ANALIZA DANYCH EKSPERYMENTALNYCH Z WYKORZYSTANIEM DATA MINING

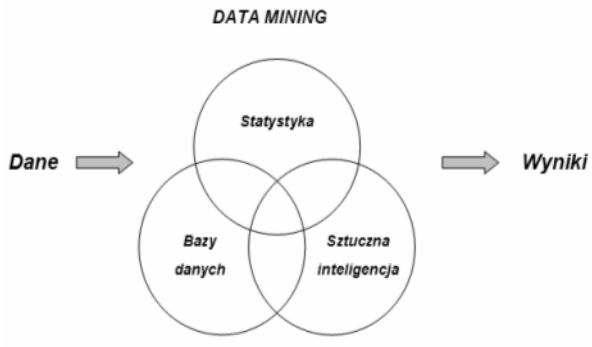
Jędrzej TRAJER, Monika JANASZEK

Streszczenie: W artykule przedstawiono koncepcję analizy z wykorzystaniem metod *Data Mining* (drążenia danych). Nakreślono założenia i metodykę analizy danych z wykorzystaniem zaawansowanych technologii informacyjnych: sztucznej inteligencji, statystyki i baz danych. Przedstawiono też przykład analizy danych eksperymentalnych (cechy fizyczne) w identyfikacji produktu roślinnego, w oparciu o opracowany system komputerowy.

Słowa kluczowe: badania eksperymentalne, analiza danych, *Data Mining*, sztuczna inteligencja.

1. Wprowadzenie

Tworzenie wiedzy naukowej w tradycyjnym podejściu polega na poszukiwaniu, za pomocą metod naukowych, prawidłowości i cech rządzących rzeczywistością oraz na ich naukowym wyjaśnieniu [3]. Metody naukowe realizowane są zwykle w czterech krokach: obserwacja, sformułowanie hipotezy wyjaśniającej (model), prognostyczne użycie hipotezy oraz weryfikacja hipotezy. Kluczowym etapem umożliwiającym opis i sformalizowanie zjawiska są badania empiryczne. Obserwacja zjawisk (ich parametrów) pozwalają naukowcom formułować różne przypuszczenia, co do ich przyczyn i charakteru. Wyniki badań eksperymentalnych stanowią podstawę do formalnego opisu (modelu), który zawsze trzeba zweryfikować, aby uznać za obowiązujący w określonych warunkach. Takie postępowanie wymaga dużego doświadczenia, które pozwala wysunąć określone hipotezy. Następnie nasze przypuszczenia nakreślają plan prowadzący do rozwiązania. Inne podejście, bazujące na analizie danych z wykorzystaniem zaawansowanych technologii informacyjnych: sztucznej inteli-gencji, statystyki i baz danych, prezentują metody bardziej znane w literaturze pod angielską nazwą *Data Mining*, niż polską drążenie danych (rys.1). W przypadku drążenia danych, do poszukiwania zależności między zmiennymi i ich opisu, klasyfikacji obiektów lub odkrywania reguł, wykorzystuje się przede wszystkim technologie informacyjne i wiedzę ekspertów. Uzyskane wyniki stanowią przesłankę do wyciągania wniosków i formułowania hipotez wyjaśniających oraz do ich weryfikacji [6]. Wyniki pochodzące z drą-



Rys.1. Technologie informacyjne *Data Mining*

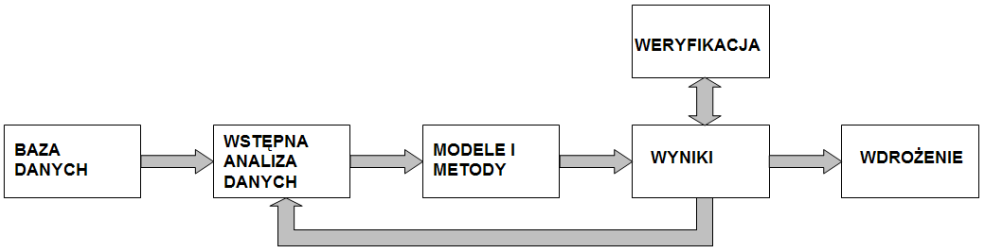
żenia danych dostarczają nam czegoś nowego o badanej rzeczywistości, co wcześniej nie było znane i nie było ku temu żadnych przesłanek, w takim znaczeniu jest to, więc odkrywaniem wiedzy. Kluczowym wymogiem w osiągnięciu celu analizy jest znajomość i właściwe wykorzystanie metod drążenia danych. Metody te zastosowane wspólnie mogą dostarczyć wiedzy niemożliwej do osiągnięcia w inny sposób.

2. *Data Mining* w procesie wydobywanie wiedzy z danych

Data Mining (dążenie danych) jest procesem interaktywnym, który wymaga, aby w procesie wnioskowania można było wykorzystać intuicję i wiedzę ekspertów w połączeniu z wydajnością obliczeniową nowoczesnych technologii komputerowych. Strategia rozwiązywania problemów dokonywana jest zgodnie z etodyką przedstawioną w następujących etapach [2]:

- sformułowania problemu ze względu na charakter analizy (opis zależności, klasyfikacja, poszukiwanie rozwiązań według przyjętego kryterium itp.),
- zebranie i wybór danych,
- przygotowanie danych do analizy,
- wybór metody lub opracowanie modelu i dokonanie analizy,
- ocena wyników,
- ewentualne dokonanie korekty w analizie,
- wybór rozwiązania i jego wdrożenie.

Sposób postępowania podczas analizy z wykorzystaniem technologii informacyjnych prezentuje rysunek 2.



Rys. 2. Etapy informatycznego projektu *Data Mining*

Przetwarzanie danych zgodnie z powyższą metodyką prowadzi do odkrywania wiedzy, która sprowadza się do znalezienia i opisu związków między danymi reprezentującymi różne zmienne. Narzędzie to polecane jest szczególnie do analizy dużych zbiorów danych wielowymiarowych, pochodzących z badań eksperymentalnych, w tych dziedzinach, w których brak jest sformalizowanych opisów zjawisk.

Cechy charakterystyczne analizy z wykorzystaniem *Data Mining* to:

- skuteczność w dużych zbiorów danych o wielowymiarowej strukturze,
- wykorzystywanie zaawansowanych technologii informacyjnych: bazy danych, statystyka, sztuczna inteligencja,
- interaktywna analiza poparta wiedzą ekspercką,
- często brak jasno sformułowanego celu analizy.

Duże znaczenie ma wykorzystanie w analizach sztucznej inteligencji, która stanowi narzędzie alternatywne względem tradycyjnych systemów komputerowych, umożliwiając analizę zjawisk niemożliwych do zbadania w inny sposób. Działanie narzędzi sztucznej inteligencji oparte jest o analogie do procesów zachodzących w przyrodzie, a sposób rozwiązania jest zawsze oparty o zaimplementowany dla danej metody jeden schemat przetwarzania sygnałów. Niebezpieczeństwo wyciągnięcia błędnych wniosków z analizy wynika przede wszystkim z:

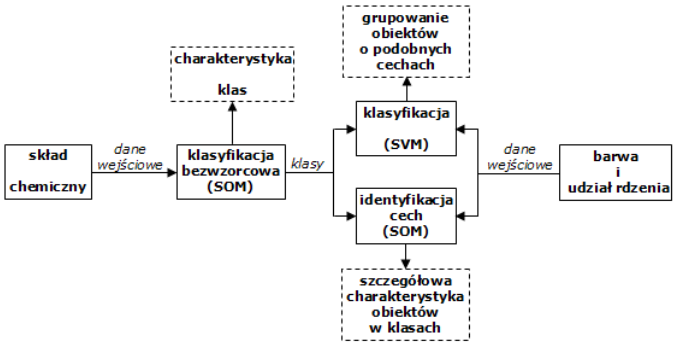
- przyjęcia błędnych założeń,
- złego przygotowania danych do analizy,
- błędnego wyboru analizy lub użycia złej metody obliczeniowej,
- braku weryfikacji lub walidacji uzyskanych wyników.

3. Przykład zastosowania *Data Mining* w analizie danych eksperymentalnych

Wiedza, inwencja i pomysłowość autora decydują o możliwościach wykorzystania różnych metod, które mogą tworzyć oryginalne i nowe rozwiązanie w postaci projektu informatycznego *Data Mining*. W tworzeniu projektu można uwzględnić następujące zalecenia, które mogą przyczynić się do znalezienia rozwiązania:

- w celu lepszego zrozumienia badanego systemu (zjawiska) empirycznego (dotyczy to w szczególności systemów złożonych) warto dokonać próby utworzenia modelu relacyjnego, opisującego strukturę tego systemu i relacje między obiektami, ukierunkuje to program całości badań i wskaże kolejność realizacji jego etapów,
- należy sprawdzić, czy występujący fakt lub prawidłowość nie są już opisane i wyjaśnione przez istniejącą lub analogiczną teorię, jeśli tak jest to istnieje możliwość wykorzystania analogicznych rozwiązań lub ich zaadaptowania,
- cel badań należy określić w taki sposób (sprowadzając go na przykład do rozwiązania odpowiedniej klasy zadania), aby można było w łatwo wybrać i wykorzystać dostępne narzędzia analizy.

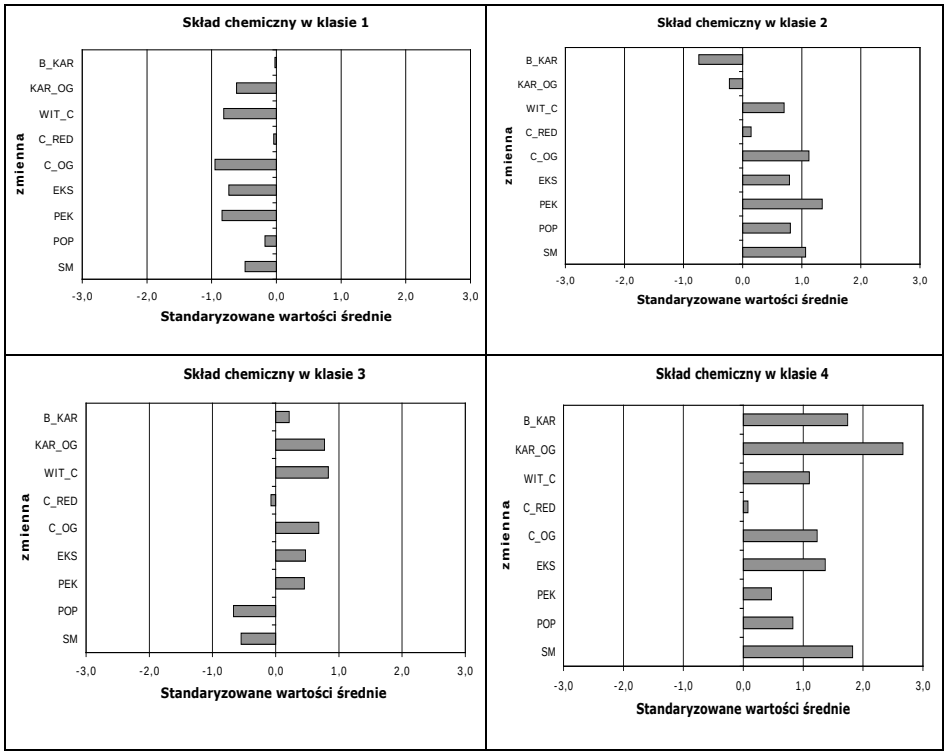
Wykorzystanie istniejących metod wymaga uwzględnienia ich specyfiki i ograniczeń, a także spełnienia założeń poczynionych w analizie.



Rys. 3. Diagram przepływu danych w informatycznym systemie klasyfikacji marchwi

Przykładem takim może być diagnostyka produktów rolniczych, które ze względu na duże zróżnicowanie muszą podlegać ciągłym ocenom w procesie dystrybucji i dalszego ich

wykorzystania. Celem prezentowanej analizy było stworzenie systemu klasyfikacji marchwi, wspomagającego proces identyfikacji cech korzeni i ich ocenę ze względu na przydatność przetwórczą [1]. Założono, iż punktem wyjścia do oceny przydatności przetwórczej marchwi będzie klasyfikacja bezwzorcowa, której kryterium stanowić będą cechy składu chemicznego (dane z badań składu chemicznego). Umożliwiło to pogrupowanie przypadków podobnych, które następnie można było zidentyfikować pod względem przydatności przetwórczej z wykorzystaniem klasyfikacji wzorcowej. W prezentowanym przykładzie postawiono cel dodatkowy: diagnostyka powinna przebiegać szybko i na podstawie łatwo mierzalnych cech, jakimi mogą być parametry barwy *R, G* i *B* (dane z komputerowej analizy obrazu). Badano, z zastosowaniem metod *Data Mining*, czy istnieją związki pomiędzy barwą i składem chemicznym. Wyniki z tych badań pozwoliły na opracowanie systemu, wspomagającego proces identyfikacji cech pojedynczych korzeni marchwi w oparciu o metody klasyfikacji. Schemat tego systemu przedstawiono na rys. 3.



Rys. 4. Zróżnicowanie składu chemicznego wyodrębnionych czterech klas

W realizacji systemu wykorzystano narzędzia drążenia danych zawarte w pakiecie *STATISTICA* firmy StatSoft Inc. [5] oraz *SAS* firmy SAS Institute Inc. [4], w szczególności wykorzystano do klasyfikacji bezwzorcowej sieci Kohonena (*SOM*), a do klasyfikacji wzorcowej metodę wektorów nośnych (*SVM*).

Pojedyncze skupienie stanowiło w dalszych etapach klasę obiektów, którą wykorzystano jako wzorzec. Dla każdego skupienia wyznaczono średnie wartości zmiennych, charakteryzujących skład chemiczny i odniesiono do średnich zawartości oznaczonych substancji w całym zbiorze danych. Przedstawione w ten sposób informacje umożliwiły dokonanie szczegółowej charakterystyki czterech klas, rys. 4.

Informacje o przydzieleniu obiektów do klas wykorzystano jako wzorzec dla sieci SVM, którą zaimplementowano w programie *STATISTICA*. Jej zadaniem była prawidłowa klasyfikacja korzeni marchwi do wyodrębnionych wcześniej klas, ale tylko na podstawie danych uzyskanych z cyfrowych obrazów tych korzeni, czyli barwy i udziału rdzenia. Dane wprowadzane do sieci podzielono na dwa zbiory: uczący i testowy. Zbiory te zawierały odpowiednio 75% i 25% obserwacji. Na wejścia sieci wprowadzono ciąg uczący, w którego skład weszły wyróżniki barwy po standaryzacji. Uczenie modelu SVM poprzedzono dziesięciokrotną walidacją krzyżową, w celu doboru optymalnych wartości parametrów uczenia sieci *SVM*. Proces uczenia sieci obejmował tysiąc iteracji, a warunkiem zatrzymania uczenia było osiągnięcie zmiany wartości błędu na poziomie 0,001. Trafność klasyfikacji danych w zbiorze uczącym wyniosła 91,43%, a w zbiorze testowym 94,29%, co świadczy o pozytywnie zweryfikowanym modelu.

Zmienne dotyczące barwy wprowadzone na wejścia sieci *SVM* wykorzystano również w procesie uczenia sieci Kohonena, która posłużyła jako narzędzie do wizualizacji danych i identyfikacji cech pojedynczych korzeni marchwi. Zmienną wyjściową była klasa przypisana w analizie skupień. Neurony w warstwie radialnej sieci utworzyły mapę cech o wymiarach 15×15. Dane podzielono na zbiory uczący i testowy, które zawierały odpowiednio 75% i 25% obserwacji. Sieć zaimplementowano w programie *STATISTICA*, a jej uczenie przebiegało według klasycznego algorytmu Kohonena. Po zakończeniu uczenia neuronom w warstwie radialnej nadano etykiety metodą *k* – najbliższych sąsiadów. Po nauczeniu sieci Kohonena poprawnego rozpoznawania struktury prezentowanych danych oznaczono na mapie obszary, w których skupiły się obiekty podobne pod względem składu chemicznego, bar-wy i udziału rdzenia, na rys. 5. pokazano mapę rozkładu cech składu chemicznego. Proponowany informatyczny system klasyfikacji marchwi ma charakter otwarty i może być rozbudowywany.

Rys. 5. Mapa rozkładu cech składu chemicznego w wyodrębnionych

W zależności od potrzeb możliwe jest wprowadzenie do niego dodatkowych danych, charakteryzujących cechy korzeni. System umożliwia wprowadzanie informacji o różnym

charakterze na podstawie, których można określić podobieństwo między obiektami, pochodzącymi z różnych partii surowca, różnych obszarów uprawnych lub będącymi przedstawicielami różnych odmian.

Struktura oraz liczba skupień w systemie może ulegać zmianie w miarę wprowadzania do niego nowych danych o korzeniach różnych odmian i pochodzących z różnych lat zbiorów. Zdefiniowanie wzorców surowca, odpowiadających określonym rodzajom przetworów, umożliwiłoby wykorzystanie proponowanego systemu nie tylko do różnicowania obiektów, ale do bezpośredniej oceny ich przydatności przetwórczej przez porównanie cech tych obiektów z wzorcem.

4. Podsumowanie

Przedstawiona koncepcja *Data Mining* stanowi nowe podejście w analizowaniu zjawisk i polecana jest w tych dyscyplinach naukowych, w których formalny opis zjawisk przysparza trudności. Dotyczy to w szczególności nauk przyrodniczych i ekonomicznych, które opierają się głównie na badaniach eksperymentalnych, a zasób danych, wykorzystywanych do analizy jest szeroki. Skuteczne przetwarzanie dużych zbiorów danych w celu wydobycia z nich wiedzy jest procesem interaktywnym, w którym wymaga się odpowiedniej współpracy pomiędzy wiedzą eksperta a wykorzystaniem zaawansowanych technologii informacyjnych. Takie podejście umożliwia dokonywanie odkryć naukowych, często z pozornie wyeksploatowanych danych, o czym świadczy nazwa tej informacyjnej technologii.

Literatura

1. Janaszek M.: Identyfikacja cech korzeni marchwi jadalnej z wykorzystaniem komputerowej analizy obrazu. Rozprawa doktorska, Wydział Inżynierii Produkcji SGGW, Warszawa, 2008.
2. Larose D.T.: Odkrywanie wiedzy z danych. PWN, Warszawa, 2006.
3. Pabis S.: Metodologia nauk empirycznych. Wydawnictwo Uczelniane Politechniki Koszalińskiej, Koszalin, 2009.
4. SAS Institute Inc. SAS 9.1 Companion for Windows. Cary, NC, USA: SAS Publishing, SAS Institute Inc, 2004.
5. StatSoft Inc. STATISTICA (data analysis software system), version 9.0. www.statsoft.com, 2009.
6. Tadeusiewicz R.: Data Mining jako szansa na relatywnie tanie dokonywanie odkryć naukowych poprzez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych. Statystyka i Data Mining w badaniach naukowych. StatSoft, Kraków, 2006.

Dr hab. inż. Jędrzej TRAJER, profesor SGGW
Dr inż. Monika JANASZEK
Katedra Podstaw Inżynierii
Wydział Inżynierii Produkcji
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
02-787 Warszawa, ul. Nowoursynowska 161
tel.: (22) 59 346 17
e-mail: jedrzej_trajer@sggw.pl