UNIVERSITAT DE BARCELONA

Departament de Física Aplicada i Òptica

Programa de Micro i Optoelectrònica Física

Bienni 1994-96

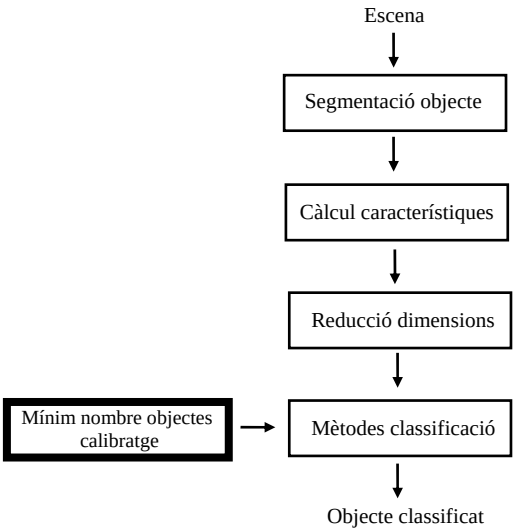
DISSENY D'UN PROTOCOL NUMÈRIC PER A LA CLASSIFICACIÓ INVARIANT D'IMATGES APLICANT TÈCNIQUES MULTIVARIANTS

Memòria presentada per optar al títol de doctor en Ciències Físiques

Directors:
Dr. Arturo Carnicer González
Dr. Ignacio Juvells Prades

Jordi-Roger Riba Ruíz

Barcelona, maig de 2000



4. Mínim nombre d'objectes de calibratge

En aquest capítol se suposa que les *m* característiques discriminants calculades per cada objecte han estat transformades en un nombre *m** més reduït de variables.

Un problema que cal afrontar quan es porta a terme un procés supervisat de classificació o reconeixement de patrons és com seleccionar adequadament el nombre d'objectes del conjunt de calibratge. D'una banda, si se n'escullen pocs, pot ser que les diferents classes no quedin prou definides (a llavors el procés de classificació pot no tenir suficients garanties d'èxit), mentre que si es treballa amb més objectes dels necessaris, augmenta la complexitat dels càlculs i el temps de computació. Per tant, cal arribar a un compromís entre agafar-ne massa o pocs.

El fet de trobar, en cada problema concret, el mínim nombre d'objectes de calibratge és un problema matemàticament molt complex, ja que depèn de molts factors, com ara:

- El nivell d'informació que aporta cada una de les característiques, és a dir, el grau d'adequació de les característiques per descriure els objectes en qüestió.
- El nombre de variables utilitzades per descriure els objectes (dimensionalitat del problema).
- Si els objectes es descriuen mitjançant característiques o variables.

- El grau de dispersió que presenten els objectes d'una mateixa classe.
- El fet de disposar, en el conjunt de calibratge, d'elements prou representatius de cada una de les classes definides en el problema de classificació analitzat.
- El tipus de distribució probabilística que segueix cada classe.
- El grau de separació o de d'encavalcament entre les diferents classes.

Hi ha molts autors que han abordat aquest problema des del punt de vista de la matemàtica estadística. Per a informació sobre aquest tema, vegeu [Kan71], [Fuk71], [Fol72], [Jai78], [Rau79], [Fuk89], [Rau91], [Koz91] i [Sim96].

En revisar la literatura especialitzada sobre aquest problema, un s'adona que molts dels autors es veuen obligats a fer les hipòtesis següents, les quals són necessàries per poder efectuar un tractament matemàtic del problema:

- La independència lineal de les diferents variables del problema. Normalment s'assumeix que les variables són linealment independents, fet que no té perquè complir-se a la pràctica.
- La funció de densitat de probabilitat que segueix cada conjunt. Se sol assumir una distribució normal multivariant, però en la pràctica això no té perquè ser així, sinó que cada conjunt pot tenir una funció densitat de probabilitat pròpia i força diferent de la gaussiana.
- Molts autors redueixen el seu estudi teòric a la separabilitat entre dues classes, mentre que en els problemes reals normalment hi ha més de dues classes.

Com que la majoria de problemes reals no tenen perquè complir les hipòtesis explicades anteriorment, això fa perdre potència als resultats teòrics i, per tant, s'intueix que és difícil poder conèixer *a priori* quin és el mínim nombre d'objectes de calibratge per poder abordar el problema amb garanties d'èxit.

4.1. Càlcul *a priori* del mínim nombre d'objectes de calibratge

L'objectiu d'aquest apartat és poder disposar d'unes eines (fórmules matemàtiques) que ens permetin calcular *a priori*, és a dir, amb uns coneixements mínims de les dades que tenim i pràcticament sense haver de fer cap estudi previ, el nombre mínim d'objectes de calibratge necessaris per poder afrontar amb garanties d'èxit un problema determinat de

classificació. En aquest apartat, se segueixen bàsicament els resultats aportats per Kanal i Chandrasekaran [Kan71], Foley [Fol72] i Jain i Chandrasekaran [Jai82].

Foley [Fol72] va estudiar el cas de dues classes i va suposar les dades normalment distribuïdes i linealment independents. Amb aquestes restriccions va arribar a les conclusions següents:

1. Com més alta sigui la relació entre el nombre n_c d'objectes per classe i el nombre m^* de variables, millors seran els resultats. Per tant, cal que el quocient n_c/m^* sigui gran. Aquesta idea és congruent amb la hipòtesi prèviament publicada per Kanal i Chandrasekaran [Kan71], els quals afirmen que si no es coneix res sobre l'estructura probabilística fonamental de les dades, cal agafar un quocient n_c/m^* gran.

2. Per poder assegurar que la taxa d'errors de classificació sigui raonablement propera a la taxa òptima d'error, cal que es compleixi que cada classe contingui com a mínim tants objectes de calibratge com el nombre de variables linealment independents multiplicat per tres:

$$mínim(n_c) \geq 3.m^*. \tag{4.1}$$

   A més, Foley assegura que aquests resultats es poden generalitzar al cas de variables que segueixin distribucions raonablement diferents a la gaussiana.

En canvi, segons Jain i Chandrasekaran [Jai82], el criteri de Foley és massa feble. Segons ells, la classe amb menor nombre d'objectes ha de tenir, com a mínim, tants objectes de calibratge com cinc vegades el nombre de variables, és a dir:

$$mínim(n_c) \geq 5.m^*. \tag{4.2}$$

Si s'agafa el mateix nombre d'objectes de calibratge per a cada classe, aquest criteri condueix a la fórmula següent:

$$n \geq 5.c.m^*, \tag{4.3}$$

on n és el nombre total d'objectes de calibratge i c és el nombre de classes.

Queda vist que no hi ha unanimitat entre autors diferents. A més, els criteris anteriors no tenen en compte dependències entre variables, distribucions diferents de la gaussiana, etc. La conseqüència de tot això és que es fa difícil poder donar una fórmula per determinar *a priori* el mínim nombre d'objectes de calibratge necessari per poder afrontar amb èxit un procés de classificació de patrons.

Tot i això explicat anteriorment, el criteri aportat per Jain i Chandrasekaran pot servir com a nivell guia o de referència. La majoria dels autors està d'acord que com major sigui el valor del quocient *n/m**, millors seran els resultats.

De fet, en alguns casos, a la pràctica podrien ser aproximadament certes les dues hipòtesis en les quals es basen la majoria dels autors (variables independents i distribució gaussiana), si es compleixen alhora les dues condicions següents:

1. No s'ha de treballar directament amb les característiques, sinó que és preferible treballar amb les variables (components principals, variables canòniques, etc. Vegeu el capítol 3), ja que aquests poden suposar-se com a variables linealment independents.
2. Si és possible disposar d'un conjunt prou gran d'objectes i si d'aquests se n'agafa aleatòriament un subconjunt, els valors de les variables dels objectes d'aquest subconjunt probablement seguiran una distribució no molt allunyada de la gaussiana.

Si les dues hipòtesis anteriors es compleixen, segurament serà vàlida la fórmula 4.3., aportada per Jain i Chandrasekaran.

4.2. Càlcul *a posteriori* del mínim nombre d'objectes de calibratge

En aquest apartat s'estudien dues tècniques *a posteriori*, la finalitat de les quals és determinar el mínim nombre d'objectes de calibratge necessari per poder afrontar amb garanties un procés de classificació.

4.2.1. Tècnica basada en la distància computada (DC).

Aquesta tècnica *a posteriori* va ser proposada per Lana i Fernández [Lan94] i aquí se n'utilitzarà una versió una mica modificada. Per aplicar-la és convenient poder disposar d'un nombre elevat d'objectes de calibratge.

La distància computada o DC es defineix com:

$$DC = \sum_{i=1}^{c} \sum_{j=1}^{m^*} (\bar{x}_{ij} - \bar{y}_{ij})^2 , \qquad (4.4)$$

on $\bar{x}_{ij}$ és la *j*-èssima component del valor mitjà corresponent a la classe *i*-èssima d'un subconjunt dels objectes de calibratge, mentre que $\bar{y}_{ij}$ correspon a la *j*-èssima component del valor mitjà corresponent a la classe *i*-èssima del total dels objectes de calibratge (suposem *c* classes i m^* variables per objecte).

L'operativa d'aquest mètode és la següent:

1. Es porten a terme, opcionalment (però és recomanable fer-ho si en el problema es té previst aplicar aquestes tècniques), les tècniques de reducció de dimensions descrites en el capítol 3 sobre la matriu que conté tots els objectes de calibratge disponibles i es calculen els valors mitjans $\bar{y}_i$ de cada classe. Els $\bar{y}_i$ representen els valors mitjans de les dades totals de calibratge, en l'espai definit o bé per les característiques, o bé per les variables.
2. Del conjunt total d'objectes de calibratge, se n'agafa aleatòriament un subconjunt amb un nombre reduït d'elements. Tot seguit s'opera idènticament al pas 1 anterior i es calculen els valors mitjans $\bar{x}_i$ de cada classe. Els $\bar{x}_i$ representen els valors mitjans del subconjunt de dades de calibratge utilitzat, en l'espai definit o bé per les característiques, o bé per les variables.
3. Es repeteix iterativament el pas 2, tot reduint progressivament el nombre d'objectes de calibratge.
4. Aplicant la fórmula 4.4. es calcula la distància computada entre el vector $\bar{y}_i$ i els vectors $\bar{x}_i$ obtinguts en els passos 2 i 3.
5. Es fa un gràfic de la distància computada contra el nombre d'objectes de calibratge de cada subconjunt. Si apareix un augment abrupte del pendent, el nombre mínim d'objectes de calibratge requerit és el que precedeix al canvi. Aquest gràfic i la interpretació del nombre mínim d'objectes de calibratge que s'ha de tenir en consideració es veuen en la figura 4.1.

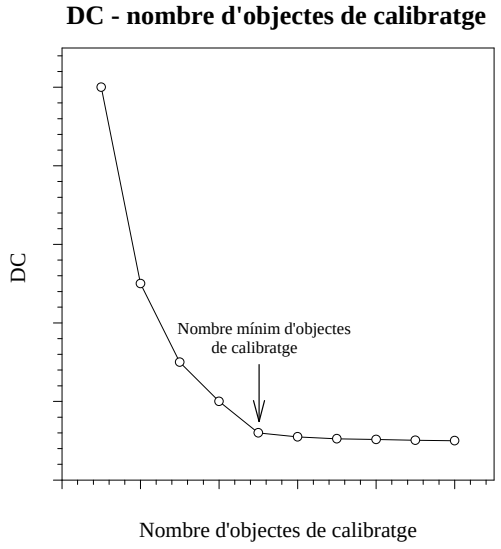


Figura 4.1.

4.2.2. Tècnica basada en l'error de classificació proporcionat per tècniques simples de classificació

Aquesta tècnica *a posteriori* és una proposta nova d'aquest treball. Per aplicar-la, és convenient poder disposar d'un nombre elevat d'objectes de calibratge.

L'operativa d'aquest mètode és la següent:

1. S'aplica el mètode més adient de reducció de dimensions dels explicats en el capítol 3 sobre la matriu que conté tots els objectes de calibratge disponibles. Després es classifiquen (utilitzant algun dels mètodes de classificació explicats en el capítol 5, especialment els dels apartats 5.2.1. i 5.2.4. per ser mètodes ràpids de càlcul i d'eficiència àmpliament contrastada) tots els objectes de calibratge disponibles, avaluant-se el nombre d'objectes classificats erròniament.
2. Del conjunt total d'objectes de calibratge, s'agafa aleatòriament un subconjunt amb un nombre reduït d'elements. Tot seguit es classifiquen tots els objectes de calibratge disponibles, avaluant-se el nombre d'objectes classificats erròniament.
3. Es repeteix iterativament el pas 2, tot reduint progressivament el nombre d'objectes de calibratge.
4. Es fa un gràfic de la taxa d'errors de classificació enfront del nombre d'objectes de calibratge de cada subconjunt. Si apareix un augment abrupte del pendent, el nombre mínim d'objectes de calibratge requerit és el que precedeix al canvi. Aquest gràfic i la interpretació del nombre mínim d'objectes de calibratge es veuen en la figura 4.2.

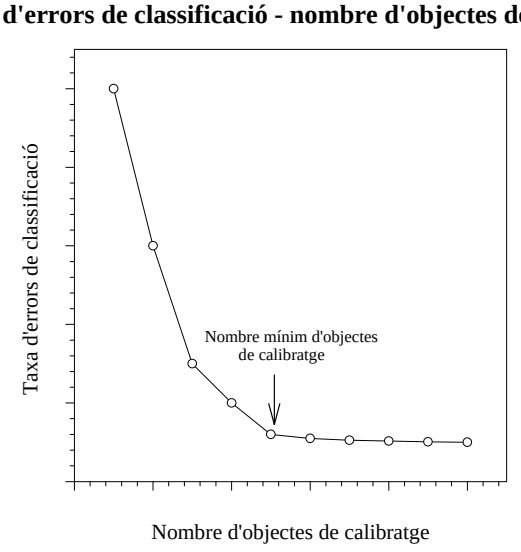


Figura 4.2.