УДК 681.3.06:621.391

ПРОГРАММЫ ТАКСОНОМИИ (ГРУППИРОВКИ) ОБЪЕКТОВ

Ю.Д.Григорьев, В.И.Котюков

Программа автоматической группировки объектов на классы написаны на " α - языке" (для ЦВМ "М-220" и "БЭСМ-4") и на АЛГОЛе (для ЦВМ "БЭСМ-6").

I. Постановка задачи оптимальной группировки. Считается заданным числовое описание в p-мерной системе признаков $X=\{x_1,\ldots,x_p\}$ исходных N объектов. Необходимо разбить N объектов на K непересекающихся таксонов (подмножеств, групп) так, чтобы минимизировать при этом критерий

$$D=\sum_{l=1}^{K}\sum_{i=1}^{N_l}\rho_{i0}(l),$$

где N_l - число объектов в l -м таксоне, а $\rho_{i0}(l)$ - расстояние i -го объекта l -го таксона до объекта l -го таксона, наиболее близкого к "центру тяжести" этого таксона. Заметим, что минимизация усредненного внутригруппового разброса D эквивалентна максимизации усредненного межгруппового разброса. Величина, обратная среднему внутригрупповому разбросу значений X_ν, будет определять информативность признака X_ν.

Программы включают в себя предварительную нормировку исходных данных по каждому признаку X_ν в соответствии с его среднеквадратичным отклонением

$$x'_{i\nu}=x_{i\nu}/\sigma_\nu.$$

Группировку можно осуществлять и в любом подпространстве $X^* \subseteq X$.

Теоретические основы метода изложены в статье В.И.Котикова "О некоторых задачах таксономии объектов" в сборнике "Вычислительные системы", Новосибирск, 1972, вып. 50.

2. И н с т р у к ц и я к п о л ь з о в а н и ю п р о г р а м м о й . Введем следующие обозначения основных параметров:

N – число объектов,

p – число признаков,

K1 – минимальное число групп ($K1 \geq 1$),

K2 – максимальное число групп ($K2 \leq N$),

ω – целое число, заключенное в пределах 1–30 (чем больше ω, тем больше вероятность того, что достигнутый локальный минимум критерия D совпадает с глобальным, но при этом и больше машинное время решения задачи),

массив $\lambda[1:p]$ указывает на то, нужно или нет при группировке учитывать тот или иной признак (если $\lambda[\nu] \geq 0{,}6$, то X_ν учитывается, если $\lambda[\nu] \leq 0{,}4$, то нет),

массив x[1:N, 1:p] – основной исходный массив чисел.

В программе на АЛГОЛе вместо символа ω используется K3, а вместо λ[1:p] – LL[1:p].

Программа последовательно выполняет оптимальную таксономию на число групп K, начиная от K1 и кончая K2.

В программе (строки отмечены звездочкой) для каждого из массивов перфорируются его истинные числовые границы, то есть используются статические массивы.

Суммарная величина указанных массивов не должна превышать объема оперативной памяти машины, при этом для α – программы (см.стр. 74) определяющую роль играют массивы x[1:N, 1:p] и r[1:N, 1:N], а для программы на АЛГОЛе (см. стр. 75) – массив x[1:N, 1:p].

Сама программа занимает не более 500 ячеек памяти. Программы не используют память магнитных лент и барабанов.

Время трансляции на "М-220" около 10 минут, а на "БЭСМ-6" менее 10 секунд.

Время решения определяется в основном величинами N и ω.

В в о д . Исходная информация вводится в следующей последовательности:

N, p, K1, K2, ω, λ[1:p], x[1:N; 1:p]

В ы в о д .

1. Сначала выдаются дисперсии значений каждого признака X_ν.

2. Затем для каждого K (числа групп) выдается:

а) величина K;

б) величина критерия D для получившегося оптимального разбиения N объектов на K групп;

в) служебное число;

г) для каждой группы, начиная с 1-й, указывается номер исходного объекта, который является "эталоном" данной группы ("центральный" объект);

д) для каждого исходного объекта, начиная с 1-го, указывается номер группы, в которую он попал;

е) для признаков $\{X_\nu\}$ выдаются величины, обратные их информативности.

Контрольный пример.

1) N = 8; 2) p = 2; 3) K1 = 2; 4) K2 = 3; 5) ω = 3; 6) λ[1] = 1; λ[2] = 1; 7) массив x[1:8, 1:2] : 2,0; 2,0; 3,0; 2,0; 3,0; 1,0; 5,0; 7,0; 6,0; 7,0; 5,0; 6,0; 7,0; 5,0; 8,0; 6,0.

В ы д а ч а :

1) дисперсии: 3, 9; 5, 3;
2) величина K:2;
3) величина D: 2,62;
4) служебный параметр (любое число);
5) эталоны: 5; 2; 0;
6) разбиение по группам: 2; 2; 2; 1; 1; 1; 1; 1;
7) величины, обратные информативности признаков: 3,1; 2,2;
8) пропуск;
9) величина K: 3;
10) величина D: 1,42;
11) служебный параметр (любое число);
12) эталоны: 7; 2; 4;
13) разбиение по группам: 2; 2; 2; 3; 3; 3; 1; 1;
14) величины, обратные информативности признаков: 1,53; 1,31.

Программа на " α-языке"

```
Начало целый  N , p , K1, K2, k , i  j , l , δ , ω;
*   веществ f0 , f1 , f2 , μ; целый массив Э0, Э [I :K2],
*  to , t[ I :N];  массив n[ I :N], λ , σ , ε[I:p],
*  x[ I:N, I:p], r[ I:N, I:N];
   процедура выбор; начало целый  α , β , γ , q;
   веществ    A,B; Э[  ]:=0; β:=0;
   A1:  B:=RAND; γ:=B x (N-I ) + I; A:=0;
   для  α:=I,...,k  цикл если  γ = Э[α]  то A:= I;
   если  A< 0.5 то {β:=β + I; Э [β]:=γ};
   если  β < k  то на A1; конец;
   процедура группировка; начало  целый     α , β , γ , q;
   веществ  A,B;
   для  α:= I,...,N цикл  {γ:= I; β:= Э[I ]; A:=r[α,β];
   для  q:= I,...,k цикл  {β:= Э[q]; если  r[α,β]< A то
    {γ:=q; A:=r[α,β]}}; t[α]:=γ}; f2:=0;  Э[  ]:=0;
   для  β:= I,...,k  цикл { n[ ]:= I0000;
   для  α:= I,...,N цикл если  t[α]=β  то {n[α]:=0;
   для  γ:= I,...,N цикл если  t[γ]=β  то  n[α]:=
    n[α] + r[α,γ]};  γ:= I;  A:=n[I ];
   для  α:= 2,...,N цикл если  n[α] < A то{γ:=α; A:=n[α]};
   Э[β]:=γ;
   для  α:= I,...,N цикл если  t[α]=β  то
    f2:=f2 + r[α,γ]};  конец;
   ввод  (N,p,KI,K2,ω, λ,x);
    σ[ ]:=0; для i:= I,...,p  цикл {μ:=0;
   для  j:= I,...,N цикл  μ:=μ + x[j,i];  μ:=μ/N;
   для  j:=I,..., N цикл  σ[i]:=σ[i] + (μ - x[j,i])↑2;
    σ[i]:=σ[i]/N};  вывод (σ, истина );
   для  i:=I,..., p цикл  σ[i]:= sqrt (σ[i]);  r[ ,]:=0;
   для  i:=I,..., (N- I) цикл
   для  j:=(i+I )..., N  цикл  {δ:=0;
   для  l:=I,..., p цикл  r[i,j]:=r[i,j] + (λ [l] x
    (abs(x[i,l]-x[j,l])))/σ[l];  r[i,j]:=r[i j]/p;
    r[j,i]:=r[i,j]}; вывод (δ, истина);
    для  k:=KI,..., K2 цикл  {δ:=k x ω;  f0:= I00000;
   вывод  (k , истина );
```



```
   для  i:= I,...,  δ цикл  {f1:= I00 000; выбор;
   B:   l:=0; группировка; если  abs(f1-f2)> 0.0000I
   то  {f1:=f2;  l:= I}; если  l> 0.5 то на B;
   если  f1< f0 то  {f0:=f1;  j:=i;  Э0[ ]:= Э[ ];
   to[ ]:=t[ ]}}; вывод  (f0, истина ); вывод  (j, истина);
   вывод  ( Э0, истина ); вывод  (to, истина );
   ε[ ]:=0;  для l:= I,...,  p цикл
   для  i:= I, ... , k  цикл {δ:= Э0[i];
   для  j:= I, ... , N цикл  если to[j]=i  то
   ε[l]:=ε[l] + (abs(x[δ,l]-x[j,l]))/σ[l]};
   вывод  (ε, истина );  δ:=0;
   для  i:= I,...,4 цикл  вывод (δ, истина ));
конец *
```

Программа на АЛГОЛе

```
   Beqin integer N,P,KI,K2,K3;:input (N,P,KI,K2,K3 );
   beqin integer k,i,j,L,D,z,QQ,PP;
   real FF,F1,F2,M, U1,U2 ;  integer array EQ ,
*  E [ I:K2 ],TQ,T[ I :N];  array NN [ I :N] , LL,S,
*  EE [I :P] , X[I :N , I :P];
   real procedure  RAND ;  beqin real U;
   U:=UI + U2 ;  UI :=U2 ;  if U > 4 then U:=U-4 ;
   U2 :=U;  RAND:=U/4  end;
   real procedure R(i,j); integer i,j;
   beqin integer k; real  RR;  RR:=0;
   for k:= I step I  until P do  RR:=RR+(LL[k] x (abs(
   X[i,k]-X[j,k])));  RR:=RR/P;  R:=RR end;
   procedure  выбор ;  beqin integer AL BE GA Q;
   real  A B;
   for k:=1 step I until K2  do  E[i]:=0; BE:=0;
   AI:   B:=RAND;  GA:=entier (B x (N-I )+ I );  A:=0;
   for AL:=I  step I until k do if GA=E[AL] then A:=I ;
   if A < 0,5 then beqin  BE:=BE +I ;  E[BE]:=GA end;
   if  BE < k  then goto  AI ;  end;
   procedure  группировка ; beqin integer AL,BE,
   GA,Q; real A,B;
   for AL:=I  step I until N do beqin GA:=I ;  BE:=E[I ];
```

```
A:=R(AL BE); for Q:= I step I until k do
begin BE:=E[Q]; if R(AL,BE) < A then
begin GA:=Q; A:=R(AL,BE) end end; T[AL]:=GA
end; F2:=0; for i:= I step I until K 2 do E[i]:=0;
for BE:= I step I until k do begin
for i:= I step I until N do NN[i]:=10 4;
for AL:= I step I until N do if T[AL]=BE then
begin NN[AL]:=0; for GA:= I step I until N do
if T[GA]=BE then NN[AL]:=NN[AL] + R(AL,GA) end;
GA:= I; A:=NN[ I];
for AL:= 2 step I until N do if NN[AL] < A then
begin GA:=AL; A:=NN[AL] end; E[BE]:=GA;
for AL:= I step I until N do if T[AL]=BE then
F2 :=F2 + R(AL,GA) end end;
input (LL);
output ('T', ' признаки', 'E', LL, '/');
input (X);
output ('T', ' исходные данные', 'E', X, '/');
UI := 3.I4I59269 ; U2 := 0.542IOI887;
for i:= I step I until P do S[i]:=0;
for i:= I step I until P do begin M:=0;
for j:= I step I until N do M:=M + X[j,i]; M:=M/N;
for j:= I step I until N do S[i]:=S[i] + (M-X[j,i])12;
S[i]:=S[i]/N end;
output ('T', 'дисперсии','E',S,'/');
for i:= I step I until P do S[i]:=sqrt(S[i]);
for i:= I step I until P do
for j:= I step I until N do X[j,i]:=X[j,i]/S[i];
for k:=KI step I until K 2 do begin D:=k x K 3; FF:=10 5;
output('T', ' число таксонов', 'E', K, '/');
output('T', ' число группирований', 'E', D, '/');
for QQ:= I step I until D do begin F I:=10 5; выбор;
B I; L:=0; группировка ; if abs(F I-F2 )>10- 4 then
begin FI :=F 2; L:= I end; if L>.5 then doto BI ;
if FI< FF then begin FF:=FI ; PP:=QQ;
for z:= I step I until k do EQ [z]:=E[z];
for z:=k+I step I until K2 do EQ[z]:=0;
for z:= I step I until N do TQ[z]:=T[z] end end;
```



```
output('T', ' показатель качества', 'E', FF);
output('T', 'номер оптимального группирования', 'E', PP, '/');
output('T', 'эталонные объекты', 'E', EQ, '/');
output('T', 'разбиение объектов по таксонам', 'E', TQ, '/');
for i:=I step I until P do EE[i]:=0;
for L:= I step I until P do for i:= I step I until k do
begin D:=EQ[i]; for j:= I step I until N do
if TQ[j]=i then EE[l]:=EE[l]+(abs(X[D,l]-X[j,l])) end;
output('T', 'информативность признаков', 'E', EE, '///');
end end end
```

Поступила в ред.-изд.отд.
24 ноября 1972 года