

# Przemysł wyrachowany

**Dogłębna analiza danych nie jest domeną jedynie sprzedaży czy marketingu. Może być bardzo użyteczna także w przemyśle, w tym przypadku występuje jednak więcej zmiennych i parametrów.**

**Tomasz Demski**

Współczesne produkty i procesy produkcyjne stają się coraz bardziej skomplikowane, rosną także wymagania odnoszące się do ich jakości. Zarówno podczas projektowania produktu, jego wytwarzania, jak i korzystania z niego przez użytkownika gromadzone są bardzo du-

dywać, które produkty prawdopodobnie będą wadliwe, aby zaoszczędzić na końcowych etapach procesu. Oczywiście, że wyniki działania systemu muszą być dostępne natychmiast, tak abyśmy mieli czas i możliwość korzystania z wyników analizy.

## ANALIZA DANYCH W PRZEMYŚLE

Dane dotyczące procesów technologicznych zawierają zazwyczaj bardzo dużo – setki lub nawet tysiące zmiennych, co stanowi ich specyfikę. Wynika ona z tego, że dane generowane są przez urządzenia automatyki przemysłowej. Zapisują one mnóstwo parametrów, których wartości często nie mają związku z wytwarzaniem w danej chwili produktem, ale mogą być decydujące dla innego produktu. Duża ilość zmiennych występuje także w przypadku analizy danych o procesach wsadowych, w których zmiennymi są wyniki pomiarów parametrów procesów dokonane w różnym czasie.

## Parametrów wiele

Można wskazać kilka specyficznych cech zgłębiania danych w przemyśle. Pierwsza z nich to bliski związek ze statystycznym sterowaniem jakością procesów (SPC). Bardzo często zgłębianie danych stosujemy jako rozwinięcie i uzupełnienie SPC. Stąd często tego typu zastosowania określa się nazwą Quality Control data mining (lub QC data mining). Nakłada to na systemy analityczne stosowane w przemyśle wymóg integracji z narzędziami do tradycyjnego SPC, jak karty kontrolne, analiza zdolności procesu i analiza niezawodności.

Na etapie projektowania użyteczna jest współpraca z aplikacjami wspomagającymi projektowanie (CAD). Kolejnym wyróżnikiem data mining dla przemysłu jest wymóg automatycznego reagowania na zmiany zachodzące w analizowanych zbiorach danych – możliwie na bieżąco. Jako ilustrację rozważmy system, który przed zakończeniem wieloetapowego procesu technologicznego ma przewi-

dywać, które produkty prawdopodobnie będą wadliwe, aby zaoszczędzić na końcowych etapach procesu. Oczywiście, że wyniki działania systemu muszą być dostępne natychmiast, tak abyśmy mieli czas i możliwość korzystania z wyników analizy. Dane dotyczące procesów technologicznych zawierają zazwyczaj bardzo dużo – setki lub nawet tysiące zmiennych, co stanowi ich specyfikę. Wynika ona z tego, że dane generowane są przez urządzenia automatyki przemysłowej. Zapisują one mnóstwo parametrów, których wartości często nie mają związku z wytwarzaniem w danej chwili produktem, ale mogą być decydujące dla innego produktu. Duża ilość zmiennych występuje także w przypadku analizy danych o procesach wsadowych, w których zmiennymi są wyniki pomiarów parametrów procesów dokonane w różnym czasie.

stosowana w takich obszarach, jak projektowanie i doskonalenie produktu, sterowanie i optymalizacja procesu produkcyjnego czy też analiza reklamacji i niezawodności.

Na etapie projektowania data mining może dotyczyć kwestii związanych z klientem, np. identyfikacji potrzeb klientów i prognozowania popytu. Ponadto możemy badać zależności między projektami produktów, portfelem produktów i potrzebami klientów. Z kolei w przypadku etapu wytwarzania są to głównie: statystyczne sterowanie jakością procesu, przewidywanie problemów z jakością, wykrywanie ich przyczyn czy utrzymanie maszyn (np. planowanie przeglądów i remontów tak, aby uniknąć awarii). Jest to również sterowanie przebiegiem procesów, wykrywanie przyczyn i związków pomiędzy parametrami tych procesów, a także podsumowanie wielowymiarowych danych.

Weźmy dwa przykłady praktycznego i udanego zastosowania data mining w przemyśle zaczerpnięte z literatury światowej (m.in. „Data Mining for Design and Manufacturing. Methods and Applications”, Kluwer Academic Publishers 2001, „Predictive data mining. A practical guide”, Morgan Kaufman Publishers 1998 czy „Mastering data mining”, John Wiley & Sons 2000).

## Sterowanie procesem

W pewnym procesie technologicznym wykorzystywano duży zbiornik przechowujący surowce dla tego procesu. Proces był obserwowany przez operatorów, którzy korygowali jego ustawienia, tak aby zmniejszyć lub zwiększyć stopień wypełnienia zbiornika. Występowały przy tym dwa dość oczywiste zagrożenia – w zbiorniku mogła znaleźć się zbyt mała ilość surowca. Z drugiej strony, groziło przepełnienie zbiornika. W pierwszym przypadku cały proces technologiczny musi zostać zatrzymany i uruchomiony ponownie. Procedura taka jest bardzo kosztowna i niebezpieczna. Nato-

miast w przypadku przepełnienia i wylania się zawartości zbiornika występuje duże zagrożenie dla bezpieczeństwa pracy.

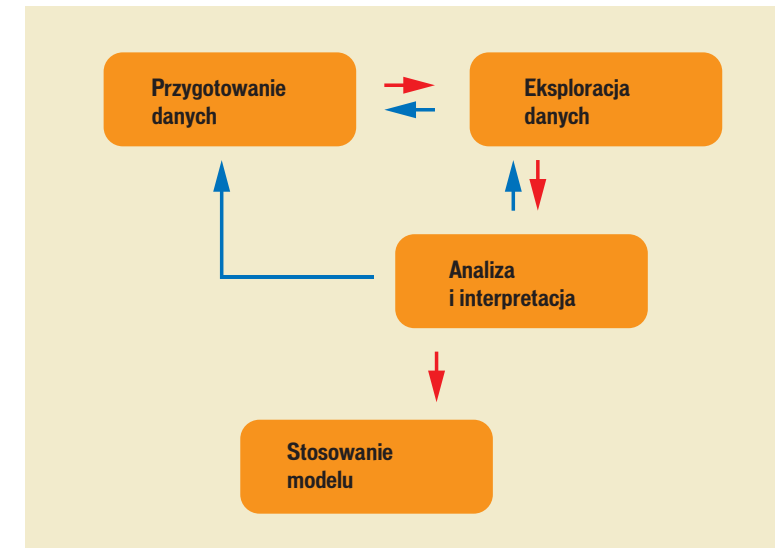
Operator ma do dyspozycji wiele ustawień, tj. wartości zmiennych sterujących, ale faktycznie na ilość materiału w zbiorniku wpływa tylko jedna z nich. Celem analizy było przewidzenie wartości zmiennej sterującej, przy której zostanie zachowana optymalna ilość surowca w zbiorniku. Parametry opisujące stan procesu i zbiornika oraz wartość zmiennej sterującej zapisywane są co 30 s. Jako wzorzec odpowiedniego ustawienia zmiennej sterującej przyjęto działania kilku różnych operatorów. Niestety, optymalnych ustawień nie udało się wyznaczyć teoretycznie.

Potrzeba jednak pozostała, bowiem umiejętność zamodelowania działania operatora, by później dało się je naśladować, była bardzo potrzebna. Formalny, komputerowy model ma wiele zalet. Przykładowo, w przypadku odejścia doświadczonego pracownika tracimy jego doświadczenie, a dopóki nowy operator nie nabierze doświadczenia ryzyko wystąpienia problemów jest znacznie większe. Przy sterowaniu automatycznym będzie (a przynajmniej powinno być) mniej wahań losowych, a cały proces powinien być stabilniejszy.

W praktyce zmiany były dokonywane przez operatora na podstawie jego wyczuć i doświadczenia. Większość z nich to zmiany małe i nieistotne. W związku z tym zdecydowano, że przewidywana będzie zmiana poziomu zmiennej sterującej po 3 min. Jakość uzyskanych wy-

## Sieci przemysłowe

Jednym z podstawowych narzędzi stosowanych w eksploracji danych są sieci neuronowe. Przyjmuje się, że jest to już praktycznie standardowy element wszystkich profesjonalnych pakietów służących do zaawansowanej statystycznej analizy danych. Takie programy potrafią same automatycznie dobrać najlepszą architekturę sieci neuronowej (zależnie od określonego zadania), określić jej złożoność i użyć ją (na podstawie zewnętrznych zbiorów danych). Sieci neuronowe zdają egzamin zwłaszcza w przypadku procesów produkcyjnych, w szczególności w obszarze kontroli jakości, bowiem mamy tam do czynienia z dużą liczbą zmiennych, odnośnie do których nie zawsze istnieją explicite reguły rządzące ich zachowaniem. Prościej i efektywniej jest więc, biorąc zebrane historyczne dane, nauczyć odpowiednich reakcji czy odpowiedzi sieć neuronową niż rekonstruować reguły.



ków zdecydowanie polepszyło stosowanie średnich ruchomych i wartości trendów zamiast surowych wartości parametrów. Pozwoliło to na wyeliminowanie losowych wahań parametrów. Ponadto w analizie uwzględniono tylko te obserwacje, w których zmienna sterująca została znacząco zmodyfikowana – drobne, wykonywane przez człowieka korekty były mylące i niedokładne.

Ważnym elementem modelu było wykrywanie trendów w wielkości strumienia surowca wpływającego i wpływającego ze zbiornika. Dwa kluczowe parametry wstępnego przekształcenia danych stanowiła liczba punktów wykorzystywanych przy obliczaniu średniej ruchomej oraz wielkość zmiany zmiennej sterującej uznawana za istotną. Optymalizację tych dwóch parametrów analizy połączono z optymalnym doбором liczności próby uwzględnianej w analizie. Na przykład za małą wartość progowej zmiany powodowała uzyskiwanie rozwiązań faworyzujących niewykonywanie żadnych zmian.

W wyniku analizy uzyskano rozwiązanie o mniejszej i większej złożoności. Pomimo tego, że rozwiązanie o większej złożoności nieco lepiej przewidywało rzeczywiste zmiany, do stosowania wybrano prostsze rozwiązanie, ze względu na jego zgodność ze standardami przedsiębiorstwa. W wyniku analizy uzyskano zaskakująco dobry wynik w postaci prostego i skutecznego modelu. Działanie modeli zbadano dla oryginalnych, zapisywanych co 30 s danych, a model spisał się dobrze, pozwalając uniknąć zarówno przepełnienia, jak i opróżnienia zbiornika.

## W drukarni

W drukarni R. R. Donnelley występował tajemniczy problem polegający na pojawianiu się serii rys na walcu wy-

korzystywanym przy drukowaniu rotograviurowym. Na wydrukach problem objawiał się jako kolorowe linie przecinające cały wydruk. Problem zaczął występować przy drukowaniu z szybkością ponad 300 mb/min. Celem analizy danych w tym przypadku było zminimalizowanie częstości występowania problemu.

Stosowana technologia powodowała, że każda przerwa w drukowaniu i ponowne uruchamianie procesu były bardzo kosztowne. Ponadto wystąpienie rysy powodowało marnotrawstwo matryc oraz dużych ilości papieru i farby drukarskiej. Usunięcie wady walca zajmowało średnio półtorej godziny, a w tym czasie cały proces był zatrzymany. Ponieważ terminy drukowania zazwyczaj są bardzo napięte, każde opóźnienie skutkowało dodatkowymi kosztami nadgodzin.

Przed rozpoczęciem projektu nie gromadzono żadnych danych. Na początku należało podjąć decyzję, jakie informacje mają być zbierane i zapisywane w bazie danych. Początkowo zdecydowano, że dla procesów poprawnych i wadliwych gromadzone będą dane m.in. o: wilgotności, temperaturze farby, lepkości farby, odczynie farby, napięciu i rodzaju papieru. Ostatecznie zestaw zbieranych informacji uzyskano na podstawie konsultacji z ekspertami, tworzenia kolejnych modeli i wybierania zmiennych istotnie wpływających na wystąpienie problemu.

Jako metodę modelowania zastosowano różne algorytmy drzew decyzyjnych. Ostatecznie w wyniku zastosowania dogłębnej analizy danych uzyskano zestaw reguł, które można było zastosować przy ustawianiu procesu produkcyjnego.

Wdrożenie reguł wydobytych z danych za pomocą data mining pozwoliło zmniejszyć liczbę wystąpień problemów w ciągu roku z 538 do 21.

## Czym data mining różni się od tradycyjnych metod statystycznych?

- Analiza dużych zbiorów danych
- Nastawienie na praktyczne wyniki i zastosowania, a nie na budowę lub sprawdzanie teorii
- Korzystanie z istniejących danych, na których zwrócić badacz ma niewielki wpływ
- Ocena modelu na podstawie próby testowej, a nie na podstawie wskaźników statystycznych

Za „duże” uznajemy takie zbiory danych, których człowiek nie jest w stanie objąć i wykorzystać bez pomocy komputera i specjalistycznego oprogramowania. Bardzo często w praktyce spotykamy się z sytuacją, gdy danych jest „za dużo”, a głównym zadaniem we wnioskowaniu z danych jest odsianie bezużytecznej informacji. Ważną częścią tradycyjnego badania statystycznego jest zaplanowanie doświadczenia, które da nam informacje podlegające właściwej analizie. W data mining mamy do czynienia z inną sytuacją: zazwyczaj analizujemy istniejące dane, gromadzone zwykle do innych celów niż analiza danych, i które, w pewnym sensie, są zbierane „przy okazji”.

Przykładami typowych źródeł danych będą informacje z systemu automatyki przemysłowej (którego głównym celem jest sterowanie produkcją), systemu rejestrującego reklamacje zgłaszane przez klientów – służącego przede wszystkim do wspomaganie rozwiązywania problemów klientów itp.

W data mining wykorzystuje się narzędzia pochodzące z trzech dziedzin – technologii bazodanowej (gromadzenie, udostępnianie i przetwarzanie danych), statystyki oraz uczenia maszyn i sztucznej inteligencji. W procesie data mining możemy wyróżnić cztery zasadnicze etapy: (1) Przygotowanie danych, (2) Eksploracyjna analiza danych, (3) Właściwa analiza danych (budowa i ocena modelu lub odkrywanie wiedzy), (4) Wdrożenie i stosowanie modelu. Warto zwrócić uwagę, że powyższe etapy nie przebiegają liniowo, jeden za drugim. Bardzo często na kolejnym etapie okazuje się, że powinniśmy wrócić do wcześniejszego (por. rysunek powyżej). Na etapie przygotowania danych decydujemy, z jakich informacji będziemy korzystać w analizie, pobieramy odpowiednie dane, sprawdzamy ich poprawność i dokonujemy odpowiednich przekształceń, aby zapewnić zgodność danych pochodzących z różnych źródeł.

Celem eksploracji danych jest poznanie ogólnych własności analizowanych danych: rozkładów jedno- i wielowymiarowych cech i podstawowych związków między zmiennymi. Wynikiem takiej wstępnej analizy jest wykrycie nietypowych przypadków. Po wykryciu odstających przypadków powinniśmy podjąć decyzję, jak będziemy z nimi postępować. Podczas eksploracji uzyskujemy również informacje, czy potrzebne i użyteczne będą jakieś przekształcenia oryginalnych danych. Przykładowo, w wyniku eksploracji danych może okazać się, że klasy zmiennej jakościowej występują tak rzadko, iż należy je połączyć z innymi.

Na etapie eksploracji danych bardzo często wykonujemy wstępną selekcję zmiennych, aby w dalszych analizach uwzględnić tylko te właściwości obiektów, które są istotne (np. wpływająca na zmienną zależną). W razie wykrycia niejednorodności danych możemy pogrupować wszystkie przypadki (obiekty analiz) w jednorodne grupy i właściwą analizę wykonywać osobno dla grup. Etap właściwej analizy danych rozpoczynamy od wstępnego doboru metod odpowiednich do rozwiązania problemu. Przy wyborze metody należy kierować się rodzajem problemu, wielkością zbioru danych, dopuszczalną złożonością modelu oraz wymaganiami odnośnie do możliwości interpretacji modelu.

Po wykonaniu analiz oceniamy, czy uzyskane wyniki są zadowalające. Kluczową sprawą jest, czy uzyskana informacja jest użyteczna z praktycznego punktu widzenia. Zazwyczaj wykorzystujemy więcej niż jedną technikę analizy danych. Istnieje wiele różnych metod oceny modeli i wyboru najlepszego z nich. Często stosuje się techniki bazujące na porównawczej ocenie modeli (competitive evaluation of models), polegającej na stosowaniu poszczególnych metod dla tych samych zbiorów danych, a następnie wybraniu najlepszej z nich lub zbudowaniu modelu złożonego.

Techniki oceny i łączenia modeli (uważane często za kluczową część predykcyjnej eksploracji danych) to m.in.: agregacja modeli (głosowanie i uśrednianie: bagging), wzmacnianie (nazywane też losowaniem adaptacyjnym i łączeniem modeli, boosting), kontaminacja modeli (stacking, stacked generalizations) i metauczenie (meta-learning). Obszerne omówienie wskaźników jakości modeli i sposobów ich porównania znajduje się w ww. publikacjach źródłowych.

Reguły te nie wyjaśniły, dlaczego pojawiają się problemy, ale pozwoliły zmniejszyć częstość ich występowania. Łączny czas przestojów przed wprowadzeniem reguł przekraczał 800 godz. rocznie, a po ich zastosowaniu spadł do 30 godz. w ciągu roku. Doświadczenia uzyskane w drukarni, gdzie przeprowadzono oryginalny

projekt, zostały przeniesione do innych zakładów. Chociaż same modele należało dostosować do każdej drukarni, to sposób rozwiązania problemu był ten sam. ▶

Tomasz Demski jest specjalistą w firmie StatSoft Polska sp. z o.o.