Milan Meloun: Chem. Listy 91, 266 - 269 (1997).

Sedmero tajemství ukrytých v datech a sedmero kroků v analýze dat

Milan Meloun, Univerzita Pardubice, 532 10 Pardubice

Počítače jsou všude, na úřadech, v továrnách, v prodejnách ale i doma. Vedle televizoru, ledničky, pračky a telefonu se stává počítač součástí vybavení rodiny zvláště, když má doma studenta. Na vysokých školách je řada předmětů postavena na práci s počítačem. Praxe si totiž dovednosti na počítači žádá a finančně je i náležitě oceňuje, a to prozíraví studenti dobře vědí. Mladí lidé si vybírají především ty školy, specializace a předměty, kde se pracuje s počítačem.

1. Počítač není pouze psací stroj

Dnešní počítač není "cvičeným robotem", automaticky vykonávajícím jednoduché, programem řízené operace. Není ani psacím strojem, na který bývá žel často degradován. Programy - říkejme jim raději software - jsou dnes již takové komplexnosti, že jimi i začátečník dokáže záhy nakreslit složitý graf, napsat článek, sestavit si vizitku z různých druhů písma, vypočítat příklad nebo sestavit telefonní seznam. Umožňují *tvořivou, kreativní činnost*, která je pro každého zajímavá a především přitahuje mladé lidi. Přirozená studentská soutěživost žene k co nejhezčímu protokolu, plnému komplikovaných obrázků, vyrobených a do textu zařazených čistě elektronicky, bez nůžek a lepidla.

Studenti rádi tvoří a předvádějí umění počítačových triků, statistické a matematické znalosti ale také vtip a vkus v textovém a grafickém editoru. Využívají široké palety matematických a typografických znaků, elegantního zařazování grafů, diagramů a obrázků, tabulek přetažených z tabulkového procesoru. Nezapomeňme, že textový editor je opatřen i kontrolorem pravopisu, gramatiky a slovníkem synonym a antonym, které pomohou text jazykově „vybrousit", vyjádřit myšlenky tím nejvýstižnějším výrazem. Textový editor obsahuje i řadu stylů, které usnadní psaní dopisů, faxů, faktur, protokolů, formulářů, všech dokumentů s předepsanou strukturou a vzhledem. Čtenář vždy dá přednost dobře vyhlížejícímu textu, který ho více vtáhne do svého obsahu. Již pouhé zvýraznění písma je působivější než holý text z psacího stroje.

2. Sedmero tajemství ukrytých v datech

Počítače se "živí" daty. Z dat "vysají" zákonitost, zisk, penále nebo úrok, obecně říkáme *informaci*. V chemických laboratořích a chemických provozech počítače monitorují data o výrobě, procesu či kvalitě sloučeniny a automaticky zapisují veškerá měření z přístrojů do databanky. V datech je uložena informace, a to třeba o vlivu hnojiva na výnosy plodiny, o vlivu krmiva na přírůstky dobytka, o vlivu kyselého deště na kyselost půdy, o obsahu chemikálie v produktu, atd. Obsah a míra čistoty sloučeniny, její vlastnosti, důkaz její kvality jsou důležité informace, které mnohdy není tak jednoduché z dat vyextrahovat. Vedle střední hodnoty obsahu sloučeniny ve vzorku, obsahu léčiva v krvi, fibrinogenu v krevní plasmě, obsahu škodlivé látky, atd. jsou zajímavé i odlehlé, vybočující hodnoty, typ rozdělení a homogenita výběru, a konečně i nezávislost prvků ve výběru. Závislé prvky totiž prozrazují poruchu na měřícím zařízení, kinetický děj v systému nebo zfalšovaná data.

Bohatou informaci obsahují i vícerozměrná data, je však daleko složitější ji z dat

vyextrahovat. Představme si, například, 160 aut charakterizovaných svými 12 ukazateli jako jsou značka auta, model auta, spotřeba, objem válců, hmotnost, akcelerace, výkon, rozměry, atd. Seskupovací analýza vyhledá skupiny, shluky uvnitř 160 aut s přihlédnutím ke všem těmto 12 ukazatelům a odhalí podobnosti a rozdíly mezi nimi. Podobně může biolog klasifikovat např. 40 jedinců polétavých mšic dle 19 ukazatelů a odhalit počet dominantních druhů. Data absorbanční matice 50 spekter, měřených při 40 vlnových délkách v sobě ukrývají informaci o počtu barevných částic v roztoku, o jejich koncentraci a jejich fyzikálních konstantách. Přes 660 podobných úloh z rozličných oborů přírodních a ekonomických věd, (**B** biologická, biochemická a farmakologická data, **C** chemická a fyzikální data, **E** environmentální, potravinářská a zemědělská data, **H** hutní a mineralogická data, **S** sociologická a ekonomická data), přináší nedávno publikovaná sbírka exaktního zpracování dat v úlohách [1].

Analyzovaná data odkrývají svá tajemství, mají však i své "vrtochy" anomálie a rušivé poruchy, tak jak je příroda či výrobní proces do dat vložily. Tyto bariéry velmi komplikují vlastní analýzu dat. Je nutné proto předem zkoumat zvláštnosti dat interaktivní průzkumovou analýzou a ověřovat základní předpoklady o datech. Která tajemství, fakta, poruchy či problémy a anomálie budeme na počítači interaktivně monitorovat? *Jak vyextrahujeme z dat co největší objem informace* ?

(1) V datech jsou skryté problémové hodnoty. Odhalíme hrubé chyby, systematické chyby, odlehlé hodnoty, extrémy. Musíme rozhodnout, zda odstraníme neobvyklé hodnoty z další analýzy či je ponecháme nebo je opravíme.

(2) Nezávislost dat znamená, že prvky analyzovaného výběru nejsou spojeny žádným skrytým vztahem a byly získány nezávisle, bez ovlivnění člověkem, přístrojem, bez ovlivnění postupem odběru dat.

(3) Soubor obsahuje chybějící data. Pak je třeba upravit tabulku dat, která má "díry" tak, aby data přesto poskytla co nejspolehlivější výsledky. Soubor může však obsahovat i málo dat. Když je získání většího počtu dat drahé či obtížné, je třeba na malý výběr aplikovat Hornův postup pivotů. Tak odhadneme objektivní míry polohy a rozptýlení, a to bodové i intervalové.

(4) Průzkum v datech provádíme rozličnými grafickými pomůckami. Do dat se lze podívat různými diagnostikami průzkumové analýzy a odhalit symetrii rozdělení, druh rozdělení, lokální koncentraci dat, homogenitu dat, anomálie a velikost šumu.

(5) Efektivní analýza dat pozná zvláštní hodnoty, které jsou v datech velmi vlivné. Vlivné body totiž významně ovlivňují hledané parametry. Vlivným bodům je třeba proto věnovat zvláštní péči.

(6) Filtrování vlivu jednotlivých proměnných. Data monitorují výsledek, který je složen z vlivů několika proměnných. Parciální grafy dokáží odfiltrovat působení právě zvolené proměnné a vyříznout graf jejího působení z vícerozměrného prostoru. Je proto užitečné se podívat, jak je tento systém v datech monitorován.

(7) Parametry tvořených modelů, jež dostatečně popisují data, mají definovaný význam a musí vyhovovat svou velikostí, znaménkem, tj. fyzikálním smyslem.

3. Sedmero kroků v analýze dat

Na školách jsou počítačové učebny, počítačová sít propojuje počítače mezi sebou a Internetem pak se světem. Studenti se učí zpracovávat data a extrahovat z nich maximální množství informace. Vedle programů Microsoft Office, které umožní napsat protokol či diplomovou práci (Word), zpracovat složitá data v tabulce a grafu (Excel), roztřídit bázi dat (Access) nebo nakreslit obrázky na plakát, blánu či diapozitiv (Power-Point) je cílem především zpracování experimentálních dat. Toto zpracování se dnes provádí t. zv. *interaktivní analýzou dat.* Existuje k tomu řada programů, které tvoří náplň nových vědních oborů jako chemometrie, biometrie, ekonometrie, medicínská statistika, obchodní statistika, statistika pro sociology,

psychology, atd. Nový přístup k analýze dat se objevuje v poslední době v řadě monografií, jmenujme alespoň jednu, na které jsou postavena licenční studia ale i řádné studium chemometrie u nás [2]. Kniha vyšla dříve anglicky ve dvou dílech [3] a [4].

Počítačová analýza experimentálních dat spočívá obvykle z provedení sedmi obecných kroků analýzy:

1. krok: Načtení a příprava dat

V datech lze objevit trendy a skryté zákonitosti a tajemství, která bychom v databázových nebo spreadsheetových programech nezjistili. Soubory dat lze agregovat, přidávat, spojovat, editovat, transponovat, setřiďovat dle proměnných. Statistický software čte a zapisuje matice dat tak, že přebírá a předává soubory dat ze software nebo do jiných software jako jsou Excel, dBASE, Lotus 1-2-3, Sylk, ale také čte data a zapisuje data do pevného, volného formátu a formátu tabulky ASCII souborů. Načítá komplexní struktury hierarchických souborů, opakující se data, smíšené soubory.

2. krok: Flexibilní formátování prezentačních tabulek

Kvalitní tabulková zpráva pro publikace a prezentace umožní jasné a efektní zobrazení i nejkomplikovanějších analýz dat formou prezentačních tabulek. Účelné přepočítání a přeformátování tabulek pro revidovaná data slouží ke kondenzování výsledků vícenásobných odpovědí, flexibilní analýze matic dat i s případně chybějícími hodnotami a k dokonalému ovládání struktury svých tabulek. Do jedné tabulky lze umístit řádky, sloupce a vrstvy. Dokonalý vzhled tabulek umožní dokonalou kontrolu výsledného vzhledu tabulky. Mezi desítkami stylů předvolených tabulek lze měnit šířku sloupců, šířku a styl řádek, typ a barvu písma, kreslit rozličné čáry, zarovnat text vpravo, vlevo či text centrovat, je zde i možnost přidání poznámek k tabulce a zaokrouhlovat čísla na určený počet desetinných míst. Pivotované vícerozměrné tabulky umožňují zaměnit řádky, sloupky a vrstvy pomocí přesunu přes ikonu a pohyb mezi vrstvami kliknutím na ikonu.

3. krok: Dynamická grafika k diagnostickému prohlížení dat

Nejširší výběr grafických znázornění poskytuje informativnější pohled na data. Zahrnuje koláčové, spojnicové a sloupkové grafy, rozptylové souřadnicové grafy, vrstevnicové grafy, rozsáhlé mapy pro kontrolu kvality. Objektově orientované grafy oživí data. Retransformace dat umožní specifické znázornění a poskytne nejlepší představu o datových relacích. Při rotaci jsou zřetelné úhly, úrovně a interakce v trojrozměrném (3D-) grafu. Snadno lze změnit popisy, symboly, vzory, barvy a typy čar. Je k dispozici přes 120 typů diagnostických grafů z oblasti statistiky a řízení jakosti, která prozradí o datech důležitou informaci. Grafy se mohou generovat samostatně nebo jako součást analýzy, zprávy. Rychlá integrální a vizuální statistická analýza je výkonou pomůckou.

4. krok: Interaktivní diagnostický přístup v průzkumové analýze dat

Kromě základních popisných charakteristik, frekvencí a kontingenčních tabulek obsahuje software obvykle několik desítek různých měr a statistik, které podstatně rozlišují hranice analýzy za běžný statistický popis. Diagnostické grafy odhalí statistické zvláštnosti v datech, konstrukci empirického rozdělení výběru, porovnání tohoto rozdělení s normálním rozdělením, vyšetření chování řady statistik na různých částech výběru. Na jejich základě je třeba rozhodnout jakým způsobem budeme postupovat při další analýze. Je-li v datech odhaleno asymetrické rozdělení, je třeba data transformovat za účelem přiblížení se k normalitě. Po každé transformaci je vyčíslen průměr, rozptyl a asymetrický interval spolehlivosti. Bohatá statistická metodologie v modulech umožňuje práci s číselnými a kategorizovanými daty a poskytuje úplný systém

prostředků pro analýzu dat.

5. krok: Vysvětlení souvislostí v datech

Analýza dat umožní hlouběji proniknout do nitra dat a porozumět více souvislostem, které jsou v datech ukryty, odhalit vztahy a závislosti v datech. Precizní regresní techniky se nabízejí v případech regrese, kdy nelze užít klasické metody nejmenších čtverců: vážené nejmenší čtverce, dvoustupňová metoda nejmenších čtverců, metoda racionálních hodností nebo ortogonální regrese se nabízí tam, kde je třeba dát některým pozorováním více váhy, nebo obě proměnné jsou zatíženy šumem či skrytým vztahem mezi proměnnými. Pomáhá také zvládnout korelace mezi prediktorem a chybami, které se často vyskytují v datech závislých na čase.

K odhalení souvislostí, které jsou ve vícerozměrných datech ukryty se užívají techniky pro klasifikaci dat. Faktorová analýza identifikuje skupiny proměnných a jejich zátěží, které vysvětlují celkové chování. Ve výzkumu chování zákazníků lze odhalit názory na kvalitu produktu, které se vztahují k trvanlivosti, dostupnosti a prospěšnosti produktu. Seskupovací analýza vyhledá skupiny, shluky uvnitř dat a odhalí podobnosti a rozdíly mezi daty. Biolog může klasifikovat skupiny živočichů a rostlin. Ve výzkumu trhu dokáže ekonom odhalit společné rysy lidí, kteří zakoupili určitý produkt.

Diskriminační analýza odvodí pravidla pro začlenění pozorování do vytvořené skupiny.

6. krok: Testem ke správnému rozhodnutí

Korektní statistické závěry na základě statistického testování umožní lépe rozhodovat, i když jsou k dispozici jen malé výběry dat nebo podskupiny. Chceme-li zjistit, zda existuje mezi proměnnými nějaký vztah, díváme se zpravidla nejdříve na hodnoty dosažené hladiny významnosti. Je-li získání rozsáhlého souboru dat nemožné nebo příliš nákladné, je možné plánovat malé výběry a přesto neztratit k výsledkům důvěru. Nijak tím nebudeme zaostávat za konkurencí, která má více prostředků, se kterými může uskutečnit rozsáhlejší studie. Není třeba slučovat kategorie aby byly splněny předpoklady tradičních testů, a tak mnohdy ztratit původní informaci. Software umožní ponechat v analýze i málo zastoupené kategorie, tak jak vyplynuly z povahy experimentu. Na otázku, kdy jsou tradiční testy spolehlivé, neexistuje jasná odpověď, protože ověřování teoretických předpokladů je v praxi nemožné. I když pracujeme s rozsáhlými soubory, budou určité situace volat po exaktním testu. Přes 30 exaktních testů nabízí správný statistický software, který odpovídá struktuře našich dat a pokrývá celé spektrum problémů s malými i velkými množinami neparametrických a kategorizovaných dat. Zahrnují jednovýběrový, dvouvýběrový a K-výběrový test pro nezávislé ale i závislé výběry, testy dobré shody, testy nezávislosti v kontigenčních tabulkách $m \times n$ a testy měr asociace.

7. krok: Prezentace výsledků

Výstupní tabulky můžeme formátovat, snadno doplnit hlavičky, fonty, barvy, velikosti a další parametry tabulek. Vzniklé tabulky lze snadno přenášet do dalších protokolů, zpráv. Velká flexibilita při formátování tabulek nabízí pohodlí při psaní protokolu.

4. Analýza dat při kontrole jakosti

Na analýze dat je postavena nejenom výzkumná práce ale především kontrolní činnost pracovníků kontrolních laboratoří a zkušeben kontroly kvality. Prosazuje se názor, že člověk, který daty disponuje, je také zodpovědný za získání informace z nich a za jejich další využití. Na stále větším počtu pracovišť záleží, zda je pracovník schopen kvalifikovaně rozhodovat na základě objektivní analýzy informací, zda je schopen na základě naměřených dat účinně zachovávat stejnou kvalitu. Důležité je také, zda je schopen tuto kvalitu doložit. Chce-li manažer v podniku dosáhnout jistého stupně excelence, musí sám ovládat paletu rozmanitého software

interaktivní analýzy dat a rozpoznat včas náznaky nepříznivých jevů a vztahů a umět odhalit i zdroje ztrát. To se týká manažerů zdravotnických, veterinárních a vodohospodářských laboratoří, potravinářské a zemědělské inspekce, chemických, potravinářských, farmaceutických a zemědělských výrob. Je ovšem také pracovní náplní pracovníků kontroly životního prostředí všech odvětví průmyslu, energetiky a zemědělství, technologů, pracovníků řízení jakosti.

5. Formy průběžného vzdělávání

Důkazem aktuální potřeby počítačových metod statistiky a nové interaktivní analýzy dat v naší odborné veřejnosti je více než 750 absolventů odborných kurzů a seminářů, pořádaných v posledních třech letech na Univerzitě Pardubice. Absolventi se rekrutují z oblasti sledování životního prostředí, soukromých i státních laboratoří, zkušeben a výzkumných ústavů stejně jako z průmyslu důlního, hutnického, železářského, textilního, plastikářského, chemického, potravinářského, farmaceutického - abychom vyjmenovali ty nejznámější. K zájmu o tento druh kurzů přispívá jejich poněkud netradiční pojetí. Po hodině přednášky následuje vždy hodinové procvičování na počítači, takže po týdnu získá účastník konkrétní praktické dovednosti s náročným softwarem. Na velkém počtu vyřešených úloh z praxe se naučí diagnostikovat data, extrahovat maximální množství informace z dat, což činí týdenní kurz zvláště kvalitním a pro začátečníka atraktivním. Pro náročnější zájemce o práci manažerskou je k dispozici i dvouleté licenční studium, ve kterém se vedle 18 počítačových předmětů objevují i předměty jako je psychologie osobnosti manažera, akreditace a certifikace kontrolní laboratoře a formy grafické prezentace a počítačová typografie. Zvláštní kategorii průběžného vzdělávání tvoří celostátní semináře "Analýza dat", které již od roku 1990 pořádá společnost TriloByte dvakrát do roka v Lázních Bohdaneč. Zásadou je, že vybraní lektoři zde vyučují originálním způsobem užitím vlastních publikací. Zde se každoročně prezentují novinky software ve statistickém zpracování dat a především v interaktivní analýze dat. Zde také můžeme vidět poslední verze světově známých paket S-Plus, Statistica, SPSS, Systat, STATGRAPHICS, SOLO, NCSS, MathSoft, Statistical Science, Microcal a dalších, kterým úspěšně konkurují české produkty ADSTAT, WinPlot a QC-Expert. ADSTAT je nejrozšířenějším statistickým softwarem u nás, a proto jsou pardubické týdenní kurzy postaveny především na něm. V licenčním studiu je důraz kladen na nejlepší sofware z USA.

6. Závěr

Seznámení s interaktivní analýzou dat v řádné výuce studentů, týdenních kurzech a v licenčním studiu formou novinek software z celého světa je obsahem globální koncepce výuky průběžného vzdělávání vysokoškoláků a zvyšování kvalifikace českých pracovišť, jejímž cílem je pozvednout je na evropskou úroveň nejen cenami, ale také přístupem, zodpovědností, vědomostmi a v neposlední řadě sebevědomím.

Doporučená literatura:

1. M. Meloun, J. Militký: STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT - SBÍRKA ÚLOH S DISKETOU, Nakladatelství Univerzita Pardubice 1996.
2. M. Meloun, J. Militký: STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT, Nakladatelství PLUS Praha 1994, EAST PUBLISHING Praha 1998.
3. M. Meloun, J. Militký, M. Forina: CHEMOMETRICS FOR ANALYTICAL CHEMISTRY, Volume 1. PC-AIDED STATISTICAL DATA ANALYSIS, Ellis Horwood Chichester 1992.
4. M. Meloun, J. Militký, M. Forina: CHEMOMETRICS FOR ANALYTICAL CHEMISTRY,

Volume 2. PC-AIDED REGRESSION AND RELATED METHODS, Ellis Horwood Chichester 1992.