ZASTOSOWANIE TECHNIK DATA MINING DO ODKRYWANIA RELACJI DIAGNOSTYCZNYCH W DANYCH OPISUJĄCYCH PRZEBIEG HISTORII EKSPLOATACJI MASZYN

Mariusz GIBIEC
Katedra Robotyki i Mechatroniki
Akademia Górniczo-Hutnicza w Krakowie, 30-059 Kraków, al. Mickiewicza 30,
tel. 012 6343505, mgi@agh.edu.pl

Streszczenie

W pracy przedstawiono przykład wykorzystania wybranych technik Data Mining do odkrywania relacji diagnostycznych w danych z rejestratora przebiegu eksploatacji górniczego kombajnu ścianowego. Wykorzystując metody grupowania określono ilość grup w danych oraz zweryfikowano ich związek ze stanem technicznym urządzenia na podstawie protokołów serwisowych. Zbudowano modele klasyfikujące wyróżnione stany techniczne urządzenia wykorzystując metody drzew klasyfikujących. Analizując działanie mechanizmu klasyfikującego drzew w postaci reguł odkryto relacje diagnostyczne opisujące przyczyny zmian stanu technicznego rozważanego urządzenia.

Słowa kluczowe: relacje diagnostyczne, klasyfikacja stanu technicznego, Data Mining, analiza danych.

DATA MINING TECHNICS APPLICATION TO DIAGNOSTIC RELATIONS DISCOVERING IN HISTORIC DATA OF MACHINERY EXPLOITATION

Summary

In this research an example of Data Mining techniques application to diagnostic relations discovering from data recorder of exploitation parameters of mining cutter-loader was presented. Using clustering methods the number of clusters in data was determined. Their correlations with technical condition of machinery was verified basing on servicing documentation. Classification trees methods were used to build models classifying listed technical conditions. Performance of their classification system in form of rules was investigated. Basing on these rules diagnostic relations describing reasons of technical condition changes were discovered.

Keywords: Diagnostic relations, technical condition classification, Data Mining, data analyses.

1. WSTĘP

Niektóre maszyny i urządzenia, na etapie ich wytwarzania, wyposażane są w różnego rodzaju czujniki i systemy pomiarowe. Stanowią one najczęściej elementy systemu sterowania lub systemu monitorowania ich pracy. Ich zadaniem jest zapewnienie bezpiecznej eksploatacji. W przypadku złożonych instalacji objętych długoterminową gwarancją celem producenta jest także monitorowanie poprawności eksploatacji pod kątem zgodności z warunkami gwarancji. Takie informacje gromadzone są w dużych bazach danych i poddawane są analizie tylko w przypadku zaistnienia awarii. Jednak ich zawartość może także służyć do określania stanu technicznego urządzenia oraz przyczyn jego zmian. Ponieważ w trakcie eksploatacji gromadzone są wielkie ilości danych ich analiza wymaga użycia narzędzi informatycznych. Uzasadnionym wydaje się zastosowanie metod Data Mining. Data Mining jest procesem automatycznego odkrywania znaczącej, pożytecznej, dotychczas nieznanej i możliwie pełnej wiedzy zawartej w dużych bazach danych, wiedzy ujawniającej ukryte własności monitorowanego procesu. Wiedza ta przyjmuje postać reguł, prawidłowości, tendencji i korelacji. Następnie jest ona przedstawiana przygotowanemu do jej spożytkowania użytkownikowi w celu rozwiązania stojących przed nim problemów i podjęcia istotnych decyzji. Proces odkrywania wiedzy wykorzystuje metody, algorytmy i techniki z wielu dziedzin takich jak statystyka, hurtownie danych, rozpoznawanie obrazów, sieci neuronowe, zbiory rozmyte i przybliżone oraz techniki wizualizacji komputerowej. W niniejszej pracy wybrane techniki Data Mining wykorzystano do odkrywania relacji diagnostycznych na podstawie danych z rejestratora pracy kombajnu górniczego. W pierwszym etapie dokonano grupowania danych w celu określenia jakie uszkodzenia lub stany pracy kombajnu znajdują swoje odzwierciedlenie w danych. Weryfikacji uzyskanych grup dokonano w oparciu o protokoły serwisowe. Zawierają one opis czynności serwisowych wykonanych bezpośrednio po okresie, w którym zarejestrowano dane. Określone w ten sposób związki grupa danych – stan techniczny stanowią podstawę do budowy modeli

klasyfikujących stan techniczny. Zastosowane metody drzew klasyfikacyjnych pozwalają na interpretację działania mechanizmu klasyfikacji w postaci reguł *jeżeli…to….* W ten sposób mogą zostać pozyskane relacje diagnostyczne z danych opisujących historię eksploatacji.

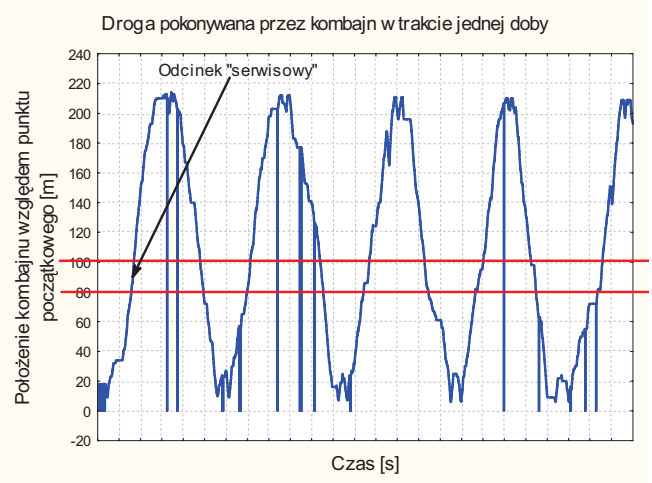
2. ANALIZOWANE DANE

Przedmiotem badań są dane zbierane podczas pracy kombajnu górniczego. Jest to płytkozabiorowy dwuramieniowy kombajn ścianowy z elektrycznym napędem posuwu do wybierania pokładów węgla. Przystosowany jest do pracy w zmiennych warunkach górniczo-geologicznych. Stosowany jest do dwukierunkowego, bezwnękowego urabiania i ładowania węgla w ścianowych systemach eksploatacji pokładów nachylonych do 35° w przypadku pokładów podłużnych, a w pokładach porzecznych do 20° (po wzniesieniu) lub 15° (po upadzie). Za posuw odpowiedzialne są silniki elektryczne zasilane napięciem z falownika w zakresie częstotliwości od 0 do 100Hz. W obszarze do 50Hz regulacja następuje przy stałym momencie, po przekroczeniu progu 50Hz zachowana jest stała moc. Dzięki zastosowaniu przemiennika częstotliwości regulacja prędkości posuwu kombajnu odbywa się automatycznie i bezstopniowo w zależności od obciążenia silników elektrycznych napędów posuwu oraz ramion urabiających. Kombajn ten współpracuje z systemem sterowania i diagnozowania (w czasie rzeczywistym) pracy kombajnów węglowych o dużej mocy, wyposażonych w układy napędowe składające się z maksymalnie 6 silników elektrycznych. Wszystkie funkcje sterownicze i diagnostyczne możliwe są dzięki informacjom dostarczonym z czujników służących do pomiaru:

- prądów obciążenia silników elektrycznych,
- temperatury uzwojeń i łożysk silników oraz mechanizmów,
- ciśnienia w obwodach hydraulicznych ciągnika i hamulców,
- ciśnienia wody chłodzącej,
- prędkości posuwu kombajnu i położenia kombajnu w ścianie, przez pomiar przebytej drogi.

Diagnostyka oparta na opisywanym systemie może przebiegać na dwa sposoby. Sposób pierwszy polegać na bieżącym przetwarzaniu danych pomiarowych i generowaniu komunikatów o stanie kombajnu. Jednak możliwa jest współpraca z układem pomiarowym, który zapisuje parametry związane z pracą kombajnu na przestrzeni 24 godzin z rozdzielczością 1 sekundy. Dane gromadzone są w arkuszach o rozmiarach 127 na "*n*", przy czym 127 stanowi ilość zmiennych uzyskanych z różnych czujników, a druga wartość "*n*" opisuje ilość przypadków i zmienia się w zależności od długości pracy kombajnu w danym dniu.

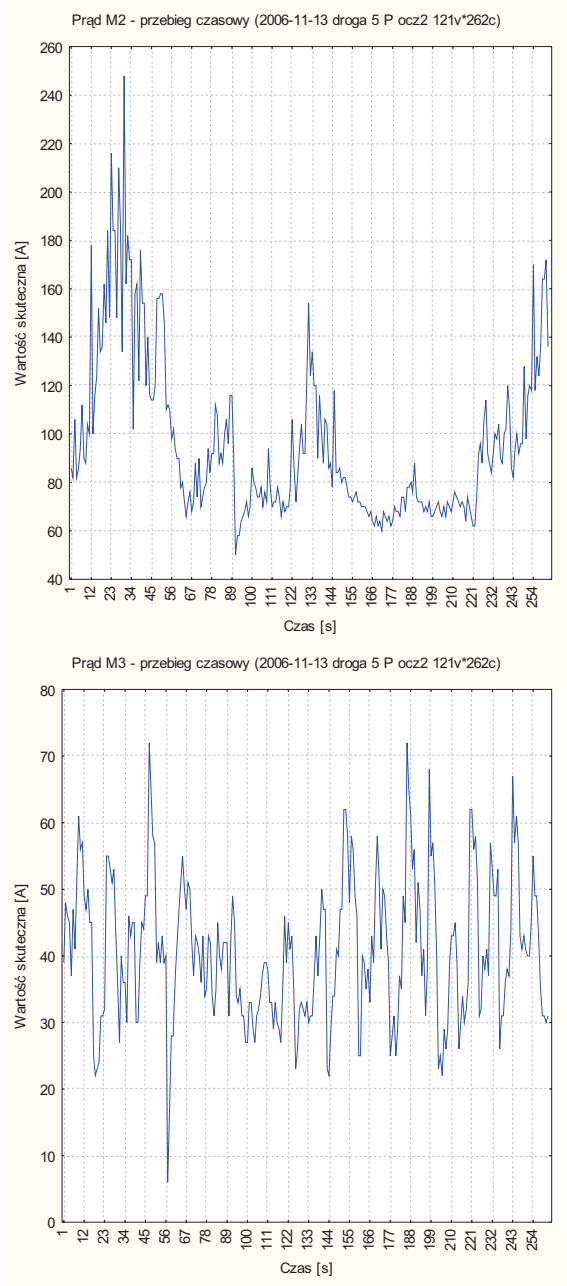
Ponieważ kombajn pracuje w środowisku o losowych parametrach, o dużym wpływie na charakter przebiegów rejestrowanych przez czujniki, zaproponowano wybór takiego odcinka pracy gdzie warunki te są porównywalne. Przeprowadzona analiza danych pozwoliła na wybór odcinka „serwisowego". Składa się on z kilku sekcji pokonywanych w każdym przejeździe, z taka samą stałą prędkością i bez występowania przeciążeń. Przykładowy przebieg zmian położenia kombajny „w ścianie" pokazano na rysunku 1.



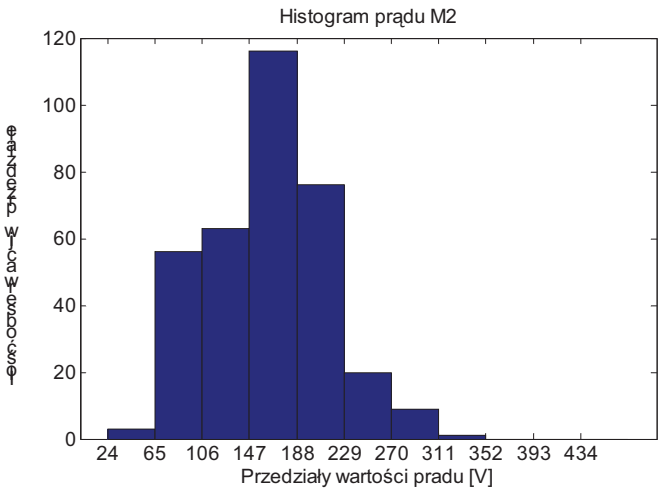
Rys. 1. Położenie kombajnu w funkcji czasu z wyszczególnionym odcinkiem serwisowym

Duża ilość zmiennych, jak również licznie występujące błędy były powodem kolejnego etapu filtracji danych. Na podstawie wiedzy uzyskanej od specjalistów, zajmujących się badanym zagadnieniem, zdecydowano się na wyselekcjonowanie następujących zmiennych: "*PradM1*", "*PradM2*", "*PradM3*", "*PradM4*", "*CzestPrzem*", "*PredkSilnPrzem*", "*PradSilnPrzem*", "*MocSilnPrzem*", "*MomentSilnPrzem*", "*ACPrzem*", "*DCPrzem*". Dodatkowo stwierdzono potrzebę dodania nowej zmiennej "*Roznica*", której wartość równa jest różnicy po miedzy zmiennymi "*PradM3*" i "*PradM4*". Na rysunku *2* przedstawiono przykładowe wykresy zmiennych "*PradM2*", "*PradM3*".

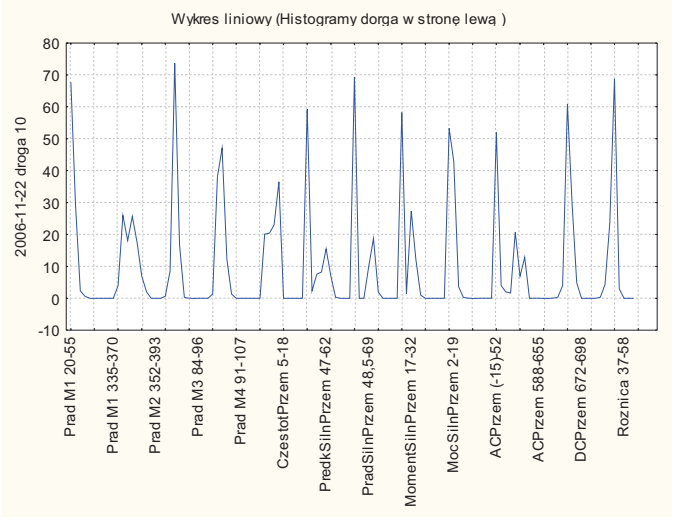
Ponieważ uzyskane charakterystyki wykazują dużą zmienność nawet dla kolejnych przejazdów zaproponowano zastosowanie metody ilościowego badania powstałych odcinków danych w postaci analizy histogramów. Histogramy te znormalizowano z powodu różnej ilości zebranych danych dla tego samego odcinka w rożnych przejazdach. Na rysunku 3 pokazano przykładowy histogram dla zmian prądu silnika napędu. Histogramy dla wszystkich zmiennych połączono w jeden wektor, stanowiący nową charakterystykę pojedynczego przejazdu. Wykres takiego wektora przypomina wyglądem widmo sygnału (rysunek 4). Tak przekształcone dane stanowią podstawę dalszych analiz.



Rys. 2. Przebiegi zmiennych ”*PradM2*”, ”*PradM3*”



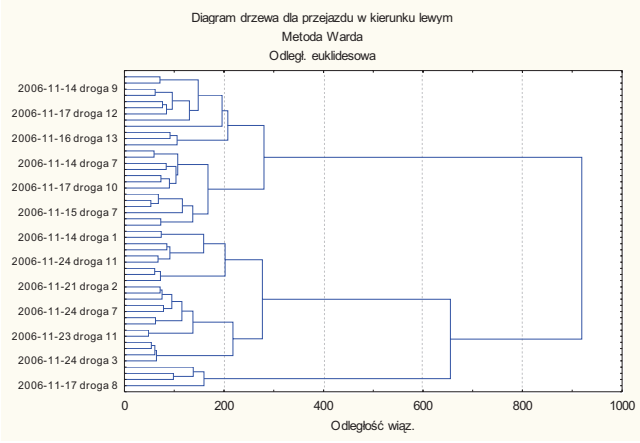
Rys. 3. Histogram zmian prądu silnika napędu



Rys. 4. Charakterystyka połączonych histogramów

3. GRUPOWANIE DANYCH

W pierwszym etapie przystąpiono do określenia ilości grup występujących w danych.. Ilość grup sugeruje ile stanów technicznych urządzenia można wyróżnić na podstawie zgromadzonych danych. W analizach założono brak informacji o ilości grup – posiadane dane pomiarowe nie zawsze są skorelowane z informacjami na temat stanu technicznego. Do grupowania zaproponowano metodę Warda. Jest to metoda z grupy hierarchicznych metod aglomeracyjnych. Tego typu metody pozwalają na określenie tzw. hierarchii drzewkowej elementów analizowanego zbioru. Drzewo połączeń otrzymuje się poprzez krokowe łączenie w podzbiory operacyjnych jednostek taksonomicznych. Na wstępie przyjmuje się, że każdy element zbioru stanowi taką jednostkę. W utworzonej macierzy odległości między jednostkami wyszukuje się najmniejszego elementu pośród leżących poza przekątną. Jest to odległość aglomeracyjna, minimalna w sensie „lokalnym”. Wskazane przez nią jednostki zostają połączone tworząc nową jednostkę. Następnie korygowana jest macierz odległości i procedura jest powtarzana. Warunkiem stopu jest uzyskanie jednej jednostki taksonomicznej. Metoda Warda różni się od pozostałych metod sposobem szacowania odległości między jednostkami taksonomicznymi. Wykorzystuje ona analizę wariancji – zmierza do minimalizacji sumy kwadratów odległości dowolnych dwóch skupień, które są tworzone na każdym etapie aglomeracji. Zaletą wybranej metody jest brak konieczności arbitralnego definiowania ilości grup w analizowanych danych oraz o 40% lepsza efektywność wykrywania prawdziwej struktury danych niż w innych metodach [2]. Poniżej zaprezentowano wyniki grupowania dla wektora wejściowego utworzonego z połączonych histogramów wszystkich zmiennych.



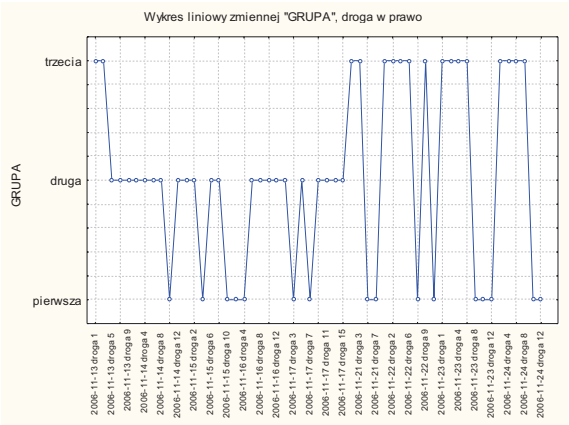
Rys. 5. Wyniki grupowania metodą Warda

Dokonując przecięcia drzewa Warda na poziomie odległości wiązania równej 400 uzyskujemy trzy grupy danych. Z wykresu można odczytać które przypadki należą do tej samej grupy. Na podstawie tego wyniku dokonano grupowania metodą k-średnich na trzy grupy. Metoda ta różni się od metod aglomeracji założeniem określonej ilości grup w danych. Jej celem jest utworzenie zadanej ilości możliwie odmiennych skupisk danych. W pewnym sensie metoda ta jest odwrotnością analizy wariancji. Działanie rozpoczyna od losowo wybranych skupisk, a następnie przenosi elementy zbioru między skupiskami tak by zapewnić minimalizację zmienności wewnątrz skupisk i maksymalizację zmienności pomiędzy nimi.

W analizowanym przykładzie celem zastosowania metody k-średnich było określenie jakie przypadki tworzą poszczególne grupy. Ich przynależność określa nowa zmienna klasyfikująca ”GRUPA”.

Wykorzystując dostępne protokoły serwisowe dokonano analizy zmian wartości zmiennej klasyfikującej w odniesieniu do przeprowadzonych czynności serwisowych.

Na rysunku 6 przedstawiono wykresy zmian wartości nowo powstałej zmiennej względem kolejnych przejazdów przez odcinek serwisowy.



Rys. 6. Wartości zmiennej klasyfikującej w kolejnych przejazdach

Na podstawie otrzymanego opisu czynności serwisowych, przeprowadzonych w rozpatrywanym okresie, odpowiednim grupom przyporządkowano stany maszyny. Jeżeli zmienna grupa przyjmuje wartość *Pierwsza* to można zakwalifikować stan urządzenia jako poprawną pracę kombajnu, wartość *Druga* powinna odwzorowywać zbiór przypadków w których występowały uszkodzenia. Opinia ta oparta jest na zapisie wystąpienia w dniu 17.11.2006r. uszkodzenia koła napędowego lewego napędu, a wartość *Druga* pojawiła się w dniu 16.11.2006r. i zmieniła się w następnym dniu po wznowieniu pracy kombajnu po naprawie. Według opisu zawartego w części serwisowej wymiany uszkodzonego koła napędowego lewego ciągnika dokonano w dniu 18.11.2006r. W przypadku wartości *Trzecia* można stwierdzić, że powinna ona również odwzorowywać poprawną pracę urządzenia, lecz jest to inny stan w porównaniu do wartości *Pierwsza*, ponieważ podczas postoju w dniu 20.11.2006 wykonano czynności serwisowe oraz wymiany uszkodzonych części. „Wymieniono przekaźniki czasowe PC zastąpiono je przekaźnikami RTx-410, przekaźnik K1 zastąpiono przekaźnikiem CI-4. Stwierdzono nieprawidłowości działania zabezpieczeń silników AMP wynikające z fizycznych uszkodzeń przewodów przekładników pomiarowych”[10]. Jednocześnie można przyjąć, że wymienione wyżej uszkodzenia nie znalazły odzwierciedlenia w analizowanych danych.

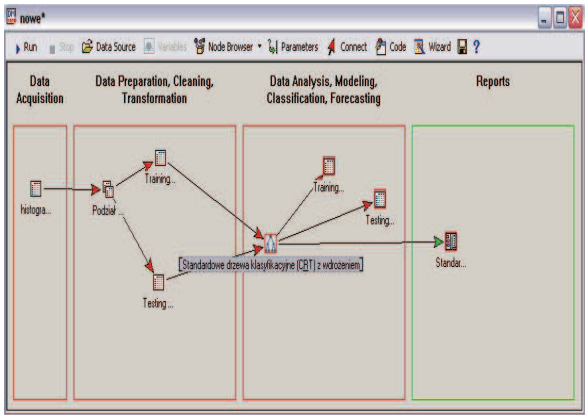
4. KLASYFIKACJA STANU

Dysponując utworzonymi w sposób opisany wcześniej przykładami klasyfikacji stanu technicznego przystąpiono do budowy modeli klasyfikujących. Dla wszystkich metod dokonano podziału zgromadzonych przykładów na próbę uczącą i testową. Ponieważ przykłady do dalszych analiz są wybierane losowo określono tylko przybliżoną ilość przykładów testowych na 30% wszystkich danych wejściowych.

Pierwszą z rozważanych metod klasyfikacyjnych jest model oparty na metodzie ”*Standardowych drzew klasyfikacyjnych C&RT*”. Graficznym wynikiem podziału zbioru danych jest drzewo. Powstaje ono w skutek rekurencyjnego podziału zbioru obserwacji A na *n* rozłącznych podzbiorów A_1, A_2, A_3, …, A_n. Budowa modelu ma na celu wypracowanie podzbiorów maksymalnie jednorodnych z punktu widzenia wartości zmiennej zależnej. W kolejnych etapach budowy modelu analizowane są wszystkie predyktory i wybierany jest ten, który umożliwia najlepszy podział węzła, co ma prowadzić do powstania najbardziej homogenicznego podzbioru [9].

Każdy model drzewa rozpoczyna się od całego zbioru obserwacji. Zbiór ten ulega podziałowi na dwa (drzewa binarne) lub więcej (drzewa dowolne) podzbiory. Powstałe węzły nazywane są węzłami potomkami (ang. child nodes), a wydzielone zostały z tzw. węzła macierzystego (ang. parent node).

Jeżeli w następnym etapie nie nastąpi kolejny podział węzła potomka, staje się on węzłem końcowym lub inaczej zwanym liściem. Jednak, że jeśli w drugim etapie węzeł potomek ulega kolejnemu podziałowi, staje się węzłem macierzystym dla danego etapu, a nowo powstałe węzły nazywane są potomkami. Metoda ta umożliwia odkrywanie pewnych reguł występujących pomiędzy zmiennymi w badanym zbiorze danych wejściowych. Opracowany w środowisku Statistica projekt Data Mining pokazano na rysunku 7.



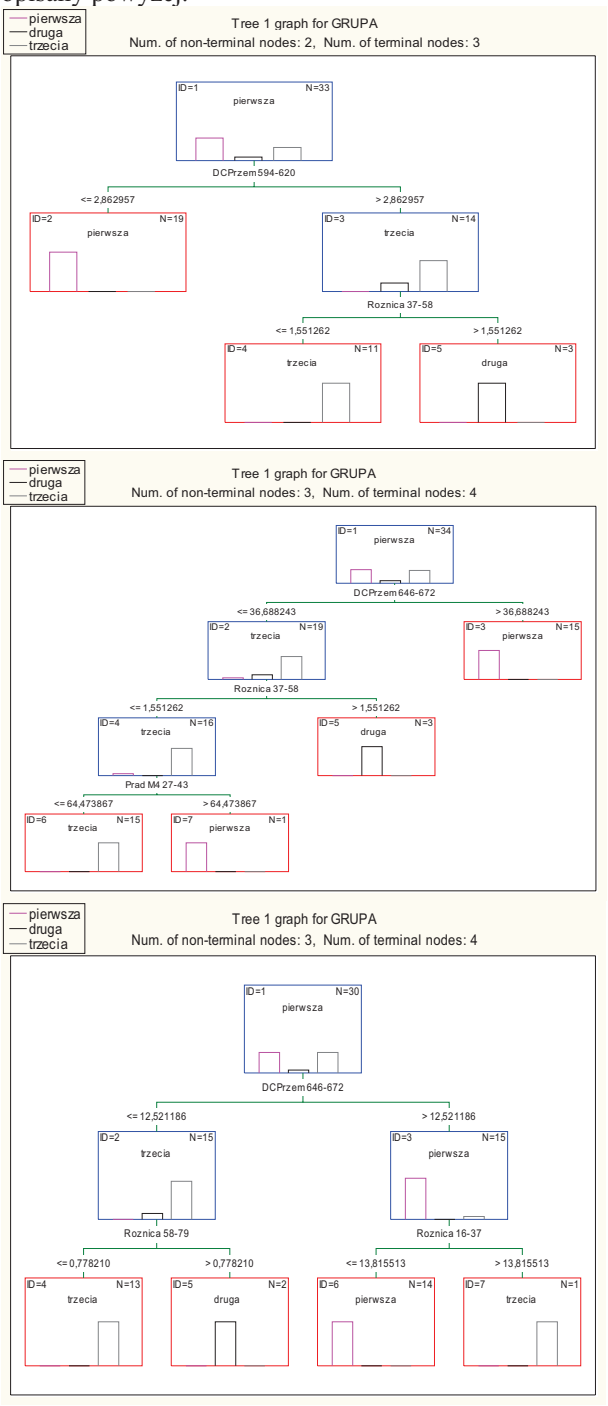
Rys.7. Model klasyfikacyjny oparty na standardowych drzewach klasyfikacyjnych C&RT

Wyniki budowy modeli klasyfikujących w postaci drzew klasyfikacyjnych przedstawiono na rysunkach 8-10. Z powodu, iż przypadki w zbiorze uczącym wybierane są w sposób losowy otrzymano kilka drzew. Przy każdorazowym uruchomieniu tworzenia modelu otrzymano nowe drzewo, które w różnym stopniu odwzorowuje relacje występujące w danych. Poniżej przedstawione są tylko te drzewa, dla których błąd dopasowania był najmniejszy.

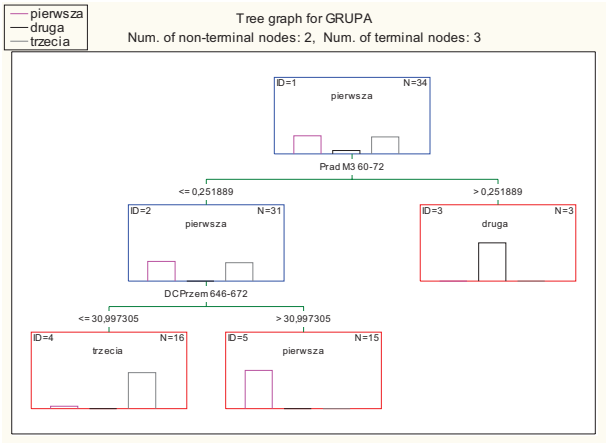
Na przykładzie pierwszego drzewa można dokonać określenia następujących reguł opisujących zależności pomiędzy opracowanymi danymi i stanem technicznym kombajnu:

1. węzeł nr 2: „jeśli zmienna DCPrzem594-620 przyjmuje wartość mniejszą bądź równą 2,862957, to stan techniczny należy do pierwszej grupy";
2. węzeł nr 4: „jeśli zmienna DCPrzem594-620 przyjmuje wartość większą od 2,862957 i zmienna Roznica37-58 przyjmuje wartości mniejsze bądź równe 1,551262 , to stan techniczny należy do trzeciej grupy";
3. węzeł nr 5: „jeśli zmienna DCPrzem594-620 przyjmuje wartość większą od 2,862957 i zmienna Roznica37-58 przyjmuje wartości większe od 1,551262, to stan techniczny należy do drugiej grupy";

Następny model został zbudowany na podstawie metody standardowej klasyfikacji CHAID. Drzewa tego typu wykorzystują wyniki testu *Chi-kwadrat* jako kryterium podziału węzła. Umożliwiają także podział na więcej niż dwie kategorie w jednym węźle. Metoda ta wykorzystuje jakościowe zmienne wejściowe zatem konieczne było utworzenie kategorii do których zakwalifikowano procentowe wartości opracowanych histogramów. Uzyskane drzewo klasyfikacyjne zaprezentowano na rysunku 11. Realizujące mechanizm podziału reguły diagnostyczne mogą zostać odczytane w sposób opisany powyżej.



Rys. 8-10. Przykładowe wyniki C&RT



Rys.11. Drzewo klasyfikacyjne typu CHAID

Ponieważ wszystkie modele klasyfikujące są obarczone pewnym błędem istotny jest wybór najlepszego z nich. W tym celu bada się jego zdolność do generalizacji wykorzystując zbiór przykładów testowych. Kryterium oceny w dostępnym oprogramowaniu jest niezgodność procentowa określająca procent błędnych klasyfikacji danego modelu dla tej samej próbki testowej. Wyniki dla testowanych modeli zestawiono w tabeli 1. Wartości te można wykorzystać do określenia stopnia ufności wobec relacji diagnostycznych odkrytych za pomocą danego modelu.

Tabela 1. Porównanie niedokładności klasyfikacji

	Summary Goodness of Fit (Sieci o radialnych funkcjach bazowych z wdrożeniem (klasyfikacja Observed variable: GRUPA		
	1 Chi-square statistic	2 G-square statistic	3 Percent disagreement
Testing_PMML_CRBF9(RBF-120:3:1-SS.EX	1,000	1,000	
Testing_PMML_CMLP7(MLP-120:3:1-BP96	49,50000	37,73394	55,00000
Testing_PMML_CECHAID6(ExhaustiveCHAI	1,50000	7,23546	15,00000
Testing_PMML_CCHAID5(CHAIDModelPr	1,50000	7,23546	15,00000
Testing_PMML_CTrees4(TreeModelPred)	1,000000	2,772589	5,000000

5. PODSUMOWANIE

W rozważanym przykładzie producent kombajnu górniczego udostępnił tylko krótki 10-cio dniowy zapis danych opisujących eksploatację kombajnu górniczego. Sposób rejestracji danych wymusił dokonanie wstępnego przetworzenia danych na potrzeby metod Data Mining. Wykorzystano dwa typy technik Data Mining. W pierwszym etapie określono ile stanów technicznych urządzenia znajduje odzwierciedlenie w danych pomiarowych. Ponieważ nie wszystkie zmiany stanu technicznego muszą implikować zmiany wartości mierzonych wielkości weryfikacji ilości zmian stanu technicznego dokonano w oparciu o zapisy w protokołach serwisowych. Wykorzystując tak utworzone związki dane – stan zbudowano modele klasyfikujące w postaci drzew decyzyjnych. Ich zastosowanie daje możliwość odkrywania relacji diagnostycznych w zgromadzonych danych w postaci czytelnych i zrozumiałych reguł. Wiarygodność odkrytych reguł określono na podstawie dokładności klasyfikacji zastosowanych modeli. W przypadku rozważanego systemu sterowania i monitorowania pracy kombajnu ścianowego przedstawione metody pozwalają na budowę systemu diagnostyki, w postaci reguł, bez dodatkowych inwestycji sprzętowych.

LITERATURA

[1] Grabiński T., Sokołowski A.: *The Effectiveness of Some Signal Identification Procedures*, *Signal Processing: Theories and Applications*, North-Holland Publishing Company, EURASIP, 1980.
[2] Gibiec M.: *Soft Computing tools for machine diagnosing,* Journal of Theoretical and Applied Mechanics. 3, vol. 42: 483 – 501, 2004.
[3] Hand D., Mannila H., Smyth P.:*Principles of Data Mining*, MIT Press, Cambridge. Tłum pol. Eksploracja Danych, WNT, Warszawa 2005.
[4] Kantardzic M.: Data Mining: *Concepts, Models, Methods and Algorithms*, Wiley-Interscience, Hoboken NJ 2003.
[5] Larose D.: *Data Mining Methods and Models*. Wiley-Interscience, Hoboken NJ 2006.
[6] Wang X. Z., *Data Mining and Knowledge Discovery for Process Monitoring and Control*, Springer-Verlag London 1999.
[7] Sohn H., Worden K., Farrar C. R.: *Statistical damage classification under changing environmental and operational conditions,* Journal of Intelligent Material Systems and Structures, 13 561-574, 2002.
[8] Skormin V. A., Popyack L. J., Gorodetski V. I., Araiza M. L., Michel J. D.: *Applications of cluster analysis in diagnostics-related problems*, in: Proceedings of the 1999 IEEE Aerospace Conference, Vol. 3, Snowmass at Aspen, CO, USA,, pp. 161-168, 1999.
[9] Wang K.: *Intelligent condition monitoring and diagnosis systems, a computational intelligence approach.* ISSN: 0922-6389, IOS Press, 2003.

Dr inż. Mariusz GIBIEC jest adiunktem w Katedrze Robotyki i Mechatroniki AGH. Jego zainteresowania dotyczą zastosowań metod eksploracji danych oraz sztucznej inteligencji (sieci neuronowych i zbiorów rozmytych) w diagnostyce technicznej. Jest autorem prac nad wykorzystaniem powyższych technik w systemach monitorujących do realizacji zadań filtracji, predykcji oraz klasyfikacji stanu maszyn.