SOFTWARE STATISTICA ANEB CO SE VYŽADUJE OD MODERNÍCH STATISTICKÝCH PROGRAMŮ

Miloš Uldrich, Tomáš Jurczyk

Anotace

Rostoucí množství sbíraných dat a vývoj nových technologií s sebou přináší změny v přístupu k datům. Mění se i pohled na samotné softwary, jež s daty pracují.

Obstojí tradiční přístupy k vyhodnocení dat? Jaké jsou současné trendy v oblasti nástrojů pro zpracování dat? A co bude dál?

Trendy v oblasti nástrojů pro analýzu dat Vám ukáže přímo v softwaru konzultant analytické platformy Statistica – Ing. Miloš Uldrich.

To vše se zaměřením na analýzu dat v oblasti medicíny.

Klíčová slova

pracovní workflow, vizualizace dat v reálném čase, kolektivní inteligence, Big Data

Úvod

Rostoucí množství sbíraných dat a vývoj nových technologií s sebou přináší změny v přístupu k datům. Mění se i pohled na samotné softwary, jež s daty pracují. Stejně tak, jako se mění operační systémy, nástroje na sdílení a úpravu obrázků, webové aplikace, se kterými pravidelně pracujete, mění se i softwary na zpracování dat. Dalo by se namítnout, že pokud využívám statistické techniky, které jsou pořád stejné, tak změna není nutná, nebo je dokonce spíše nežádoucí.

Statistické přístupy (například v oblasti regresních metod, zpracování časových řad apod.) prochází stále vývojem a vznikají upravené a lépe použitelné metody.

Velký význam pro modely typu „Co se stane, když...“ mají analytické přístupy založené na umělé inteligenci, které umožňují počítačovému softwaru učit se z historických dat. Jsou to zejména algoritmy založené na strojovém učení, které jsou v posledních letech velmi oblíbené a prochází též neustálým vývojem.

Velmi oblíbené jsou i výpočetní modely inspirované chováním lidského mozku – Neuronové sítě.

Řada přístupů, zejména v oblasti testování statistických hypotéz, se ale nemění. Rozsáhlé softwarové balíky, které jsou na trhu řadu let, se proto nemění revolučně, jako někteří jejich nepřímí „kolegové“ z oblasti softwaru, ale mění se evolučně. Přidávají nové funkce a možnosti, ale pokud někdo potřebuje (nebo vzhledem k typu dat musí) používat „tradiční“ přístupy, má je tam.

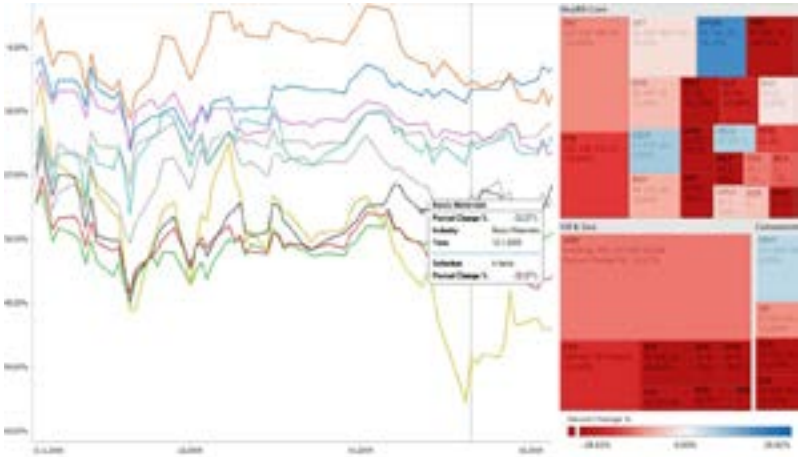
Nároky uživatelů na software však změnou prochází a to i u uživatelů, kteří využívají statistiky, u jejichž zrodu stála jména jako: Carl Pearson, Frank Yates, Francis Galton, Carl Friedrich Gauss, Jacob (Jacques) Bernoulli a mnoho dalších. Pojďme se na aktuální trendy v oblasti analytických software podívat.

Pracovní workflow

Rozhraní softwarových nástrojů se neustále vyvíjí. Hlavním trendem není již pouze přívětivé a přehledné „klikací“ rozhraní, ale možnost si celý pracovní postup srozumitelně zaznamenat. Dříve se tento postup realizoval pomocí skriptovacího jazyka, který konkrétní software nabízel (ať už to byl jazyk vlastní nebo standardní). Dnešní moderní analytické platformy obvykle obsahují pracovní plochu, která umožní analytický proces i bez znalosti programovacího jazyka definovat, spravovat a jednoduše spouštět. Tato myšlenka není nijak nová, ale v posledních letech je velmi aktuální, neboť potřeba data zpracovávat (načítat, agregovat, čistit a analyzovat) je stále větší.

V analytických rozhraních softwarů na zpracování dat jsou k dispozici objekty (též nazývané jako uzly), které vynášíme na plochu. Každý uzel reprezentuje nějakou metodu, graf, datovou transformaci, prostě funkcionalitu software. Uzly spojujeme v pořadí, v jakém bychom logicky analýzu prováděli, začínáme od zdroje dat, poté řešíme transformace čištění a spojování dat z různých zdrojů, nakonec využíváme statistické metody pro zisk důležitých informací z dat. Celé pracovní workflow je následně uloženo a pojmenováno. Celý proces lze kdykoliv jednoduše spustit a získat automaticky výsledek. Hlavní přednost je v úspoře času, pokud se data na vstupu mění, ale také ve snadné správě, například při drobné změně v nějakém uzlu (přidání kritéria, změna proměnné) máme k dispozici okamžitě všechny výsledky, které bychom jinak museli všechny „proklikat“ znovu. Poslední předností je srozumitelnost, dobře vytvořené a popsané workflow bude srozumitelné (na rozdíl od skriptu) i po delší době a navíc i pro uživatele, kteří sami workflow nevytvořili. Odpadá tak problém, kdy ve firmě je jeden specialista, který umí programovat, ale ostatní nedokáží jeho skript spravovat, opravovat nebo rozšiřovat.

Obrázek 1 – Stat. 1: Ukázka rozhraní Statistica Workspace



Obrázek 2 – Stat. 2: Ukázka modulu pro vizualizaci dat v reálném čase

Vizualizace dat v reálném čase

Vizualizace výsledků nebo vstupních měření je s analýzou dat úzce spjata. Zde je na první pohled znát vývoj posledních let nejvíce. Vzhled grafů prošel značným vývojem, stejně jako vzhled většiny obsahu internetu. Trendem posledních několika let je bouřlivý rozvoj modulů pro vizualizaci dat v reálném čase. Grafy se v moderních analytických platformách sdružují v tzv. Kontrolních panelech (Dashboard) a jsou interaktivní. Lze tedy do nich klikat a interaktivně řídit, co se bude zobrazovat.

Mnoho organizací dnes požaduje, aby aktuální data a výsledky analýz byly dostupné uživatelům v reálném čase a pracovníci mohli vidět aktuální vývoj a rozhodovat se

Obrázek 3 – Stat. 3: Mnoho úhlů pohledu na stejná data

na základě skutečně aktuálních dat. Jednoduše, rychlé rozhodnutí bez nutnosti čekat dlouhou dobu na výsledek zadané analýzy může poskytnout společnosti konkurenční výhodu. Tento přístup se od tradičních statických analytických výstupů liší. Vše, co na výstupech kontrolních panelů uvidíte, je skutečný aktuální pohled na data, pokud změníte (v databázi, nebo Excel tabulce) jednu hodnotu, na grafech se to okamžitě projeví. Výstavba těchto grafů není těžká a rozhraní fungují zpravidla na principu „Táhni a pusť". Na panelech se kombinují různé grafické a sumarizační techniky najednou. To umožňuje na jednom monitoru vidět současně mnoho úhlů pohledu a odhalit trendy a odlehlé hodnoty, které by jinak zůstaly skryty. Právě kombinace mnoha úhlů pohledu na jednom „obrázku" je velkým přínosem poslední doby.

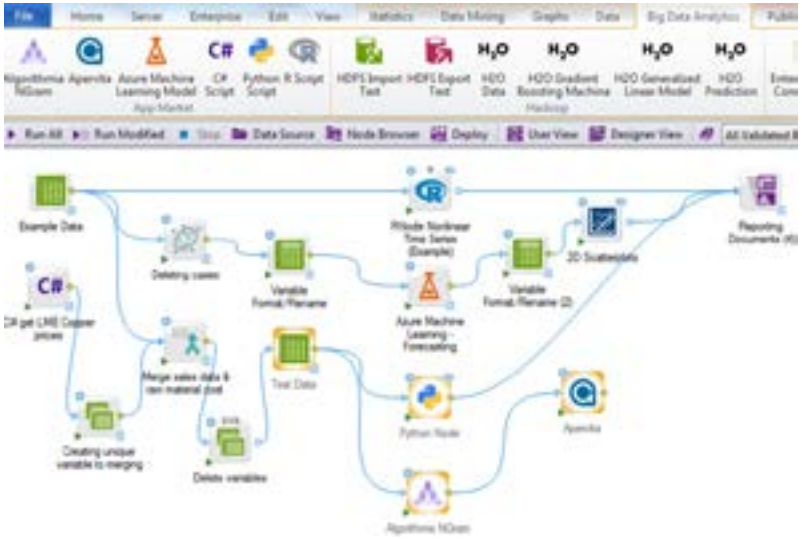
Dalším krokem v evoluci analýz v reálném čase je pak nejen zobrazení dat v reálném čase, ale i výpočet predikcí a modelů v reálném čase. Jako příklad takovéto analytiky uveďme reálný připad University of Iowa, kde přímo na operačním sále pomáhá operatérovi analytický predikční model, který na základě průběhu operace předpovídá pravděpodobnost pooperačních potíží a poskytuje doporučení, jak se těmto potížím vyhnout. Nutno poznamenat, že takováto aplikace přímo zachraňuje životy a šetří obrovské náklady na pooperační komplikace. Ano, i takto může vypadat analýza dat.

Kolektivní inteligence

Pojem Kolektivní inteligence je obecně definován jako schopnost skupiny najít větší množství, anebo kvalitnější řešení nějakého problému, než její jednotliví členové. S rostoucím množstvím dat, které můžeme sledovat prakticky všude, vznikají nároky na hardware, výpočetní algoritmy, ale (ačkoli to na první pohled nemusí být zcela zřejmé) je zde tlak také na množství analytiků, tedy specialistů, kteří budou z dat získávat informace. Cílem firem a moderních analytických softwarů je tedy také co největší zjednodušení práce a kooperace těchto specialistů, včetně zapojování lidí, kteří nejsou nutně specialisty na analýzu dat, ale data znají dobře. Další cestou pak může snaha využívat znalosti (například modely) z externích zdrojů.

Je vidět, že tyto požadavky se již netýkají analytických či statistických metod, týkají se spíše organizace práce a otevřenosti a flexibility analytické platformy, kterou firma používá. Nutností bývají funkcionality jako centrální serverové uložiště všech dat, analýz a modelů, možnost nastavit přístupová oprávnění k jednotlivým objektům (zaměstnanec na pobočce banky jistě nesmí mít přístup k modelu, který slouží ke skórování klientů, kteří za ním přijdou) nebo třeba „verzování" a záloha dokumentů.

Hodí se jistě také rozhraní, které pomůže sdílení postupů mezi uživateli (například ideální je zmíněné prostředí pracovní plochy s jednotlivými funkcionalitami v podobě uzlů). Analytik může také předpřipravit analýzy nebo celé aplikace pro manažery a konzumenty výstupů z analýz, kteří se potřebují podle výsledků rozhodovat, ale nemusí se již zabývat tím, co je na pozadí výpočtu. Zde je nutné mít jakýsi obal nad analytickými postupy, který zpřístupní uživatelům jen několik málo nastavení a zbytek zůstane skryt – i třeba vůči možným úpravám.



Obrázek 4 – Stat. 4: Ukázka kolektivní inteligence v softwaru Statistica

Typickým zjednodušením práce je také možnost využívat v rámci jedné platformy i další programovací jazyky (typicky se používají jazyky R, Python, C). Zjednodušení je v tom, že specialista na analýzu dat pracuje v tom, co je mu blízké nebo co je nejvhodnější, může také využít skripty, které naprogramoval v jiných jazycích (můžete si to představit tak, že tento specialista vytvoří v pracovní ploše uzel založený na těchto jiných programovacích jazycích). Uživatel se tedy nemusí přeučovat či přizpůsobovat.

V dnešní době již existují tržiště s předpřipravenými modely pro konkrétní úlohy, které si můžete koupit a poté používat pro vaše účely. Výhodou je, že nemusíte mít ve své firmě specialistu, který bude modely složitě vyvíjet, využijete tedy expertní znalosti z externího zdroje, což bude s rostoucím množstvím dat čím dál častější praxe. Dalším směrem vývoje analytické platformy je tedy i komunikace s těmito tržišti v rámci jedné platformy. Jak můžete vidět, už dávno nejsou moderní analytické platformy pouze o metodách a algoritmech.

Příprava analytického souboru

Příprava dat je nezbytná součást analytické práce. Data v syrové podobě nemusí mít takovou vypovídací hodnotu. Prvním krokem před vlastní analýzou je zajištění přesnosti záznamu, kde se snažíme zkontrolovat (validace dat) správnost jednotlivých dat. Sleduje se počet chybějících hodnot, duplicitní záznamy, kontrolují se jednotlivé varianty znaku, srovnatelnost jednotlivých proměnných v různých tabulkách apod. K odhalení těchto nepřesností v datovém souboru slouží celá řada technik, jež mají analytické softwary implementovány. Po aplikaci vhodných technik, překódování a odstranění chybných záznamů, máme k dispozici datový soubor, který je již použitelný pro další analýzu.

Dnešní softwary musí umět víc než jen překódovat tabulku a identifikovat odlehlé hodnoty. Největší výzvou pro analytické softwary je aktuálně reagovat na všechny nově vznikající datové zdroje. Řešení musí být schopna načíst data z nových moderních datových zdrojů, sloučit je, provést jejich čistění a uložit je zpět do databáze. Druhou velkou výzvou posledních let je objem dat, který v některých oblastech diametrálně narostl a v budoucnu bude narůstat dále.

Big Data, Cloud

Schopnost analyzovat tzv. „Velká data" (Big Data) je dnes základním argumentem pro nákup komerční analytické platformy. Velká data jsou dnes prakticky ve všech oblastech včetně medicíny. Moderní softwary se díky novým technologiím zvládnou připojit na strukturovaná, částečně strukturovaná i nestrukturovaná data z libovolných zdrojů jako jsou například NoSQL DB, Big Data úložiště založená na Hadoop a různá Cloud úložiště. Informace z blogů, webových článků, lékařských systémů a dalších zdrojů se kombinují s demografickými a regionálními daty. Analytické softwary dnes musí být schopné z těchto nových zdrojů získat data a přeměnit je na informace, které mají pro zadavatele hodnotu.

Analytika přímo v datových zdrojích (Edge analytics)

Dovolte zmínit také důležitý trend velkých dat a to analytiku přímo v datových zdrojích nebo na místě, kde jsou data sbírána (například v datovém zdroji Hadoop nebo v místě kudy proudí data například ze senzorů). Místo přesouvání dat z datových zdrojů jsou data analyzována přímo v datových zdrojích, což má výhodu nejen v tom, že náročné přesouvání dat odpadá, ale je zde i aspekt bezpečnosti.

Kontakty

Ing. Miloš Uldrich
e-mail: *milos.uldrich@quest.com*

Mgr. Tomáš Jurczyk, Ph.D.
Autoři jsou odbornými konzultanty softwaru Statistica (QUEST software).
Web: *http://statistica.io*