitir comparar uma hipótese formulada pelo utilizador (a de que existiria um possível grupo *interessante* a que foi dado o nome provisório Grupo Nº 1 com a sugestão automática dada pela análise *cluster* de que existiria um grupo de municípios, a que foi dada a identificação provisória de Grupo Nº 2. Como

se pode verificar, o biplot permite ver que os fechos convexos desses grupos (obtidos por métodos diferentes) quase coincidem.
A **figura 4.4.** ilustra outro aspecto da funcionalidade do sistema: a apresentação e estudo dos grupos definidos até um certo instante. Neste caso, estão definidos quatro grupos: o Grupo Nº 1, definido pelo utilizador, e os grupos de números 2, 3 e 4 correspondentes a uma partição sugerida por análise *cluster*.
Tendo-se verificado que o Grupo Nº 1 (obtido manualmente) e o Grupo Nº 2 (obtido automaticamente), praticamente coincidem, torna-se oportuno pedir ao sistema que sugira uma caracterização. Essa sugestão aparece na parte superior da figura nº 7, tendo da forma: $(UR= 1) \wedge (1 \leq Dimensão \leq 17)$.
Trata-se de municípios de carácter rural em que a Dimensão da exploração agrícola se situa entre *1 e 17 hectares*: pequenas explorações, face ao máximo registado de 264. Esta sugestão de interpretação – obtida automaticamente – é consistente com a caracterização que uma inspecção visual do biplot permite detectar.
Feito isto, convém comparar o significado desta descrição – formada por todos os municípios para os quais $\{ i \in I: UR(i) = 1 \} \cap \{ i \in I: 1 \leq Dimensão (i) \leq 17\}$, com o significado formal do grupo que o utilizador detectou.
Na **figura 4.5.** vê-se o fecho convexo da descrição automática, verificando-se que, coincide praticamente com o fecho convexo do grupo a caracterizar. Isto é: a descrição automática é, de grande qualidade neste caso.
Convém, agora, tentar descobrir que posição ocupariam, no biplot definido pelas variáveis ecológicas, as variáveis inactivas, formadas pelas votações nos partidos. Isso está feito na **figura 4.6.**, depois de, por questões de legibilidade, se ter eliminado os símbolos identificadores dos municípios.
Verifica-se que as votações do PCP (75, 76, 79, 88) ocupam praticamente a mesma posição na parte inferior do biplot. Estão, pois, associadas, à grande DIMENSÃO da exploração agrícola e a trabalhadores assalariados agrícolas (TRAB1).
As votações no PSD, CDS e AD nesses mesmos períodos ocupam, na parte superior do biplot, uma posição muito bem definida e também quase invariante com o tempo, associada a municípios com grandes valores da variável RELIGIÃO percentagens elevadas de pequenos proprietários e rendeiros agrícolas (ISO1). Verifica-se, finalmente, que as votações no PS definem uma trajectória que começa na parte inferior do biplot para PS75 (associada a trabalhadores do terciário) e vai subindo em direcção aos trabalhadores da indústria (TRAB2) em eleições posteriores.
Como se verifica, estes resultados que o biplot apresenta de forma geométrica e legível, têm um significado facilmente compreensível por um conhecedor não muito sofisticado da vida política portuguesa.
Nas **figuras 4.7** a **4.14.** pode observar-se um biplot em que as variáveis activas são, agora, as votações nos partidos e coligações participantes nas eleições de 1975 , 1976, 1979 e 1988.
No biplot da **figura 4.7.** verifica-se que essas variáveis formam três cones. Um, à esquerda, formado pelas votações nos partidos e coligações de direita (PSD, CDS e AD), outro no canto SE formado pelas votações no PCP e o terceiro cone, muito mais aberto (denotando correlações mais fracas entre as votações no PS), ocupando uma posição angular intermédia.
Realizando uma classificação automática dos municípios e visualizando uma partição com três classes, vê-se na **figura 4.7** que a essas classes ou grupos de

municípios correspondem, quase perfeitamente, às votações na direita (PSD, AD, CDS); às votações no PCP e às votações no PS.
Na **figura 4.8.** pode ver-se o mesmo resultado, depois de se ter eliminado os pontos representativos dos municípios e sobrepondo como variáveis suplementares as variáveis ecológicas.
Ao Grupo Nº1 desta nova partição (votação nos partidos da direita) ficam associadas as seguintes variáveis ecológicas: Religião, ISO1. Às votações do PCP estão associadas as TRAB1 e DIMENSÃO da exploração agrícola, e às votações no PS, municípios em que se verificam grandes valores de TRAB2 e TRAB3.
A **figura 4.9.** ilustra outra função do sistema: a sugestão de um conjunto de regras para classificar os grupos 1, 2, 3 da partição, usando as variáveis activas, depois de discretizadas. Assim, na parte superior da **figura 4.9.** vê-se que o sistema sugere a seguinte regra para o Grupo Nº 1:

$$(0 \leq PCP \leq 5) \wedge (10 \leq PS76 \leq 24) \vee (0 \leq PCP \leq 10) \wedge (25 \leq PS76 \leq 37).$$

Visualizando esta sugestão de interpretação – veja agora a **figura 4.10** – verifica-se que o respectivo fecho convexo coincide praticamente com o do grupo que pretendíamos caracterizar: a votação na direita.
A **figura 4.11.** ilustra a utilização de uma funcionalidade importante quando se trata de formular hipóteses acerca das estruturas dos dados e comparar essas hipóteses com os grupos obtidos em diversas análises: a definição interactiva e visualização de queries.
Suponhamos, por exemplo, que se desejava ver no biplot presente a localização do conjunto de municípios para os quais $(UR= 1) \wedge (1 \leq Dimensão \leq 17)$. A definição obtém-se construindo interactivamente, com o auxílio do sistema, a intersecção seguinte: $\{ i \in I: UR(i)= 1\} \cap \{ i \in I: 1 \leq Dimensão (i) \leq 17\}$.
A lista de municípios que integra esta intersecção é apresentada no lado direito da janela de definição. Definida a expressão, esta pode ser visualizada quer pelo respectivo fecho convexo quer pintando a lista dos indivíduos que a formam com uma cor especificada pelo utilizador. Veja na **figura 4.12** o resultado.
Constata-se, como já se havia verificado com o anterior biplot, que a grande maioria desses municípios vota à direita; observe-se que a maior densidade do conjunto $\{ i \in I: UR(i)= 1\} \cap \{1 \leq Dimensão (i) \leq 17\}$ coincide com a votação na AD, CDS e PSD, nos anos 75, 76,79 a 88.
Finalmente, as **figuras 13** e **14**, ilustram outra função útil: a possibilidade de projectar sobre uma direcção arbitrária definida pelo analista, tanto os indivíduos como as variáveis. Assim, na **figura 4.13**, definida uma direcção pelo centro dos municípios que votam à direita e pelo centro dos municípios que votam à esquerda vê-se que o sistema projectou sobre essa direcção as variáveis activas. Estabelece-se, deste modo, uma relação de ordem entre as variáveis: num dos extremos projectam-se as variáveis que são as votações nos partidos de direita, ao centro as variáveis que representam votações no PS e no outro extremo, as votações no PCP.
A **figura 4.14.** ilustra a mesma ideia mas agora para os municípios. À direcção referida, que define a perspectiva esquerda-direita, corresponde uma relação de ordem sobre os municípios. Correlacionando esta relação de ordem com as variáveis activas, estas aparecem ordenadas segundo o valor dessa correlação. Veja na **figura 4.14** a lista ordenada das variáveis segundo esse critério. Em VAIRINHOS (2003, e

2003a) podem ver-se outras aplicações deste sistema a dados de estudos pós eleitorais em Espanha e ao estudo de dados multivariados em PSICOLOGIA.

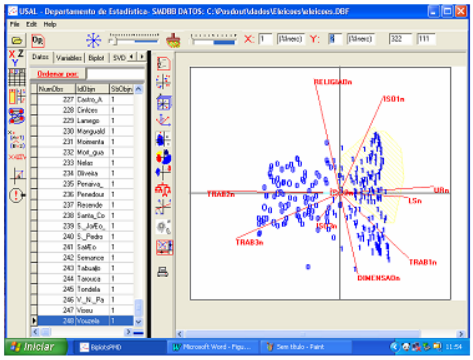

Figura Nº 4.1 – as variáveis activas deste biplot são as variáveis ecológicas e os municípios estão representados pelos valores de UR

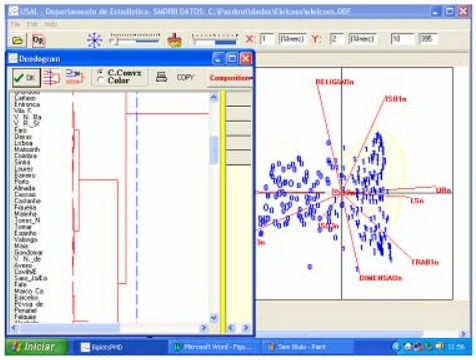

Figura Nº 4.2 –Análise cluster dos municípios. A árvore correspondente foi cortada ao nível indicado pelo tracejado vertical do lado esquerdo.

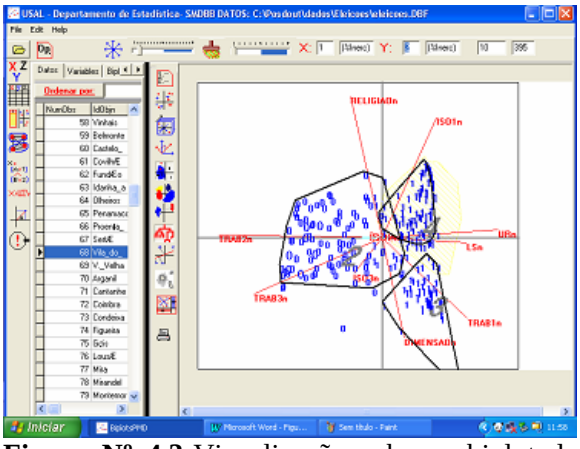

Figura Nº 4.3-Vizualização sobre o biplot da partição definida pelo corte da árvore indicado na figura 4.2

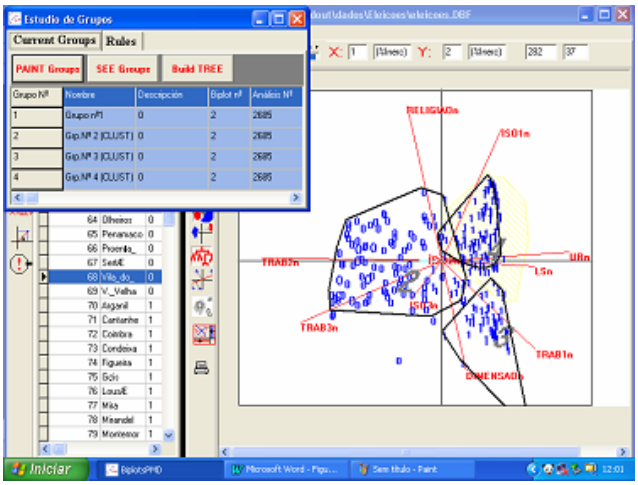

Figura Nº 4.4-Apresentação dos grupos definidos até este instante, para selecção e estudo.

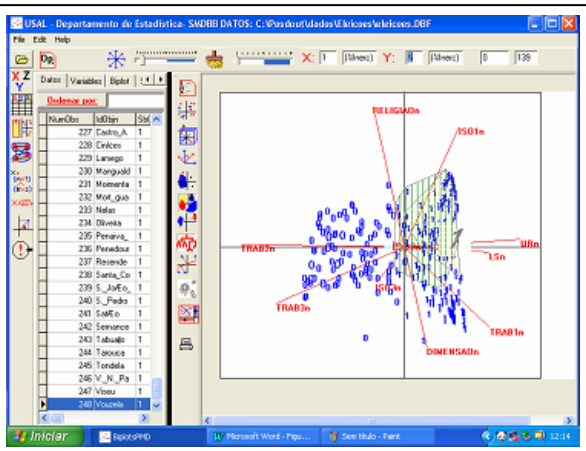

Figura Nº 4.5 – Fecho convexo da descrição sugerida automáticamente pelo sistema para o Grupo nº1: coincide com o grupo a descrever.

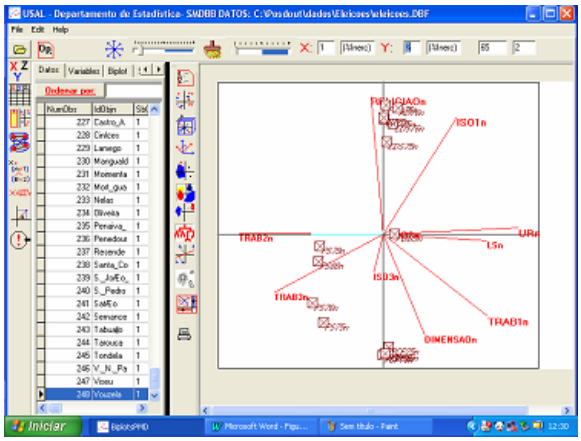

Figura Nº 4.6- Os municípios foram eliminados do gráfico e foram sobrepostos como suplementares as variáveis não activas votação nos partidos.

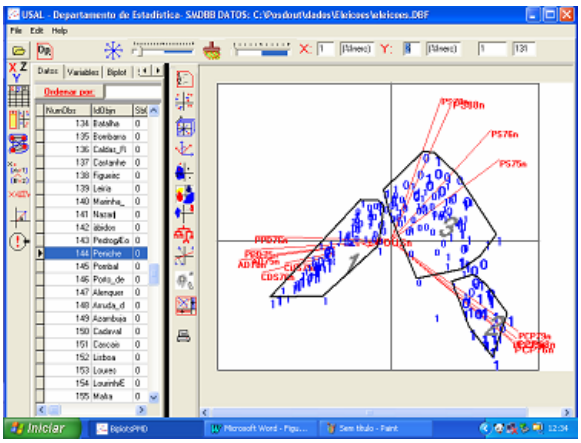

Figura Nº 4.7- Biplot formado pelas variáveis representando as votações nos partidos, com uma partição dos municípios em 3 grupos representando os municípios que votaram à direita, no PCP e no PS.

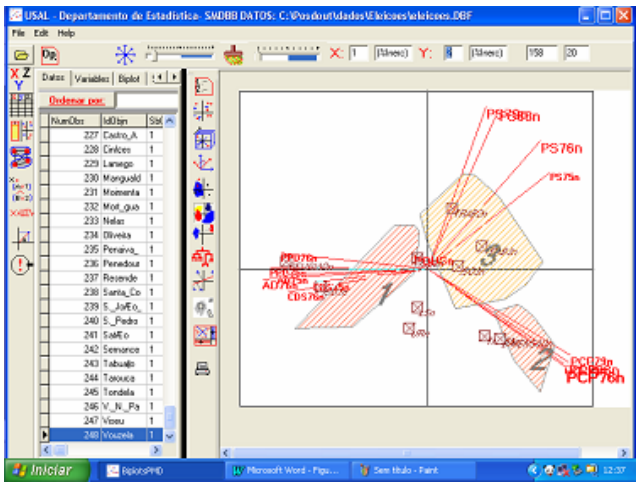

Figura Nº 4.8 – O mesmo biplot da figura 4.7 mas em que agora aparecem como suplementares as variáveis ecológicas, não activas em relação a este biplot.

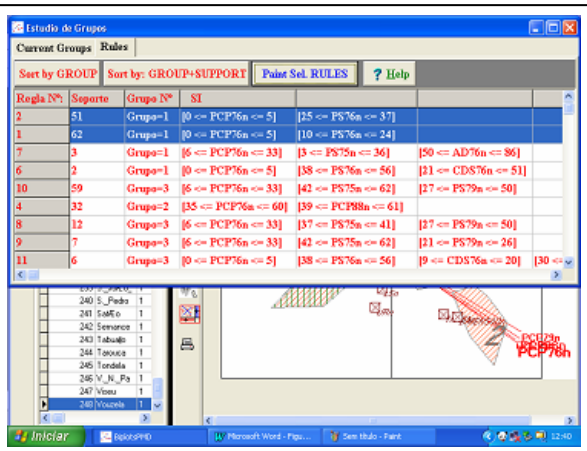

Figura Nº 4.9-Regras de classificação geradas automáticamente para a partição dos municípios em três grupos, sugerida por análise *cluste*r.

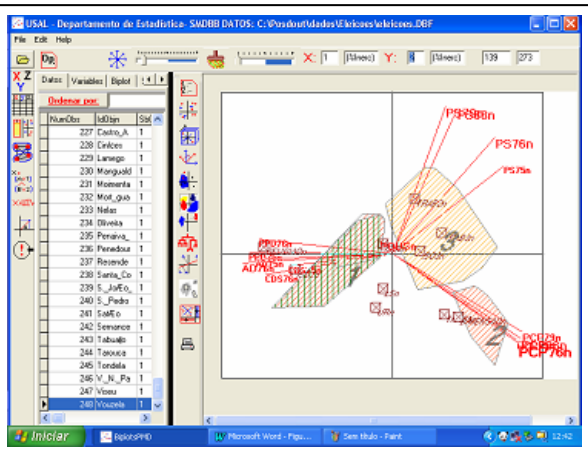

Figura Nº 4.10 - Sobreposição no biplot do fecho convexo dos municípios que satisfazem a regra encontrada automaticamente para o grupo nº 1 da partição. Quase coincidência com a extensão desse grupo

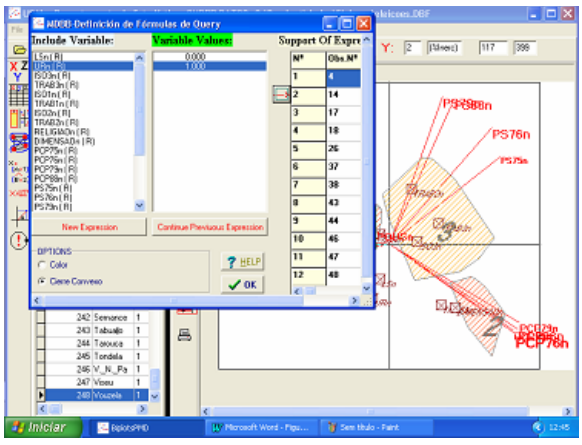

Figura Nº 4.11- Definição interactiva de uma expressão conjuntiva, por intersecção de átomos.

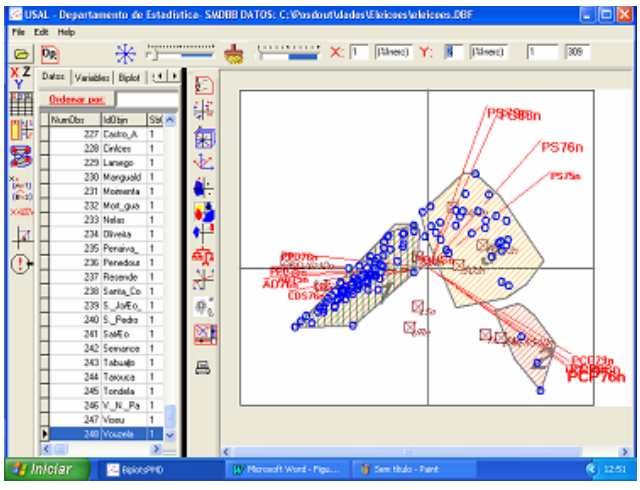

Figura Nº 4.12- Vizualização do fecho convexo da descrição auomática sugerida pelo sistema para o Grupo Nº1 do primeiro biplot. A maior densidade dos municípios que satisfazem votam à direita.

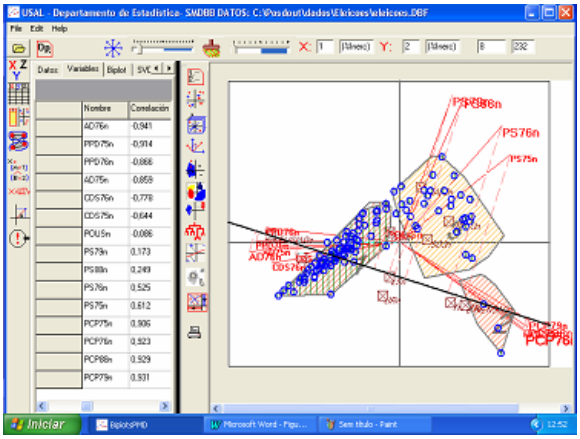

Figura 4.13- Definida uma recta pelos centros de dois grupos, pode-se sobre ela projectar as variáveis activas.

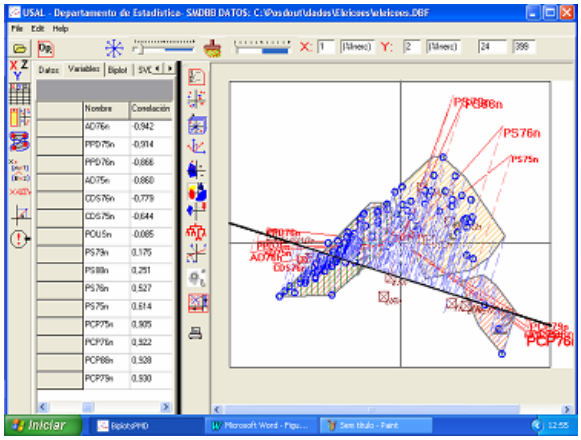

Figura Nº 14- Definida uma recta pelos centros de dois grupos, pode-se sobre ela projectar os indivíduos activos, correlacionando-se a relação de ordem correspondente com as variáveis activas. variáveis activas.

5. CONCLUSÕES.

Neste trabalho procurou-se mostrar a adaptabilidade do conceito de biplot a núcleo de um sistema de *data mining*, mostrando-se que esse método satisfaz as condições gerais de incrementabilidade, escalabilidade e interpretabilidade exigíveis a essas técnicas.

A construção de sistemas visuais de *data mining* centrados em biplots justifica-se, ainda, pelo papel unificador desse conceito no conjunto de técnicas de análise de dados multivariados e pelo facto de os resultados principais gerados por essas técnicas se poderem representar por configurações de marcadores de observações e variáveis em biplots.

A construção de sistemas capazes de gerar sugestões de interpretação, exige a formulação matemática do conceito de interpretação e a criação de um corpo teórico para as actividades de interpretação, sobre o qual basear algoritmos com essa capacidade.

Foram referidas investigações nesse sentido, desenvolvidas no Departamento de Estatística da Universidade de Salamanca. Estas investigações conduziram à criação de um sistema experimental de *data mining* centrado em biplots, assente numa formulação teórica do problema de interpretação – mostrando que tal formulação é possível e tem interesse prático.

A aplicação desse sistema experimental a dados de várias origens é encorajante, tendo-se verificado que as sugestões de interpretação geradas são em geral ajustadas à realidade subjacente aos dados.

BIBLIOGRAFIA

Anderson, R. (1990). *Cognitive Psychology and its Implications.* Freeman.

Benzécri, J. (1973). *L'Analyse des Données*, Dunod.

Benzécri, J. (1992). *Correspondence Analysis Handbook*, Marcel Dekker.

Blázquez Zaballos, A. (1998). *Análisis Biplot Basado en Modelos Lineales Generalizados*, Tesis Doctoral, Universidad de Salamanca.

BORGELT, C. & KRUSE, R. (2001). *Graphical Models for Data Analysis and Mining.* John Wiley.

Bradu, D. & Gabriel, K. (1978). *The Biplot as a Diagnostic Tool for Models of Two-Way Tables*, *Technometrics*, **20** (**1),** 47-68.

Díaz-Leno, M. S. (1995). *Los Métodos Biplot como Herramienta de Diagnóstico en la Modelización de Datos Multidimensionales*, Doctoral, Universidad de Salamanca.

Fayyad, U.; Piatestsky-Shapiro, G; Smyth,P & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* Mit Press.

Gabriel, K. R. (1971). *The Biplot Graphic Display of Matrices with Application to Principal Component Analysis*, *Biometrika*, **58** (**3**), 453-467.

Gabriel, K. R. (1995a). *Biplot Display of Multivariate Categorical Data with Comments on Multiple Correspondence Analysis*, In: W. Krzanowski, (Ed.), *Recent Advances in Descritive Multivariate Analysis*, Oxford Science Publications, p. 190 – 226.

Gabriel, K. R. (1995b). *Manova Biplots for Two - Way Contingency Tables*, In: W. Krzanowski, (Ed.), *Recent Advances In Descritive Multivariate Analysis*, Oxford Science Publications, p. 227 – 268.

Gabriel, K. R.; Galindo, M. P. & Vicente-Villardón, J.L. (1998). *Use of Biplots to Diagnose Independence Models in Three-Way Contingency Tables*, In: J. Blasius & M. Greenacre.(Ed.), Visualization of Categorical Data, Academic Press., p. 391 – 404.

Gabriel, K. R. (2002). *Goodness of Fit of Biplots and Correspondence Analysis*, Biometrika, **89** (**2**), p. 423-436.

Galindo, M. P. (1985). *Contribuciones a la Representación Simultánea de Datos Multidimensionales*, Tesis Doctoral, Universidad de Salamanca.

Galindo, M.P. (1986). *Una Alternativa de Representación Simultanea: Hj-Biplot*. Qüestión **10 (1)**, p. 12 – 23

Golub, G.H.; Van Loan, C. F; *(1983, 1989, 1996) Matrix Computations. Third Edition. The Johns Hopkins University Press.*

Gower , J. C. *(1995) A general theory of biplots. In Krzanowski, Wojtek, J. edit (1995) Recent Advances in Descriptive Multivariate Analysis, pp 283-303* Oxford.

Gower, J. & Hand, D. (1996). *Biplots*, Chapman & Hall

Greenacre, M. (1984). *Theory and Application of Correspondence Analysis*. Academic Press.

Hall, P.; Marshall, D. & Martin, R. (2000). *Merging and Splitting Eigenspace Models, IEE Transactions on Pattern Analysis and Machine Intelligence*, **22** (**9**), p. 1042-1049.

Hall, P.; Marshall, D. & Martin, R. (2002). *Adding and Subtracting Eigenspaces with Eigenvalue Decomposition and Singular Value Decomposition*, Image And Vision Computing, **20**, p. 1009-1016.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. Academic Press.

Krazanowski, W. J. (1998, 2000). *Principles of Multivariate Analysis - A User´s Perspective* (Revised Ed.). Oxford Science Publications.

Nataf, D. (1995) *Democratization and Social Settlements: The Politics of Change in Contemporary Portugal*. State University Of New York.

Sepúlveda, R. (2004). *Contribuciones al Analisis de Clases Latentes en Presencia de Dependencia Local*, Tesis Doctoral, Universidad de Salamanca.

Tukey, J. W. & Wilk, M. B. (1966). 'Data Analysis and Statistics: an expository overview'. *In: L.V. JONES (Ed.). The Collected Works of John W. Tukey . Wadsworth & Brooks/Cole*, pp:. 549-578.

Vairinhos, V. M. (2003a). *Desarrollo de un Sistema de Minería de Datos Basado en los Métodos de Biplot*, Tesis Doctoral, Universidad de Salamanca.

Vairinhos, V. M.; Galindo (2003b). *Aplicação dos Biplots a Dados de Psicologia*, Liberdade, Anuais Científicos da Universidade Independente, Nova Série (3), 2003 p. 67-84.

Vairinhos, V. M.; Galindo, M. P. (2004). *The Multivariate Data Analysis Interpretation Problem, An Essay Of Mathematical Formulation*, Submitted for Pulication.

Vicente-Villardón, J. L. (1992). *Una Alternativa a las Técnicas Factoriales Clásicas Basada en una Generalización de los Métodos BIPLOT*, Tesis Doctoral, Universidad de Salamanca.