Przegląd Medyczny Uniwersytetu Rzeszowskiego
Rzeszów 2005, 4, 385–391

Damian Mazur

Systemy informatyczne wspomagające procesy medyczne

Z Politechniki Rzeszowskiej, Zakładu Podstaw Elektrotechniki i Informatyki

Systemy informatyczne dzięki badaniom z zakresu sztucznej inteligencji wspomagają procesy medyczne. W pracy pokazano, że pewne czynności konsultacyjne mogą być odwzorowane w postaci algorytmu matematyczno-logicznego. Tym samym możliwe było zbudowanie programów z bazą wiedzy zdolnych do samodzielnego rozwiązywania problemów, stawiania diagnoz i formułowania porad. Programy te nazywamy systemami ekspertowymi, gdyż imitują czynności intelektualne dotychczas wykonywane przez najlepszych specjalistów. Bardzo duże wolumeny danych zawierają w sobie użyteczną wiedzę, ukrytą pod postacią wzorców, trendów, regularności i wyjątków. Tradycyjne metody analizy danych tracą zastosowanie, nie będąc w stanie przetworzyć ogromnych ilości danych gromadzonych przez współczesne organizacje. W pracy przedstawiono system informatyczny oparty o technikę data mining, która jest dyscypliną łączącą takie dziedziny, jak: systemy baz danych, statystykę, systemy wspomagania decyzji, sztuczną inteligencję, uczenie maszynowe, wizualizację danych, przetwarzanie równoległe i rozproszone, i wiele innych.

Słowa kluczowe: procesy medyczne, sztuczna inteligencja, systemy ekspertowe, bazy danych

Informatics systems supporting medical processes

Medical processes are supported nowadays by informatics systems thanks to the advancement in the field of Artificial Intelligence (AI). This paper shows that certain consultation routines can be mapped into a mathematical/logical algorithm. Thus it is possible to design programs with knowledge base, able of solving problems independently, diagnosing and formulating advisory suggestions. These programs are called expert systems, as they imitate intellectual tasks which used to be performed, until now, by the best professionals. Very big data volumes include very useful knowledge, hidden in the form of patterns, trends, regularities and exceptions. Traditional data analysis methods are useless here, as they are not able to process vast amounts of data collected by contemporary organisations. This paper presents an informatics system based on data mining technique, which is the branch of informatics science, linking such knowledge areas as: database systems, statistics, support systems for decision making, artificial intelligence, machine learning, data visualisation, parallel and distributed computing, and many others..

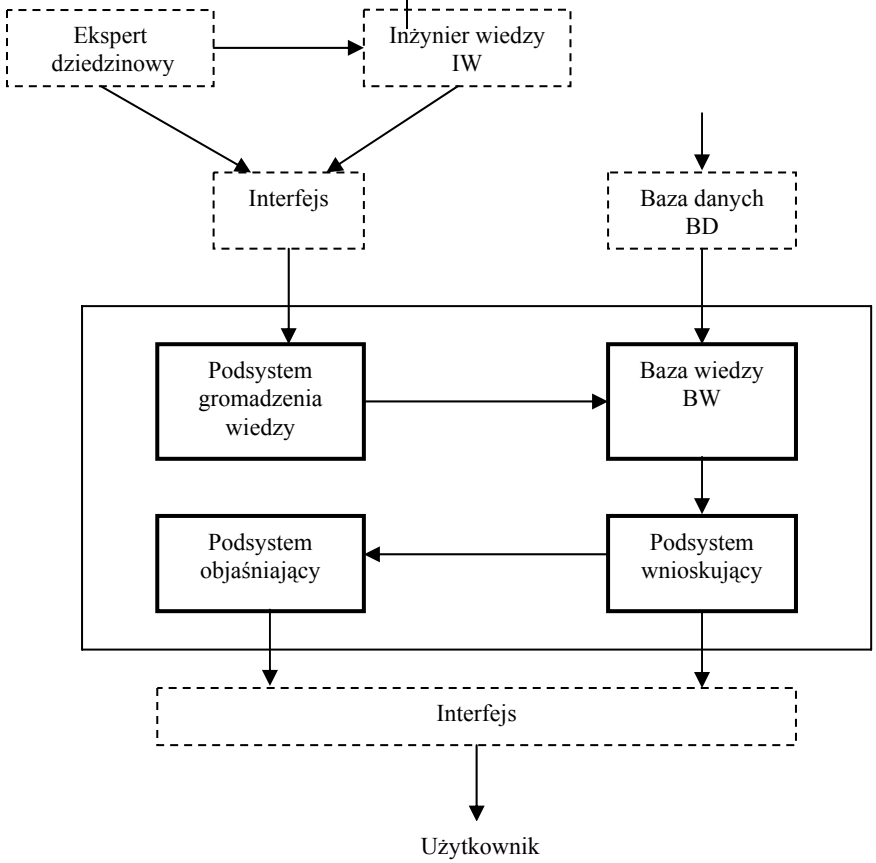
Key words: medical processes, Artifical Intelligence, experts system, data base.

WPROWADZENIE

Obserwowane w ostatnich latach upowszechnienie się systemów baz danych w praktycznie każdej dziedzinie ludzkiej działalności doprowadziło do gwałtownego wzrostu ilości danych, które są gromadzone i zapisywane w postaci cyfrowej. Systemy baz danych są obecne niemal we wszystkich dziedzinach i stanowią niezbędny element infrastruktury informatycznej. Ewolucja systemów baz danych była podyktowana z jednej strony rosnącymi wymaganiami dotyczącymi funkcjonalności baz danych, a z drugiej strony, koniecznością obsługiwania coraz większych kolekcji danych.

Danych medycznych pacjentów, wyników badań laboratoryjnych, obrazów z aparatury me-

dycznej (CT, NMR, USG, EEG, EMG, itd.) z dnia na dzień przybywa. Dane gromadzone są w różnych



RYC. 1. Schemat systemu ekspertowego

miejscach, o różnych porach, na wielu komputerach, niekoniecznie ze sobą połączonych, zapisywane na wielu platformach systemowych. Z powyższych informacji wynika, iż dane są rozproszone i nie zawierają kompletnej informacji o pacjencie. Oczywiście są one w jakiś sposób przekazywane pomiędzy oddziałami szpitala czy innymi jednostkami medycznymi, lecz informacja przekazywana jest w postaci ręcznie zapisanych zaleceń, zleceń leczenia czy dalszej diagnostyki przekazanej ustnie. Stąd propozycja ujednolicenia systemów, połączenia i dostępności danych na serwerze, wspierana przez systemy wspomagania decyzji, systemy wykorzystujące metody sztucznej inteligencji pomagające analizować i klasyfikować informacje zawarte w rozproszonych bazach danych.

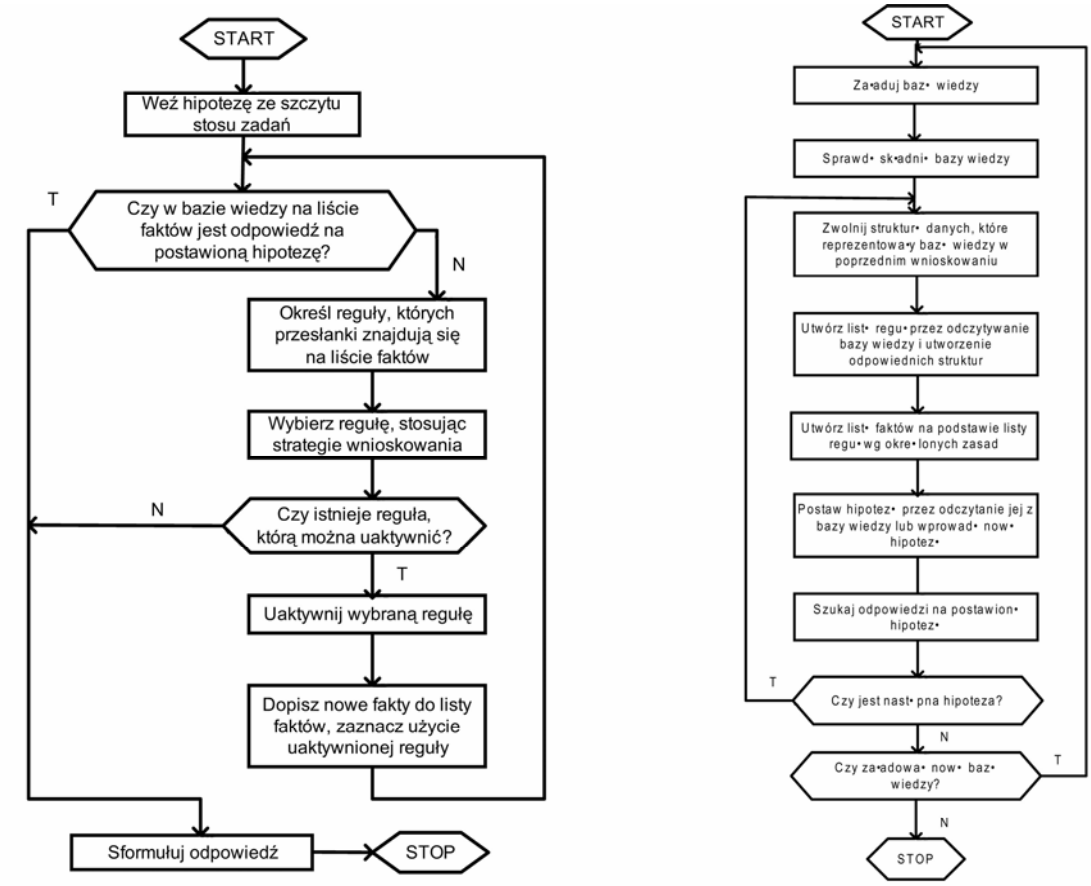
1. SYSTEM EKSPERTOWY WSPOMAGAJĄCY PROCESY MEDYCZNE

Zrąb strukturalny systemu ekspertowego tworzą moduły, które są niezbędne do wykonania najważniejszej funkcji programu, tzn. do generowania porad. Strukturę tę tworzą następujące elementy: mechanizm wnioskujący, baza wiedzy i moduł współpracy z użytkownikiem.

Mechanizm wnioskujący nadzoruje działanie systemu ekspertowego, identyfikuje wiedzę niezbędną do rozwiązania danego problemu, a następnie określa sposób przetwarzania informacji. W sensie informatycznym mechanizm ten jest zespołem rozkazów (procedur, programów), który steruje tokiem wnioskowania, łącząc informacje przekazywane od użytkownika oraz fakty zawarte w bazie wiedzy z regułami wnioskowania.

Główne elementy mechanizmu wnioskującego to układ planowania rozwiązań, egzekutor i kompilator. Układ planowania rozwiązań analizuje problem i formułuje strategię wnioskowania. Na tej podstawie jest sporządzana lista reguł przewidzianych do wykonania wraz ze wskazaniem priorytetów. Egzekutor ściąga te reguły z bazy wiedzy i wiąże je z tymi faktami, które pełnią rolę przesłanek. Wnioski cząstkowe są zapisywane na tablicy, która jest prowizoryczną bazą danych rejestrującą wszystkie pośrednie wyniki procesu wnioskowania. Tok wnioskowania jest nadzorowany przez kompilator, który sprawdza, czy rozwiązania cząstkowe są sensowne i nie-

sprzeczne. Pod względem informatycznym baza wiedzy systemu ekspertowego jest bazą danych, czyli zbiorem rekordów wraz z oprogramowaniem zarządzającym zapisem,



RYC. 2. Algorytm wnioskowania w przód i wnioskowania wstecz

przechowywaniem, dostępem, selekcją i modyfikacją tych danych. Można uznać, że cechą dystynktywną bazy wiedzy (odróżniającą ją od klasycznych baz danych) jest przyjęcie zadania jako elementarnej jednostki informacyjnej. Zarządzanie bazą wiedzy obejmuje funkcje definiowania i strukturalizacji bazy oraz przekształcania i porządkowania informacji. W praktyce, baza wiedzy może być uzupełniana dodatkowymi wiadomościami i procedurami, niezbędnymi do realizacji procesu wnioskowania.

2. OPIS TECHNIKI DATA MINING I JEJ ZASTOSOWANIE

Eksploracja danych (ang. *data mining)* to techniki, metody i algorytmy umożliwiające odkrywanie użytecznej wiedzy w bardzo dużych bazach danych [11]. Oracle Data Mining to nowy komponent serwera bazy danych, oferujący użytkownikom wiele różnych metod eksploracji danych. Użytkownicy mogą wykorzystywać takie funkcje eksploracji, jak: *klasyfikacja i predykcja, regresja, określanie ważności atrybutów, grupowanie obiektów podobnych, znajdowanie reguł asocjacyjnych, oraz eksploracja dokumentów tekstowych* [4].

Współczesne bazy danych stoją przed licznymi wyzwaniami: skalowalność, efektywność, bogata funkcjonalność, pojemność, to niezbędne cechy dobrego systemu zarządzania bazą danych.

Według corocznego raportu Winter Corporation [Wint05] rozmiar największej operacyjnej bazy danych w roku 2005 osiągnął 23 TB (Land Registry for England and Wales), podczas gdy rozmiar największej hurtowni danych przekroczył 100 TB (Yahoo!). Inne przykłady gwałtownie zwiększających się baz danych obejmują bazy danych informacji naukowych (projekt poznania ludzkiego genomu, informacje astronomiczne), dane strategiczne (informacje związane z bezpieczeństwem narodowym) oraz komercyjne kolekcje danych (dane POS, dane o ruchu internetowym, dane o transakcjach elektronicznych). Nieustanny wzrost rozmiarów baz danych jest skutkiem postępu w technikach pozyskiwania i składowania

informacji. Niestety, postęp ten nie jest równoważony przez zdolność do analizy pozyskanych danych. Gigantyczne rozmiary współczesnych kolekcji danych skutecznie uniemożliwiają jakiekolwiek próby ręcznej analizy zgromadzonych informacji. Z drugiej strony, większość repozytoriów zawiera użyteczną wiedzę ukrytą w danych pod postacią trendów, regularności, korelacji, czy osobliwości. W ostatnich latach zaproponowano wiele metod automatycznego i półautomatycznego pozyskiwania wiedzy z ogromnych wolumenów danych [3].

Metody te określa się mianem eksploracji danych (ang. *data mining*) lub odkrywania wiedzy w bazach danych (ang. *knowledge discovery in databases*). Wiedza odkryta w procesie eksploracji danych może zostać wykorzystana do wspomagania procesu podejmowania decyzji, predykcji przyszłych zdarzeń, czy określania efektywnych strategii biznesowych.

Techniki eksploracji danych można ogólnie podzielić na dwie zasadnicze kategorie. Techniki predykcyjne starają się, na podstawie odkrytych wzorców, dokonać uogólnienia i przewidywania (np. wartości nieznanego atrybutu, zachowania i cech nowego obiektu, itp.). Przykładami zastosowania technik predykcyjnych mogą być: identyfikacja docelowych grup, ocena ryzyka lub oszacowanie prawdopodobieństwa [1].

Techniki deskrypcyjne mają na celu wykorzystanie wzorców odkrytych w danych do spójnego opisu danych i uchwycenia ogólnych cech danych. Typowe przykłady technik deskrypcyjnych obejmują odkrywanie grup podobnych, identyfikację osobliwości występujących w danych.

Inny podział technik eksploracji danych jest związany z charakterem danych wejściowych. W przypadku technik uczenia nadzorowanego (ang. *supervised learning*) dane wejściowe zawierają tzw. zbiór uczący, w którym przykładowe instancje danych są powiązane z prawidłowym rozwiązaniem [2]. Na podstawie zbioru uczącego dana technika potrafi „nauczyć się" odróżniać przykłady należące do różnych klas, a zdobyta w ten sposób wiedza może być wykorzystana do formułowania uogólnień dotyczących przyszłych instancji problemu. Najczęściej spotykanymi technikami uczenia nadzorowanego są techniki klasyfikacji (drzewa decyzyjne, algorytmy bazujące na najbliższych sąsiadach, sieci neuronowe, statystyka bayesowska), oraz techniki regresji [6, 7].

Drugą klasą technik eksploracji danych są techniki uczenia bez nadzoru (ang. *unsupervised learning*), gdy algorytm nie ma do dyspozycji zbioru uczącego. W takim przypadku algorytm eksploracji danych stara się sformułować model najlepiej pasujący do obserwowanych danych. Przykłady technik uczenia bez nadzoru obejmują techniki analizy skupień (ang. *clustering*) [9], samoorganizujące się mapy, oraz algorytmy maksymalizacji wartości oczekiwanej (ang. *expectation-maximization*). Terminy „eksploracja danych" i „odkrywanie wiedzy w bazach danych" są często stosowane wymiennie, choć drugi termin posiada dużo szersze znaczenie. Odkrywanie wiedzy to cały proces akwizycji wiedzy, począwszy od selekcji danych źródłowych, a skończywszy na ocenie odkrytych wzorców. Zgodnie z tą definicją, eksploracja danych oznacza zastosowanie konkretnego algorytmu odkrywania wzorców na wybranych danych źródłowych, i stanowi jeden z etapów składowych całego procesu odkrywania wiedzy.

Na cały proces składają się: sformułowanie problemu, wybór danych, czyszczenie danych, integracja danych, transformacja danych, eksploracja danych, wizualizacja i ocena odkrytych wzorców, i wreszcie zastosowanie wzorców. Postać uzyskanych wzorców zależy od zastosowanej techniki eksploracji danych. Poniżej przedstawiono opisy najpopularniejszych technik eksploracji.

Z konieczności nie jest to lista wyczerpująca, uwzględniono tylko te metody eksploracji danych, które zostały zaimplementowane w pakiecie Oracle Data Mining.

2.1. Reguły asocjacyjne

Odkrywanie reguł asocjacyjnych polega na znalezieniu w dużej kolekcji zbiorów korelacji wiążącej współwystępowanie podzbiorów elementów. Znalezione korelacje są prezentowane jako reguły postaci X□Y (wsparcie, ufność), gdzie X i Y są rozłącznymi zbiorami elementów, wsparcie oznacza częstotliwość występowania zbioru X□Y w kolekcji zbiorów, zaś ufność reprezentuje prawdopodobieństwo warunkowe P(Y|X). Duży wysiłek badawczy włożono w opracowywanie efektywnych algorytmów odkrywania reguł asocjacyjnych. Najbardziej znane przykłady takich algorytmów to Apriori, FreeSpan oraz Eclat. Problem znajdowania wzorców sekwencji polega na znalezieniu w bazie danych sekwencji, podsekwencji występujących częściej niż zadany przez użytkownika próg częstości, zwany progiem minimalnego wsparcia (ang. *minsup*) [5]. Dodatkowo, użytkownik może sformułować ograniczenia dotyczące maksymalnych interwałów

czasowych między kolejnymi wystąpieniami elementów sekwencji.

Klasyfikacja (ang. *classification*) – jest jedną z najpopularniejszych technik eksploracji danych. Polega na stworzeniu modelu, który umożliwia przypisanie nowego, wcześniej niewidzianego obiektu, do jednej ze zbioru predefiniowanych klas. Model umożliwiający takie przypisanie nazywa się klasyfikatorem. Klasyfikator dokonuje przypisania na podstawie doświadczenia nabytego podczas trenowania i testowania na zbiorze uczącym. W trakcie wieloletnich prac prowadzonych nad klasyfikatorami i ich zastosowaniem w statystyce, uczeniu maszynowym, czy sztucznej inteligencji, zaproponowano bardzo wiele metod klasyfikacji. Najczęściej stosowane techniki to klasyfikacja bayesowska, klasyfikacja na podstawie k najbliższych sąsiadów, drzewa decyzyjne, sieci neuronowe, sieci bayesowskie, czy algorytmy SVM (ang. *support vector machines*) [8]. Popularność technik klasyfikacji wynika przede wszystkim z faktu szerokiej stosowalności tego modelu wiedzy. Klasyfikatory mogą być wykorzystane do oceny ryzyka związanego z wyznaczeniem prawdopodobieństwa przejścia z jednej grupy do drugiej. Podstawową wadą praktycznie wszystkich technik klasyfikacji jest konieczność starannego wytrenowania klasyfikatora i trafnego wyboru rodzaju klasyfikatora w zależności od charakterystyki przetwarzanych danych.

Techniką podobną do klasyfikacji jest **regresja** (ang. *regression*). Różnica między dwiema technikami polega na tym, że w przypadku klasyfikacji przewidywana wartość jest kategoryczna, podczas gdy w regresji celem modelu jest przewidzenie wartości numerycznej.

Analiza skupień (ang. *clustering*) to popularna technika eksploracji danych polegająca na dokonaniu takiego partycjonowania zbioru danych wejściowych, które maksymalizuje podobieństwo między obiektami przydzielonymi do jednej grupy i, jednocześnie, minimalizuje podobieństwo między obiektami przypisanymi do różnych grup. Sformułowanie problemu przypomina problem klasyfikacji, jednak należy podkreślić istotne różnice. Analiza skupień jest techniką uczenia bez nadzoru, stąd nieznane jest „poprawne" przypisanie obiektów do grup, często nie jest znana nawet „poprawna" liczba grup. Jeśli porównywane obiekty leżą w przestrzeni metrycznej, wówczas do określenia stopnia podobieństwa między obiektami wykorzystuje się funkcję odległości zdefiniowaną w danej przestrzeni. Zaproponowano wiele różnych funkcji odległości, do najpopularniejszych należy rodzina odległości Minkowskiego (odległość blokowa, odległość euklidesowa, odległość Czebyszewa), odległość Hamminga (wykorzystywana dla zmiennych zakodowanych binarnie), odległość Levenshteina (zwana odległością edycji), czy popularna w statystyce odległość Mahalanobisa. W przypadku, gdy porównywane obiekty nie leżą w przestrzeni metrycznej, zazwyczaj definiuje się specjalne funkcje określające stopień podobieństwa między obiektami. Specjalizowane funkcje podobieństwa istnieją dla wielu typowych dziedzin zastosowań, takich jak porównywanie stron internetowych, porównywanie sekwencji DNA, czy porównywanie danych opisanych przez atrybuty kategoryczne.

Metody analizy skupień najczęściej dzieli się na metody hierarchiczne i metody partycjonujące. Pierwsza klasa metod dokonuje iteracyjnego przeglądania przestrzeni i w każdej iteracji buduje grupy obiektów podobnych na podstawie wcześniej znalezionych grup.

Rozróżnia się tutaj metody aglomeracyjne (w każdej iteracji dokonują złączenia mniejszych grup) i metody podziałowe (w każdej iteracji dokonują podziału wybranej grupy na mniejsze podgrupy). Druga klasa metod analizy skupień to metody partycjonujące, które od razu znajdują docelowe grupy obiektów. Do najbardziej znanych algorytmów analizy skupień należą algorytmy k-średnich, samoorganizujące się mapy, CURE (ang. Clustering Using REpresentatives) i wiele innych.

3. ODKRYWANIE CECH

Wiele przetwarzanych zbiorów danych charakteryzuje się bardzo dużą liczbą wymiarów (atrybutów). Niczyjego zdziwienia nie budzą tabele z danymi wejściowymi zawierające setki atrybutów kategorycznych i numerycznych. Niestety, efektywność większości metod eksploracji danych gwałtownie spada wraz z rosnącą liczbą przetwarzanych wymiarów. Jednym z rozwiązań tego problemu jest wybór cech (ang. *feature selection*) lub odkrywanie cech (ang. *feature extraction*). Pierwsza metoda polega na wyselekcjonowaniu z dużej liczby atrybutów tylko tych atrybutów, które posiadają istotną wartość informacyjną. Druga metoda polega na połączeniu aktualnie dostępnych atrybutów i stworzeniu ich liniowych kombinacji w celu zmniejszenia liczby wymiarów i uzyskania nowych źródeł danych. Wybór i generacja nowych atrybutów może odbywać się w sposób nadzorowany (wówczas wybierane są atrybuty, które umożliwiają dyskrymi-

nację między wartościami atrybutu decyzyjnego), lub też bez nadzoru (wówczas najczęściej wybiera się atrybuty powodujące najmniejszą utratę informacji).

Większość komercyjnie dostępnych systemów eksploracji danych wykorzystuje relacyjną bazę danych tylko jako źródło danych. Analizowane dane są pobierane z bazy danych (co najczęściej wiąże się z wysokim kosztem transferu bardzo dużych wolumenów danych), a następnie przetwarzane i analizowane w zewnętrznym narzędziu. ODM dokonuje wszystkich procesów składających się na odkrywanie wiedzy wewnątrz relacyjnej bazy danych. Oznacza to, że selekcja danych, przygotowanie i transformacja danych, tworzenie modelu i wreszcie zastosowanie modelu odbywają się w środowisku systemu zarządzania bazą danych. ODM wspiera zarówno metody uczenia nadzorowanego, jak i uczenia bez nadzoru. Najpopularniejszą metodą uczenia nadzorowanego jest bez wątpienia klasyfikacja. ODM zawiera cztery algorytmy klasyfikujące: naiwny klasyfikator Bayesa, adaptatywną sieć Bayesa, algorytm indukcji drzew decyzyjnych oraz algorytm SVM. Pokrótce omówimy te algorytmy. Naiwny klasyfikator Bayesa jest bardzo prostym, a jednocześnie efektywnym w praktyce klasyfikatorem. Podstawą działania klasyfikatora jest twierdzenie Bayesa, które określa prawdopodobieństwo warunkowe hipotezy hi przy zaobserwowaniu danych D jako $P(hi|D)=P(D|hi)*P(hi)/P(D)$.

Jeśli przyjąć, że hi reprezentuje przypisanie do i-tej klasy, wówczas prawdopodobieństwo przypisania obiektu D do i-tej klasy można znaleźć na podstawie prawdopodobieństwa a posteriori $P(D|hi)$, reprezentującego prawdopodobieństwo posiadania przez D pewnych cech jeśli D rzeczywiście należy do i-tej klasy. Prawdopodobieństwo to określa się na podstawie danych zawartych w zbiorze trenującym poprzez analizę cech obiektów rzeczywiście należących do i-tej klasy, zaś dodatkowym uproszczeniem jest założenie o warunkowej niezależności atrybutów (tzn. założenie, że w ramach danej klasy rozkład wartości każdego atrybutu jest niezależny od pozostałych atrybutów).

Podstawową zaletą naiwnego klasyfikatora Bayesa jest prostota i szybkość, tworzenie i wykorzystanie modelu są liniowo zależne od liczby atrybutów i obiektów. Adaptatywna sieć Bayesa to algorytm probabilistyczny, który generuje model klasyfikacji w postaci zbioru połączonych cech (ang. *network feature*). W zależności od wybranego trybu algorytm może wyprodukować płaski model stanowiący odpowiednik naiwnego klasyfikatora Bayesa (w takim modelu każda cecha połączona będzie zawierać jeden atrybut-predyktor i jedną klasę docelową), model składający się z jednej cechy połączonej (w ramach cechy znajdzie się wiele związanych ze sobą atrybutów-predyktorów, taki model odpowiada drzewu decyzyjnemu), lub model składający się z wielu cech połączonych [10]. Przy wykorzystaniu modelu z jedną cechą połączoną algorytm może zaprezentować model w postaci prostych reguł klasyfikacyjnych.

Algorytm indukcji drzew decyzyjnych buduje model w postaci drzewa, którego węzły odpowiadają testom przeprowadzanym na wartości pewnego atrybutu, gałęzie odpowiadają wynikom testów, a liście reprezentują przypisanie do pewnej klasy. Algorytm tworzy drzewo decyzyjne w czasie liniowo zależnym od liczby atrybutów-predyktorów. O kształcie drzewa decydują takie parametry jak: kryterium wyboru punktu podziału drzewa, kryterium zakończenia podziałów, czy stopień scalania drzewa. Wielką zaletą drzew decyzyjnych jest fakt, że wygenerowany model można łatwo przedstawić w postaci zbioru reguł, co pomaga analitykom zrozumieć zasady działania klasyfikatora i nabrać zaufania do jego decyzji [12].

Algorytm SVM (ang. *Support Vector Machines*) może być wykorzystany zarówno do klasyfikacji, jak i regresji [11]. Algorytm SVM dokonuje transformacji oryginalnej przestrzeni, w której zdefiniowano problem klasyfikacji, do przestrzeni o większej liczbie wymiarów. Transformacja jest dokonywana w taki sposób, że po jej wykonaniu w nowej przestrzeni obiekty są separowalne za pomocą hiperpłaszczyzny (taka separacja jest niemożliwa w oryginalnej przestrzeni). Głównym elementem transformacji jest wybór funkcji jądra (ang. *kernel function*) odpowiedzialnej za odwzorowanie punktów do nowej przestrzeni. W przypadku regresji algorytm SVM znajduje w nowej przestrzeni ciągłą funkcję, w sąsiedztwie której mieści się największa możliwa liczba obiektów. Algorytmy SVM wymagają starannego doboru funkcji jądra i jej parametrów. Doświadczenia wskazują, że algorytmy te bardzo dobrze się sprawdzają w praktycznych zastosowaniach, takich jak: rozpoznawanie pisma odręcznego, klasyfikacja obrazów i tekstu, czy analiza danych biomedycznych. Implementacja algorytmów SVM w ODM posiada kilka interesujących cech. Oferuje między innymi mechanizm aktywnego uczenia się, którego celem jest wybór

z danych źródłowych tylko najbardziej wartościowych przykładów i trenowanie modelu tylko na wyselekcjonowanych danych wejściowych. W trakcie trenowania algorytm SVM dokonuje automatycznego próbkowania. Wybór funkcji jądra i jej parametrów następuje automatycznie. Algorytm SVM może także zostać wykorzystany do wykrywania osobliwości. Stosuje się wówczas specjalną wersję algorytmu, tzw. SVM z jedną klasą docelową, który pozwala identyfikować nietypowe obiekty. Poza klasyfikacją, regresją i wykrywaniem osobliwości ODM oferuje jeszcze jedną technikę uczenia nadzorowanego, którą jest wybór cech. Jak wspomniano wyżej, czas tworzenia modelu klasyfikacji najczęściej zależy liniowo od liczby atrybutów. Jest również możliwe, że duża liczba atrybutów wpłynie ujemnie na efektywność i dokładność modelu poprzez wprowadzenie niepożądanego szumu informacyjnego. Algorytm wyboru cech (ang. *feature selection*) umożliwia wybór, spośród wielu atrybutów, podzbioru atrybutów które są najbardziej odpowiednie do przewidywania klas docelowych. Wybór cech jest więc często wykorzystywany jako technika wstępnego przetworzenia i przygotowania danych, przed rozpoczęciem trenowania klasyfikatora.

ODM zawiera także jeden specjalizowany algorytm BLAST (ang. *Basic Local Alignment Search Tool*), służący do wyszukiwania dopasowania białek w bioinformatycznych bazach danych.

PODSUMOWANIE

Eksploracja danych to nowa i niezwykle prężnie rozwijająca się dziedzina. Wiedza odkryta w dużych wolumenach danych może być traktowana jako metadane, oferujące wzbogacony wgląd w dane. Metody eksploracji danych wymagają specjalizowanych narzędzi umożliwiających budowanie modeli, testowanie modeli, stosowanie modeli do nowych danych. Oracle Data Mining jest pierwszym systemem do tego stopnia integrującym eksplorację danych z systemem zarządzania bazą danych. Wykorzystując ODM można przeprowadzić prawie cały proces odkrywania wiedzy, począwszy od selekcji i wstępnego przetwarzania danych źródłowych, aż po wygenerowanie wzorców. Ścisła integracja technik eksploracji danych z bazą danych umożliwia wykorzystanie technik eksploracji w aplikacjach, ułatwia pielęgnację aplikacji, oferuje ogromnie wzbogaconą funkcjonalność aplikacji.

PIŚMIENNICTWO

1. Almuallin H., Dietterich T.G.: *Efficient algorithms for identifying relevant features*, in Proc. of 9 Canadian Conference on Artificial Intelligence, 38, Vancouver BC, 1992.
2. Aha D.: *Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms,* International Journal of Man-Machine Studies 36(2), 267.
3. Agrawal R., Imielinski T., Swami A.: *Mining association rules between sets of items in large databases*, Proc. of 1993 ACM SIGMOD International Conference on Management of Data,Washington D.C., May 26-28 1993, 207, ACM Press, 1993.
4. Agrawal R., Srikant R.: *Fast Algorithms for Mining Association Rules*, Proc. of 1994 International Conference on Very Large Databases VLDB, Santiago de Chile, September 12–15, 487. Morgan Kaufman, 1994.
5. Agrawal R., Srikant R.: *Mining sequential patterns*, In Proc. of the 11th International Conference on Data Engineering, Taipei, Taiwan, 1995.
6. Breiman L., Friedman J.H., Olshen R.A., Stone C.J.: *Classification and regression trees*, Wadsworth, 1984.
7. Bigus J.P.: *Data mining with neutral networks*, McGraw Hill, 1996.
8. Bolstad W.M.: *Introduction to Bayesian statistics,* Wiley-Interscience, 2004.
9. Everitt B.S., Landau S., Leese M.: *Cluster analysis*, Arnold Publishers, 2001.
10. Gidofalvi G., Pedersen T.B.: *Spatio-temporal Rule Mining: Issues and Techniques*, in Proc. of the 7th International Conference on Data Warehousing and Knowledge Discovery Da-WaK 2005, Copenhagen, Denmark, 2005.
11. Guha S., Rastogi R., Shim K.: CURE: *An Efficient Clustering Algorithm for Large Databases*, In Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 73-84, New York, 1998.
12. Provost F., Fawcett T.: *Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions*, in Proc. of the 3rd International Conference onKnowledge Discovery and Data Mining, Huntington Beach, CA, 1997.

Damian Mazur
ul. Niezapominajek 33
35-604 Rzeszów
Tel. 604 459 775
e-mail: dmazu@wp.pl