Krzysztof Świder

Dział Informatyzacji
Politechnika Rzeszowska

Bartosz Jędrzejec

Wydział Elektrotechniki i Informatyki
Politechnika Rzeszowska

Zaawansowane metody analizy danych i niekomercyjne pakiety analityczne w systemach wspomagania decyzji na potrzeby administracji publicznej

1. Wstęp

W badaniach nad systemami wspomagania decyzji szczególne miejsce zajmują metody oparte na analizie danych. Ostatnie dwudziestolecie przyniosło istotny przełom w tej dziedzinie i zaowocowało opracowaniem skutecznych metod i algorytmów odkrywania wiedzy z danych.

Tradycyjne aplikacje baz danych (formularze, raporty) są przeznaczone głównie do wspomagania codziennej (operacyjnej) działalności firm i instytucji. Możliwości analizowania danych przy użyciu takich narzędzi są ograniczone i sprowadzają się głównie do sporządzania zestawień i wykresów. Tymczasem działalność na poziomie decyzyjnym wymaga nie tylko danych, ale przede wszystkim wiedzy. Dlatego we współczesnych systemach informacyjnych coraz więcej uwagi poświęca się pozyskiwaniu wiedzy przy użyciu nowej generacji technik analizy. Interesującym oraz dynamicznym kierunkiem rozwoju współczesnych technologii baz danych jest odkrywanie wiedzy z danych (*knowledge discovery from data* – KDD), mające ścisłe odniesienie do takich pojęć, jak: hurtownie

danych, technologia OLAP oraz eksploracja danych[1]. W systemach odkrywania wiedzy z danych, określanych także jako systemy analityczne, wykorzystuje się technologię baz danych oraz wybrane metody i techniki z dziedziny informatyki (sztuczna inteligencja) i nauk matematycznych (statystyka). W takich systemach analiza, z założenia, dotyczy danych o znacznej objętości.

Współczesne metody odkrywania wiedzy z danych mogą w znaczący sposób poprawić stopień wykorzystania technik IT do wspomagania strategicznych decyzji w jednostkach administracji publicznej. Istotną przeszkodą w ich powszechnym zastosowaniu może się jednak okazać znaczny koszt zakupu i wdrożenia istniejących na rynku rozwiązań komercyjnych. W niniejszej pracy rozważono możliwości potencjalnego zastosowania współczesnych metod odkrywania wiedzy z danych na potrzeby administracji publicznej, a jednocześnie wskazano na perspektywę praktycznego użycia do tego celu w pełni profesjonalnych pakietów analitycznych rozpowszechnianych na zasadach niekomercyjnych. Intencją autorów było m.in. wykazanie, że zastosowanie w praktyce najnowszych metod KDD nie musi być domeną bogatych korporacji.

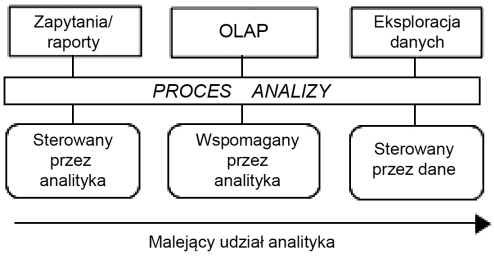
2. Odkrywanie wiedzy z danych

Stosowane obecnie techniki analizy danych można zaliczyć do jednej z następujących kategorii: (i) zapytania i raporty, (ii) analiza OLAP oraz (iii) eksploracja danych. O ile zapytania i raporty są zaliczane do klasycznych aplikacji baz danych, o tyle analiza OLAP oraz eksploracja danych reprezentują nowy kierunek rozwoju współczesnych systemów analitycznych, określany jako odkrywanie wiedzy z danych.

Podstawowe klasy technik analizy danych pokazano na rysunku 1. Zapytania i raporty są rodzajem analizy sterowanej całkowicie przez analityka, który musi sformułować konkretne zapytania bądź utworzyć aplikacje raportujące. Oznacza to, że analiza jest w pełni nakierowana na ten problem, który w danej chwili bada analityk, a uzyskane wyniki nie wykraczają poza ściśle określony, wcześniej zaplanowany, obszar.

Metody stosowane do odkrywania wiedzy z danych takie jak analiza OLAP i eksploracja danych charakteryzują się ograniczeniem wpływu analityka na proces analizy oraz uzyskiwane wyniki.

[1] J. Han, M. Kamber, *DataMining: Concepts and Techniques*, wyd. 2, Morgan Kaufmann Publishers, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo 2006, s. 2.



Rysunek 1. Spektrum technik analizy danych

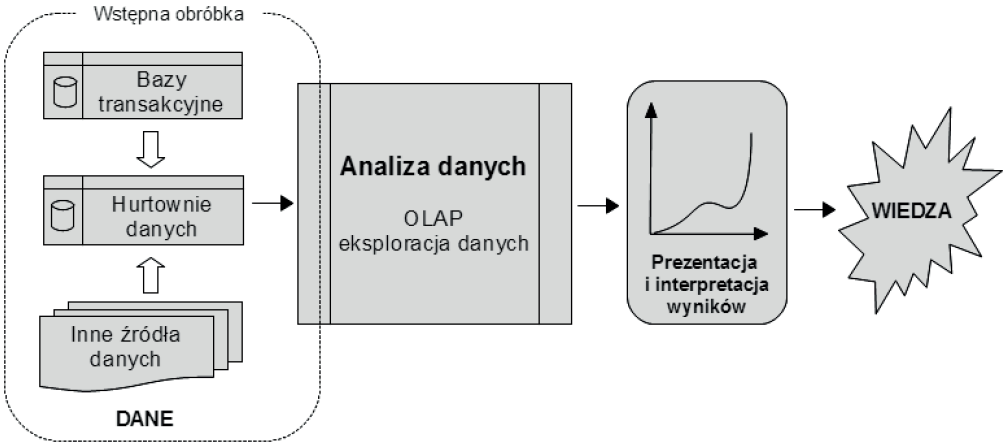
Źródło: opracowanie własne.

W przypadku przetwarzania analitycznego on-line, określanego także jako analiza OLAP (*On-Line Analytical Processing*), dane są organizowane w dużych zbiorach w taki sposób, aby mogły być efektywnie przeglądane i analizowane w odniesieniu do pewnych kategorii semantycznych nazywanych wymiarami. Typowe wymiary to: czas, produkt, odbiorca itp. Dane można przeglądać nie tylko „wzdłuż" różnych wymiarów, ale także na różnych poziomach hierarchii w ramach pojedynczego wymiaru. Na przykład wymiar odbiorca można rozpatrywać na następujących poziomach hierarchii: państwo–region–miasto–klient. Liczba wymiarów może być różna, jednak dość powszechnie dane zorganizowane wielowymiarowo przedstawia się w postaci sześcianów nazywanych kostkami OLAP lub kostkami danych (*data cubes*). Aplikacja OLAP pozwala wspomagać system podejmowania decyzji w zakresie tzw. objaśniania zachowań, przez możliwość stopniowego przechodzenia do kolejnych raportów, których zakres i postać są na bieżąco definiowane przez analityka.

Współczesne systemy wspomagania decyzji wymagają zastosowania coraz bardziej zaawansowanych narzędzi analitycznych, które umożliwiłyby konstruowanie modeli czy też odkrywanie wcześniej nieznanych wzorców i trendów na podstawie danych. Liczne prace prowadzone w tym zakresie doprowadziły do powstania nowej dziedziny zastosowań metod informatyki znanej pod nazwą „eksploracja danych" (*data mining*). Dyscyplina ta obejmuje obszerną klasę metod i technik analizy dużych zbiorów danych obserwacyjnych w celu odkrywania nieznanych wcześniej zależności i wzorców oraz formułowania ich w postaci zwięzłych podsumowań i uogólnień, które byłyby zarówno zrozumiałe, jak i przydatne dla analityka[2]. Biorąc pod uwagę realizowane funkcje oraz charakter uzyskiwanych wyników, można rozróżnić (i) eksplorację opisową (*descriptive*

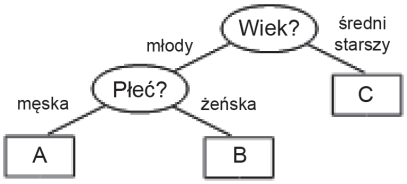
[2] D. Hand, H. Manilla, P. Smyth, *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 2005, s. 35.

mining) oraz (ii) eksplorację predykcyjną (*predictive mining*). Eksploracja opisowa służy do tego, aby uzyskać ogólny wgląd w istotne właściwości danych w analizowanym zbiorze. Natomiast eksploracja predykcyjna zakłada wykorzystanie istniejących danych do przewidywania (predykcji) nieznanych wartości w nowych (przyszłych) danych. W porównaniu ze wspomnianymi wcześniej sposobami analizowania danych, takimi jak przetwarzanie analityczne on-line, a zwłaszcza zapytania i raporty, metody eksploracji danych charakteryzują się znacznym ograniczeniem udziału analityka zarówno w formułowaniu problemu, jak i w samym przebiegu analizy. Można stwierdzić, że przebieg i wyniki analizy w znaczącym stopniu zależą od samych danych, a cechą charakterystyczną metod eksploracji danych jest to, że pozwalają one dotrzeć do wiedzy na tyle „ukrytej" w danych, że analityk nie jest w stanie z góry przewidzieć jej istnienia.



Rysunek 2. Proces odkrywania wiedzy na podstawie danych
Źródło: opracowanie własne.

Rysunek 2 przedstawia ogólny schemat wieloetapowego procesu odkrywania wiedzy na podstawie danych. Do analizy wykorzystuje się różne źródła danych, jak: bazy danych, pliki tekstowe i inne. Dane źródłowe są z reguły poddawane wstępnej obróbce i organizowane w specjalne bazy przeznaczone do analizy, nazywane hurtowniami danych. Wyselekcjonowane dane z hurtowni są poddawane analizie. Wynik stanowią różnego rodzaju modele i wzorce, zwykle prezentowane w przejrzystej formie z użyciem grafiki. Ocena i interpretacja wyników jest zadaniem analityka, a odkrywana wiedza to zależności, modele, reguły lub wzorce wynikające z danych. Przykładami są: klasyfikatory, segmenty (klastry) oraz reguły asocjacyjne.



Rysunek 3. Przykład klasyfikatora w postaci drzewa decyzyjnego
Źródło: opracowanie własne.

Typowym przykładem analizy z zastosowaniem eksploracji danych jest klasyfikacja. Zadanie polega na uzyskaniu metodą uczenia maszynowego pewnego modelu klasyfikującego, który umożliwiłby predykcję (przewidywanie) tego, do jakiej klasy należy zaliczyć analizowane obiekty. Model powstaje na podstawie zbioru danych uczących (trenujących), których przynależność do określonej klasy jest z góry ustalona, dlatego w przypadku klasyfikacji mówi się o tzw. uczeniu nadzorowanym (*supervised learning*). Każdy obiekt (przypadek) w zbiorze danych uczących jest charakteryzowany przez pewną liczbę atrybutów opisowych, a jednocześnie przypisany do określonej klasy, wskazanej przez atrybut klasyfikujący (zmienną celu). Zbiór uczący powinien zawierać możliwie reprezentatywną grupę przykładów tak, aby skuteczność klasyfikatora była zadowalająca. Jest ona sprawdzana na specjalnie przygotowanym do tego celu zbiorze testującym, po czym model spełniający oczekiwane wymagania może być użyty do klasyfikowania nowych przypadków. Na rysunku 3 pokazano intuicyjną ilustrację klasyfikatora mającego postać drzewa decyzyjnego.

Prace nad rozwojem systemów baz i hurtowni danych doprowadziły do integracji ich podstawowych funkcjonalności z algorytmami eksploracji danych. Skutkuje to m.in. redukcją kosztów oraz wzrostem efektywności eksploracji danych[3].

3. Wspomaganie administracji podatkowej

Do kluczowych zadań administracji podatkowej należy maksymalnie efektywne wykorzystanie własnych, często ograniczonych zasobów do uzyskania możliwie najwyższego stopnia przestrzegania przepisów podatkowych. Typowym postępowaniem są kontrole podatkowe, które pośrednio powodują naturalny wzrost

[3] T. Morzy, *Eksploracja danych. Metody i algorytmy*, Wydawnictwo Naukowe PWN, Warszawa 2013, s. 6.

stopnia przestrzegania prawa, a bezpośrednio generują dodatkowe wpływy z podatków. W obu przypadkach następuje zmniejszenie istniejącej różnicy pomiędzy naliczoną kwotą podatków a rzeczywistymi wpływami. Z tego powodu kontrole mają zasadnicze znaczenie dla skuteczności egzekwowania prawa podatkowego, pomagając uzyskać założone wpływy i zapewnić stabilność kraju i regionu. Zarządzanie procesem przeprowadzania kontroli wymaga podejmowania wielu ważnych decyzji. Istotnym problemem staje się opracowanie skutecznej strategii wyboru (typowania) kontroli podatkowych. Chodzi w szczególności o odpowiedź na pytania typu: „Czy kierować się wysokością zadeklarowanego podatku, czy też branżą podatnika?", „Jak rozłożyć istniejące zasoby w zakresie kontrolowania pomiędzy różne typy podatków?" itp. Niektóre typy podatku mogą wykazywać większe rozmiary korekty przypadające na poszczególne kontrole, podczas gdy inne mogą być związane z większym odsetkiem stwierdzonych nieprawidłowości. Kontrola jest procesem wieloetapowym: od wyboru przez wywiady, orzeczenie i negocjacje do pobrania, a w niektórych przypadkach wyegzekwowania podatku. Na każdym z tych etapów muszą być podejmowane decyzje, które mogą poprawić bądź pogorszyć globalne efekty kontroli.

Metody typowania kontroli mogą być wielorakie: od prostego wyboru losowego przez bardziej zaawansowany wybór na podstawie reguł aż do złożonych procedur selekcji wykorzystujących metody statystyczne i techniki eksploracji danych. Strategie mogą różnić się od siebie nie tylko w zależności od typu podatku, lecz nawet wewnątrz tego samego typu. Można np. przyjąć strategię, według której dokonuje się grupowania (segmentacji) podatników w ramach określonego typu podatku, a następnie stosuje odrębne reguły wyboru wobec każdego segmentu.

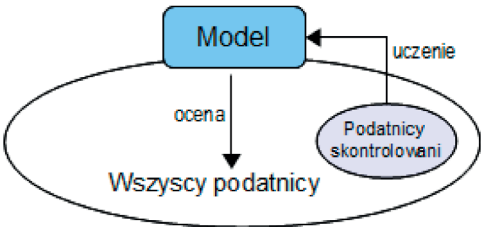
W przypadku zaawansowanych metod analitycznych opartych na technikach eksploracji danych wykorzystuje się odpowiednio przygotowane systemy hurtowni danych. Integrują one dane pochodzące zarówno ze źródeł własnych, jak i zewnętrznych i umożliwiają użycie wielu różnorodnych aplikacji, obejmujących np. analizę trendu, wykrywanie niezgodności czy przewidywanie dochodu. Umożliwia to organom kontrolującym uzyskiwanie praktycznej pomocy m.in. w zakresie takich zagadnień:

- W jaki sposób dzielić istniejące zasoby pomiędzy różne typy podatków?
- Którzy podatnicy powinni być kontrolowani w pierwszej kolejności?
- Jakich efektów należy spodziewać się po kontroli określonego typu?
- Które branże są związane z wyższym odsetkiem nieprawidłowości?

Organy podatkowe mają dostęp do ogromnych ilości danych dotyczących podatników, które mogą być wykorzystane do odkrywania wiedzy wspomagającej

procesy przeprowadzania kontroli. W trakcie selekcji można przeszukiwać dostępne źródła celem zidentyfikowania podatników o zadanym profilu. Profile są określane przez ekspertów na podstawie pojedynczego atrybutu (np. kod branży) bądź kombinacji atrybutów (np. podatnicy w określonym sektorze handlu, w których przypadku występuje zadana proporcja pomiędzy wielkością obrotu a wysokością zadeklarowanego podatku). Techniki eksploracji danych działają analogiczne, z tym że na znacznie większą skalę. Przy ich pomocy organy podatkowe mogą analizować dane pochodzące od setek tysięcy podatników w celu identyfikacji wspólnych (podobnych) atrybutów, a następnie utworzenia profili reprezentujących różne typy zachowań. Można w szczególności tworzyć profile deklaracji podatkowych z wysokim uzyskiem, aby kontrolerzy mogli skupić się na nowych deklaracjach o podobnych wartościach atrybutów. W ten sposób, dzięki zastosowaniu technik i narzędzi eksploracji danych, organy administracji podatkowej mogą użyć swoich danych do zrozumienia, przeanalizowania oraz predykcji niezgodnych z przepisami zachowań podatników.

Jak wspomniano w punkcie 2, w praktyce eksploracji danych często stosuje się metodę klasyfikacji. Otrzymywane w wyniku uczenia nadzorowanego modele mogą być następnie użyte do predykcji pewnych zjawisk i procesów. W przypadku poszukiwania właściwej strategii typowania kontroli podatników celem jest próba oszacowania (przewidywania), które z potencjalnych kontroli dają większe prawdopodobieństwo, że zakończą się korzystnym dla kontrolującego wyrównaniem podatku. W typowym przypadku modele takie generują pewną wartość liczbową, która stanowi ocenę (*score*) prawdopodobnego wyniku kontroli. Można przy tym założyć, że wysoka ocena sugeruje dużą wartość wyrównania, podczas gdy niska ocena wskazuje na brak takich przesłanek. Modele oceniające (*scoring models*) zyskały pewną popularność dzięki zdolności do operowania tysiącami atrybutów w dużych populacjach.



Rysunek 4. Model oceniający otrzymany na podstawie danych historycznych

Źródło: opracowano na podstawie: Raport Elite Analytics, http://www.spss.ch/upload/1122641565_Improving%20tax%20administration%20with%20data%20mining.pdf (data odczytu 24.11.2013).

Przykładem modułu analitycznego wspomagającego strategię wyboru podatników do kontroli jest opracowany przez Elite Analytics system *Audit Select*, w którym zastosowano model oceniający otrzymany na podstawie danych historycznych[4]. Koncepcję otrzymania oraz użycia modelu pokazano schematycznie na rysunku 4. Uczenie odbywa się z wykorzystaniem zbioru danych historycznych, zawierającego wyniki kontroli przeprowadzonych w przeszłości. Na podstawie tych danych model „poznaje" zależności pomiędzy pewnymi atrybutami charakteryzującymi podatnika a wynikami kontroli. Model oceniający systemu *Audit Select* dotyczy kontroli płatników podatku obrotowego, a do utworzenia profilu podatnika wykorzystuje pięć następujących źródeł danych: (i) dane ewidencyjne (kod branży, typ przedsiębiorstwa, adres); (ii) dane o podatku obrotowym z ostatnich 4 lat; (iii) dane o innych podatkach; (iv) dane o pracownikach i zarobkach; (v) wyniki już przeprowadzonych kontroli.

Otrzymany w wyniku uczenia model może zostać użyty do analizy nowych zeznań podatkowych oraz przypisania im wyznaczonej oceny. Ocena dostarczona przez *Audit Score* może być następnie wykorzystana przez osoby planujące kontrole.

4. Niekomercyjne pakiety analityczne

Rozwój algorytmów odkrywania wiedzy z danych łączy się z koniecznością opracowania oprogramowania, które umożliwia ich zastosowanie w praktyce. Złożoność procesu odkrywania wiedzy wymaga automatycznego i skoordynowanego przeprowadzenia wielu operacji. W fazie początkowej należy pobrać dane ze zbiorów źródłowych i, w typowym przypadku, dokonać ich wstępnego przetwarzania (*preprocessing*), selekcji oraz ewentualnego zapisu do hurtowni. Następnie należy zastosować odpowiednie algorytmy eksploracji danych, a także przeprowadzić wizualizację i interpretację wyników. W przeszłości proces ten wymagał częstokroć użycia różnych narzędzi informatycznych realizujących kolejne etapy, co w pewnym stopniu komplikowało pracę.

W ostatnim czasie powstało wiele programów, zarówno komercyjnych, jak i dostępnych na zasadach *open source*, których zadaniem jest ułatwienie użytkownikom projektowania pełnego procesu odkrywania wiedzy oraz jego późniejszą, ewentualną, modyfikację. Narzędzia te są stale rozwijane i coraz częściej nie

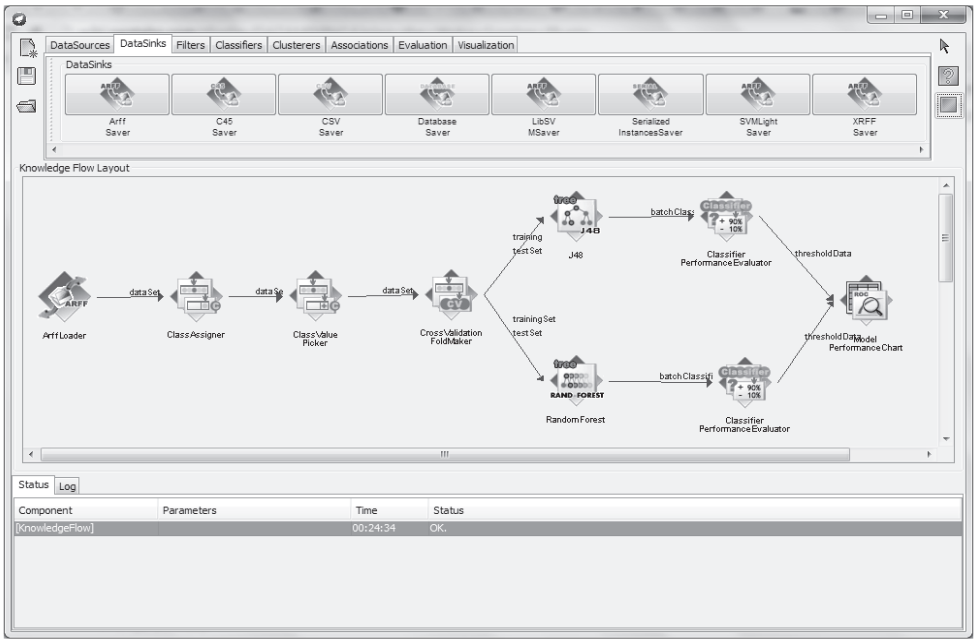
[4] http://www.spss.ch/upload/1122641565_Improving%20tax%20administration%20with%20data%20mining.pdf (data odczytu 24.11.2013).

wymagają już od użytkowników znajomości języków programowania czy specjalistycznych funkcji, ale wykorzystują interfejs graficzny i technikę „przeciągnij i upuść". Dzięki temu mogą z nich skutecznie korzystać także użytkownicy, którzy nie mają szerokiej wiedzy informatycznej. Umożliwia to stosowanie aplikacji tego typu w instytucjach czy firmach, które nie mają wydzielonego działu informatycznego, a potrzebują przetwarzania swoich zbiorów danych w celu pozyskania wiedzy użytecznej do wspomagania decyzji.

Drugim aspektem związanym z szerszym wykorzystaniem systemów eksploracji danych i odkrywania wiedzy jest koszt ich wdrożenia. Komercyjne pakiety, takie jak m.in. IBM SPSS Modeler czy Oracle Data Mining, przeznaczone dla dużych i średnich instytucji oraz przedsiębiorstw, ze względu na cenę oraz wymagania sprzętowe nie są dostępne dla mniej zasobnych organizacji. Z pomocą w takich przypadkach przychodzą rozwiązania niekomercyjne, rozpowszechniane na licencji otwartego oprogramowania, które pod względem funkcjonalności często nie odbiegają od profesjonalnych pakietów, a wydatki związane z ich wdrożeniem są znacznie mniejsze. Wśród dostępnych bezpłatnie pakietów oprogramowania wspomagających proces odkrywania wiedzy można wymienić m.in.: Orange[5], KNIME[6] oraz Weka[7]. Pakiety te można pobrać ze stron domowych projektów, a ponadto są one utrzymywane i stale rozwijane przez społeczność internetową.

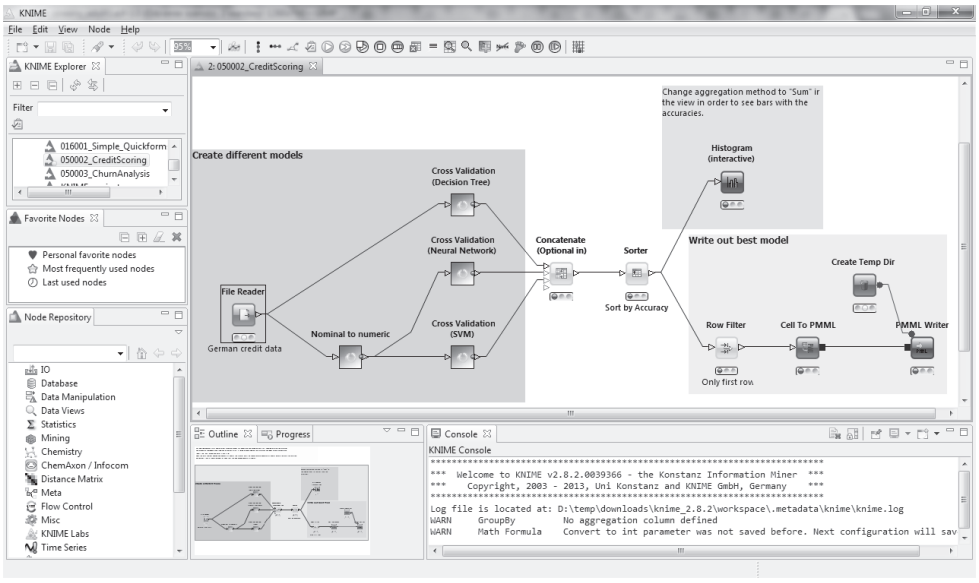
Wśród wymienionego oprogramowania największą liczbą funkcji dysponuje pakiet KNIME oraz niewiele mu ustępująca Weka, choć z tego powodu są one zalecane dla bardziej zaawansowanych użytkowników. Dzięki szerokiej palecie funkcjonalności, które udostępniają m.in. w zakresie przetwarzania danych wejściowych, raportowania oraz możliwości dodawania nowych modułów rozszerzających w języku Java, mogą być wykorzystywane do rozwiązywania najbardziej zaawansowanych problemów w dziedzinie eksploracji danych i odkrywania wiedzy. Na rysunkach 5 i 6 pokazano przykładowe projekty odkrywania wiedzy udostępnione wraz z pakietami Weka i KNIME.

[5] http://orange.biolab.si (data odczytu 24.11.2013).
[6] http://www.knime.org (data odczytu 24.11.2013).
[7] http://www.cs.waikato.ac.nz/ml/weka/index.html (data odczytu 24.11.2013).



Rysunek 5. Przykład użycia pakietu Weka w procesie odkrywania wiedzy

Źródło: opracowanie własne.



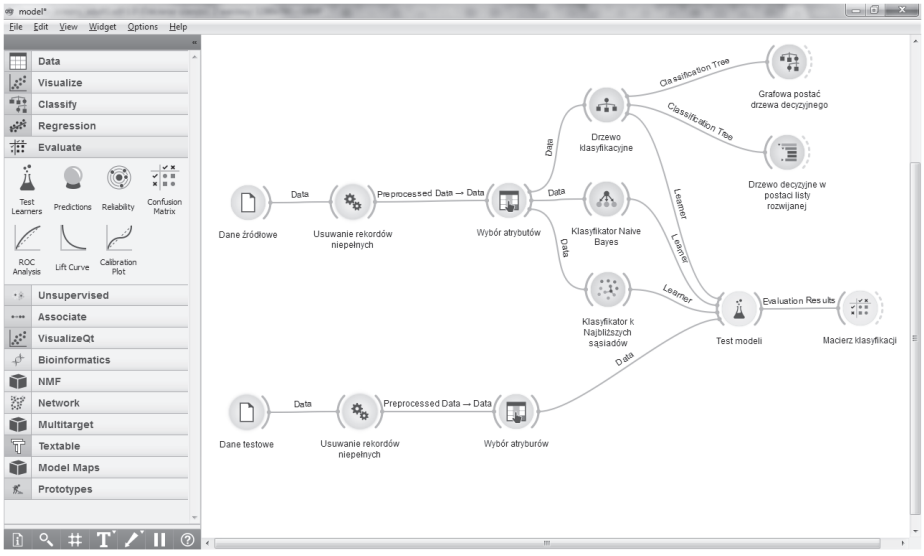
Rysunek 6. Przykład zastosowania pakietu KNIME

Źródło: opracowanie własne.

Pod względem przejrzystości interfejsu oraz zapewnienia odpowiedniej funkcjonalności najlepszym rozwiązaniem, zarówno dla osób początkujących, jak i bardziej zaawansowanych użytkowników, wydaje się pakiet Orange. Oprogramowanie Orange zostało opracowane oraz zaimplementowane w języku Python w Laboratorium Bioinformatyki na Uniwersytecie w Lubljanie[8]. Ma ono budowę modułową i dostarcza wielu funkcji przydatnych w procesie odkrywania wiedzy. Do dyspozycji użytkownika oddano również wiele tzw. wtyczek, które otwierają możliwości użycia pakietu Orange w dziedzinie bioinformatyki, analizy dokumentów tekstowych, sieci itp. Modułowa architektura pozwala na tworzenie własnych elementów w języku Python i wdrożenie ich w środowisku Orange, co daje szerokie możliwości w dostosowaniu pakietu do własnych potrzeb.

5. Predykcja dochodów ludności

Na rysunku 7 pokazano przykład typowego zadania w dziedzinie eksploracji danych, jakim jest problem doboru algorytmu klasyfikacji danych oraz jego prawidłowych parametrów, zrealizowanego przy pomocy oprogramowania Orange.



Rysunek 7. Przykład zastosowania pakietu Orange w klasyfikacji danych

Źródło: opracowanie własne.

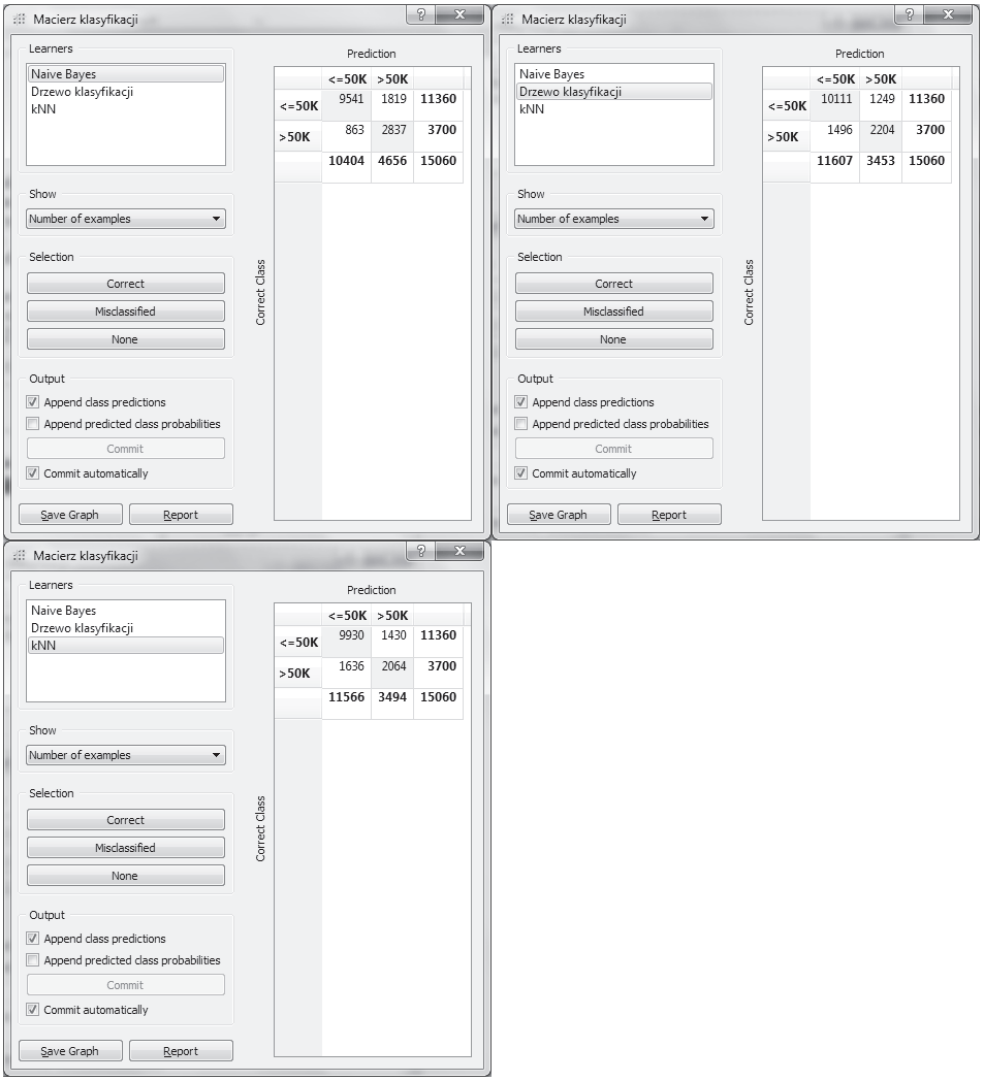
[8] J. Demšar, T. Curk, A. Erjavec, *Orange: Data Mining Toolbox in Python*; „Journal of Machine Learning Research" 2013, vol. 14, August, s. 2349–2353.

W przykładzie wykorzystano rzeczywiste dane spisowe ze Stanów Zjednoczonych[9]. Na źródłową bazę danych składają się 48 842 rekordy opisane 14 atrybutami, które zostały wybrane z danych zebranych przez rządową agencję odpowiedzialną za spis ludności w USA (*Bureau of the Census*) w 1994 r. Poszczególne wartości atrybutów przechowują dane o osobach powyżej 17 roku życia, m.in. takie jak: wiek, płeć, poziom edukacji, stan cywilny, rodzaj zatrudnienia itp., oraz dodatkowo informację o liczbie osób o takich parametrach w bazie spisu. Atrybutem klasyfikującym jest informacja, czy dana grupa osób osiągnęła dochód powyżej progu 50 000 USD. Do testowania algorytmów klasyfikacyjnych użyto dodatkowego zbioru zawierającego 16 281 rekordów, które stanowią 33% podstawowego zbioru danych uczących.

W pokazanym na rysunku 7 przykładzie wykorzystano wiele elementów (*widgets*) dostępnych w pakiecie Orange. W pierwszej kolejności użyto elementu *File* z kategorii *Data*, który posłużył do pobrania danych uczących oraz testowych z plików typu CSV (*Comma Separated Values* – wartości rozdzielone przecinkiem). W kolejnym kroku usunięto z danych (element *Preprocess*) rekordy, które zawierały wartości niepełne, redukując tym samym zbiór danych uczących do 30 162, natomiast zbiór testowy do 15 060 rekordów. Następnie dokonano selekcji atrybutów do analizy (element *Select Attributes*) oraz powiązano dane uczące z trzema typami algorytmów klasyfikacji (drzewem klasyfikacyjnym, naiwnym klasyfikatorem Bayesa oraz metodyką *k* najbliższych sąsiadów), które są dostępne jako elementy kategorii *Classify*. Następnym etapem w opisywanym procesie było sprawdzenie dokładności, z jaką zbudowane modele klasyfikują dane. W tym celu połączono elementy reprezentujące poszczególne rodzaje klasyfikacji oraz dane testowe z modułem testującym *Test Learners*, dostępnym w grupie *Evaluate*. Wyniki przeprowadzonych testów powiązano z macierzą błędów (inaczej nazywaną macierzą klasyfikacji) *Confusion Matrix*, która wskazuje rozbieżności pomiędzy wartościami rzeczywistymi a zwróconymi przez model. Na rysunku 8 pokazano macierze klasyfikacji dla trzech testowanych algorytmów.

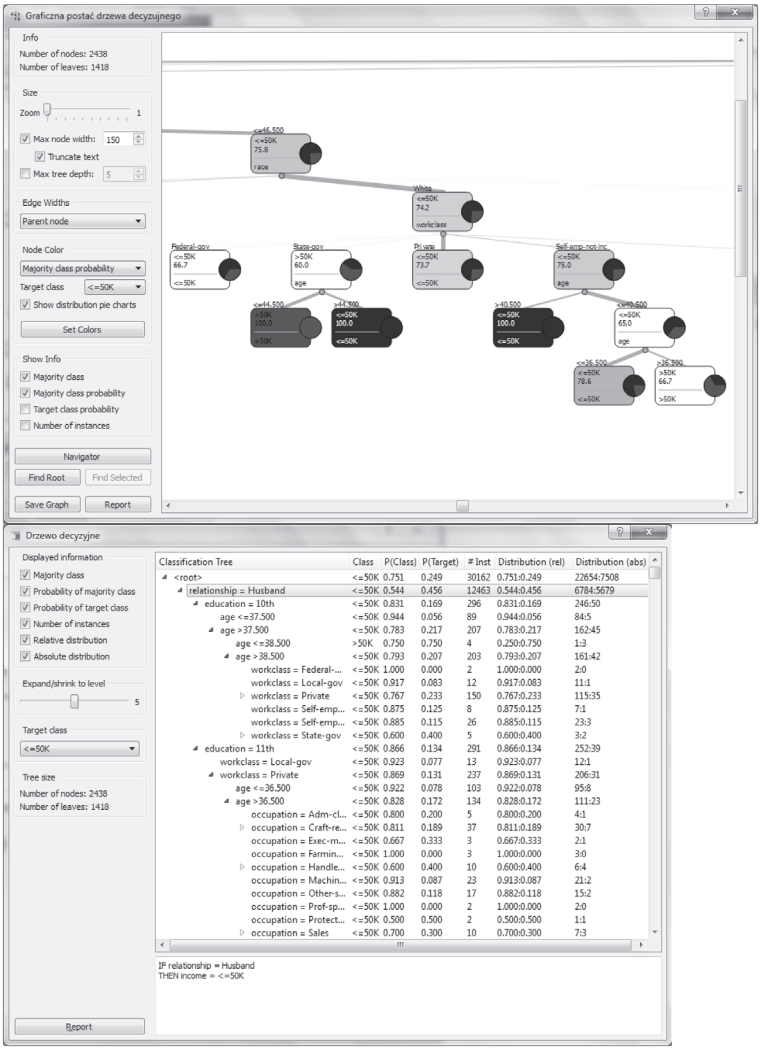
Testy wykazały, że najwyższą dokładność klasyfikacji uzyskał algorytm naiwnego klasyfikatora Bayesa – 82,19% (12 378 przypadków prawidłowo sklasyfikowanych), a pozostałe algorytmy odpowiednio: drzewo klasyfikacyjne 81,77% (12 315 przypadków) i k najbliższych sąsiadów 79,64% (11 994 przypadki). Na potrzeby prezentacji możliwości pakietu w dziedzinie wizualizacji drzew klasyfikacyjnych wykorzystano dodatkowo elementy *Classification Tree Graph* oraz *Classification Tree Viewer*, których działanie pokazano na rysunku 9. Wizualizacja może odbywać się na dwa sposoby – za pomocą grafowej reprezentacji drzewa lub w formie rozwijanej listy wierzchołków.

[9] http://archive.ics.uci.edu/ml/datasets/Adult (data odczytu 24.11.2013).



Rysunek 8. Wyniki testowania dla trzech różnych algorytmów: naiwny klasyfikator Bayesa, drzewo klasyfikacji oraz metoda *k* najbliższych sąsiadów

Źródło: opracowanie własne.



Rysunek 9. Wizualizacja drzew klasyfikacyjnych w pakiecie Orange

Źródło: opracowanie własne.

Warto podkreślić fakt, że cały proces realizacji przedstawionego eksperymentu wymagał jedynie wybrania z menu właściwych elementów – akcji, ustawienia odpowiednich parametrów wejściowych (np. wskazania ścieżki do plików danych) oraz połączenia ich w odpowiedniej kolejności i utworzenia wymaganego przepływu danych. Pokazany przykład prezentuje tylko wybrane możliwości pakietu Orange w dziedzinie odkrywania wiedzy. Dostępnych jest także wiele innych algorytmów, z których warto wymienić: grupowanie, generowanie reguł asocjacyjnych oraz regresję.

Podsumowując, można stwierdzić, że pakiet Orange dzięki dużej funkcjonalności, jak również prostocie użytkowania może stanowić solidną podstawę do projektowania procesów odkrywania wiedzy na podstawie danych oraz wizualizacji uzyskanych wyników. Z kolei w przypadku bardziej zaawansowanych systemów, wymagających wielu działań na danych, oraz specjalistycznych metod raportowania dobrym wyborem będzie oprogramowanie KNIME oraz Weka.

6. Podsumowanie

W pracy rozważono problematykę zastosowania zaawansowanych metod odkrywania wiedzy na podstawie danych w systemach przeznaczonych do jednostek administracji publicznej. Scharakteryzowano różne klasy stosowanych obecnie metod analizy danych, ze szczególnym uwzględnieniem technik eksploracji danych, stanowiących ważny element w wieloetapowym procesie odkrywania wiedzy z danych. Analiza danych pochodzących z różnorodnych i coraz bardziej zasobnych źródeł stanowi obiecujący kierunek rozwoju współczesnych systemów informacyjnych w jednostkach administracji publicznej. Może w tym pomóc duża dostępność i względna prostota użytkowania w pełni profesjonalnych pakietów analitycznych rozpowszechnianych na zasadach niekomercyjnych.

Bibliografia

1. Demšar J., Curk T., Erjavec A., *Orange: Data Mining Toolbox in Python*, „Journal of Machine Learning Research" 2013, vol. 14, August.
2. Han J., Kamber M., *DataMining: Concepts and Techniques*, wyd. 2, Morgan Kaufmann Publishers, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo 2006.
3. Hand D., Manilla H., Smyth P., *Eksploracja danych*, Wydawnictwo Naukowo-Techniczne, Warszawa 2005.
4. Morzy T., *Eksploracja danych. Metody i algorytmy*, Wydawnictwo Naukowe PWN, Warszawa 2013.

Źródła sieciowe

1. http://archive.ics.uci.edu/ml/datasets/Adult (data odczytu 24.11.2013).
2. http://orange.biolab.si (data odczytu 24.11.2013).
3. http://www.cs.waikato.ac.nz/ml/weka/index.html (data odczytu: 24.11.2013).
4. http://www.knime.org (data odczytu 24.11.2013).
5. http://www.spss.ch/upload/1122641565_Improving%20tax%20administration%20with%20data%20mining.pdf (data odczytu 24.11.2013).

* * *

Using modern KDD methods along with non-commercial analytical packages to support decisions in public administration

Summary

The paper concerns the potential perspectives of the use of knowledge discovery from data (KDD) methods together with non-commercial analytical tools in public administration. The main categories of data analysis techniques were identified and characterised with a special focus on the predictive models for data driven decision support. Some non-commercially available packages supporting the overall KDD process were enumerated and characterised in order to show their professional capabilities available not only for big companies.

Keywords: public administration, data analysis, knowledge discovery from data, data warehouse, data mining, prediction, scoring models