VÍCEROZMĚRNÁ STATISTICKÁ ANALÝZA DAT V LABORATOŘI

MILAN MELOUN

Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice, Česká republika, Email: milan.meloun@upce.cz

V technické praxi se vedle informací, obsažených v náhodném skaláru >, vyskytují i vícerozměrné informace, obsažené v náhodném vektoru $\boldsymbol{x}$ s m složkami x_1, ..., x_m. Příklady vícerozměrných informací jsou (a) vyjádření vlastností produktů jako jsou potraviny, oleje, slitiny, atd. pomocí řady různých analytických metod, (b) hodnocení spekter pomocí poloh a ploch absorpčních pásů sloužící k charakterizaci a identifikaci chemických sloučenin, (c) sledování složení surovin, produktů, odpadů, v závislosti na čase nebo na místě výskytu, (d) regulace jakosti na základě různých procesních proměnných, (e) stanovení charakteristiky produktu na základě měření souvisejících proměnných, např. spekter (vícerozměrná kalibrace).

Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných, $y = w_1 x_1 + ... + w_m x_m$. *Zdrojová matice*, tj. matice výchozích dat (popisující např. řadu aut) obsahuje proměnné v m sloupcích (např. obsah motoru, výkon, spotřeba paliva, hmotnost vozu, zrychlení, výška, šířka, délka, atd.) a objekty v n řádcích (např. auta různých výrobců), na nichž jsou tyto proměnné (vlastnosti) měřeny. Protože měřené proměnné mají různé jednotky, a často se řádově liší, bývá zdrojová matice před zpracováním ještě upravována, *škálována*, a to buď (a) *centrováním*, kdy se od prvků sloupce odečte jejich sloupcový aritmetický průměr, nebo (b) *standardizací* čili *normováním*, kdy se prvky centrovaných sloupců vydělí svou sloupcovou směrodatnou odchylkou.

Metrické proměnné se vyskytují ve čtyřech škálách:

(a) *Proměnné v absolutní škále* mají na škále přirozený počátek a jediné měřítko, např. obsah uhlíku v %, rychlostní konstanta.

(b) *Proměnné v poměrové škále* mají zachován podíl hodnot charakteristik $c = x_2/x_1$, např. vztah vůči standardní sloučenině, vztah vůči jevu s definovaným nulovým počátkem, parametr F v Hammettově rovnici.

(c) *Proměnné v intervalové škále* mají zachován podíl rozdílů $c = x_2 - x_1$. Jedná se o poměrovou škálu s přirozeným počátkem pro obě srovnávané hodnoty, např. poměr absorbancí indikátoru, vztažený na absorbanci nulové linie.

(d) *Proměnné v rozdílové škále* jsou vztahovány k různému počátku, např. hodnoty časových škál, stáří, atd.

Nemetrické proměnné se vyskytují ve dvou škálách:

(a) *Proměnné v ordinální škále* mají svou hodnotu danou pořadím v neklesající posloupnosti proměnných dle nějakého kritéria, např. počet atomů chloru v molekule, žebříček umístění, pořadové číslo.

(b) *Proměnné v nominální škále* jsou nejméně informativní. Obsahují kód, např. barvu kódem 1 až 16.

(c) *Proměnné v alternativní (binární) škále* vyjadřují rovnost či nerovnost vůči nějakému kritériu. Mají binární charakter, relaci můžeme popsat dvojicí 1 (ano), 0 (ne).

Třídu nebo *shluk* chápeme jako množinu objektů se společnými nebo alespoň blízkými proměnnými, znaky (např. auta typu BMW). Blízkost či podobnost objektů posuzujeme na základě *míry blízkosti* či *vzdálenosti objektů* v *m*-rozměrném prostoru proměnných.

Mírou podobnosti dvou objektů či proměnných x_i a x_j může být *párový korelační koeficient r*. Objekty jsou si tím podobnější, čím je párový korelační koeficient větší. V případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient.* Podobnost binárních nebo nominálních proměnných vyjadřují různé koeficienty asociace. Před vlastní vícerozměrnou statistickou analýzou je třeba provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje (a) posoudit *podobnost objektů* pomocí rozptylových diagramů a symbolových grafů, (b) nalézt *vybočující objekty*, resp. jejich proměnné, (c) stanovit, zda lze použít předpoklad *lineárních vazeb,* (d) ověřit *předpoklady o datech* (normalita, nekorelovanost, homogenita). Jednotlivé techniky pro stanovení vzájemných vazeb se dále dělí podle toho, zda se hledají struktury v proměnných nebo v objektech:

(1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza* a *analýza hlavních komponent*.

(2) Hledání struktury *v objektech* v metrické škále: *shluková analýza*.

(3) Hledání struktury *v objektech* v obou škálách: *vícerozměrné škálování.*

(4) Hledání struktury *v objektech* v nemetrické škále: *korespondenční analýza.*

(5) Většina metod vícerozměrné statistické analýzy umožňuje *zpracování lineárních vícerozměrných modelů*, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných.

(a) *způsob měření vzdáleností mezi objekty:* i když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá *euklidovská metrika*, která je přirozeným zobecnením běžného pojmu vzdálenosti;

(b) *volba vhodné shlukovací procedury*, dle zvoleného způsobu metriky. Metody shlukování jsou

Metoda průměrová (Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy mezishlukovou vzdáleností objektů se rozumí vzdálenost dvou objektů, z nichž každý patří do jiného shluku.

Metoda centroidní (Centroid): vzdálenost shluků se počítá jako euklidovská vzdálenost jejich centroidů, tj. průměrů proměnných v jednotlivých shlucích.

Metoda nejbližšího souseda (Nearest): kritériem pro spojování shluků je minimum z možných mezishlukových vzdáleností objektů.

Metoda nejvzdálenějšího souseda (Furthest): počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů.

Metoda mediánová (Median): jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné "váhy", které centroidní metoda dává různě velkým shlukům.

Nehierarchické shlukové metody: *metoda typických bodů* (Seeded), kdy uživatel na základě svých věcných znalostí určí, které objekty mají být "typickými" představiteli nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů.

Diagram shluků se objeví pouze v případě, že jsme zadali hodnoty původních proměnných a nikoli matici vzdáleností. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. Objeví se také "obkroužení" objektů v jednotlivých shlucích. Dendrogram je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích. Cílem je nalézt shluk jako množinu podobných objektů s podobnými proměnnými. Podobnost objektů posuzujeme na základě vzdálenosti (míry) objektů v *m*-rozměrném prostoru (vzdálenost Euklidovská, Manhattanská, Minkovského a Mahalanobisova), párového korelačního

koeficientu a koeficientu asociace (Sokalův-Michelenerův, Russelův-Raoův a Hamanův): čím je vzdálenost shluků či objektů větší, tím menší je jejich podobnost. K rychlému posouzení podobnosti slouží grafy exploratorní analýzy vícerozměrných dat: profily, polygony, tváře, křivky, stromy, sluníčka a hvězdičky. Strukturu a vazby mezi proměnnými vystihují metody snížení dimensionality, metoda hlavních komponent (PCA) a metoda faktorové analýzy. Důležitou pomůckou je rozptylový diagram, který zobrazuje objekty, rozptýlené v rovině prvních dvou hlavních komponent. Graf komponentních vah porovnává vzdálenosti mezi proměnnými x_i a x_j, kde krátká vzdálenost značí silnou korelaci. Dvojný graf pak kombinuje oba předchozí grafy. Objekty lze seskupovat do shluků hierarchicky dle předem zvoleného způsobu metriky (průměrově, centroidně, nejbližším sousedem, nejvzdálenějším sousedem, medianově, mezi těžišti a průměrnou vazbou) a nehierarchicky dle uživatelem vybraných objektů-představitelů. Výsledkem je dendrogram.

Tato práce vznikla za podpory vědeckého záměru MSM253100002.

LITERATURA:

1. Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.
2. Seber G. A. F.: *Multivariate Observations*. Wiley, New York 1984.
3. Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis,* Ellis Horwood, Chichester 1992.
4. Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
5. Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
6. Meloun M., Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha, 2002.