Clasificarea automata folosind un model probabilistic

Conf.dr. Emil STAN
Academia de Politie "Alexandru Ioan Cuza" Bucuresti

O cerinta naturala a vehicularii volumelor mari de date, proprii sistemelor informatice, este sistematizarea (ordonarea si gruparea) acestora. Lucrarea propune o procedura iterativa de clasificare care este indicat sa se utilizeze atunci când numarul de valori pe care le iau criteriile de clasificare este relativ mic si prin urmare este posibil ca multe valori obtinute din masurarea obiectelor de clasificat sa se repete.
Cuvinte cheie: *clustering, probabilitate, similaritate, multimi cilindru.*

Introducere

Spre deosebire de clusteringul prin corelatie care construieste iterativ clustere compacte, tehnica prezentata aici construieste clustere din centrele multimilor, unul la fiecare moment de timp si, prin urmare, elimina multe din calculele implicate de recalcularea distantelor pe care alte metode de clasificare le cer. De asemenea, în loc de a lua în consideratie distributiile de probabilitate prin estimarea parametrica a matricei de covarianta ca în tehnica clasificarii prin proiectie [1], aceasta tehnica lucreaza cu estimari ale distributiei .

Procedura de clasificare se bazeaza pe urmatoarele axiome de lucru:

1. Unitatile cu aceleasi vector masura sau vectori masura similari apartin unei aceeasi categorie (clasa, cluster)
2. Când probabilitatea unui vector devine prea mica în comparatie cu vecinii sai, atunci vectorul masura sta pe frontiera între doua clustere .

Clustering-ul este efectuat prin definirea si ordonarea centrelor multimilor prin importanta. Se grupeaza toti vectorii masura care sunt suficient de similari fata de primul centru de multime si îl aditionam la multimea centrelor. Continuam clustering-ul prin gruparea împreuna a tuturor vectorilor masura care sunt suficient de similari fata de multimea centrelor, acum largita.

Procedura continua pâna când nu mai sunt vectori masura destul de similari fata de prima multime centru. Apoi, putem clasifica în jurul multimii centru alaturate nefolosite s.a.m.d.

Prezentarea modelului matematic

Fie S= (g1 , g2, g3,…,gm) sirul de date , Li= li1, li2,..,liNi, multimea de valori a criteriului Xi, i=1,2..,N. Spatiul masura G va fi L1×L2×…×LN, ζ va fi o metrica pe G si P distributia de probabilitate observata empiric. P este definit prin: $P(g) = \frac{card(S^{-1}(g))}{m}$.

Multimile centru posibile vor fi multimi singleton {g}, g∈G. Se considera G ordonata prin probabilitatile observate: multimea cu cea mai mare probabilitate la început si multimea cu cea mai joasa probabilitate la urma. Totusi, cu o astfel de ordine, unele multimi centru vor fi prea jos specificate în pachete mici, izolate avînd, relativ, probabilitate înalta comparata cu vecinatatile, dar probabilitati joase global. Astfel de multimi centru, de probabilitate relativ înalta local si joasa global, se vor ordona înaintea multimilor centru de probabilitate ridicata relativ-local si probabilitate înalta-global.

Prin urmare, ordonarea pentru elemente din spatiul masura G, obtinut din probabilitatea observata empiric P, trebuie bazata pe proprietati mai globale decât ale probabilitatii lui g. Astfel este posibil ca ordonarea sa fie bazata pe asocierea multimilor cilindru care caracterizeaza multimile singleton {g} pentru g din G . Multimile cilindru Eij,Eij ⊂ G sunt definite prin:

Eij = {∂ = (∂1, ∂2, …, ∂N) | ∂I = lij}, i=1,2..N; j=1,2..Ni.

Se observa ca pentru orice g∈G, exista sirul de întregi δ1(g),..,δN(g) definit prin : δi(g)=j

daca si numai daca g∈ Eij astfel încât $\{g\} = \bigcap_{i=1}^{N} E_i \delta_i(g)$.

Întrucât fiecare multime singleton {g} este echivalenta cu o intersectie adecvata a multimilor cilindru, iar multimile cilindru reflecta proprietati mai globale decât {g}, este rezonabil sa bazam ordonarea pe asociatia între multimile cilindrice adecvate .

Folosim coeficientul V ca masura a asociatiei. Pentru doua submultimi A si B ale lui G, V este definit prin:

$$V(A,B) = \frac{P(A\cap B)P(A^C\cap B^C) - P(A^C\cap B)P(A\cap B^C)}{[P(A)P(A^C)P(B)P(B^C)]^{1/2}}$$

Vom arata:

1. V(A,B) = V(B,A)
2. $V(A,B^C) = -V(A,B)$
3. daca $P(A\cap B) = P(A)P(B)$ atunci V(A,B)=0
4. V(A,B) = 1 daca si numai daca $P(A\cap B^C) = P(A^C\cap B) = 0$
5. V(A,B)= -1 daca si numai daca $P(A\cap B) = P(A^C\cap B^C) = 0$
6. $-1 \le V(A,B) \le 1$

Proprietatile (2) si (3) sunt obtinute prin substitutie

$$V(A,B^C) = \frac{P(A\cap B^C)P(A^C\cap B) - P(A^C\cap B^C)P(A\cap B)}{[P(A)P(A^C)P(B^C)P(B)]^{1/2}} =$$

$$= -\frac{P(A\cap B)P(A^C\cap B^C) - P(A^C\cap B)P(A\cap B^C)}{[P(A)P(A^C)P(B)P(B^C)]^{1/2}} = -V(A,B)$$

Presupunem $P(A\cap B) = P(A)P(B)$. Atunci,

$$V(A,B) = \frac{P(A)P(B)P(A^C)P(B^C) - P(A^C)P(B)P(A)P(B^C)}{[P(A)P(A^C)P(B)P(B^C)]^{1/2}} = 0$$

Pentru proprietatile (4), (5), (6) facem notatiile

a= $P(A\cap B)$

b= $P(A\cap B^C)$

c= $P(A^C\cap B)$

d= $P(A^C\cap B^C)$

Atunci: $V(A,B) = \frac{ad-bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}$

Prin ridicare la patrat :

$$\{V(A,B)^2[a^2(bc+cd+bd)+b^2(ac+ad+cd)+c^2(ab+ad+bd)+d^2 *$$

$$*(ac+ab+bc)] + 2(V(A,B)^2+1)abcd\} + \{V(A,B)^2-1)(a^2d^2+b^2c^2)\} = 0$$

Rezulta V(A,B)≤1 , adica proprietatea (6).

Pentru proprietatea (4) substituim

$V(A,B)^2 = 1$ *în relatia anterioara* :

$$a^2(bc+bd+cd)+b^2(ac+ad+cd)+c^2(ab+ad+bd)+d^2ac+bd+bc)+$$

$$+4abcd = 0$$

Cum toti termenii sunt pozitivi, trebuie sa avem toti termenii zero pentru ca suma sa fie zero. Deoarece A si B sunt submultimi proprii, rezulta ca a=d=0 sau c=b=0

Daca a=d=0, atunci V(A,B) = -1. Daca b=c=0 atunci V(A,B) =1

Daca $V^2(A,B) = 1$ *fie* $a = d = 0$ *fie* $b = c = 0$

Invers daca $P(A \cap B) = P(A^C \cap B^C) = 0$ atunci a=d=0 si V(A,B)= -1. De asemenea, daca $P(A^C \cap B) = P(A \cap B^C) = 0$, atunci c=b=0 si deci V(A,B) =1. Deci, V(A,B)=1 daca si numai daca $P(A^C \cap B) = P(A \cap B^C) = 0$ iar V(A,B)= -1 daca si numai daca $P(A \cap B) = P(A^C \cap B^C) = 0$.

Fie f functia care stabileste ordinea fiecarui g în ordonarea multimilor centru {g}. Noi definim f prin:

$$f(g) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} V(E_{id_i(g)}, E_{jd_j(g)}). Secvente\ B = \{g^{i1}, g^{i2}, .., g^{im} \mid g^{in} \in S(\{1,2,..,m\})\}$$ n=1,2,..,m

este construita astfel încât:

$f(g^{i_j}) \geq f(g^{i_n})$ pentru orice $j \leq n$.

Sunt trei parametri care guverneaza procesul de clustering: k numarul maxim de celule (clustere) în partitie, $\underline{\varepsilon}$ parametrul depasirii probabilitatii, $\underline{\alpha}$ parametrul depasirii distantei.

Procesul de clustering iterativ

Fie primul cluster H0 = ∅. Presupunem am definit Hn. Definim Hn+1, n<k.

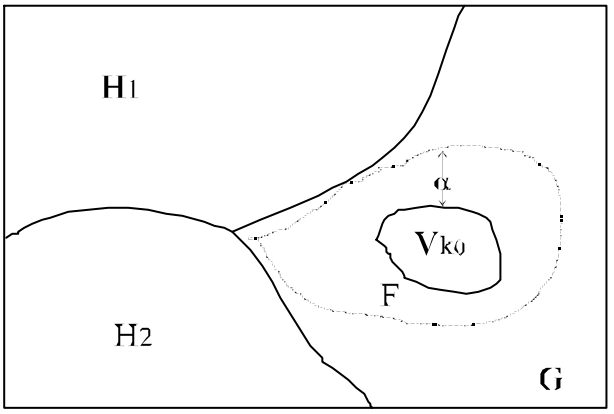
Fie $t_{n+1} = \min\{j \mid g^{i_j} \in S(\{1,2,..,m\}) - \bigcup_{k=0}^{n} H_k\}$

Astfel $i_{t_{n+1}}$ este cel mai mic indice al acelor g^{i_j} din secventa B care n-au fost înca inclusi în $\bigcup_{i=1}^{n} H_i$ si $\{g^{i_{t_{n+1}}}\}$ este cel mai important centru de multime nefolosit continut în G. Construim clusterul Hn+1 în jurul multimii centru $V = \{g^{i_{tn+1}}\}$. Construim pe Hn+1 succesiv, formînd V1,V2,..,VK0,.. pâna când nu mai este nimic care poate fi aditionat la multimea curenta.

Descriem cum Vi sunt construiti iterativ. Presupunem Vk0 a fost definit.

Fie $F = \{g \in S(\{1,2,..,m\}) - \bigcup_{k=1}^{n} H_k, 0 < dist(g,Vk0) \leq \alpha\}$, unde $dist(g,Vk0) = \min_{\bar{g} \in V_{k0}} (g, \bar{g})$

F este multimea tuturor masurilor din secventa S care n-au fost incluse în alte clustere $\bigcup_{k=0}^{n} H_k$ sau multimea prezenta Vk0 si care sunt la o distanta mai mare decât 0 si mai mica sau egala decât α.



Daca F=∅, atunci procedura clustering este terminata pentru clusterul (n+1) si definim Hn+1=Vk0.

Daca gasim $g \in F$ care este destul de similar lui Vk0, atunci aditionam acel g la Vk0.

"Destul de similar" este determinat prin urmatoarele conditii:

1) $P(g) \geq P(\bar{g}), \forall \bar{g} \in F$

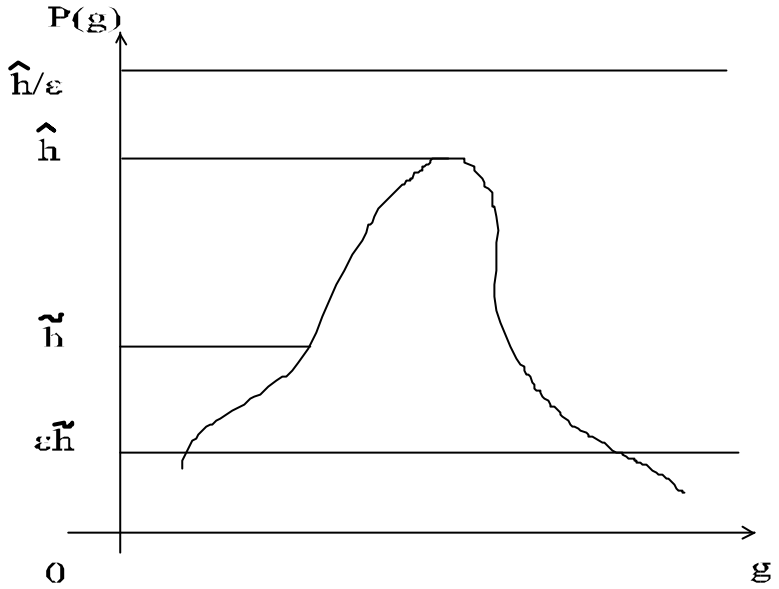
2) $P(g) \geq \varepsilon \min_{\bar{g} \in Vk0} P(\bar{g})$

3) $P(g) \leq \frac{1}{\varepsilon} \max_{\bar{g} \in Vk0} P(\bar{g})$

Daca exista un g∈F satisfacând aceste conditii, punem Vk0+1= Vk0 ∩{g); daca nu exista un astfel de g , atunci procedura de determinare a clusterului n+1, este terminata si Hn+1=Vk0.

Conditia (1) înseamna ca din masuratorile neclasificate înca, cu o distanta mai mica de ∝ fata de Vk0, ł vom considera numai pe cele cu probabilitatea cea mai mare . Conditiile (2) si (3) restrictioneaza atentia noastra numai la acele elemente care au probabilitati nu prea diferite fata de acelea considerate în cluster.

Diferentele extreme uzual indica o granita extrema pentru cluster.



$\hat{h} = \max_{g \in V_{k_0}} P(g)$

x elemente în V_{k_0}

$\tilde{h} = \min_{g \in V_{k_0}} P(g)$

Dar nu este sigur daca $\bigcup_{k=1}^{k} H_k = G$. Pentru aceasta $\forall g \in G - \bigcup_{k=1}^{k} H_k$, fie r(g) cel mai mic indice astfel încât:

$$dist(g,Hr(g)) \le dist(g,Hk), \forall k \in K.$$

Dupa ce fiecarui astfel de g i s-a atasat un astfel de indice, ei se includ în submultimea Hr. Am obtinut astfel o partitie a lui G cu k celule.

Concluzii

În ultima vreme aplicatiile practice din ce în ce mai numeroase ale clasificarii automate (întâlnita uneori si sub denumirea de **cluster analysis**) au condus la tipuri particulare de date statistice care caracterizeaza multimea vectorilor de clasificat. Functiile de clustering existente în software-ul statistic desigur ca au un grad de generalitate foarte mare si nu surprind forma particulara a modelului asigurat de baza de date. Procedura care a fost prezentata poate fi utilizata cu succes în cazul unor multimi mari de obiecte de clasificat, analizate prin prisma mai multor criterii cu un numar finit de valori. Astfel, pentru un anumit criteriu o valoare specificata poate fi întâlnita relativ frecvent.

Bibliografie

1. M.R.Anderberg,"Cluster analysis for Applications", Academic Press, New York, 1988

2. Emil Stan,"Clasificarea prin proiectie", *Informatica Economica,***4,** nr.3(15), 2000

3. Emil Stan,"Modelarea dinamica a sistemelor", Editura RDA, Bucuresti,1998