



## DATA MINING W STEROWANIU PROCESEM (QC DATA MINING)

*Tomasz Demski, StatSoft Polska Sp. z o.o.*

### Wprowadzenie

Sterowanie i optymalizacja jakości to dziedziny, w których zastosowanie zgłębiania danych (*data mining*) może przynieść szczególnie duże korzyści, zwłaszcza jeśli połączymy je z metodami monitorowania parametrów procesów za pomocą kart kontrolnych oraz z technikami planowania doświadczeń (*DOE*). Korzyści te wypływają m.in. z tego, że koszty jakości mogą być ogromne (przykładowe analizy kosztów można znaleźć np. w [1] i [2]).

Warto zauważyć, że obecnie najskuteczniejsza strategia zapewnienia jakości, tj. metodyka Sześć Sigma, kładzie bardzo duży nacisk na zbieranie danych i ich analizę – credo Sześć sigma brzmi: „Zmienić możemy tylko to, co mierzymy” (zob. [2]).

Zgłębianie danych w sterowaniu jakością (QC data mining) od zwykłego zgłębiania danych (*data mining*) odróżnia między innymi konieczność reagowania na zmiany w danych na bieżąco. Jako ilustrację rozważmy system, który przed zakończeniem wieloetapowego procesu technologicznego ma przewidywać, które produkty prawdopodobnie będą wadliwe, aby zaoszczędzić na końcowych etapach procesu. Oczywiście jest, że wyniki działania systemu muszą być dostępne natychmiast, tak abyśmy mieli czas i możliwość skorzystać z wyników analizy. Zauważmy, że system *STATISTICA Data Miner* może automatycznie przeliczać projekty przy każdej zmianie danych (szczególnie łatwo można to osiągnąć, stosując zestaw *STATISTICA Data Miner* i *SEWSS*), a źródło danych może stanowić baza danych (tzn. nie ma konieczności importowania danych).

Innym wyróżnikiem QC data mining jest konieczność stosowania metod typowych dla sterowania jakością, takich jak karty kontrolne, analiza zdolności procesu, planowanie doświadczeń itp. Dla osiągnięcia optymalnego wyniku analizy te muszą być zintegrowane z narzędziami typowymi dla data mining.

Specyfika danych dotyczących procesów technologicznych polega na tym, że zazwyczaj tworzone są one przez urządzenia automatyki przemysłowej. Zapisują one zazwyczaj mnóstwo parametrów, które często nie mają żadnego wpływu na wytwarzany w danej chwili produkt, ale mogą być decydujące dla innego produktu. Ponadto w wielu dziedzinach produkcji zmiany zachodzą bardzo szybko – czas życia produktów i okres

stosowania konkretnej technologii ciągle się zmniejsza. W związku z tym bardzo często będziemy potrzebować narzędzia tworzącego modele typu czarna skrzynka – na ich zrozumienie nie będziemy mieli po prostu czasu. Modele muszą radzić sobie z dużą liczbą danych nie wpływających w żaden sposób na zmienną wyjściową i łatwo adaptować się do zmienionych technologii i nowych produktów. Takim właśnie problemem się zajmujemy.

## Przykład projektu QC data mining

Rozważmy proces technologiczny podzielony na cztery etapy. Naszym celem będzie wykrycie, już po trzech etapach, tych partii, dla których liczba wadliwych elementów na koniec procesu będzie zbyt duża.

Mamy 760 obserwacji następujących danych:

- ♦ identyfikator partii i typu produktów,
- ♦ liczba defektów dla każdej partii,
- ♦ dla etapu 1: 2 zmienne jakościowe i 87 zmiennych ciągłych,
- ♦ dla etapu 2: 121 zmiennych ciągłych,
- ♦ dla etapu 3: 341 zmiennych ciągłych.

Łącznie mamy 553 predyktory, które mogą wpływać na liczbę defektów. Przy takiej liczbie potencjalnych zmiennych opisujących proces tradycyjne podejście do czyszczenia danych i ich analizy praktycznie nie jest możliwe do zrealizowania: nakład pracy potrzebny na samo przejrzanie zmiennych jest ogromny, a przecież jest to dopiero początek. Początkowy fragment danych widzimy na rys. 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Typ	Partia	e1operat	e1sid	E1T0	E1TF	e1p0	e1pf	e1vf	e1pin1	e1pin2	e1pin3	e1pin4	e1pin5	e1pin6
1	A	1	104	3	22.79	27.77	100.36	107.6233	100	0.413		0.384	0.639	3.425	987.672
2	A	2	104	3	22.8	27.69	100.23	108.026	100	0.227		0.597	0.288	2.603	981.649
3	A	3	104	3	21.74	26.77	100.22	108.6312	100	0.308		0.541	0.909	2.491	983.564

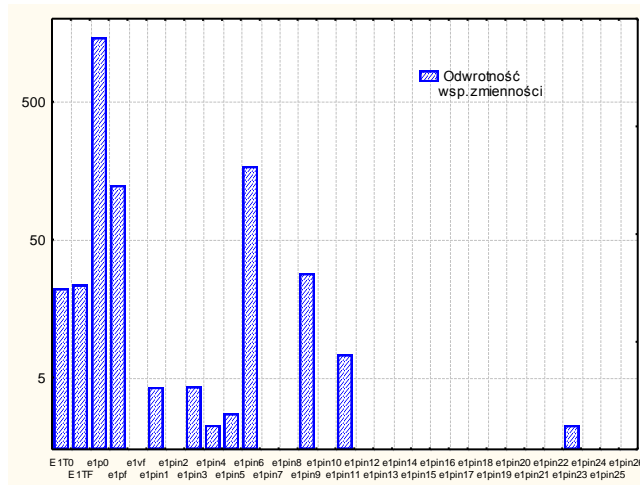
  

	15	16	17	18	19	20	21	22	23	24	25	26	27	28
	e1pin6	e1pin7	e1pin8	e1pin9	e1pin10	e1pin11	e1pin12	e1pin13	e1pin14	e1pin15	e1pin16	e1pin17	e1pin18	e1pin19
1	987.672			112.476		71.53064	0	0	0	0	0	0	0	0
2	981.649			117.471		75.38282	0	0	0	0	0	0	0	0
3	983.564			121.156		87.98189	0	0	0	0	0	0	0	0
4	996.023			116.692		64.87412	0	0	0	0	0	0	0	0
5	981.931			116.432		68.1526	0	0	0	0	0	0	0	0
6	992.668			121.260		67.82235	0	0	0	0	0	0	0	0
7	996.929			121.802		56.88557	0	0	0	0	0	0	0	0

rys. 1.

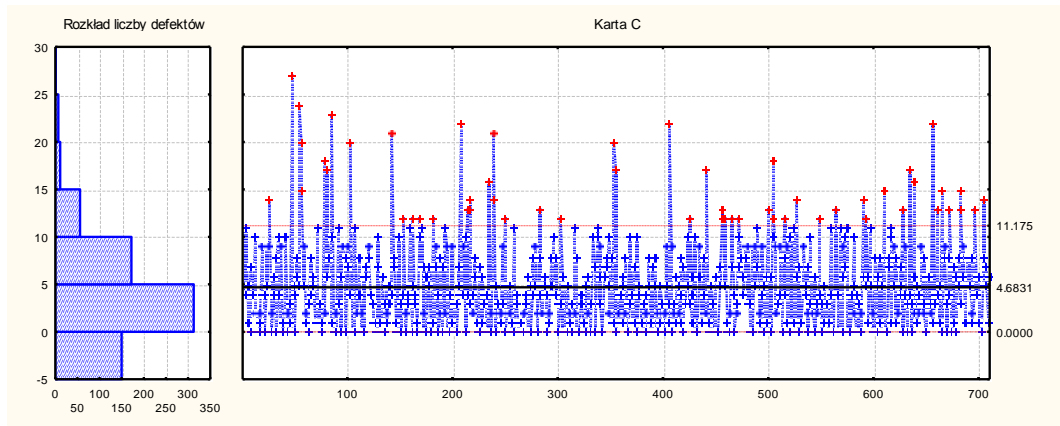
Jakość naszych danych możemy zilustrować za pomocą wykresu przedstawiającego stosunek średniej do odchylenia standardowego (odwrotność współczynnika zmienności) dla pierwszych 30 predyktorów (rys. 2). Jak widać, mamy mnóstwo zmiennych, dla których w ogóle nie występuje żadna zmienność w danych; takie zmienne to np. *e1vf*%. Zwróćmy uwagę, że oś Y na tym wykresie ma skalę logarytmiczną. Wartości stosunku

średniej do odchylenia standardowego tylko dla tych 30 predyktorów zmienia się od około 2 do ponad 1200.



rys. 2.

Zobaczmy najpierw, czy liczba defektów w partii jest stabilna statystycznie. Użyjemy w tym celu standardowej karty kontrolnej, konkretnie karty c. Po umieszczeniu w przestrzeni roboczej *STATISTICA Data Miner* naszego źródła danych jako badaną zmienną wybieramy liczbę defektów, następnie z przeglądarki obiektów wybieramy węzeł *Attribute Charts (C, U, Np, p)*. Uzyskamy następującą kartę:

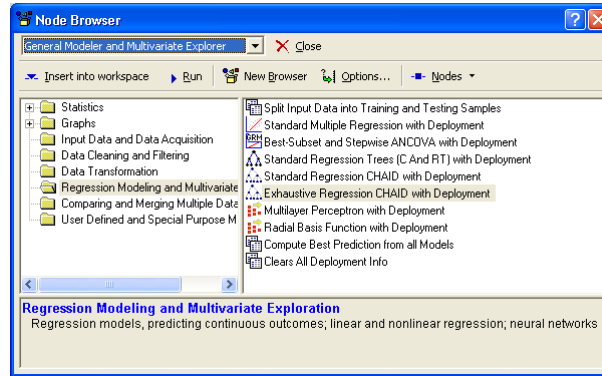


rys. 3.

Jak widać, nasz proces jest nieuregulowany – mamy sporo sygnałów o rozregulowaniu. Spróbujemy teraz przygotować model, który będzie przewidywał liczbę defektów dla każdej partii.



Nasze zagadnienie jest problemem regresyjnym i skorzystamy z metod z grupy *General Modeler and Multivariate Explorer* (zob. rys. 4), dlatego nowy projekt tworzymy poleceniem *Build Your Own Project* z menu *Data Mining - General Modeler and Multivariate Explorer*.

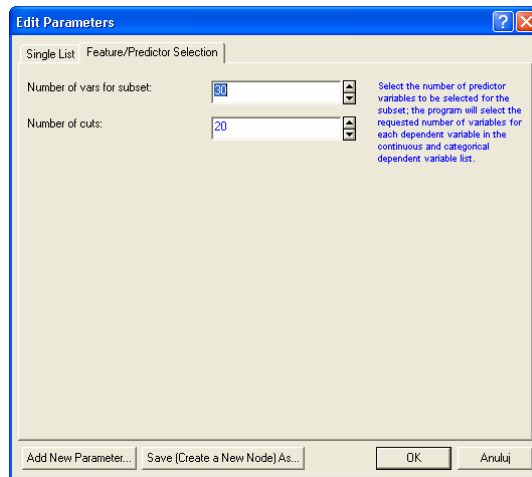


rys. 4.

Po utworzeniu nowego obszaru roboczego wstawiamy do niego źródło danych i wskazujemy zmienne. Podstawą całej naszej analizy będzie węzeł *Feature Selection and Variable Screening*, który wybierze tylko zmienne wpływające na liczbę defektów. Zauważmy, że zastosowana procedura wyboru zmiennych nie zakłada typu zależności między predyktorami a opisywaną zmienną (w szczególności zależność ta nie musi być liniowa). Ponadto metoda ta jest bardzo szybka. Więcej informacji na ten temat można znaleźć w [3].

Procedura wyboru zmiennych pozwala m.in. na określenie liczby zmiennych wybieranych jako najbardziej prawdopodobne predyktory i liczby cięć (*number of cuts*). Im większa liczba odcięć, tym bardziej nieliniowe zależności jest w stanie wykryć stosowana przez nas procedura. Ponieważ przyjmujemy, że nasz model może być silnie nieliniowy, ustalimy liczbę cięć na 20.

Do modelowania użyjemy między innymi sieci neuronowych. Sieci neuronowe są wrażliwe na dużą liczbę zmiennych nic niewnoszących do modelu, dlatego zmniejszymy liczbę wybieranych predyktorów z 50 do 30. Ustawienia parametrów dla węzła *Feature Selection and Variable Screening* widzimy na rys. 5.



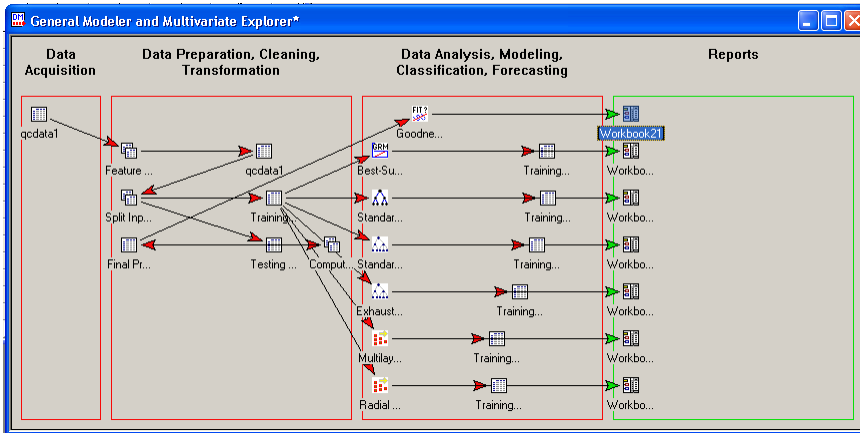
rys. 5.

Po wykonaniu procedury wyboru zmiennych w obszarze roboczym pojawi się nowe źródło danych. Zauważmy, że jest ono tylko „wirtualnym” zbiorem danych (jest to coś w rodzaju odsyłacza lub wskaźnika do oryginalnych danych) – program nie kopiuje oryginalnych danych, co mogło by się wiązać z przesyłaniem dużych ilości danych.

Do oceny jakości naszego modelu zastosujemy standardowy w data mining sposób postępowania, tzn. podzielimy nasze dane na zbiór uczący (dla którego zbudujemy model) i testowy (na którym sprawdzimy zgodność modelu z danymi). Zwróćmy uwagę, że dane o procesie mają charakter sekwencji czasowej. W takim przypadku jako próbę testową wybiera się najpóźniejsze dane. W naszym przypadku jako zbiór testowy wykorzystamy 230 ostatnich przypadków ze zbioru danych, a wcześniejsze będą stanowiły zbiór uczący.

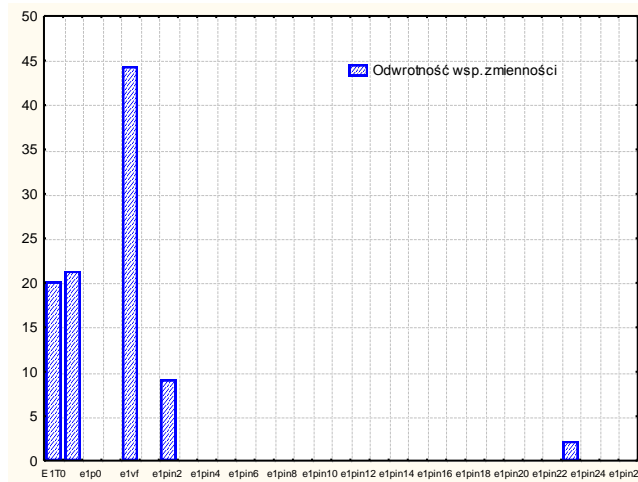
Po utworzeniu zbioru testowego i uczącego uruchamiamy przeglądarkę węzłów i wybieramy wszystkie metody analityczne z grupy *General Modeler and Multivariate Explorer* (zob. rys. 4), za wyjątkiem regresji liniowej (ponieważ stosujemy metodę *GRM*, która jest jej uogólnieniem). Jako prognozę liczby defektów użyjemy średniej prognoz wszystkich tych modeli. Taką prognozę uzyskamy, stosując węzeł *Compute Best Prediction from all Models* dla próby testowej (na etapie oceny modelu) lub dla nowych danych (na etapie stosowania modelu).

Po uruchomieniu projektu do arkusza z prognozowanymi liczbami defektów podłączamy węzeł *Goodness of Fit*, aby ocenić jakość naszego modelu (kompletny projekt przedstawiono na rys. 6). Syntetycznym wskaźnikiem jakości dopasowania jest średni błąd bezwzględny – dla naszego modelu złożonego wynosi on około 2,75. Jeśli popatrzymy na kartę kontrolną dla liczby defektów (rys. 3), to zauważymy, że rozregulowania odpowiadają liczbie defektów przekraczającej 11, a więc nasz model, przynajmniej na pierwszym etapie doskonalenia procesu, powinien okazać się użyteczny.



rys. 6.

Możemy teraz zastosować przygotowany przez nas projekt dla innych danych, dotyczących innego produktu wytwarzanego na tej samej linii technologicznej. Do porównania danych użyjemy wykresu przedstawiającego stosunek średniej do odchylenia standardowego dla pierwszych 30 predyktorów (rys. 7). Jeżeli porównamy ten wykres z wykresem dla pierwszego pliku danych (rys. 2), to zauważymy, że nasze dane są zupełnie inne (aczkolwiek mają taką samą strukturę). W przypadku procesów przemysłowych dosyć często zdarza się, że zmiana produktu wiąże się z dramatyczną zmianą tego, jakie czynniki wpływają na cechy produkowanego obiektu, a nawet zmierzeniem zupełnie innych cech.



rys. 7.

Aby zastosować nasz projekt do nowych danych, wystarczy wstawić źródło danych do naszego projektu i połączyć je z pierwszym węzłem używanego projektu. Po wykonaniu projektu uzyskamy model dla nowych danych. Średni błąd bezwzględny dla nowych danych wynosi około 2,45.



## Literatura

1. Bank J., 1997, Zarządzania przez jakość, Gebtehner & Ska.
2. Harry M., Schroeder R., 2001, Six sigma. Wykorzystanie programu jakości do poprawy wyników finansowych, John Willey & Sons.
3. *STATISTICA Data Miner* dokumentacja, StatSoft Inc. 2002.