



Nowa wersja systemu data mining, PASW MODELER 13 (dawniej CLEMENTINE) wnosi cały szereg zmian nie tylko w zakresie interfejsu, ale także dostępnych technik przygotowania i analizy danych. Zmiany w obrębie interfejsu dostarczają użytkownikowi nowe rozwiązania funkcjonalne, które w istotny sposób wpływają na wygodę pracy oraz przejrzystość projektowanych strumieni analitycznych. Znacząco poszerzone zostały w PASW MODELER 13 możliwości dotychczasowych węzłów, a także pojawiły się nowe. Pozwalają one na ograniczenie czasu potrzebnego na przygotowanie danych i ich analizę. Poniżej znajdziecie Państwo wybrane nowości PASW MODELER 13.

zmiany w interfejsie programu

Komentarze wyświetlane w oknie obszaru roboczego (w bezpośrednim sąsiedztwie węzłów). Komentarz może być związany z konkretnym węzłem, bądź opisywać fragment strumienia danych. Jego rozmiar i położenie mogą być modyfikowane przez użytkownika. Po ustaleniu tych parametrów komentarz staje się tłem dla znajdujących się w strumieniu węzłów.

Ta forma opisu jest niezwykle przydatna w dużych rozwiązaniach, w których użytkownik nie ma możliwości objęcia całego strumienia danych w widocznym fragmencie obszaru roboczego.

Narzędzie do tekstowego przeszukiwania węzłów bieżącego strumienia danych. Wyszukiwać węzły można poprzez szereg kategorii takich jak: etykieta, kategoria węzła, słowa kluczowe, nazwę zmiennej tworzonej w węźle i inne.

Wprowadzenie kategorii grupujących węzły w zakładce Modelowanie (Zautomatyzowane, Klasyfikacja, Segmentacja, Asocjacja). Dzięki temu rozwiązaniu dotarcie do pożądanego algorytmu jest znacznie prostsze. Paleta węzłów została również uporządkowana w zakresie typów węzłów. Powstały dwie nowe zakładki, EKSPORT oraz PASW STATISTICS, które zawierają wydzielone węzły eksportu danych (dawniej znajdujące się w zakładce Wyniki) oraz węzły integrujące program z pakietem statystycznym PASW STATISTIC (dawniej SPSS).

W oknach dialogowych pojawił się również pomocny przycisk pozwalający na maksymalizację rozmiaru okna oraz na powrót do oryginalnych wymiarów.

nowe węzły – operacje na danych

Nowy węzeł Auto Przygotowanie, który pozwala na wykonanie szeregu automatycznych operacji na danych zbioru wejściowego. Wiele z przekształceń węzła Auto Przygotowanie można odtworzyć z pomocą szeregu węzłów dostępnych we wcześniejszych wersjach programu, jednak możliwość skompresowania ich do postaci pojedynczego elementu strumienia danych może być niezwykle przydatna, zwłaszcza w sytuacji, gdy rezultat trzeba uzyskać szybko.

Funkcjonalność kategoryzacji optymalnej opartej o algorytm MLDP (Minimum Length Description Principle) w węźle Kategoryzacja. Poszerza ona możliwości w zakresie kategoryzacji danych.

Opcja balansowania, w **węźle Ważenie, wyłącznie zbioru uczącego**. W połączeniu z nowymi opcjami węzłów modelowania pozwala to wyliczyć ocenę (prawdopodobieństwo) na podstawie próby przeznaczonej do testowania, która nie została zważona. Taka ufność charakteryzuje się słabszymi wynikami niż wyliczana na podstawie zbioru uczącego, jednak ma znacznie silniejszy związek z analizowanym zjawiskiem.

Możliwość złożonego losowania w węźle Losowanie. Nowością jest losowanie w zespołach, co pozwala na wybór wszystkich rekordów spełniających wartość wprowadzonego klucza. Np. kiedy chcemy wylosować 20% klientów do analizy koszykowej, ale tabela zagregowana jest do poziomu rachunku, algorytm wybierac będzie pojedynczych klientów (20%) oraz wszystkie przypisane im rachunki. Z kolei opcja losowania **po warstwach**

pozwała np. losować po 20% klientów z oddziałów, jeżeli jako warstwę stratyfikującą wskażemy oddział. W wyniku otrzymamy zbiór, w którym z każdego oddziału wybranych zostaje 20 % klientów. Możliwe jest łączenie tych dwóch opcji, określanie wielkości próbki i wielkości warstw przy pomocy procentów i liczności, wyliczanie wagi w oparciu o mechanizm losowania i kilka innych operacji.

Nowy węzeł Zespolenie. Pozwala wykorzystać kilka konkurencyjnych modeli klasyfikacyjnych, budowanych na przykład przy pomocy różnych technik. W węźle budowana jest jedna, wspólna ocena w oparciu o wszystkie dostarczone do węzła scoringi. Węzeł Zespolenie pozwala zatem połączyć informacje płynące z wielu modeli jednocześnie poprzez automatyczne przeliczenie prawdopodobieństw z modeli. Metod łączenia ocen różnych modeli jest kilka, od najprostszej, głosowania, po bardziej złożone.

Nowy węzeł Agregacja RFM. Dzięki niemu pojawiła się możliwość szybkiego przygotowania danych do analizy RFM (Recency Frequency Monetary). Po wskazaniu identyfikatora jednostki, zmiennej mówiącej o dacie ostatniego kontaktu i wartości (w tym miejscu możemy wykorzystać różne informacje w zależności od podejścia i celu analizy), węzeł dokona wyliczenia trzech nowych zmiennych (wymiarów analizy RFM) czyli Świeżości, Częstości i Wartości związanej z jednostkami. Węzeł ma również więcej opcji związanych np. z możliwością wyliczania daty ostatniego kontaktu i wcześniejszych kontaktów, określania interesujących nas wartości związanych z wymiarem Monetarnym (np. warunkowy wybór kwot branych pod uwagę w analizie – nie mniejszych niż 5 PLN).

W związku z węzłem Agregacja RFM pojawił się również węzeł Analiza RFM – pozwala on przeprowadzić proces segmentacji na podstawie wprowadzonej przez analityka informacji o liczbie przedziałów, na które podzielone zostaną wymiary segmentacyjne.

nowe węzły – algorytmy

Auto Klasyfikacja, Auto Predykcja, Auto Grupowanie – to w zasadzie nie algorytmy, ale dodatkowe funkcjonalności polegające na możliwości wygenerowania w ramach jednego węzła wielu nowych modeli. Wynik daje możliwość porównania tych modeli przy pomocy wielu statystyk i wskaźników oraz wygenerowania ich.

Lista reguł – algorytm klasyfikacyjny, umożliwiający łączenie obserwacji do zadanych wcześniej zbiorów. Praca z nim ma charakter interaktywny. Analityk może zdać się na algorytm, bądź zdefiniować własne warunki podziału wartości predyktorów. Takie podejście jest szczególnie cenne przy segmentacji, w której mamy narzucone pewne główne kryteria (np. wiek, płeć, dochód), ale w ich obrębie chcemy poszukiwać specyficznych grup klientów. Technika umożliwia połączenie podejścia eksperckiego i analitycznego.

SLRM (Self Learning Response Model) – technika pozwalająca na douczanie modelu na podstawie rzeczywistej informacji od kontaktowanych klientów. W połączeniu z narzędziem PES możliwa jest automatyczna aktualizacja modelu.

Model Coxa – technika przydatna do przewidywania czasu wystąpienia zjawiska (np. kiedy klient odejdzie) na podstawie innych zmiennych.

GenLin (ogólny model liniowy) – szereg technik regresyjnych, ujętych pod wspólnym modelem, realizujących różne dopasowania funkcji (np. gamma, poissona).

Dyskryminacyjna – technika analizy dyskryminacyjnej.

Sieci Bayesa – dwie odmiany algorytmu działającego zgodnie z logiką Bayesowską: Naive Bayes (TAN) oraz Markov Blanket.

SVM (Support Vector Machine) – algorytm klasyfikacyjny oparty o procedurę wyznaczania wektorów wspierających w wielowymiarowym zbiorze danych. Algorytm sprawnie pracujący na zbiorach danych zawierających dużą liczbę predyktorów.

KNN (K Nearest Neighbor) – technika klasyfikacyjna umożliwiająca predykcję na podstawie wartości najbliższych sąsiadów analizowanego przypadku. Jest to pierwsza technika pamięciowa (memory-based reasoning) dostępna w programie. Stanowi logiczne uzupełnienie dostępnych obecnie technik, gdyż spełnia wszystkie kryteria algorytmu klasy drążenia danych (można ją wykonywać na dużych zbiorach danych, jest techniką bezpośrednio wywodzącą się z zagadnień sztucznej inteligencji).

Dodatkowym, ciekawym i przydatnym elementem jest **analiza czułości** – element graficzny pojawiający się w generowanych w programie modelach. Pokazuje ona na histogramie czułość (ważność) predyktorów dla analizy. Czułość jest miarą znormalizowaną, więc histogram łatwo interpretować w kategoriach procentowych.

Kompletna Platforma Analiz Predykcyjnych

Zmiana nazewnictwa oprogramowania firmy SPSS wynika z repozycjonowania produktów, a w obszarze funkcjonalnym z pełnej integracji oferowanych rozwiązań. Obecnie PASW MODELER 13 oferuje możliwość pełnej integracji z trzema innymi rozwiązaniami SPSS: PASW COLLABORATION & DEPLOYMENT SERVICES (dawniej PES – PREDICTIVE ENTERPRISE SERVICES), PASW DATA COLLECTION (dawniej SPSS DIMENSIONS) oraz PASW STATISTICS (dawniej SPSS STATISTICS). Szerszą informację na ten temat można uzyskać z materiału poświęconego Korporacyjnej Platformie Analiz Predykcyjnych. Tutaj warto zauważyć, iż obecnie użytkownik posiadający licencję na program PASW® STATISTIC ma możliwość

skorzystania z szeregu uzupełniających funkcjonalności dostępnych dotychczas wyłącznie w pakiecie statystycznym, takich jak specyficzne procedury i techniki statystyczne, wizualizacja danych przy pomocy tabel, czy opcji przekształcania danych poprzez język poleceń).

nowości technologiczne

Jedną z kluczowych zmian w PASW MODELER 13 jest krótszy czas przekształceń danych i wyliczania modeli. Znaczący wpływ na skrócenie czasu przygotowania strumieni oraz ich wykonania ma rozszerzenie zakresu optymalizacji dostępnych w programie przekształceń do postaci kodu SQL (pushback do bazy). Obecnie ogromna większość węzłów i wykorzystywanych funkcji jest konwertowana do postaci zapytań. W przypadku najpopularniejszych baz (Oracle, MS SQL, IBM DB2, Teradata) funkcjonalności mechanizmu zostały rozbudowane o elementy dla nich specyficzne. Powoduje to zarówno ograniczenie wolumenu danych przenoszonych pomiędzy repozytorium a systemem analitycznym, jak również wykorzystanie zazwyczaj znacznie wydajniejszego serwera bazodanowego.

PASW MODELER 13 został wyposażony w nowy mechanizm integracji programu z zewnętrznymi modułami analitycznymi. Dotychczasowy interfejs CEMI został rozbudowany o nowy mechanizm – CLEF, który umożliwia wykorzystywanie zewnętrznych aplikacji i wprowadzanie ich w postaci nowych funkcjonalności interfejsu programu lub nowych węzłów (mechanizm CEMI pozwalał wyłącznie na definiowanie nowych węzłów). CLEF znacznie rozszerza również zakres komunikacji z zewnętrzną aplikacją.

SPSS POLSKA

ul. Raclawicka 58
30-017 Kraków
tel./faks 012.636.96.80
tel./faks 012.636.07.91
tel./faks 012.636.45.35
e-mail: info@spss.pl
www.spss.pl
www.analizadanych.pl
www.webmining.pl