PROBLEMES D'ANALISI MULTIDIMENSIONAL EN INFORMATICA

RAMON PUIGJANER

SPERRY UNIVAC I FACULTAT D'INFORMATICA DE LA UNIVERSITAT POLITECNICA DE BARCELONA

RESUM

Després de presentar diversos mètodes d'anàlisi multidimensional (mesures de similaritat, anàlisi factoral, anàlisi de correspondències, mètodes d'agrupament, mètodes de descripció per grafs i mesures de consistència de les classes) s'exposen alguns problemes del mon informàtic on l'anàlisi de dades multidimensionals pot aportar llum a la seva interpretació.

Entre altres s'esmenten el de la caracteritzacio de la càrrega d'un sistema informàtic en vistes a la confecció de models o a la planificació de la capacitat, el de la determinació del conjunt de treball (working set) d'un programa paginat, el de la restructuració dels programes en un medi paginat per millorar el seu comportament, i el de l'organització de las bases de dades en vistes a reduir el nombre d'accessos quan son accedides a l'atzar en un entorn transaccional.

PROBLEMES D'ANÀLISI MULTIDIMENSIONAL EN INFORMATICA

RAMON PUIGJANER

SPERRY UNIVAC I FACULTAT D'INFORMÀTICA DE LA UNIVERSITAT POLITECNICA DE BARCELONA

1.- INTRODUCCIÓ

En un centre de procés de dades que utilitzi un sistema informàtic es presenten nombrosos problemes de recollir mesures i processar-les, si es vol tenir un bon coneixement del seu comportament. Exemples d'aquesta mena de problemes son :

- en l'estudi per metodes analítics o de simulació dels models de sistemes informàtics calen dades com entrada del model i dades per a la seva posterior calibració i validació.
- en l'estudi de les estratègies d'assignació d'un sistema de base de dades cal definir i mesurar indicadors del seu comportament.
- per a la gestió eficient d'un centre de procés de dades es precís coneixer la càrrega que ha de tractar i les seves variacions i fluctuacions.

En tots aquests casos i en d'altres també, un cop s'han recollit les mesures, el seu tractament requereix l'ús d'eines estadístiques. Moltes d'aquestes mesures de sistemes informàtics tenen algunes característiques en comú, com son que el fenòmen observat depen de molts factors, la influència dels quals no es pot considerar de forma separada i, d'altra banda, que les mostres recollides son molt grans (p.e. la traça d'un programa pot tenir milions d'observacions).

En aquest treball es revisen tècniques existents d'Anàlisi de Dades Multidimensionals per al tractament de mostres d'aquestes característiques. En general son tècniques d'estadística descriptiva (o geomètrica) que permeten el tractament de dades multidimensionals.

Al punt 2 d'aquest treball es presenten algunes tècniques d'anàlisi de dades multidimensionals d'ús freqüent i al punt 3 alguns problemes que plantegen els sistemes informàtics i possibles tractaments amb les eines de l'apartat 2.

2.- MÈTODES D'ANÀLISI DE DADES MULTIDIMENSIONALS

Els problemes que se'ns plantejaran son, com hem intuit, de tractar un gran nombre d'observations d'un conjunt important de variables. Per tant, representem per R^p un espai euclidià de p dimensions que sigui el conjunt de tots els vectors de p dimensions de nombres reals. Sigui n el nombre de vectors (o punts) observats a la nostra mostra. Si p es igual a 1 ó 2, es a dir si no més s'observen una o dues variables hi ha tot un conjunt de tècniques estadís-

tiques clàssiques que ens poden donar una descripció acurada de la mostra : histogrames, tests de l'ajust a una distribució, correlació i la seva significació, anàlisi de regresió d'una variable respecte l'altre, etc. En general quan no més hi han una o dues variables la representació dels punts sobre una recta o un pla pot permetre inituir les classes, que es produiran, per simple inspecció i validar la seva qualitat per les tècniques suara esmentades.

Quan aquesta representació no es possible ($n \geq 3$) en calen eines per poder determinar una representació sintètica dels conjunts de dades multidimensionals. Per això en els apartats d'aquest punt estudiarem les mesures de similaritat entre vectors, les representacions geomètriques planes o - projeccions, les descripcions per agrupament, les descripcions per grafs i les mesures de consistència.

2.1.- Mesures de similaritat.
Admetem que la informació bàsica que utilitzarem es la taula X que agrupa totes les observacions de las nostres variables, que podem considerar com un conjunt de n vectors definits a R^p que defineixen les p variables de cada observació o com un conjunt de p vectors definits a R^n que reuneixen les n observacions de cada variable.

Per analitzar un conjunt de dades multidimensionals, la primera cosa que ens cal es una mesura de la seva similaritat o proximitat, es a dir un mètode de representar les observacions d'acord amb les seves semblances mútues. Amb altres paraules, ens cal poder dir que "aquestes dues observacions estan pròximes l'una de l'altra" o que "una observació s'assembla més a aquesta que a aquella altra".
Això exigeix la definició d'una distància d o mesura de similaritat definida en l'espai de les observacions R^p. Entre les distàncies que s'acostumen a definir quan les dades son nombres reals que representen variables contínues (o al menys numèriques), hi han les que s'esmenten a continuació :

2.1.1.- Distàncies quadràtiques.
Son totes aquelles associades a una mètrica caracteritzada per una matriu M de dimensió p x p, simètrica i definida positiva, que defineix un producte intern en R^p :

$$\forall\ \underline{x}_i ,\ x_l \quad R^p \qquad \langle \underline{x}_i , \underline{x}_l \rangle_M = \underline{x}_i^T M \underline{x}_l$$

on $\underline{x}_i^T$ transposat de $\underline{x}_i$ i la distancia d_m vé definida per :

$$d^2_M (\underline{x}_i , \underline{x}_l) = \langle (\underline{x}_i - \underline{x}_l), (\underline{x}_i - \underline{x}_l) \rangle_M =$$

$$= (\underline{x}_i - \underline{x}_l)^T \; M \; (\underline{x}_i - \underline{x}_l)$$

Alguns casos particulars d'aquestes distàncias son :

. *La distància euclidiana on M es la matriu unitària.*

. *La distància euclidiana ponderada on M es una matriu diagonal amb pesos assignats a cada variable (valors reals positius).*

. *La distància euclidiana normalitzada on M es una matriu diagonal amb les inverses de las variànces de cada variable calculades a partir de la taula X.*

Aquesta distància representa que quan més varia la variable j, menys significativa es la seva diferència entre dues observacions.

2.1.2.- Distància de chi quadrat.
Representem per $x_{i.}$ *i per* $x_{.j}$ *la suma de les files i les columnes de la taula X, es a dir :*

$$\forall\ i = 1, \ldots, n \qquad x_{i.} = \sum_{j=1,p} x_{ij}$$

i

$$\forall\ j = 1, \ldots, p \qquad x_{.j} = \sum_{i=1,n} x_{ij}$$

i $x_{..}$ *la suma de tots els elements de X*

$$x_{..} = \sum_{i=1,n} \sum_{j=1,p} x_{ij} = \sum_{i=1,n} x_{i.} = \sum_{j=1,p} x_{j.}$$

Aleshores definirem aquesta distància per

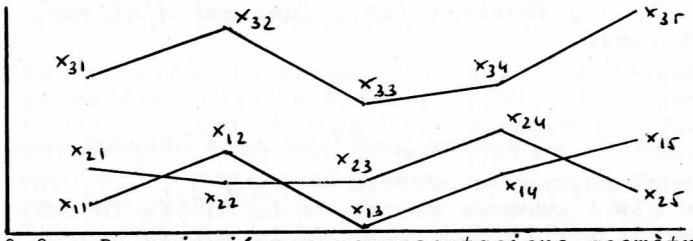
$$d^2(\underline{x}_i, \underline{x}_l) = \sum_{j=1,p} \frac{x_{..}}{x_{.j}} \left(\frac{x_{ij}}{x_{i.}} - \frac{x_{lj}}{x_{l.}} \right)$$

L'interès d'aquesta distància vé del fet que té en compte el "perfil" de les observacions (files de la taula X) i no les diferències individuals de cada variable.

$$d^2(\underline{x}_i, \underline{x}_l) = 0 \leftrightarrow \frac{x_{ij}}{x_{i.}} = \frac{x_{lj}}{x_{l.}} \quad \forall_j = 1,..p$$

D'acord amb les característiques del problema, pot ser mes interessant trobar dues observacions que tinguin el mateix perfil que no pas que tinguin valors pròxims de totes les variables.

Per exemple, a la figura 1, les observacions 1 i 2 estan pròximes des del punt de vista d'una distància euclidiana, mentre que la 1 i la 3 ho estan des del punt de vista d'una distància de chi el quadrat.



2.2.- Descripció per representacions geomètriques. Aquestes mètodes consisteixen en escollir en l'espai de p dimensions de les dades un subespai de dimensió reduida (normalment 2,3,...) de manera que la projecció del conjunt de dades en aquest subespai de baixa dimensió conservi tanta informació com sigui possible del conjunt original de p dimensions. Hi han tants mètodes de projecció com nocions de "bona" representació. A continuació exposarem el principi de l'Anàlisi en Components Principals (ACP) i de l'Anàlisi de Correspondències de Benzecri(12), sense que això signifiqui menyspreu d'altres mètodes, com l'Algorisme de Persecució de la Projecció de Friedman i Tukey (23) i l'Anàlisi Discriminant d'Anderson (5).

Sigui X la nostra taula de dades i $E=\{\underline{x}_i/i=1,\ldots,n\}$ el conjunt de n punts a R^p que s'ha d'analitzar. Suposem que l'espai R^p té una funció distància quadràtica d_M i que s'assignen ponderacions p_i a les n observacions (Hi ha casos en que aquestes ponderacions es prenen totes iguals, però n'hi ha d'altres en que apareixen de manera natural, com és el cas d'observacions repetides). La suma de les ponderacions valdrà la unitat

$$\sum_{l=1}^{n} p_i = 1$$

Aleshores, tal com hem dit, el problema de l'ACP es trobar un subespai de R^p de dimensió reuida de manera que la pro-

jecció de E sobre aquest subespai sigui una "bona" representació de E. Generalment es procura trobar subespais de 2, 3, 4 ó 5 dimensions i les observem en les seves diferents seccions transversals planes.

Més precisament, el problema de l'ACP es pot enunciar de la següent manera :

Trobar un subespai $E_k \subset R^p$ $(dim\ (E_k) = k < p)$

tal que

$$\sum_{l=1}^{n} p_i \ d^2_M \ (\underline{x}_i, \ E_k) \quad sigui\ mínim$$

A continuació exposem el teorema que ens permet resoldre el problema concret sense la demostració que es pot trobar a Anderson (5) o a Benzecri (12), junt amb d'altres detalls i precisions.

Teorema :

L'espai E_k que respon al nostre problema està generat per la base ortogonal segons la mètrica M $(\underline{u}_1, \underline{u}_2, \ldots)$, tal que els $\underline{u}_m$ son els k vectors propis de la matriu VM associats als k valors propis mes grans de la matriu suara esmentada, on V es la matriu de variança - covariança de E.

$$\forall m,\ m = 1, \ldots k \qquad VM\underline{u}_m = \lambda_m \ u_m$$

A més

$$\lambda_1 + \lambda_2 + \ldots + \lambda_k = C - \sum_{l=1}^{k} p_i \ d_M \ (x_i, \ E_k)$$

que és l'expressió que s'ha de maximitzar, on C es una constant.

Observem que aquest teorema ens proporciona un espai E_k, però no ens diu que aquesta solució sigui única. De fet, el nombre de solucions es pot deduir del nombre de les possibles ordenacions equivalents dels valors propis de VM (Per exemple si E C R^3 es esfèric hi haurà un nombre infinit de plans on projectar-lo i representar-lo de forma equivalent)

Malgrat que des del punt de vista teòric el problema està resolt, des del punt de vista pràctic convé aprofondir un xic més, calculant las components dels vectors - associades als eixos determinats pels vectors propis retinguts (que es denominen factorials o principals per analogia amb la mecànica). Aquest càlcul es duu a terme

facilment a partir dels vectors principals, la matriu de la mètrica i els valors de les observacions.

Un altre aspecte important es la mesura de la qualitat de la representació de E per E_k que ve donada per :

$$\sum_{j=1}^{k} \lambda_j \quad / \quad \sum_{j=1}^{p} \lambda_j$$

Generalment aquesta expressió posada en tant per cent es denomina "percentatge" o "part" de la informació conservada per E_k.

Per concloure, hem vist doncs que el nostre problema es redueix bàsicament a manipulacions sobre la matriu VM, que és simètrica i de dimensió p x p (la necessitat de memoria es doncs de p (p-1)/2), la principal de les quals es l'extracció dels k valors propis més grans i dels seus vectors propis associats.

Una extensió interessant de l'ACP es l'anàlisi en correspondències degut a Benzecri (12). Aquí considerarem el nostre conjunt de dades com constituint dues taules, la ja coneguda E de les observacions i la taula F dels paràmetres (la mateixa d'abans, si es vol, però considerada per columnes en lloc de per files). En l'ACP hem vist que a més de la taula de dades s'han de proporcionar dues entrades més : el sistema de ponderació $\{p_i / i = 1,\ldots n\}$ i una distància quadràtica, associada amb una matriu M sobre R^p. En l'anàlisi per correspondències, aquestes dues entrades s'escolleixen de manera que les dues ACP sobre els conjunts E i F d'observacions i de paràmetres, es puguin deduir una de l'altra i interpretar simultàniament.

Per això, comencem transformant la taula de dades inicial X de la manera següent.

. Suposem que $x_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} = 1$, si no ho fos

transformarîem totes les x_{ij} en $x_{ij}/x_{..}$

. $\forall i = 1,\ldots n$ i $\forall j = 1,\ldots p$ calcularem

$$f_j^i = \frac{x_{ij}}{x_{i.}} \qquad i \qquad f_i^j = \frac{x_{ij}}{x_{.j}}$$

on $x_{i.} = \sum_{j=1}^{p} x_{ij}$ i $x_{.j} = \sum_{i=1}^{n} x_{ij}$

Aquestes transformacions son bastant naturals quan les dades de X son comptatge d'aconteixements o correspondències o quan les dades son booleanes. Aleshores les x_{ij} es poden interpretar com freqüències empíriques, les $x_{i.}$ i $x_{.j}$ com distribucions marginals empíriques i les f_j^i i f_i^j com freqüències condicionals empíriques.

Aleshores podem fer una ACP del conjunt de n punts de R^p de les freqüències condicionals empíriques $\{f_j^i\}$ fent servir com ponderacions les freqüències marginals $\{x_{i.}\}$ i com distància quadràtica la que està associada a una mètrica inversa de les freqüències marginals $\{x_{.j}\}$ (la qual cosa ens duu a una distància de χ^2 entre observacions de $\{f_j^i\}$). Fem d'altra banda l'ACP simètrica, es a dir dels p punt de R^n de les freqüències condicionals empíriques $\{f_i^j\}$ fent servir com ponderacions les freqüències marginals $\{x_{.j}\}$ i com distància quadràtica la que està associada a una mètrica inversa de les freqüències marginals $\{x_{i.}\}$.

Aquestes dues ACP condueixen a components principals associades als mateixos valors propis i que estan relacionades. En conseqüència :

- La descripció dels dos conjunts E i F requereix una sola anàlisi on treballem a l'espai de menor dimensió.
- Les components principals trobades constitueixen bones representacions dels conjunts E i F que a més es poden interpretar simultàniament i conjunta.

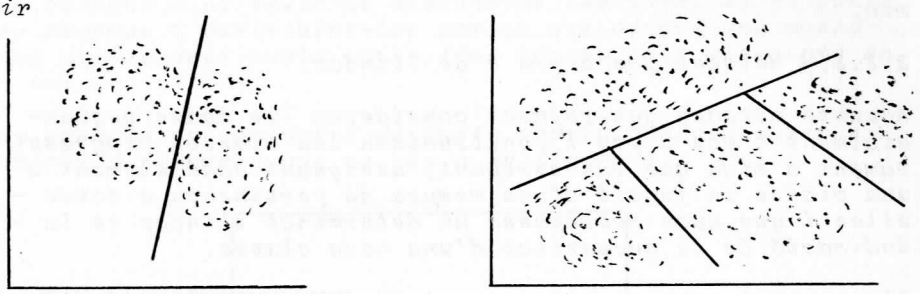
2.3.- Descripció per agrupament.

Els mètodes anteriors descriuen conjunts de dades multidimensionals mitjançant representacions geomètriques que posen de manifest usualment algunes propietats de les poblacions observades. En aquest apartat presentarem mètodes que fan particions del conjunt de dades. Per precisar la noció d'una "bona"partició cal introduir una funció criteri sobre totes les possibles particions del conjunt E d'observacions. El diferents mètodes d'agrupament difereixen o pel criteri de qualitat que s'intenta satisfer o pel procediment que condueix a la solució que satisfà un criteri determinat.

Donat el conjunt de dades E a R^p i una distància d a R^p, els criteris més freqüents es basen en la dispersió de les classes de la partició, es a dir en la suma de les distàncies dels punts de cada classe al seu centre de gravetat.

A la figura 2 es presenten dos exemples de particions naturals. A l'exemple 2a la partició en dues classes (P_1, P_2) es una bona partició per al criteri de dispersió suara proposat. A l'exemple 2b la partició natural en 4 classes co-

rrespondria a optimitzar un criteri força difícil de definir



2a 2b

Figura 2

El primer mètode que podriem emprar per trobar la partició P de E en k classes tot optimitzant una funció criteri donada es la de calcular W (P) per a totes les possibles P, i triar la millor. Per poc nombroses que siguin les observacions de E aquest tractament es fa excessivament car. Altres mètodes eviten l'enumeració completa introduint alguna informació a priori o porten a solucions que només son localment òptimes. Els mètodes d'agrupament els podem classificar de manera senzilla, en tres famílies.

2.3.1.- Mètodes iteratius.

Aquests mètodes defineixen unes classes inicials i després les milloren iterativament. Aquesta millora s'aconsegueix agregant els agrupaments al voltant d'algun centroide característic.

Sobre aquest esquema bàsic s'han desenvolupat nombrosos mètodes diferents segons el centroide, la distància i el criteri de convergència que es facin servir. Respecte del centroide trobem els que utilitzan el centre de gravetat com l'ISODATA de Ball i Hall (7), o un subconjunt de la classe com els núvols dinàmics de Diday (17), (18) i les k-mitjanes de McQueen (34).

2.3.2.- Mètodes per manipulacions sobre la matriu de distàncies.

A partir de la matriu de similaritat o distàncies entre tots els elements que s'hagin de classificar, un "bon" agrupament hauria de ser de manera que la matriu de similaritat reordenada després de l'agrupament estigués el més a prop possible d'una forma diagonal de blocs, es a dir que als blocs de la diagonal principal hi hagués valors grans de la similaritat (o distàncies petites) i als altres els valors fossin petits (o distàncies grans). A partir d'aquesta senzilla observació, McCormick et al (33), han desenvolupat el "Bond Energy Algorithm" que es un mètode eficient per permutar les files i les columnes de la matriu -

de similaritat fins trobar una forma diagonal satisfactòria.

2.3.3.- Mètodes pas a pas o de llindar.

Aquests mètodes generalment consideren les dades seqüencialment o pas a pas i construeixen les classes progressivament a mida que van arribant, assignant cada element a una classe en funció d'una mesura de pertinença a totes elles i que quan sobrepassa un determinat llindar es la indicació de la necessitat d'una nova classe.

Alguns punts encara no prou estudiats son l'elecció del llindar que condiciona el nombre de classes, la inicialització de las característiques de cada classe, l'ordre de tractament de les dades i la possibilitat de treballar amb l'ajut d'un "professor".

Algorismes dins d'aquesta classe, més utilitzats però en robòtica i autoperentatge, es poden trobar a Aguilar Martin et al (3) i a López de Màntaras (29).

2.4.- Descripció per grafs.

El que en l'apartat 2.2. hem anomenat representacions "geomètriques" eran descripcions gràfiques que calia interpretar com mapes : les nocions de "proximitat" tant individual com global eren raonables. En aquest paragraf presentarem les representacions per grafs en que les nocions de nodes, arestes i camins es faran servir per interpretar l'estructura de les dades.

Tots els mètodes d'aquest apartat treballen sobre taules de distàncies D, es a dir, el conjunt E que s'ha d'analitzar es caracteritza per totes les distàncies entre les parelles dels seus elements, independentment que aquestes distàncies siguin les dades reals o s'hagin obtingut a partir d'alguna representació o calculat d'alguna manera especial.

Dient-ho d'una altra manera, aquests mètodes no fan servir directament la representació multidimensional del conjunt que s'ha d'analitzar com un conjunt de punts en un espai de p dimensions. El que es considera explícitament es l'estructura de distàncies d'aquest conjunt. No obstant això, la visió multidimensional està subjacent en tots els mètodes basats en la distància i en moltes aplicacions les dues visions es consideren conjuntament.

2.4.1.- Mètodes d'agrupament jeràrquic.

L'agrupament jeràrquic es una representació de les dades mitjançant un arbre jeràrquic o dendograma. Proporciona també particions aniuades de les dades, raó per la qual es denomina agrupament jeràrquic.

L'esquema general d'aquests algorismes consisteix en anar buscant a la taula de distàncias les parelles de punts més propers i substituint-los per un que tingui una massa suma de les dels punts units fins reunir tots els punts en un sol.

Per exemple un conjunt E de punts podria quedar representat pel dendograma que es veu a la figura 3.

La partició en classes s'aconsegueix fixant una distància x que sigui superior als grups formats a distàncies inferiors, com es veu també a la figura 3.

La dificultat major resideix en amagatzemar la matriu D quan el conjunt es gran i en la determinació de la distància que ens permet separar les classes.

En d'altres algorismes la partició jeràrquica es fa de dalt a baix per successives particions del conjunt E.

2.4.2.- Mètodes de partició de grafs.

El coneixement de la taula de distàncies D permet assignar una estructura de graf al conjunt E dels objectes a analitzar. Pot considerar-se com un graf complet valorat per les distàncies.

Si E es molt gran, hi han mètodes que permeten construir un graf incomplet sobre E de manera que les arestes existents portin tanta informació com sigui possible sobre el graf de distàncies complet. Dos d'aquests mètodes son :

La construcció de l'arbre d'extensió mínim d'E (Minimum Spanning Tree, Zahn (42)) : es un subgraf (es a dir, connex i tenint tots els punts d'E com nodes) en forma d'arbre (es a dir, sense cicles) i d'entre tots ells el que té una llargada mínima.
Mètodes que redueixen la mida dels conjunts de nodes i d'arestes del graf complet inicial. (Milgram (35)) .

2.5.- Mesures de la consistència de les classes.

Dels paragrafs precedents es dedueix facilment que els algorismes de classificació i les seves variants poden ser molt nombrosos i, puix que sovint condueixen a agrupaments sensiblement diferents, cal disposar de criteris i mesures que ens permetin avaluar la qualitat de l'agrupament obtingut tenint en compte l'ús a que es destinará i la interpretabilitat de les classes generades.

2.5.1.- Experiència.

El contrast de les classificacions obtingudes amb l'experiència adquirida per les persones responsables o que tinguin un bon coneixement del conjunt a classificar, pot permetre una estimació, malgrat la seva subjectivitat, de la qualitat relativa de les classificacions obtingudes.

2.5.2.- Densitat.

Per mesurar la homogeneitat de les classes es poden definir, en l'espai de les observacions centrades i reduides, hiperparalelepípedes rectangulars que continguin tots els punts de cada classe, essent aleshores la densitat de cada classe el nombre d'observacions que hi ha en l'esmentat hiperparalelepípede per unitat d'hipervolum.

2.5.3.- Radis de classe i distància entre centres de classe.

Aquestes dues mesures i la seva comparació ens proporcionen una idea de la possibilitat de que dues classes puguin tenir una intersecció dels seus àmbits sempre i quan admetem que tenen una forma aproximadament hiperesfèrica.

2.5.4.- Agregació.

Aquestes mesures extretes de l'anàlisi canònica (Romeder (38)) ens donen la distància mitjana entre tots els punts de la classe segons diferents mètriques.

Si per calcular aquesta distància agafem diferents matrius per definir la distància obtindrem diferents mesures que tindran significats diferents i la comparació de les quals, entre elles i amb el radi de la classe, ens permetrà tenir una idea de la forma de la classe.

Per un exemple d'utilització de distàncies d'auesta mena vegeu Puigjaner (37).

2.5.5.- Distància generalitzada de Mahalanobis.

Aquesta mesura extreta també de l'anàlisi canònica preten, prèvia comparació de que la dispersió dins de cada classe es la mateixa, verificar que els centres de les diferents classes son significativament diferents.

2.5.6.- Anàlisi de la generació de classes.

La realització d'una anàlisi en components principals a cadascuna de les classes obtingudes ens permet observar si -

llur estructura es la mateixa o hi han diferents orientacions dels eixos principals, es a dir si aquests están explicats o no per les mateixes variables. (Barcelo i Puigjaner, (9)).

3.- APLICACIÓ DELS MÈTODES D'ANÀLISI DE DADES MULTIDIMENSIONALS A PROBLEMES INFORMATICS.

En el paragraf anterior hem presentat algunes tècniques d'anàlisi de dades multideimensionals, l'ús de les quals ha d'anar precedit d'una anàlisi del problema a fi de ser capaços de descriure'l amb precisió, d'escollir l'estratègia de mostreig, de seleccionar els paràmetres importants etc. Cal ser conscients de que la construcció de la taula X, que en l'apartat anterior hem pres com punt de partida, presenta problemes d'ordre pràctic, que, en molts casos, no es poden subestimar.

3.1.- Caracterització de la càrrega.

D. Ferrari (21) assenyala que el problema de descriure amb precisió i de caracteritzar la càrrega d'un sistema (o d'un centre de càlcul) es un pas important tant en el camp de la seva gestió com en el de l'avaluació del seu comportament, i preten un millor coneixement del tipus de recursos que necessiten les diferents parts de la càrrega. Aquest coneixement es necessari per enfocar diferents problemes pràctics com per exemple, la preparació d'un benchmark, la planificació de la capacitat d'un sistema informàtic, obtenir de dades per construir models, etc.

La mesura de la càrrega es realitza generalment de la manera següent : durant algun periode representatiu d'activitat per a cada treball (o programa) es recull informació sobre als requeriments de temps, espai, perifèrics, etc. Es a dir es creen tantes mostres estadístiques com variables s'observin. Normalment l'eina que es fa servir per a aquesta extracció es el sistema de comptabilitat del nostre computador.

Els treballs clàssics sobre caracterització de la càrrega es feien estudiant per separat cadascuna de las variables o com a màxim mitjançant una anàlisi per regressió per posar de manifest l'influència d'un conjunt de paràmetres sobre una variable privilegiada. L'anàlisi de dades multidimensionals ens permet enfocar totes les possibles interaccions entre els paràmetres, caracteritzant cada treball (o programa) per una sequència de variables tal com :
(instant d'arribada, temps de CPU, memòria, nombre d'I/O sobre disc, nombre de cintes, etc.),
que podem considerar com un vector de nombres reals. En conseqüència el problema de la caracterització de la càrrega cau de ple en l'anàlisi de dades multidimensionals i el que pretenem es trobar agrupaments de treballs (o progra-

mes) que necessitin recursos similars.

A partir d'aquest punt els diferents treballs observats diferexen tant en la forma de triar i codificar les variables com pels algorismes de classificació que es fan servir. Així per exemple a Agrawala et al (2) es fa servir una variant del mètode de les k-mitjanes i a Artis (6) una variant del mètode Isodata. A Puigjaner (37) hi trobem una comparació entre mètodes d'agrupament jeràrquics i no jeràrquics i a Barcelo i Puigjaner (10) la seva comparació amb mètodes pas a pas. Un altre enfocament interessant es el que podem trobar a Biondi (13) on cada variable es classifica en un cert nombre de classes equiprobables donant a cada punt una codificació binària amb nomes un 1 a cada variable. A continuació la determinació de las classes es fa mitjançant l'aplicació d'algorismes derivats de l'anàlisi en correspondències.

Finalment un altre treball interessant es el de Serazzi - (41) que preten lligar la classificació obtinguda amb caracteritzacions externes del programa (dificultat, tipus d'aplicació, llenguatge de programació, etc.)

La conclusió de totes aquestes aplicacions de l'anàlisi - de dades multidimensionals es que encara cal definir criteris suficientment fiables per poder triar entre les classificacions que obtenim pels diferents mètodes, que poden portar a resultats discordants.

3.2.- Comportament de l'adreçament dels programes.

Per ajudar a comprendre com els programes referencien la memòria, la presència de "localitats", Denning (15) i (16) pot ser la característica més important del comportament - d'un programa. Això significa que el temps d'execució d'- un programa es pot dividir en diferents fases durant les quals les referències del programa es concentren en algun conjunt d'adreces determinat. El concepte de localitat de referències a memòria es de gran importància per dissenyar sistemes de memòria virtual i es fa servir explícitament en diverses polítiques de gestió de memòria d'aquests sistemes. Malgrat la seva senzillesa intuitiva, el concepte de localitat és difícil de quantificar a causa de la seva relativitat : el mateix programa pot tenir diferents estructures de localitat quan es consideren diferents nivells de detall.

La construcció d'un model a posteriori com observació física del comportament de programa ens porta de bell nou a un cas de tractament d'un nombre de dades a partir de les quals cal escatir la seva significació de manera resumida.

Dins d'aquest enfocament del problema es interessant el - treball de Schroeder (40) en que la seqüència de conjunts de treball (working sets, Denning, (15)), s'ha obtingut mostrejant-los a intervals de llargada constant (mesurada

en nombre de referències) i on també ho es la finestra - que ens permet determinar els conjunts de treball. Cada conjunt de treball es pot representar mitjançant el vector :

$$(x_{i1} \quad x_{i2} \ldots\ldots x_{iN})$$

on N es el nombre total de pàgines a l'espai virtual i x_{ij} es igual a 1 si la j-ena pàgina està referenciada al i-è conjunt de treball i igual a 0 en cas contrari.

En conseqüència ens trobem davant d'un problema en que l'aplicació de l'anàlisi de correspondències es adequat - car la seva esctructura es de dades booleanas (sequència de vectors a $\{0,1\}^N$) i ens interessa trobar les localitats o règims estacionaris durant els quals el programa treballa sobre conjunts de classes determinats i quines - son les pàgines que caracteritzen aquests conjunts de treball.

En el treball esmentat s'arriba a l'agrupament de cinc - classes de conjunts de treball i de sis classes de pàgines i la caracterització de cada classe de conjunts de - treball per les pàgines de les diferents classes amb probabilitats determinadas de participació.

3.3.- Reestructuració de programes .

L'objectiu de les tècniques de reestructuració de programes es de millorar l'eficiència de la paginació mitjançant una disposició adequada dels programes a la memòria virtual.

Per això, els programes es parteixen en varis blocs que venen determinats pel programador i, per tant, son generalment els mòduls lògics que composen el programa. Alesho- res el problema es redueix a aplicar aquests mòduls a les pàgines virtuals a fi de minimitzar el nombre de falles - de pàgina que es produeixen durant l'execució. Per aconseguir-ho cal que els blocs que s'accedeixen freqüentment junts estiguin a la mateixa pàgina.

Hi han nombrosos mètodes proposats a la literatura (Ferrari, (20),)21), Masuda et al, (32) Achard et al, (1) etc.) i tots tenen en comú que parteixen de la matriu de similaritat (o de disimilaritat) entre els blocs. Es a dir cada bloc està caracteritzat pel vector d'indexos de similaritat amb tots els altres. Aquesta mesura de similaritat varia d'un autor a l'altre i acostumen a fonamentar-se en la frequència en que els blocs han estat referenciats - junts o en la noció de referències crítiques, que son les que fan apareixer una falla de pàgina per una determinada disposició del programa.

Un cop obtinguda la matriu de similaritat ens cal procedir a l'agrupament dels mòduls, per a la qual cosa podem fer

servir qualsevol dels mètodes exposats als paragrafs 2.3 i 2.4.

Cal però tenir en compte a més la restricció del tamany de la pàgina; fins a l'actualitat no hi han mètodes que resolguin aquest problema de manera òptima.

3.4.- Assignació i reorganització de bases de dades.

En les grans bases de dades implantades en jeràrquies de memòries, es important distribuir els registres en grups de manera que cada transacció requereixi accedir a tan pocs d'aquests grups com sigui possible. Per això cal que els grups d'enregistrements, que es referencien freqüentment per les mateixes transaccions, estiguin identificats per poder ajuntarlos. Per coneixer com les transaccions d'usuaris referencien els enregistraments, s'ha de recollir aquells que s'han accedit per cada transacció durant algun periode representatiu d'activitat. Un enregistrement pot representar-se aleshores per un vector,

$$(x_{i1} \quad x_{i2} \quad \ldots x_{iM})$$

on x_{ij} es igual a 1 si la j-ena transacció fa servir el i-é enregistrament i igual a 0 en cas contrari, essent M el nombre total de transaccions.

Aquesta representació té el mateix aspecte del presentat a 3.2 : les unitats d'adreçament son registres en lloc de pàgines d'una banda i d'altra les unitats d'activitat elemental son transaccions en lloc de conjunts de treballs.

Hi han diversos treballs (Gorenstein i Galati, (24), Hoffer i Severance, (26), Flory et al, (22)) que plantejen el problema de diverses maneres, totes elles encoratjadores però els resultats obtinguts son encara força parcials.

4.- CONCLUSIONS

L'objectiu d'aquest treball, com deiem, es introduir algunes tècniques d'estadística descriptiva des del punt de vista de problemes multidimensionals i la seva aplicació a problemes d'anàlisi i millora del comportament de sistemes informàtics.

Aquestes tècniques no representen treballs closos, ans al contrari, es tracta d'una àrea de treball oberta a la collaboració d'informàtics i estadístics per definir el problema com la descripció d'un conjunt finit d'elements; per caracteritzar aquestes dades per variables significatives i manejables; per pensar quan direm que dos elements de dades estan propers o son similars; i per col·laborar en la

selecció del mètode de descripció i suggerir adaptacions o millores al problema concret.

Finalment, i no per això la tasca menys important, l'informàtic ha d'interpretar els resultats obtinguts que hauran d'orientar l'estadístic i a ell mateix cap a noves recerques i millores dels mètodes.

5.- BIBLIOGRAFIA

(1) M.S.ACHARD, J.Y.BABONNEAU, G.MORISSET.
Automatic and General Solution to the Adaptation of Programs in the Paging Environment.
IRIA LABORIA Research Report n. 196 Nov.1977

(2) A.K.AGRAWALA, J.M.MOHR, R.M.BRYANT.
An Approach to the Workload Characterization Problem. Computer Jun.1976 pp 18-32.

(3) J.AGUILAR MARTIN, M. BALSA, R. LPEZ DE MANTARAS.
Estimation Recursive d'une Partition. Exemples d'Apprentissage et Auto-apprentissage dans R^N et I^N:
QUESTIIO Set. 1981 pp 150-172.

(4) M.R.ANDERBERG.
Cluster Analysis for Applications.
Academic Press 1973.

(5) T.W.ANDERSON.
Introduction to Multivariate Statistical Analysis.
Wiley 1958.

(6) H.P.ARTIS.
A Technique for Determining the Capacity of a Computer System.
Proc CPEUG Nov. 1976 pp 150-162

(7) G.H.BALL, D.J. HALL.
A Clustering Technique for Summarizing Multivariate Data.
Behavioral Sciences Vol. 12 n. 2, 1967, pp 153-155.

(8) J.BARCELO, R. PUIGJANER.
Techniques for Computer Modeling and Workload Characterization.
Proc. ECOMA-7 Oct. 1979 pp 268-284

(9) J.BARCELO, R.PUIGJANER.
Workload Modeling by Clustering Techniques : a Rafinement Procedure and its Uses as Input to Simultate a Computer.
IV Meeting Euro Working Group on Operational Research and Computer Science. Des. 1981.

(10) J.BARCELO, R. PUIGJANER.
Workload Characterization Self Learning Adaptive Method.
Raport de Recerca FIB 82-05 May 1982.

(11) J.L.BAER, G.R.SAGER.
Dynamic Improvement of Locality in Virtual Memory Systems.
IEEE-TSE Vol. 2 n. 1 Mar 1976 pp 54-62

(12) J.P.BENZECRI et al.
L'Analyse des données
Dunod 1973.

(13) J.BIONDI.
Description de la Charge d'un Systeme Informatique et Application a l'Evaluation des Performances.
RAIRO-Informatique

(14) P.BURGEVIN, J.LEROUDIER.
Characteristics and Models of Program Behavior.
Nat Conf. ACM Oct 1976 pp 344-350.

(15) P.J.DENNING
The Working set Model for Program Behavior.
CACM Vol 11 May 1968 pp 323-333.

(16) P.J. DENNING
Working set : Past and Present.
CACM

(17) E. DIDAY
La Methode des Nuées Dynamiques. Revues de Statistique Appliquée Vol. 19 n. 2, 1971, pp 19-34.

(18) E. DIDAY
Organisation en classification Automatique et Reconnaissance des Formes.
RAIRO Nov. 1972 pp 61-95.

(19) E. DIDAY
Classification Automatique Sequentielle pour Grands Tableaux.
RAIRO Mar. 1975 pp 1-29.

(20) D.FERRARI.
The improvement of Program Behavior. Computer Jul-Aug. 1972 pp 18-24.

(21) D.FERRARI.
Computer Systems Performance Evaluation
Prentice Hall 1978.

(22) A.FLORY, J.GUNTHER, J.KOULOUMDJIAN.
Data Base Reorganization by Clustering Methods.
Information Science Vol. 3 pp 59-62

(23) J.H.FIEDMAN, J.W.TUKEY.
A Projection Pursuit Algorithm for Exploratory. Data Analysis.
IEEE-TC 23 n.9 Set. 1974, pp 881-890

(24) S.GORENSTEIN, G.GALATI.
Data Base Reorganization for Storage Hierarchy.
IBM Research Report n. RC5063. Yorktown Heights Oct. 1974.

(25) J.A.HARTIGAN.
Clustering Algorithms. Wiley 1975.

(26) J.A.HOFFER, D.G.SEVERANCE
The USE of Cluster Analysis in Physical Data Base Design.
Proc. Very Large Data Bases. Conf. Ed.D.S. Kerr, 1975 pp 69-86.

(27) B.W.KERNIGHAN.
Optimal Sequential Partitions of Graphs J ACM Vol 18 n. 1 Jan 1971 pp 34-40

(28) L.LEBART, J.P.FENELON.
Statistique et Informatique Appliquées
Dunod 1975.

(29) R.LOPEZ DE MANTARAS.
Algorismes d'Aprenentatge en Reconeixement de Formes : Aplicació a la Robótica.
Tesi Doctoral. FIB Des 1981.

(30) S.A.MAMRAK, P.D.AMER.
A feature Selection Tool for Workload Characterization.
Int. Conf. on Computer Performance, Measurement, SIGMETRICS/CMQ VIII Nov. 1977.

(31) A.W. MADISON, A.P.BATSON.
Characteristics of Program Loralities.
CACM Vol 19 n. 5 May 1974 pp 285-294

(32) T.MASUDA, H.SHIOTA, K.NGUCHI, T.OHKI.
Optimization of Program Organization by Cluster Analysis.
IFIP - 74 pp 261-265

(33) J.McCORMICK, P.J.SCHWEITZER, T.W.WHITE.
Problem Decomposition and Data Reorganization by a Clustering Technique.
Operation Research Vol. 20 n.5 Set. 1972 pp 993-1009

(34) J.McQUEEN
Some Methods for Classification and Analysis of Multivariate Observation.
5th Berkeley Symp. Math Stat and Prob. Vol. 1 n. 1 1967 pp 281-297.

(35) M.MILGRAM, B.DUBUISSON, B.VACHON.
A Computationally Eficient Clustering Algorithm.
IEEE-TSMC Vol. 7 n.2 Feb. 1977. pp 99-104

(36) D.F.MORRISON.
Multivariate Statistical Methods.
McGraw Hill 1973

(37) R. PUIGJANER
Comparación de Algoritmos y algunos Refinamientos a la Caracterización de la Carga por Métodos de Agrupamiento.
Treball presentat per al Concurs Oposició d'-Arquitectura de Computadors de la FIB. No publicat Des. 1981.

(38) J.M.ROMEDER
Mèthodes et Programmes d'Analyse Discriminante Dunod 1973.

(39) A. SCHROEDER.
Analyse d'un Melange de Distributions de Probabilité de Meme Type.
Revue de Statitique Appliquée Vol. 24 n.1 1976 pp 39-62.

(40) A. SCHROEDER.
A Statistical Aproach to the Study of Program Behavior.Via Reference String Analysis.
Research Report IRIA-LABORIA n. 240 1977

(41) G.SERAZZI
A Functional and Resource - oriented Procedure for Workload Modeling.
Proc. PERFORMANCE'81 North Holland 1981