Prediktivní řízení kvality výroby

s podporou dataminingových technik

Lenka Blažková

Takzvané metody dolování informací z dat bývají nejčastěji spojovány s aplikacemi v bankovnictví, pojišťovnictví, telekomunikačních společnostech či marketingu. Jejich uplatnění je však opodstatněné také v průmyslu. V tomto článku se zaměříme na porovnání čtyř dataminingových postupů aplikovaných v řízení kvality potravinářského výrobního procesu. Pro názornější představu zvolíme modelový příklad řízení kvality při výrobě popcornu.

Eliminace ztrát dříve, než k nim dojde

Zabývejme se nyní hledáním prediktivního modelu, který dokáže odhalit riziko poklesu kvality výroby ještě dříve, než dojde k vlastní výrobě. Oddělení řízení kvality má díky takovému modelu možnost otestovat, zda konkrétní nastavení regulovatelných vstupních parametrů výrobního procesu povede k získání potraviny dosahující požadovaného standardu kvality, či nikoli. Jsou-li výsledky simulace nepříznivé, mohou být včas a beze ztrát, jež produkce nekvalitní potraviny představuje, učiněna potřebná opatření. Dataminingové prediktivní metody mohou navíc pomoci odhalit skryté vztahy mezi kvalitou výroby a vstupními parametry, které mohou být klíčové pro kontrolu a řízení kvality.

Modelový příklad

Kvalitní popcorn získáme optimálním nastavením parametrů výrobního procesu. Sledovány jsou veličiny jako například teplota, tlak a obsah oxidu uhličitého, hmotnost výrobní dávky, tlak vzduchu, plnění kyslíku atd. Rozhodujícím kritériem je ukazatel kvality, kde za hraniční považujeme hodnotu 0,15. Přesáhne-li hodnota ukazatele kvality tuto mez, nesplňuje vyrobený popcorn standardní požadavky na kvalitu. Řízení kvality se soustředí výhradně na regulovatelné parametry, ačkoli kvalita popcornu pochopitelně závisí i na parametrech neregulovatelných.

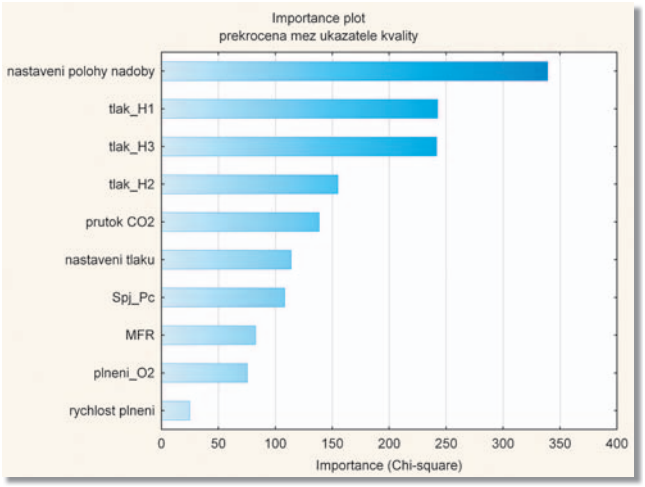
Kde tedy začít? Prvním krokem libovolné analýzy je stanovení parametrů, které nejvíce ovlivňují sledovaný ukazatel kvality, v řízení kvality totiž zejména tyto regulovatelné parametry hrají zásadní roli. K jejich identifikaci zpravidla užíváme graf důležitosti vstupních parametrů (obr. 1). Redukce počtu vstupních proměnných vede k jednodušší interpretaci výsledného modelu a také samotné uvedení do praxe je snazší, pokud měníme nastavení pro menší počet parametrů.

Data mining

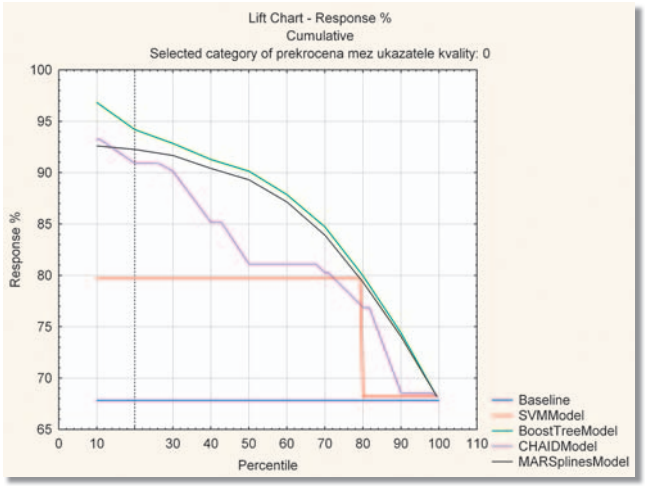
Pro řešení optimalizační úlohy jsme zvolili tyto čtyři klasifikační metody: boosted trees, chaid trees, support vector machines a MAR splajny. Zmíněné metody patří k dataminingovým postupům a jako takové bývají součástí pokročilých balíků statistických softwarů. Prediktivní modely jsou vyvíjeny přímo v konkrétním softwaru. Vybrané modely určené pro nasazení lze následně pohodlně exportovat v různých formátech v závislosti na typu použitého softwaru (např. ve formátu skriptů PMML, C+) a implementovat je jako součást BI. Pracovníci kontroly kvality mají tak vždy k dispozici automatické aktuální výstupy ze simulací, které slouží jako senzory ohlašující potenciální riziko.

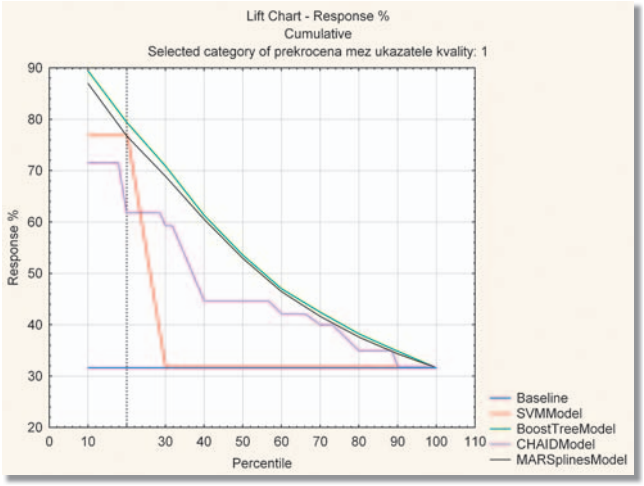
Každá z těchto technik vytváří model, který případy na základě vstupních dat (zvolených regulovatelných parametrů) klasifikuje do dvou skupin. V ideálním případě by jedna skupina obsahovala pouze případy, kdy ukazatel kvality nepřekročí kritickou mez, a druhá případy, kdy došlo k překročení kritické hodnoty. Vytvořené modely jsou vždy zatíženy chybou, proto se v obou skupinách obecně budou vyskytovat chybně

Obr. 1: Graf důležitosti vstupních parametrů



Obr. 2: Lift Chart





Obr. 3: Lift Chart

klasifikované případy. Nejlepším modelem bude ten, pro který bude míra chybně klasifikovaných případů nejmenší. Pro samotné budování prediktivních modelů jsou používána historická (trénovací) data, účinnost získaných modelů a jejich srovnání je provedeno na testovacím vzorku dat, který nebyl při vytváření modelů použit. Účinnost prediktivních modelů bývá posuzována nejčastěji na základě různých typů grafů, jako je například lift chart. V našem případě máme k dispozici dva grafy (obr. 2) – první posuzuje úspěšnost klasifikace do skupiny bez překročení hraniční meze pro ukazatel kvality, druhý posuzuje úspěšnost klasifikace do skupiny, kdy k překročení hraniční hodnoty došlo. Jednotlivé modely jsou porovnávány se základní úrovní (base line), která odpovídá situaci, kdy nepoužijeme pro klasifikaci žádný model (v grafu označena modře). Efekt, jaký v porovnání s náhodným zařazením přinese konkrétní klasifikační model, ukazuje vždy příslušná křivka (čím více se křivka blíží pravému hornímu rohu, tím lépe model klasifikuje). Jako nejefektivnější se jeví modely boosted trees a MAR splajny. Nejhůře se jeví metody chaid trees a support vector machines. Pro praktickou implementaci volíme tedy buď boosted trees anebo model MAR splajnů.

Principy jednotlivých technik

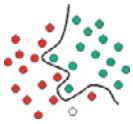
Základem pro chaid i boosted trees je tzv. rozhodovací strom, který může mít obecně několik úrovní. Ve vstupním uzlu dostává strom všechna trénovací data, v každé úrovni je dále dělí podle hodnot některého ze vstupních parametrů na podskupiny. Konkrétní vstupní parametr a kritérium pro dělení na podskupiny jsou přitom voleny tak, aby vznikly skupiny co nejvíce homogenní (v případě klasifikace se tedy snažíme vytvářet podskupiny, v nichž má většina případů stejnou klasifikaci). V dělení pokračujeme tak dlouho, až vznikne jakási stromová struktura pravidel pro dělení, která dává jak pro trénovací, tak i pro testovací data dostatečně dobrou klasifikaci.

Chaid trees

Tato metoda je jedním z nejstarších stromových klasifikačních algoritmů. Automaticky je vytvářena stromová struktura, kdy se data každého uzlu mohou dále dělit i do více než dvou podskupin. Pro stanovení nejlepšího dalšího dělení je využíván statistický chí-kvadrát test, který dal metodě název.

Boosted trees

Boosting je poměrně nový algoritmus. Základní myšlenka je vytvořit posloupnost jednoduchých klasifikačních stromů. Tyto na první pohled primitivní klasifikátory samy o sobě vykazují značnou nepřesnost při klasifikaci případů. Proto jsou jim přiřazeny váhy v závislosti na počtu špatně klasifikovaných případů. Boosting kombinuje i několik stovek jednoduchých klasifikátorů, a umožňuje tak vytvořit model silný, jehož predikce vychází z „hlasování“ jednoduchých stromů.



Obr. 4: Support vector machines

Support vector machines

Hodnoty vstupních parametrů jednotlivých pozorování ze souboru trénovacích dat jsou nahlíženy jako souřadnice bodů ve vícerozměrném prostoru. Body, které odpovídají případům, kdy ukazatel kvality nepřekročil hraniční hodnotu, obarvíme červeně a zbylé body modře. Nyní se snažíme najít ideální oddělovač červených a modrých bodů. V rovině to bude křivka, v prostoru plocha a dál sice naše představivost většinou nesahá, ale matematika má prostředky, jak takový vícerozměrný oddělovač najít a popsat. A právě tento oddělovač je hledaný prediktivní klasifikační model support vector machines.

MAR splajny

Český název zní vícerozměrné adaptivní regresní splajny. Ačkoli je v názvu obsažena regresní analýza, lze tuto techniku aplikovat i na klasifikační úlohy. Jedná o neparametrickou modelovací proceduru založenou na tzv. bázových funkcích. Trénovací data jsou použita pro stanovení koeficientů jednoduchých bázových funkcí. Podle hesla „rozděl a panuj” jsou vstupní pozorování nejprve rozdělena podle hodnot vstupních parametrů na menší podskupiny a pro každou z těchto podskupin jsou hledány koeficienty vlastní regresní rovnice. Výsledný model je kombinací těchto jednodušších modelů.

Obecně mají uvedené metody za cíl najít interakce mezi vstupy výrobního procesu (použité výrobní stroje, sledované fyzikální a chemické parametry atd.) a jeho výsledkem, jímž je buď kvalitní, nebo nekvalitní výrobek. Při použití stromových algoritmů je výhodou i interaktivní vizualizace jednotlivých kroků dělení, která poskytuje názorný obrázek důležitosti nastavení jednotlivých parametrů a umožňuje interaktivní přizpůsobení modelu v souladu s expertními zkušenostmi kontrolního pracovníka. Doporučení, která pro výrobu vyplývají z jednotlivých modelů, se nejlépe interpretují (a obhajují) také právě u stromových algoritmů.

Skrytý potenciál

Model natrénovaný a ověřený na historických datech, u nichž známe kvalitu výstupu, můžeme aplikovat na aktuální data, a zajistit tak efektivní kontrolu kvality výroby. A to v reálném čase. Pro aktuální či zvažované nastavení vstupních parametrů výstupy simulace predikují, jaký výsledek můžeme očekávat. Výrobní společnosti jsou silnou konkurencí nuceny optimalizovat vlastní výrobní proces. Hledají proto sofistikované a ekonomicky výhodné způsoby řízení a kontroly kvality výroby pomocí modelování a simulací. Metody demonstrované v tomto článku jsou příkladem postupů, které mohou přispět k výraznému zlepšení kvality a zvýšení efektivity výroby. Ačkoli je prokázáno, že moderní algoritmy přinášejí úspory a vyšší efektivitu, nasazování v praxi je jen postupné a masivní využití i v menších společnostech je zatím bohužel jen hudbou budoucnosti. ■

Autorka je odbornou konzultantkou firmy StatSoft CR.