3. Хиерархиски класификации

3.1 Префинетост на поделбата на една класификација

Барања што се поставуваат при испитување на една статистичка појава:

- Формирање структура на множеството *S* во однос на една погруба класификација K_g заради стекнување на прва основна информација (макро-информација) за структурата на тоа множество;
- Формирање структура на множеството *S* во однос на една пофина класификација K_f заради стекнување не само покомплетна информација за структурата на множеството *S* туку и информација за структурата на неговите поодделни делови (микро-информација);

Секоја класа K_f е целосно содржана во една и само една класа K_g.

$\langle K_f, K_g \rangle$ претставува <u>хиерархија</u>, при што хиерархискиот ранг од K_g е повисок од хиерархискиот ранг на K_f.

k / N може да се користи како <u>мерка за префинетост на поделбата</u> (*k* – број на класи; *N* – број на елементи на множеството *S*)

3.2 Подредување на класификациите според префинетоста на поделбата

Нека е *K* релација на еквиваленција која ја дефинира класификацијата *K* од посматраното множество *S* и нека е $K(S)$ множество од сите можни класификации од тоа множество.

$G(K)$ претставува графот на релацијата на еквиваленција на класиф. *K*:

$$G(K) = \{\langle x, y \rangle | x \in S \wedge y \in S \wedge xKy\}$$

Инклузијата во $S \times S$ овозможува да се дефинира еден поредок во $K(S)$

Ако важи $k \le k'$, тогаш $G(K) \subset G(K')$ односно

$$\forall x, y \,[x \in S \wedge y \in S \wedge xKy \Rightarrow xK'y]$$

Ако $k \le k'$ тогаш велиме дека класификацијата *K* е попрефинета од K'.

За секој пар класификации $\{K, K'\}$ одговара т.н. <u>минорант</u> $K \wedge K'$ и т.н. <u>мајорант</u> $K \vee K'$ коишто се дефинирани на следниов начин:

$$G(K \wedge K') = G(K) \cap G(K')$$

$$G(K \vee K') = G(K) \cup G(K')$$

3.3 Хиерархиска класификација

Ако од секој елемент од посматраното множество $S = \{e_1, e_2, \ldots e_N\}$ образуваме посебна класа, ќе добиеме <u>класификација од нулто ниво</u>:

$$K_{(0)} = \langle A_{01}, A_{02}, \ldots A_{0N} \rangle \qquad \text{каде } A_{0i} = \{e_i\} \qquad i \in \{1, \ldots N\}$$

Ако N класи од $K_{(0)}$ ги групираме во k_1 нови класи:

$$A_{11} = A_{01} \cup A_{02} \cup \ldots \cup A_{0t_1}$$

$$A_{12} = A_{0t_1+1} \cup A_{0t_1+2} \cup \ldots \cup A_{0t_2}$$

$$\vdots$$

$$A_{1k_1} = A_{0t_{k_1-1}+1} \cup A_{0t_{k_1-1}+2} \cup \ldots \cup A_{0t_{k_1}}$$

ќе добиеме <u>класификација од прво ниво</u> $K_{(1)} = \left\langle A_{11}, A_{12}, \ldots A_{1k_1} \right\rangle$.

Со групирање на овие k_1 класи во k_2 нови класи $(k_2 < k_1)$:

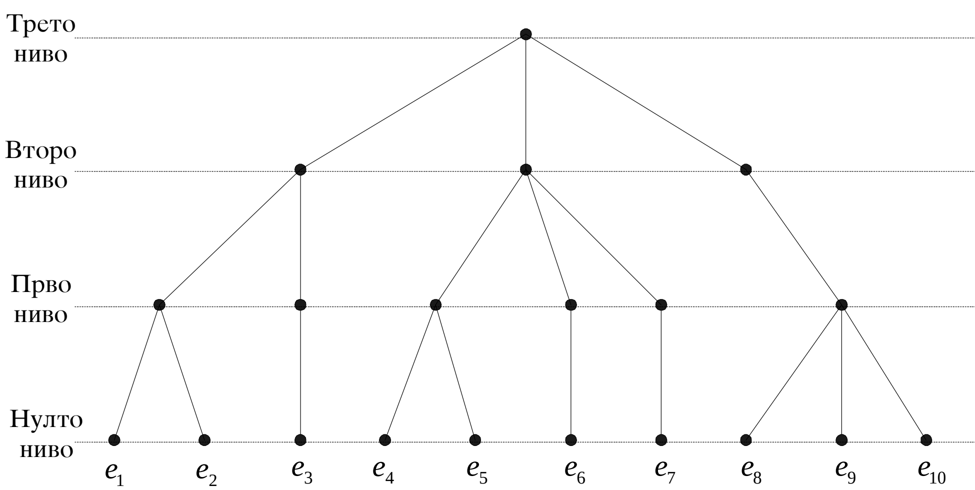
$$A_{2i} = A_{1t_{i-1}+1} \cup A_{1t_{i-1}+2} \cup \ldots \cup A_{1t_i} \qquad i \in \{1, \ldots k_1\}$$

ќе добиеме <u>класификација од второ ниво</u> $K_{(2)} = \left\langle A_{21}, A_{22}, \ldots A_{2k_2} \right\rangle$.

Ваквото групирање може да се врши сѐ дотогаш додека не се добие класификација што содржи само една класа составена од сите елементи на множеството S, т.е. $K_{(p)} = S$

Низата $K_{(0)}, K_{(1)}, \ldots K_{(p)}$ претставува <u>ланец на класификации</u>. Неговата прва алка е подредено множество од сите единечни делови од S, а последната е множество од една класа составена од сите подредени елементи на множеството S. Во првиот случај велиме дека поделбата е <u>дискретна</u> а во вториот случај е <u>груба</u>.

Пример:



Дефиниција: Ако сите елементи од една класа на било кое ниво се наоѓаат секогаш во исти класи од повисоките нивоа, за таквиот ланец на класификации велиме дека претставува <u>хиерархиска класификација</u>. Истата ја означуваме:

$$\Re(K) = \{K_{(0)}, K_{(1)}, \ldots K_{(p)}\} \quad \text{каде} \quad p \leq N - 1$$

p се нарекува <u>должина на ланецот</u> на хиерархиската класификација $\Re(K)$.

$$\forall r \left[r \in \{1, \ldots p-1\} \Rightarrow K_{(r)} \neq K_{(r+1)} \right]$$

Должината на ланецот не може да биде поголема од $N-1$.

3.4 Класификација од *r*-та поделба

Начини на кои може да се врши хиерархиската класификација:

- Сукцесивно групирање (агрегација);
- Сукцесивно разложување (дезагрегација);

Со поделбата на секоја класа од $r-1$-та поделба се добива класификација од *r*-та поделба. Ваквата поделба може да се продолжи се́ додека не добиеме единечна поделба на множеството *S* т.е. поделба чија секоја класа содржи само еден елемент од множеството *S*.

$$K_r = K_{(p-r)} \qquad r \in \{0,1,\ldots p\}$$

каде *p* е должина на ланецот на хиерархиската класификација $\Re(K)$.

$\Rightarrow$ Класификацијата од *r*-та поделба идентична е со класификацијата од (*p*-*r*)-то ниво

Точката *S* од нултото ниво се нарекува <u>корен</u> а точките од последната поделба се нарекуваат <u>терминали</u> на графикот на хиерархиската класификација. Останатите точки од грфикот се нарекуваат <u>јазли</u>, а отсечките што спојуваат две точки од графикот се нарекуваат <u>гранки</u>.

3.5 Структура на хиерархиската класификација

Ако над елементите од множеството *S* се мери обележјето *X* и ако $\Re(K)$ е една хиерархиска класификација на елементите од множеството *S* тогаш може да се одреди структурата од *X* за множеството *S* во однос на класификацијата $\Re(K)$.

x_{0i} - вредности на обележјето *X* кај елементот e_i, $i \in \{1,\ldots N\}$

X – збир на сите вредности од обележјето за сите елементи од *S*

$$X = \sum_{i=1}^{N} x_{i0}$$

Ако со x_{ri} го означиме збирот на вредностите од *X* од сите елементи од класата A_{ri}, структурата на обележјето *X* за множеството *S* во однос на класификацијата $\Re(K)$ ќе биде:

$$\{X, S, K_{(r)}\} = \begin{array}{|c|c|c|c|c|} \hline A_{r1} & A_{r2} & A_{r3} & \ldots & A_{rk_r} \\ \hline x_{r1} & x_{r2} & x_{r3} & \ldots & x_{rk_r} \\ \hline \end{array}$$

Или во облик на вектор: $\{X, S, K_{(r)}\} = \left\langle x_{r1}, x_{r2}, x_{r3}, \ldots x_{rk_r} \right\rangle$

Вкупно има *p*+1 такви структури кои шематски можеме да ги прикажеме на следниов начин:

$$\{X,S,\Re(K)\}=\left|\begin{array}{l} x_{p1} \\ x_{p-1,1},x_{p-1,2},\dots x_{p-1,k_{p-1}} \\ \vdots \\ x_{11},x_{12},\dots x_{1k_1} \\ x_{01},x_{02},\dots x_{0N} \end{array}\right.$$

Збирот на вредностите на елементите од сите класи на било кое ниво е секогаш ист, т.е.

$$\sum_{i=1}^{N} x_{0i} = \sum_{i=1}^{k_1} x_{1i} = \dots = \sum_{i=1}^{k_{p-1}} x_{p-1,i} = x_{p_1} = X$$

3.6 Претставување на повеќедимезионална класификација во вид на хиерархиска класификација

Да претпоставиме дека имаме *n* класификации од множеството *S*

$$K_1 = \langle A_{11}, A_{12}, \dots A_{1k_1} \rangle$$
$$K_2 = \langle A_{21}, A_{22}, \dots A_{2k_2} \rangle$$
...
$$K_n = \langle A_{n1}, A_{n2}, \dots A_{nk_n} \rangle$$

Ако елементите од секоја класа од K_1 ги класираме според класификацијата K_2, ќе добиеме дводимензионална класификација $K_1 \times K_2$. Ако потоа елементите од секоја класа од $K_1 \times K_2$ ги класираме според класификацијата K_3, ќе добиеме тродимензионална класификација $K_1 \times K_2 \times K_3$. Воопшто, со развивање на елементите од секоја класа од $r-1$-димензионална класификација ќе добиеме *r*-димензионална класификација.

$$B_{i_1} = A_{1i_1}$$

$$B_{i_1,i_2,\dots i_r} = A_{1i_1} \cap A_{2i_2} \cap \dots \cap A_{ri_r}$$

каде $j \in \{1,\dots r\}$ $\quad i_j \in \{1,\dots k_j\}$ $\quad r \in \{1,\dots n\}$

$$\forall j, i_j \left[j \in \{1,\dots r\}, i_j \in \{1,\dots k_j\} \Rightarrow B_{i_1,i_2,\dots i_{r-1},i_r} \subseteq B_{i_1,i_2,\dots i_{r-1}} \right]$$

Низата од класификации $\{K_1, K_1 \times K_2, \dots K_1 \times K_2 \times \dots \times K_n\}$ претставува една хиерархиска класификација.

Ова претставува <u>специјален случај</u> на хиерархиска класификација. Со оглед на тоа што елементите од секоја класа на *r*-тата поделба се класираат според една иста класификација K_{r+1}, следува дека на секоја класа од *r*-тата поделба одговараат K_{r+1} поткласи од $r+1$-та поделба.

Оваа особина во општ случај ја нема една хиерархиска класификација.

3.7 Линеаризација на хиерархиската класификација

Нека се $B_{i_1,i_2,\ldots i_r}$, $r \in \{1,\ldots n\}$ класи од r-тата поделба од една хиерархиска класификација.

Ако ги означиме со: $B_{i_1,i_2,\ldots i_r} = B_{\underbrace{i_1,i_2,\ldots i_r,0,\ldots 0}_{n}}$

Секоја класа од секоја поделба ќе има n индекси при што секој индекс е едноцифрен број.

Ако ги подредиме класите од $\Re(K)$ според големината на соодветните вредности од r, ќе добиеме <u>линеаризирана хиерархиска класификација</u> од множеството S.

Може да се забележи дека линеаризираната хиерархиска класификација не претставува една класификација бидејќи исти елементи можат да се наоѓаат во различни класи. Сепак, во пракса ваквиот начин на приккажување често се употребува бидејќи на тој начин се избегнува употребата на дводимензионална шема на хиерархиската класификација, чиј формат може да биде многу голем.

Ако над елементите од множеството S го мериме обележјето X и ако со x_{ijk} го означиме збирот на вредностите од X на сите елементи кои припаѓаат на класата B_{ijk} , горната линеаризирана хиерархиска класификација ќе претставува <u>линеаризирана хиерархиска структура</u>.