ODKRYWANIE WIEDZY W BAZACH DANYCH JAKO PROCES IDENTYFIKACJI MODELI DIAGNOSTYCZNYCH[1]

Dominik WACHLA

Politechnika Śląska, Katedra Podstaw Konstrukcji Maszyn
ul. Konarskiego 18A, 44 - 100 Gliwice, fax: (32) 237-13-60, e-mail: dwachla@kpkm.polsl.pl

Streszczenie

W referacie zaprezentowano problem zastosowania nowej dziedziny inżynierii wiedzy, jaką jest odkrywanie wiedzy w bazach danych w zakresie identyfikacji ilościowych modeli diagnostycznych. Przedstawiono genezę tej dziedziny, scharakteryzowano ją jako interaktywny i iteracyjny proces, a także wymieniono zadania oraz metody, jakimi się ta dziedzina posługuje. Szczególną uwagę zwrócono na metody odkrywania zależności funkcyjnych. W dalszej części referatu przedstawiono przykładowe zastosowania do zadań odkrywania statycznych zależności przyczynowo-skutkowych i „diagnostycznych" oraz zależności dynamicznych. W podsumowaniu dokonano analizy uzyskanych wyników oraz przeprowadzono dyskusję dotyczącą wprowadzenia zmian pozwalających zastosować te metody w praktyce.

Słowa kluczowe: bazy danych, odkrywanie wiedzy w bazach danych, modele diagnostyczne

KNOWLEDGE DISCOVERY IN DATABASES AS A PROCESS OF IDENITIFICATION OF DIAGNOSTICS MODELS

Summary

The paper deals with the problem of application of the new knowledge engineering domain, which is the knowledge discovery in databases (KDD), for the identification of quantity diagnostics models. The origin of the domain was presented. Then KDD was characterized as an interactive an iterative process. The tasks and methods used were specified, as well. The special attention was paid to the methods of discovering functional dependencies. The exemplary applications to tasks of static cause-and-effects and "diagnostics" dependencies and also dynamics dependencies were presented in the further past of the paper. The analysis of the obtained results was included in the summary, which discusses some changes in the KDD methodology allowing to put these methods into practice.

Keywords: databases, knowledge discovery in databases, diagnostics models

1. WPROWADZENIE

Jednym z obszarów zainteresowań diagnostyki technicznej maszyn jest pozyskiwanie wiedzy o relacjach zachodzących pomiędzy cechami stanu diagnozowanych obiektów, cechami ich wejść oraz cechami wyjść. Najczęściej wiedza o tych relacjach jest przedstawiana za pomocą modeli, w szczególności modeli ilościowych, które umożliwiają bezpośrednie wnioskowanie o cechach stanu maszyny. Identyfikacja tych modeli wiąże się z przygotowaniem odpowiedniego zbioru danych, który zazwyczaj jest pozyskiwany poprzez czynne eksperymenty diagnostyczne na stanowiskach badawczych lub na drodze symulacji komputerowych.

Z drugiej strony, istnieje wiele baz danych mieszczących się w obszarze zainteresowań diagnostyki technicznej maszyn, które mogą być źródłem użytecznej wiedzy diagnostycznej. Próby zastosowania klasycznych metod analizy danych, dla tego typu zbiorów napotykają przeszkody związane przede wszystkim z wielkością jak również z brakiem i niepoprawnymi wartościami zgromadzonych danych.

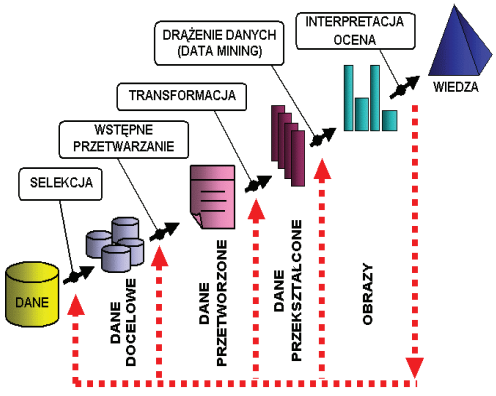
[1] Przedstawione wyniki uzyskano w trakcie badań częściowo finansowanych przez Komitet Badań Naukowych w ramach grantu promotorskiego Nr 4 T07B 059 26.

Naprzeciw tym problemom wychodzą metody rozwijane w ramach dziedziny zwanej odkrywaniem wiedzy w bazach danych (ang. Knowledge Discovery in Databases, KDD). Krótkie omówienie tych metod wraz z prezentacją przykładowych zastosowań w diagnostyce maszyn jest przedmiotem niniejszego referatu.

2. ODKRYWANIE WIEDZY W BAZACH DANYCH

Odkrywanie wiedzy w bazach danych jest nową dziedziną inżynierii wiedzy. Jej szybki rozwój jest związany z rosnącą z każdym rokiem liczbą baz danych oraz wielkością gromadzonych w nich danych, sięgającą w niektórych przypadkach kilku terabajtów. Analiza tak dużych zbiorów danych za pomocą konwencjonalnych metod statystycznych oraz metod uczenia maszynowego jest niemożliwa. Przyczyną takiego stanu rzeczy jest przede wszystkim nieprzystosowanie wymienionych metod do analizy tak wielkich zbiorów danych. W związku z tym zaistniała potrzeba opracowania nowych i/lub przystosowania istniejących już metod analizy danych. Odpowiedzią na to zapotrzebowanie są metody rozwijane w ramach odkrywania wiedzy w bazach danych.

Odkrywanie wiedzy w bazach danych jest procesem o iteracyjnym i interakcyjnym charakterze (rys. 1.). Stosowanie tego procesu wymaga wielu umiejętności oraz podejmowania różnych decyzji, co jest związane z posiadaniem wiedzy dziedzinowej o rozpatrywanym problemie. Na proces odkrywania wiedzy w bazach danych skład się wiele powiązanych ze sobą etapów (rys. 1.).



Rys. 1. Etapy procesu odkrywania wiedzy w bazach danych [3].

Wśród etapów procesu KDD kluczową rolę odgrywa etap drążenia danych (ang. Data Mining, DM). Etap ten jest poprzedzony wyborem zadania drążenia danych oraz metody drążenia danych. Wybór zadania drążenia danych określa apriori wynik procesu odkrywania wiedzy oraz formę reprezentacji odkrytej wiedzy. Do typowych zadań DM zalicza się [1,3]:

- klasyfikację,
- aproksymację,
- odkrywanie zależności przyczynowych,
- odkrywanie zależności funkcyjnych,
- odkrywanie podobieństw,
- odkrywanie asocjacje.

Wybór zadania DM determinuje wybór określonej metody drążenia danych oraz algorytmów służących do poszukiwania w danych określonej klasy wzorców i ich parametrów tj. określonej formy reprezentacji wiedzy. Wśród metod stosowanych w DM wyróżnia się [1,3]:

- indukcję drzew decyzyjnych i reguł,
- metody minimalnoodległościowe,
- sieci neuronalne,
- sieci bayesowskie.

W obszarze diagnostyki technicznej i eksploatacji maszyn gromadzone są bazy danych, które w przeważającej większości zawierają dane liczbowe. Dla takiej klasy baz danych najbardziej odpowiednimi metodami DM są metody odkrywania zależności ilościowych łączących zadanie odkrywania zależności funkcyjnych i zadanie aproksymacji. Metody odkrywania zależności ilościowych pozwalają na odkrywanie wiedzy ilościowej w postaci modeli nieparametrycznych i/lub modeli parametrycznych. Odkrywanie modeli nieparametrycznych bazuje na zastosowaniu:

- Sieci neuronalnych,
- SVM (Support Vector Machines),
- MARS (Multivariate Adaptive Regression Splines),

zaś proces odkrywania modeli parametrycznych na zastosowaniu:

- algorytmów ewolucyjnych,
- programowania genetycznego,
- gramatyk bezkontekstowych,
- różnych metod heurystycznych.

Dodatkowo, w zależności od tego czy analizowane dane opisują własności obiektu czy procesu, odkrywana wiedza reprezentowana przez zależności ilościowe może mieć charakter:

- statyczny (np. równania algebraiczne),
- dynamiczny (np. równania różniczkowe).

Niezależnie od wyboru końcowej postaci odkrywanej wiedzy ilościowej, właściwe stadium odkrywania zależności funkcyjnych obejmuje następujące operacje:

1. automatyczne generowanie struktury modelu wskazanej klasy,
2. identyfikacji parametrów dla założonej struktury modelu.

W celu ograniczenia nakładów obliczeniowych oraz zwiększenia pewności odkrycia zależności istotnych statystycznie, przedstawiony powyżej schemat działań zostaje poprzedzony weryfikacją występowania zależności funkcyjnych za pomocą odpowiednich metod statystycznych, np. za pomocą tablic kontyngencji [4].

3. PRZYKŁADY ZASTOSOWAŃ

Badania w zakresie zastosowań metod KDD do odkrywania wiedzy w diagnostycznych bazach danych zostały zapoczątkowane przez W. Moczulskiego i J. M. Żytkowa [4]. Sformułowany został wówczas problem badawczy, dotyczący odkrywania dwóch rodzajów zależności:

1. zależności przyczynowo-skutkowych,

$$Y = f(X,U) \tag{1}$$

2. zależności „diagnostycznych",

$$X = f(Y,U) \tag{2}$$

gdzie: Y - wyjścia, X - stan obiektu, U - sterowanie.

Aktualnie prowadzone badania dotyczą rozwoju metod odkrywania zależności dynamicznych [2] oraz ich zastosowania w różnych dziedzinach [7].

3.1. Identyfikacja zależności przyczynowo – skutkowych

W zakresie badań dotyczących zastosowania metod KDD w identyfikacji zależności przyczynowo skutkowych zastosowano bazę danych, w której zgromadzone dane opisywały rozkład niewyrównoważenia wzdłuż wału wirnika. Dane pozyskano za pomocą obliczeń numerycznych przeprowadzonych za pomocą systemu programów MESWIR [4]. Symulowano działanie wirnika wyposażonego w dwie tarcze i podpartego w łożyskach ślizgowych. Określony stan techniczny wirnika był uzyskiwany za pomocą siedmiu parametrów kontrolnych tj. prędkości obrotowej, położenia dwóch tarcz na wale wirnika oraz wartości niewywag i ich położenia kątowego na tarczach. Rejestrowanymi w bazie danych wartościami były cechy geometryczne obserwowanych trajektorii przemieszczeń wskazanych 4 węzłów modelu MES wirnika (m.in. węzły w podporach łożyskowych).

Dla omówionej powyżej bazy danych poszukiwano zależności przyczynowo-skutkowych wg następującej metodologii [4]:

- Selekcja i wstępne przetwarzanie danych.
- Dekompozycja wyselekcjonowanego zbioru danych na podzbiory w celu ujawnienia regularności zachodzących między atrybutami kontrolnymi i zależnymi.
- Poszukiwanie równań o możliwie tej samej strukturze w podzbiorach danych.
- Stopniowe uogólnianie równań w podzbiorach danych do równania wielu zmiennych.

Stosując powyższą metodologię wyznaczono układ 16 równań algebraicznych opisujących rozkład niewywag wzdłuż wału wirnika [4,5].

3.2. Identyfikacja zależności diagnostycznych

Proces odkrywania zależności „diagnostycznych" można przeprowadzić w sposób pośredni, polegający na „odwracaniu" (rozwiązywaniu) ilościowych zależności przyczynowo-skutkowych, lub bezpośrednio, stosując metodologię przedstawioną w poprzednim podpunkcie z uwzględnieniem zamiany ról atrybutów tj. atrybuty niezależne stają się atrybutami zależnymi, zaś atrybuty zależne stają są atrybutami niezależnymi. Dla rozważanego problemu, zastosowano pierwszą z wymienionych możliwości, tj. zależności „diagnostyczne" identyfikowano poprzez rozwiązywanie pozyskanego w pierwszej fazie badań układu równań przyczynowo-skutkowych.

W trakcie realizacji tej fazy badań, pojawił się problem nadokreśloności i nieliniowości rozwiązywanego układu równań. Powyższy problem rozwiązano poprzez zastosowanie:

- Zmodyfikowanej metody Newtona-Raphsona rozwiązywania układów równań nieliniowych [5].
- Algorytmów genetycznych [6].

Przykładowe wyniki uzyskane dla obydwu metod zestawiono w tabeli poniżej.

Tabela 1. Błąd predykcji stanu x_1 [6].

	Algorytm genetyczny			Metoda Newtona - Raphsona		
x_1	$\hat{x}_1$	Δ	Δ [%]	$\hat{x}_1$	Δ	Δ [%]
360	361	-1	-0.28	374	-14	-3.88
90	110	-20	-22.22	92	-2	-2.01
203	182	21	10.34	197	6	2.91
90	117	-27	-30.00	124	-34	-37.62

3.3. Identyfikacja zależności dynamicznych

Aktualny rozwój metod odkrywania zależności ilościowych w bazach danych koncentruje się na zależnościach dynamicznych danych w postaci równań różniczkowych [2]. Odkrywanie zależności dynamicznych stało się możliwe dzięki wprowadzeniu do systemów i algorytmów odkrywania zależności ilościowych, modułów numerycznego różniczkowania i całkowania danych.

W prowadzonych badaniach do odkrywania zależności dynamicznych w postaci równań różniczkowych, zastosowano m.in. algorytm *LAGRANGE* [2], który pozwala na odkrywanie tego typu równań dzięki procedurze różniczkowania numerycznego.

Schemat działania algorytmu obejmuje następujące etapy [2]:

1. Wyznaczenie wszystkich pochodnych zmiennych systemowych.
2. Generowanie nowych „termów" dla zbioru zmiennych systemowych i ich pochodnych.
3. Generowanie i testowanie równań.

Z uwagi na brak danych o charakterze dynamicznych, w badaniach zastosowano dane pozyskane poprzez symulacje numeryczne układu opisanego następującym równaniem różniczkowym:

$$m\ddot{x} + b\dot{x} + cx + e(t) = 0 \qquad (3)$$

gdzie: *e(t)* - szum o rozkładzie równomiernym.

Działanie układu (3), symulowano w środowisku MATLAB/Simulink, przy czym brano pod uwagę różne kombinacje parametrów tego układu jak również różne wartości warunków początkowych i parametrów szumu *e(t)*.

Tabela 2. przedstawia przykładowe rezultaty procesu odkrywania równania (4) za pomocą systemu *LAGRANGE* dla trzech przypadków [7]:

I – dane nie zawierały zakłóceń (szumu),
II – dane zawierały zakłócenia (szum) o różnym stopniu natężenia,
III – dane zawierały zakłócenia (szum) lecz na wstępie podlegały wygładzaniu.

Tabela 2.
Przykłady odkrytych równań różniczkowych [7].

Nr	Równanie	R	S
I	$\dot{x} = -36.00x - 2.40\dot{x}$	1.000	0.0075
II	*brak*	—	—
III	$\dot{x} = -2.27 - 36.01x - 2.40\dot{x}$	0.8608	3.9906

R – korelacja, S – odchylenie standardowe

4. PODSUMOWANIE

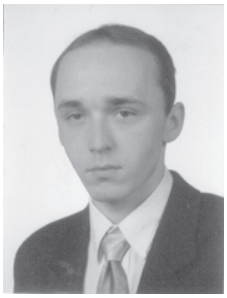
Wyniki przeprowadzonych dotychczas badań w zakresie odkrywania zależności przyczynowo-skutkowych i „diagnostycznych" potwierdzają przydatność zastosowanych metod. Wymagane jest przeprowadzenie badań dotyczących bezpośredniej identyfikacji zależności „diagnostycznych".

W przypadku metod odkrywania zależności dynamicznych w postaci równań różniczkowych pojawia się m.in. problemem numerycznego różniczkowania zaszumionych danych. Dodatkowo, proponowane w tym zakresie metody nie uwzględniają możliwości badania stabilności odkrytych równań. W związku z tym, w dalszych badaniach przewiduje się modyfikacje istniejących metod polegające m.in. na zastąpieniu równań różniczkowych równaniami różnicowymi, wprowadzeniu procedury badania stabilności odkrywanych równań, jak również zastosowania procedury walidacji krzyżowej (ang. *cross validation*) zapobiegającej nadmiernemu dopasowaniu odkrywanych równań do danych.

Weryfikacja proponowanych zmian nastąpi dla aktualnie przygotowywanej bazy danych zawierającej wyniki pomiarów parametrów procesowych agregatów pompowych pracujących w wyłączonych z eksploatacji szybach kopalnianych.

LITERATURA

[1] Cichosz P.: *Systemy uczące się*. WNT, Warszawa 2000.
[2] Džeroski S., Todorovski L.: *Discovering dynamics: From inductive logic programming to machine discovery*. Journal of Intelligence Information Systems, 4(1995), str. 89-108.
[3] Frawley W., Piatetsky-Shapiro G., Matheus C.: *Knowledge Discovery in Databases - An Overview.* KDD collection, str. 1-30. Reprinted in AI Magazine, Fall 1992.
[4] Moczulski W., Żytkow J.M.: *Discovery of diagnostic knowledge from multi-sensor data*. 15th SPIE AeroSense Meeting on Aerospace/Defense Sensing, Simulation and Control, Orlando, 16-17.04.2001.
[5] Wachla D.: *The idea of method of searching for global inverse models in machinery diagnostics*, Materiały konferencji Methods of Artificial Inteligence in Mechanics and Mechanical Engineering, AI MECH 2001, Gliwice 2001, str. 291-294.
[6] Wachla D.: *An Example of genetic algorithm application in knowledge discovery in databases*. Symposium on Methods of Artificial Intelligence in Mechanics and Mechanical Engineering, Gliwice 13 - 15.11.2002, str. 429 – 432.
[7] Wachla D.: *Identyfikacja dynamicznych modeli obiektów mechanicznych metodami odkryć wiedzy w bazach danych*. VI Krajowa Konferencja Naukowo-Techniczna Diagnostyka Procesów Przemysłowych, DPP'03, Władysławowo 15–17 września, 2003, str. 337-340.



Mgr inż. Dominik WACHLA jest słuchaczem studiów doktoranckich na Wydziale Mechanicznym Technologicznym Politechniki Śląskiej. Jego zainteresowania badawcze skupiają się wokół zastosowań metod sztucznej inteligencji w diagnostyce technicznej maszyn.

Jest laureatem Stypendium Promocyjnego Fiata (nagroda zbiorowa wspólnie z D. Sławikiem i J. Wojtusikiem).