ROZHODOVÁNÍ S PODPOROU DISKRIMINAÈNÍ ANALÝZY

Bohdan Linda

Ústav matematiky, FES, Univerzita Pardubice

Abstract:
The discrimination analysis is a multivariable statistical method with the aim to find the optimal assignment regulation and prediction, into which of k groups a watched element belongs to. Several characteristics are measured at every element. They express its features. It means, that this element is characterised by means of a random variable $\boldsymbol{X}=(X_1, X_2, \ldots, X_m)$. *The final result of this procedure is a model, that allows to predict the pertinence of the watched element to the certain group on the basis of measured values* $(x_1, x_2, \ldots, x_m)$.

Ve veøejné správì èasto potøebujeme seskupit nìjaké objekty, což mohou být napøíklad obce, okresy regiony apod., podle nìkterých znakù, které vzájemnì mohou být nesoumìøitelné, ale spoluvytváøí nìjakou vlastnost, kterou nelze pøímo mìøit – napøíklad ekonomickou úroveò regionu, výkonnost a další. Podobnì tomu je i vekonomické praxi, kde mnohdy potøebujeme jednotlivé podniky rozdìlit do nìkolika skupin podle pøímo nemìøitelného znaku. A právì zde mùže diskriminaèní analýza usnadnit èlenìní, napomoci k pøijetí správného rozhodnutí pøi zaøazení sledovaného regionu, podniku èi jiného sledovaného objektu.

Objekty, které mají nìkteré podobné vlastnosti, mùžeme z dùvodu lepšího sledování jejich vlastností sluèovat do skupin. Nebývá ale vždy možné jednoduchým zpùsobem a jednoznaènì o nìjakém prvku rozhodnout, do které skupiny patøí. Jednou z metod, která mùže v rozhodování pomoci je diskriminaèní analýza. Hlavním cílem diskriminaèní analýzy je nalézt statisticky nejvýhodnìjší zpùsob rozlišení skupin prvkù a pøedpovìdìt, do které z k skupin patøí sledovaný prvek. U každého prvku je mìøeno nìkolik znakù, vyjadøujících jeho vlastnosti, to znamená, že tento prvek je charakterizován prostøednictvím náhodné velièiny $\boldsymbol{X}=(X_1, X_2, \ldots, X_m)$. Postup zaèíná analýzou vlastností prvkù, u nichž je známá jak pøíslušnost ke skupinì, tak i hodnoty pøíslušné náhodné velièiny. Koneèným výsledkem postupu je model, který umožòuje pøedpovìdìt pøíslušnost sledovaného prvku k urèité skupinì na základì namìøených hodnot $(x_1, x_2, \ldots, x_m)$.

Diskriminaèní analýza se s úspìchem aplikuje v øadì oborù. Jedno z prvních použití bylo v archeologii. Pøi nálezech hrobù s kosterními pozùstatky jsou nalézány také kultovní pøedmìty. Na základì urèitých charakteristických vlastností lze pak nález pøiøadit k urèitému období, kultuøe, rase. Ve školské i v personální podnikové praxi se v pøijímacím øízení uplatòují soubory testù, jejichž výsledkem je bodové hodnocení uchazeèe. Po uplynutí urèitého období lze získat informaci, jak úspìšný èi neúspìšný je sledovaný jedinec. Disponujeme tedy informativním výbìrem a pokud volba testù je taková, že jejich výsledky souvisí s pozorovanou úspìšností, lze pøistoupit ke klasifikaci. To znamená, že se snažíme provést s malým rizikem omylu pøedpovìï úspìchu podle výsledku testù. Nejjednodušším pøípadem diskriminaèní analýzy je pøedpovìï pøíslušnosti k dichotomicky èlenìné množinì prvkù založené na jednorozmìrných promìnných.

Jedním ze zásadních problémù metody je pøesnost pøedpovìdi. Je-li kritériem zaøazení prvku do skupiny jen jednoduchá promìnná, pak postup diskriminaèní analýzy umožòuje rozlišit zaøazení prvku do skupiny nìkdy témìø dokonale, nìkdy èásteènì a nìkdy to

neumožòuje vùbec. Vìtší rozdíl mezi støedními hodnotami sledované velièiny ve skupinách umožòuje lepší rozlišení skupin. V jednoduchém pøípadu dichotomických skupin zpravidla není problémem rùzná variabilita ve skupinách. V pøípadì vìtšího poètu skupin musíme ale pøedpokládat rovnost rozptylù ve skupinách.

Obecný princip metody

Úkolem diskriminaèní analýzy je nalezení optimálního pøiøazovacího pravidla, tzn. pravidla, které minimalizuje pravdìpodobnost chybné klasifikace, tedy minimalizuje støední hodnotu chyby rozhodnutí. Mùže se totiž stát, že prvek, který skuteènì pochází z urèité skupiny zaøadíme do jiné skupiny.

Pøedpokládejme, že je dáno k skupin prvkù (k = 2, 3, …) a každý z prvkù je charakterizován pomocí náhodné velièiny $\boldsymbol{X} = (X_1, X_2, X_3, \ldots, X_m)$ se známým typem rozdìlení pravdìpodobností. To znamená, že prvky patøící do i-té skupiny mùžeme považovat za náhodný výbìr ze základního souboru s rozdìlením $f_i(\boldsymbol{x})$ se støední hodnotou $\boldsymbol{m_i}$ a variaèní maticí $\mathbf{S_i}$.

Cílem diskriminaèní analýzy je zjistit, do které z k skupin patøí sledovaný prvek.

Pøedpokládejme, že náhodná velièina $\boldsymbol{X}$ nabývá hodnot z R_m (m-rozmìrného reálného vektorového prostoru). Utvoøíme rozklad prostoru R_m na k množin $M_1, \ldots, M_k$ (tj. pro množiny M_i musí platit: $R_m = \bigcup_{i=1}^{k} M_i$; $M_i \cap M_j = \varnothing$; $i \neq j$). Padne-li hodnota velièiny $\boldsymbol{X}$ do množiny M_i, tvrdíme, že sledovaný prvek náleží do i-té skupiny. Problémem je nalézt takový rozklad, aby rozhodnutí o pøíslušnosti k dané tøídì bylo optimální. Abychom mohli rozhodnout o tom, zdali je rozklad optimální, potøebujeme mít k dispozici nìjakou kriteriální funkci. Takovým kritériem mùže být napøíklad velièina Z, pøedstavující ztrátu, která vznikne chybným zaøazením prvkù v dùsledku nesprávného rozkladu R_m na podmnožiny M_i. Protože zaøazení prvku chceme provádìt na základì konkrétní realizace náhodného vektoru $\boldsymbol{X}$, je logické, že velièiny Z by mìla být kromì M_i i funkcí $\boldsymbol{X}$, jinými slovy øeèeno, Z bude náhodnou velièinou. Proto pøi urèování optimálního rozkladu budeme pracovat s její støední hodnotou. Za optimální budeme považovat takový rozklad, který minimalizuje støední hodnotu ztráty Z.

Základem kritéria Z bude ztráta z_{ij}, která vznikne chybným zaøazením prvku i-té skupiny do j-té a kterou ve vìtšinì konkrétních situací dokážeme urèit. Nejdøíve urèíme støední hodnotu náhodné velièiny Z_i, pøedstavující ztrátu, když prvek i-té skupiny zaøadíme nesprávnì (tj. do kterékoliv skupiny j = 1, 2, …, m, $j \neq i$).
Platí pro ni:

$$EZ_i = z_{i1} \int_{M_1} f_i(x_1,\ldots,x_m)\,dx_1 \ldots dx_m + \ldots + z_{ik} \int_{M_k} f_i(x_1,\ldots,x_m)\,dx_1 \ldots dx_m \qquad i = 1, 2, \ldots, k.$$

Funkce EZ_i je podmínìná støední hodnota ztráty, podmínkou je, že objekt pøísluší i-té skupinì.
Nepodmínìná støední hodnota ztráty EZ (tj. støední hodnota ztráty bez ohledu na to, do které skupiny prvek patøí) je pak

$$EZ = p_1 Z_1 + p_2 Z_2 + \ldots + p_k Z_k ,$$

kde p_i je pravdìpodobnost, že objekt patøí do i-té skupiny.

Položíme-li $h_j(x_1, \ldots, x_m) = \sum_{i=1}^{k} p_i z_{ij} f_i(x_1,\ldots,x_m)$,

pak støední hodnotu ztráty mùžeme pøepsat ve tvaru

$$EZ = \sum_{i=1}^{k} \int_{M_i} h_i(x_1,..., x_m)\, dx_1 ... dx_m$$

K nalezení optimálního rozkladu mùžeme použít následující vìtu:

VÌTA: Jestliže existuje rozklad $M_1^0, M_2^0, ..., M_k^0$ takový, že pro libovolné (x_1, x_2, …, x_m)$\in M_s^0$ platí $h_s(x_1, x_2, \ldots, x_m) \le h_j(x_1, x_2, \ldots, x_m)$, pak tomuto rozkladu odpovídá nejmenší støední hodnota ztráty EZ^0.

Obvykle se v diskriminaèní analýze volí ztráty z_{ij} zjednodušeným zpùsobem takto:

$z_{ij} = 1$ pro $i,j = 1, 2, \ldots, m$ a $i \neq j$,
$z_{ii} = 0$ pro $i = 1, 2, \ldots, m$

V takovém pøípadì minimalizovat EZ znamená minimalizovat poèet chybnì zaèlenìných objektù. Výraz $g_j(x_1, \ldots, x_m)$ mùžeme vyjádøit jako

$$g_j(x_1, \ldots, x_m) = \sum_{i>1}^{k} p_i f_i(x_1,..., x_m) - p_j f_j(x_1,..., x_m),$$

Pro danou hodnotu $\boldsymbol{x} = (x_1, \ldots, x_m)$ je potom $g_s(x_1, \ldots, x_m) \le g_j(x_1, \ldots, x_m)$, $j = 1,2,\ldots,k$, právì když je

$$p_s f_s(x_1, \ldots, x_m) \ge p_j f_j(x_1, \ldots, x_m).$$

Je-li u zkoumaného prvku zjištìna hodnota $\boldsymbol{x} = (x_1, \ldots, x_m)$, je prvek zaøazen do skupiny t, pro niž je splnìna nerovnost

$$p_t f_t(x_1, \ldots, x_m) > p_j f_j(x_1, \ldots, x_m), \qquad (j = 1, 2, \ldots, k;\ j \neq t) \qquad (1)$$

V praxi obvykle pøedpokládáme, že náhodná velièina $\boldsymbol{X}$ má m-rozmìrné normální rozložení pravdìpodobností se známým vektorem støedních hodnot a se známou varianèní maticí. To znamená, že v jednotlivých skupinách M_j, $j = 1, 2, \ldots, k$ bude mít vektor $\boldsymbol{X}$ hustotu pravdìpodobnosti

$$f_j(x_1, \ldots, x_m) = (2\pi)^{-\frac{1}{2}m} \cdot \left[\det(\Sigma_j)\right]^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(x-\mu_j)' \Sigma_j^{-1} (x-\mu_j)}$$

Nerovnost (1) je ekvivalentní s nerovností

$\ln p_s + \ln f_s(x_1, \ldots, x_m) > \ln p_j + \ln f_j(x_1, \ldots, x_m)$; $j \neq s$; $j = 1, 2, \ldots, m$

Oznaèíme-li

$$D_j = -0{,}5 \ln [\det(\mathbf{S}_j)] - 0{,}5\, (\boldsymbol{x} - \boldsymbol{\mu}_j)' \mathbf{S}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) + \ln p_j ,$$

bude vztah (1) platit, právì tehdy, když $D_t > D_j$, tzn.

$$-\frac{1}{2}\ln[\det(\mathbf{S}_j)] - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_t)' \mathbf{S}_t^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_t) + \ln p_t > -\frac{1}{2}\ln[\det(\mathbf{S}_j)] - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)' \mathbf{S}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) + \ln p_j$$

Pøi praktickém provádìní diskriminaèní analýzy vypoèítáme pro všechny hodnoty náhodné velièiny $\boldsymbol{X}$ hodnoty $D_1,\ldots,D_k$. Zkoumaný prvek pøísluší té skupinì, která odpovídá nejvìtší z hodnot D_j. Hodnoty $\boldsymbol{\mu}_j$ a $\mathbf{S}_j$ zpravidla neznáme, ale použijeme jejich odhady.

Pravdìpodobnost p_j se zpravidla volí úmìrná rozsahu j-té skupiny, pokud tyto rozsahy nejsou známy, volí se $p_j = 1/k$.

Diskriminaèní analýza zatím v ekonomické a veøejno-správní praxi u nás nenalezla širší uplatnìní. Je to dáno jednak tím, že její teoretické pozadí je vzhledem k jiným statistickým disciplinám pomìrnì složité a také tím, že výpoèty jsou nároèné. Díky výpoèetní technice však v dnešní dobì tyto pøekážky ztrácejí na významu.

Literatura:

1. Andìl,J.: *Matematická statistika.* SNTL, Praha 1978
2. Antoch,J.,Vorlíèková,D.: *Vybrané metody statistické analýza dat*. Akademia Praha 1992
3. Bock,H.H.:*Automatische Klassifikation.* In:Walter E.a kol.: Statistische Methoden II. Berlin, Springer 1970
4. Bowerman,B.,OÆonnell,R.,T.: *Applied Statistics*. Irwin, USA 1997

Kontaktní adresa:

doc.RNDr. Bohdan Linda, CSc.
Ústav matematiky, FES, Univerzita Pardubice
Studentská 84
532 10 Pardubice

tel.: 466 036 020
e-mail: Bohdan.Linda@upce.cz