dr MAREK SIKORA
Instytut Informatyki Politechnika Śląska, Gliwice
Centrum Elektryfikacji i Automatyzacji Górnictwa
mgr inż. ŁUKASZ WRÓBEL
Instytut Informatyki Politechnika Śląska, Gliwice

Wyniki zastosowania algorytmów indukcji reguł do klasyfikacji stanu zagrożenia tąpaniami w kopalniach węgla kamiennego

W artykule przedstawiono wyniki prób zastosowania algorytmów indukcji reguł do wyprzedzającej klasyfikacji stanu zagrożenia tąpaniami w wyrobisku górniczym. Na początku artykułu krótko opisano sposób pozyskiwania i przekształcania danych będących przedmiotem analizy. W części teoretycznej przedstawiono schemat algorytmu indukcji reguł będących podstawą działania klasyfikatora, a także sposób dostrajania klasyfikatora umożliwiający uzyskanie lepszej dokładności klasyfikacji. W części eksperymentalnej zaprezentowano wyniki analizy danych pochodzących z dwóch wyrobisk górniczych.

1. WPROWADZENIE

Jednym z głównych zadań stacji geofizycznych w kopalniach węgla kamiennego jest ustalanie stopnia zagrożenia tąpnięciem w czynnych wyrobiskach górniczych. W celu określenia tego stopnia stosowane są, w zależności od kopalni, różnego rodzaju szczegółowe metody oceny zagrożenia (zazwyczaj są to metody: sejsmoakustyczna, sejsmologii, wierceń małośrednicowych).

Na bazie metod szczegółowych wykonuje się ocenę końcową (kompleksową), która uwzględnia oceny uzyskane przez każdą z poszczególnych metod szczegółowych oraz warunki geologiczne panujące w danym wyrobisku. Opis wykonania ocen metodami szczegółowymi oraz kompleksową wydany został się jako instrukcja postępowania wydana przez Główny Instytut Górnictwa [26].

Ponieważ dokładność obecnie obowiązujących metod zagrożenia tąpaniami daleka jest od doskonałości o czym może świadczyć duża liczba nietrafnych ocen, obserwuje się intensywny rozwój alternatywnych metod predykcji zagrożenia [6], [12], [13], [14]. Alternatywne metody oceny zagrożenia wykorzystują wysokoprzetworzone dane pochodzące z systemów pomiarowych. Metoda tomografii pasywnej jest metodą umożliwiającą cykliczne kreślenie map tomograficznych interesującego rejonu kopalni i na tej podstawie wnioskowanie o obszarach szczególnie zagrożonych. Metodami predykcji ciągłej jest metoda prognozy liniowej, która polega na przewidywaniu sumarycznej (sejsmoakustycznej i sejsmologicznej) energii jaka wydzieli się w wyrobisku w zadanym horyzoncie czasu. Przedstawiona m.in. w pracach [13], [14] koncepcja prognozy liniowej wykorzystuje aparat matematyczny stosowany w predykcji szeregów czasowych i umożliwia godzinową predykcję logarytmowanej sumarycznej energii zjawisk rejestrowanych przez wybrany geofon oraz zarejestrowanych w danym wyrobisku zjawisk sejsmicznych. Ze względu na posługiwanie się wartościami będącymi logarytmem z wartości rzeczywistych niewielkie błędy prognozy energii logarytmowanej mogą przekładać się na duże rzeczywiste błędy prognozy. Metoda poza prognozowaną wartością energii podaje także przedziały ufności dla prognozy. Inną metodą predykcji ciągłej jest metoda funkcji wskaźnikowych bazująca na probabilistycznej analizie hazardu sejsmicznego. Do estymacji wartości funkcji wskaźni-

kowych wykorzystywana jest emisja sejsmoakustyczna (dokładniej odstępy czasu pomiędzy zjawiskami). Autorzy metody stwierdzają, że wartość funkcji wskaźnikowej stanowi podstawę do oceny zagrożenia oraz określania czasu wystąpienia wstrząsów. Przeprowadzane badania eksperymentalne wskazują, że w metodzie konieczne jest dopracowanie sposobu przekładania wartości funkcji wskaźnikowej na ocenę stanu zagrożenia i możliwość wystąpienia wstrząsu. Autor metody przyjmuje, że wzrost wartości funkcji wskaźnikowej oznacza wzrost zagrożenia, ewentualnie możliwość wystąpienia wstrząsu, nie określa jednak jak wartości oraz tempo wzrostu funkcji wskaźnikowej przekładają się na konkretną ocenę zagrożenia oraz okres w jakim ocena ta będzie obowiązywać. Bez dokładnego określenia tych wartości, metoda może powodować zbyt dużo tzw. fałszywych alarmów.

Niniejszy artykuł przedstawia możliwości zastosowania metody maszynowego uczenia jaką jest indukcja reguł logicznych do utworzenia klasyfikatora umożliwiającego klasyfikację dwóch stanów zagrożenia (zagrożony, niezagrożony). Idąc za pracami prof. Kornowskiego uznajemy, że stan w którym sumaryczna energia sejsmoakustyczna i sejsmiczna przekroczą w zadanym okresie wartość 1×10^5 J uznać należy za zagrożony, jednak w opisanych eksperymentach przesunęliśmy akceptowalny zakres energii do 5×10^5 J. W ten sposób nie staramy się przewidywać wystąpienia konkretnego wstrząsu, a jedynie na podstawie gromadzonych danych pomiarowych przewidywać z zadanym wyprzedzeniem możliwość zaistnienia sytuacji niebezpiecznej. Ważnym aspektem prowadzonych przez nas prac jest również to, że proponowana przez nas metoda może wykorzystywać wszystkie dostępne w stacji geofizyki górniczej dane, w tym również wyniki prognoz generowane przez obecnie obowiązujące metody szczegółowe i metodę kompleksową.

Artykuł zorganizowany jest w następujący sposób, w kolejnym rozdział przedstawiono źródła oraz sposób przygotowania danych, które następnie poddawane są analizie. W rozdziale trzecim przedstawiono w zarysie algorytm indukcji reguł i metody polepszania zdolności predykcyjnych uzyskanego klasyfikatora. W rozdziale czwartym przedstawiono wyniki badań eksperymentalnych, wreszcie w rozdziale piątym zamieszczono wnioski.

2. PRZYGOTOWANIE DANYCH DO ANALIZY

Dane pomiarowe wykorzystane do przeprowadzonych eksperymentów pochodziły z systemu wspomagania stacji geofizyki górniczej Hestia. Hestia jest konsumentem danych dostarczanych przez system sejsmoakustyczny ARES oraz sejsmologiczny ARAMIS. Poza danymi pomiarowymi Hestia umożliwia także przechowanie danych o wykonanych wierceniach małośrednicowych, postępie ściany itd. Na podstawie gromadzonych danych, Hestia generuje zmianowe bądź dzienne prognozy dotyczące stanu zagrożenia tąpaniami. Wartości prognoz są wynikiem stosowania metod szczegółowych (sejsmologicznej, sejsmoakustycznej, wierceń małośrednicowych) oraz metody kompleksowej. Szerszą charakterystykę systemu wspomagania stacji geofizyki górniczej można znaleźć m.in. w [19].

Baza danych systemu Hestia jest relacyjną bazą danych zarządzaną przez system SQL Server 2008 firmy Microsoft; umożliwia to wykonywanie różnego rodzaju operacji agregowania danych. Do przeprowadzonego eksperymentu wybierano dane gromadzone w KWK Wesoła. W badaniach rozważano dwie ściany wydobywcze, dane agregowano w okresach godzinowych oraz zmianowych. Po agregacji danych dla każdego z rozważanych wyrobisk dostępne były następujące dane:

- zmianowa ocena wynikająca z metody sejsmologicznej (wartości: a, b, c, d),
- zmianowa ocena wynikająca z metody sejsmoakustycznej (wartości: a, b, c, d),
- informacja o tym czy zmiana jest wydobywcza czy nie,
- maksymalna sumaryczna energia (umowna) rejestrowana w czasie zmiany przez geofony przyporządkowane w systemie do rozważanego wyrobiska (dla ułatwienia dalszego zapisu przyjmijmy, że geofon z maksymalną energią sumaryczną oznaczymy przez GMax),
- maksymalna sumaryczna liczba impulsów rejestrowana przez GMax,
- odchyłka energii rejestrowanej przez Geofon GMax (sposób obliczania odchyłki był zgodny z obowiązującą metodą sejsmoakustyczną [26]),
- odchyłka liczby impulsów rejestrowanych przez Geofon GMax (sposób obliczania odchyłki był zgodny z obowiązującą metodą sejsmoakustyczną [26]),
- ocena zagrożenia według metody sejsmoakustycznej obliczona dla geofonu GMAx (wartości: a, b, c, d),
- liczba zjawisk sejsmicznych zarejestrowanych w czasie zmiany (obliczana w sumie oraz dla każdej klasy energetycznej oddzielnie),
- sumaryczna energia zarejestrowanych w czasie zmiany zjawisk sejsmicznych,
- maksymalna energia zarejestrowanych czasie zmiany zjawisk sejsmicznych.

Jeśli do wyrobiska przyporządkowanych było więcej niż jeden geofon, to w zbiorze zmiennych pojawiały się także:

- średnia energia (umowna) rejestrowana w czasie zmiany przez geofony przyporządkowane w systemie do rozważanego wyrobiska
- średnia liczba impulsów rejestrowanych przez geofony przyporządkowane do wyrobiska,
- średnia odchyłka energii rejestrowanej przez geofony,
- średnia odchyłka liczby impulsów rejestrowanych przez geofony.

Zmienną poddawaną predykcji była sumaryczna energia sejsmoakustyczna (rejestrowana przez geofon GMax) i sejsmologiczna jaka zarejestrowana została w czasie zmiany. W opisanych eksperymentach horyzont prognozy wynosił jeden, to znaczy, że przewidujemy energię z wyprzedzeniem o jedną zmianę. Aby nie przewidywać dokładnych wartości, zakres energii podzielony został na dwa przedziały poniżej i powyżej 5×10^5 J. Wartości poniżej 5×10^5 J uznano za stany bezpieczne, wartości powyżej 5×10^5 J uznano za stany niebezpieczne.

W przypadku agregacji godzinowej obliczano oczywiście wartości maksymalnej, średniej energii, odchyłek energii i impulsów itd. w okresach godzinowych, a predykcja stanu zagrożenia dotyczyła horyzontu jednej godziny. Ponieważ oceny według metod szczegółowych wykonuje się najczęściej raz na zmianę, w przypadku agregacji godzinowej wartości zmiennych, w których zwarto informacje o wynikach ocen szczegółowych zmieniały się co osiem rekordów (czyli co osiem godzin).

Na zakończenie warto odnotować jak duże były przygotowane zbiory danych oraz jaki był rozkład liczby przykładów reprezentujący stany „bezpieczny", „niebezpieczny":

- wyrobisko I – Sc508
  - agregacja zmianowa 864 rekordy, z czego 97 przyporządkowano do stanu „zagrożony",
  - agregacja godzinowa 1487 rekordy, z czego 123 przyporządkowano do stanu „zagrożony",
- wyrobisko II – SC503
  - agregacja zmianowa 1097 rekordy, z czego 188 przyporządkowano do stanu „zagrożony",
  - agregacja godzinowa 1489 rekordy, z czego 3 przyporządkowano do stanu „zagrożony".

Jak widać dane godzinowe analizowane były z znacznie krótszego czasu niż dane zmianowe; wynika to z faktu, iż w czasie pobierania danych autorom zależało na umieszczeniu w zbiorze analizowanych danych również danych o postępie ściany. Ze względu na fakt, iż postęp ściany wprowadzany jest do systemu przez operatora, informacji takich jest w bazie danych niewiele.

W przypadku danych zmianowych w analizowanym okresie w wyrobisku I zarejestrowano 563 zjawiska o energii w klasie 10^3 J, 82 w klasie 10^4 J i 10 zjawisk o energii 10^5 J. W przypadku wyrobiska II liczby te wynosiły odpowiednio: 803, 128, 6.

3. METODA INDUKCJI I DOSTRAJANIA KLASYFIKATORA REGUŁOWEGO

Technika indukcji reguł logicznych jest techniką maszynowego uczenia [7],[11],[16] realizującą paradygmat uczenia na podstawie przykładów. Indukcja reguł ma również zastosowanie w intensywnie rozwijającej się dziedzinie informatyki, jaką jest odkrywanie wiedzy w bazach danych. Reguły jako intuicyjne i łatwo interpretowalne zależności wykorzystywane są do celów opisu i klasyfikacji. Reguły logiczne reprezentowane są zazwyczaj za pomocą zależności (1)

$$\text{IF } a_1 \in V_{a_1} \text{ and} \ldots \text{and } a_k \in V_{a_k} \text{ THEN } d=v_d \quad (1)$$

Indukcji reguł dokonuje się na podstawie treningowego zbioru danych $\boldsymbol{DT}=(U, A\cup\{d\})$, w którym U jest skończonym zbiorem obiektów (rekordów) charakteryzowanym przez zbiór cech (atrybutów warunkowych) A oraz atrybut decyzyjny d. Każdy atrybut $a\in A$ traktowany jest jako funkcja $a:U\rightarrow D_a$, gdzie D_a jest zakresem atrybutu a. Konsekwencją przyjętego zapisu jest to, że w regule postaci (1) mamy $\{a_1,..,a_k\}\subseteq A$, $V_{a_i} \subseteq D_{a_i}$ oraz $v_d\in D_d$. Wyrażenie $a\in V$ nazywane jest deskryptorem warunkowym. Zbiór obiektów o identycznych wartościach atrybutu decyzyjnego nazywa się klasą decyzyjną (ozn. $X_v=\{x\in U: d(x)=v\}$).

Indukcji reguł na podstawie danych zawartych w tablicy treningowej dokonać można wykorzystując różne algorytmy generujące zarówno tzw. minimalne reguły decyzyjne [17], [22],[25] jak i wykorzystujące metodę sekwencyjnego pokrywania [7],[9],[11],[16]. Wszystkie algorytmy wykorzystują pewne miary, które decydują bądź o postaci wyznaczanej reguły, bądź o tym, które spośród wyznaczonych już reguł można usunąć lub połączyć. Miary te nazywane są miarami oceniającymi jakość reguł i ich głównym celem jest takie pokierowanie procesem indukcji i/lub redukcji, aby w wynikowym zbiorze reguł znalazły się reguły o jak najlepszej jakości. Zbiór złożony z reguł o dobrych zdolnościach uogólniania (wysoka dokładność klasyfikacji) i opisu (niewielka wyjściowa liczba reguł) jest zbiorem reguł o wysokiej jakości.

3.1. Miary oceniające jakość reguł

Każdą regułę postaci (1) znajdującą się w zbiorze *RUL* można zapisać w postaci $\varphi\rightarrow\psi$. Dowolna reguła ustala zatem dwa podziały zbioru U, każdy wyznaczany odpowiednio przez poprzednik φ (ozn. U_φ) i następnik ψ (ozn. U_ψ) reguły. Zbiór U można zatem zapisać jako sumę zbiorów $U=U_\varphi\cup U_{\neg\varphi}$ oraz $U=U_\psi\cup U_{\neg\psi}$.

Obiekt $x\in U$ rozpoznaje regułę postaci (1) wtedy i tylko wtedy, gdy $\forall i\in\{1,..,k\}\ a_i(x)\in V_{a_i}$. Obiekt $x\in U$ wspiera regułę postaci (1) wtedy, gdy ją rozpoznaje oraz $d(x)=v_d$

Tablicę kontyngencji dla reguły $r\equiv\varphi\rightarrow\psi$ przedstawia się w następujący sposób:

$n_{\varphi\psi}$	$n_{\varphi\neg\psi}$	n_φ
$n_{\neg\varphi\psi}$	$n_{\neg\varphi\neg\psi}$	$n_{\neg\varphi}$
n_ψ	$n_{\neg\psi}$	

gdzie: $n_\varphi= n_{\varphi\psi}+ n_{\varphi\neg\psi}=|U_\varphi|$ liczba obiektów rozpoznających regułę $\varphi\rightarrow\psi$; $n_{\neg\varphi}= n_{\neg\varphi\psi}+ n_{\neg\varphi\neg\psi}=|U_{\neg\varphi}|$ liczba obiektów nie rozpoznających reguły $\varphi\rightarrow\psi$; $n_\psi= n_{\varphi\psi}+ n_{\neg\varphi\psi}=|U_\psi|$ liczba obiektów należących do klasy decyzyjnej wskazywanej przez regułę $\varphi\rightarrow\psi$; $n_{\neg\psi}= n_{\varphi\neg\psi}+ n_{\neg\varphi\neg\psi}=|U_{\neg\psi}|$ liczba obiektów nie należących do klasy decyzyjnej opisywanej przez regułę $\varphi\rightarrow\psi$; $n_{\varphi\psi}=|U_\varphi\cap U_\psi|$ liczba obiektów wspierających regułę $\varphi\rightarrow\psi$; $n_{\varphi\neg\psi}=|U_\varphi\cap U_{\neg\psi}|$; $n_{\neg\varphi\psi}=|U_{\neg\varphi}\cap U_\psi|$;$n_{\neg\varphi\neg\psi}=|U_{\neg\varphi}\cap U_{\neg\psi}|$.

Dwie podstawowe miary oceniające to dokładność (2) i pokrycie reguły (3)

$$q^{acc}(\varphi\rightarrow\psi)=\frac{n_{\varphi\psi}}{n_\varphi} \tag{2}$$

$$q^{cov}(\varphi\rightarrow\psi)=\frac{n_{\varphi\psi}}{n_\psi} \tag{3}$$

Obie miary rozważane jednocześnie dają pełny obraz jakości reguły. Zgodnie z zasadą indukcji enumeracyjnej [1] przyjmuje się, że reguły o dużej dokładności i pokryciu odzwierciedlają prawdziwe zależności, które prawdziwe są również dla obiektów spoza analizowanego zbioru danych.

Łatwo wykazać, że wraz ze wzrostem dokładności maleje pokrycie reguły, stąd duża liczba prób nad zdefiniowaniem miar oceniających, które jednocześnie uwzględniają dokładności i pokrycie reguł [2],[5],[10],[20],[21]. Ocenianie jednocześnie dokładności i pokrycia reguł ma znaczenie zwłaszcza w przypadku danych niepewnych, które mogą zawierać błędy pomiarowe lub przekłamania. W tym przypadku niektóre z reguł dokładnych opisują bowiem właśnie przypadki błędne. W opisanych pracach do indukcji reguły wykorzystano miarę zaproponowaną przez Cohena (4), która swoją genezę posiada w statystyce matematycznej, a dokładniej w dwuwymiarowej analizie dyskretnej:

$$q^{Cohen}(\phi\rightarrow\psi)=\frac{nn_{\varphi\psi}+nn_{\neg\varphi\neg\psi}-n_\varphi n_\psi-n_{\neg\varphi}n_{\neg\psi}}{n^2-n_\varphi n_\psi-n_{\neg\varphi}n_{\neg\psi}} \tag{4}$$

Dla dowolnej reguły miara Cohena mierzy siłę zależności pomiędzy zdarzeniami:

- „obiekt u rozpoznaje regułę", a „obiekt u należy do klasy decyzyjnej opisywanej przez tę regułę", co odzwierciedla różnica $nn_{\varphi\psi}-n_\varphi n_\psi$;
- „obiekt u nie rozpoznaje reguły", a „obiekt u nie należy do klasy opisywanej przez regułę", co odzwierciedla różnica $nn_{\neg\varphi\neg\psi}-n_{\neg\varphi}n_{\neg\psi}$.

3.2 Algorytm indukcji reguł

Proces indukcji (tworzenia) reguły postaci (1) w oparciu o pewien zbiór danych polega na wyborze atrybutów warunkowych, które będą tworzyły deskryptory warunkowe reguły oraz na ustaleniu zakresów tych deskryptorów (czyli zbiorów V_a). Dla ustalonego atrybutu $a\in A$, zakres deskryptora może mieć jedną z trzech postaci: prostą $a=v$, gdzie $v\in D_a$, przynależnościową $a\in V$, gdzie $V\subseteq D_a$ lub nierównościową $a<v$ lub $a>v$, gdzie $v\in D_a$.

Poniżej krótko omówiono zmodyfikowaną wersję [20],[21] algorytmu MODLEM [23], która dopuszcza tworzenie reguł tzw. aproksymacyjnych, a więc takich, które mogą być niespójne ze zbiorem danych treningowych. W przypadku danych zaszumianych reguły aproksymacyjne lepiej wychwytują zależności występujące w analizowanym zbiorze danych.

Algorytm MODLEM działa w ten sposób, że dla każdego atrybutu warunkowego i dla każdej wartości tego atrybutu występującej w bieżąco rozpatrywanym zbiorze obiektów U (początkowo jest to cały zbiór treningowy U) testuje się kolejne wartości atrybutów warunkowych (uprzednio posortowane niemalejąco), szukając tzw. punktu granicznego g.

Punkt graniczny znajduje się w środku, pomiędzy dwoma kolejnymi wartościami atrybutu a (np. $v_a<g<w_a$) i dzieli bieżący zakres wartości atrybutu a na dwie części. Taki podział ustala również podział bieżącego zbioru obiektów treningowych na dwa podzbiory U_1 oraz U_2. Optymalny jest ten punkt graniczny, który minimalizuje wartość poniższego wyrażenia (5).

$$\frac{|U_1|}{|U|}Entr(U_1)+\frac{|U_2|}{|U|}Entr(U_2) \qquad (5)$$

gdzie:
$Entr(U_i)$ oznacza entropię zbioru U_i.

Jako deskryptor warunkowy wybiera się ten z dwóch przedziałów, dla którego w odpowiednich zbiorach U_1, U_2 znajduje się więcej przykładów z klasy decyzyjnej, na którą wskazuje reguła.

Deskryptor dodawany jest do deskryptorów utworzonych wcześniej i razem z nimi (w formie koniunkcji warunków) tworzy część warunkową reguły.

Jeśli dla jakiegoś atrybutu a w kolejnych krokach algorytmu wybrane zostaną dwa punkty graniczne, to utworzony deskryptor ma postać $[g_1,g_2]$. Jeśli punkt graniczny będzie jeden, to deskryptor będzie w postaci nierówności $a<g$ ($a>g$). Jeśli atrybut nie wygeneruje żadnego punktu granicznego, to atrybut nie wystąpi w części warunkowej reguły.

W klasycznej wersji algorytmu MODLEM proces tworzenia reguły kończy się z chwilą, kiedy jest ona dokładna lub dokładna „na tyle na ile jest to możliwe w analizowanym zbiorze treningowym".

Algorytm tworzy pokrycie danej klasy decyzyjnej, po wygenerowaniu reguły usuwane są ze zbioru treningowego wszystkie obiekty wspierające utworzoną regułę, a algorytm stosowany jest do pozostałych obiektów z opisywanej klasy.

Generowanie reguł dokładnych powoduje niekorzystną (zwłaszcza dla danych zaszumionych) sytuację polegającą na tym, iż reguły dokładne sa zbytnio dopasowane do danych treningowych, przez co część z nich reprezentuje zależności nieprawdziwe, które w oczywisty sposób wpływać będą zarówno na jakość uzyskiwanej klasyfikacji, jak również na uzyskaną wiedzę.

Modyfikacja algorytmu MODLEM polega na zastosowaniu miary oceny jakość reguł do oceny tworzonej na bieżąco części warunkowej reguły. Po dodaniu (lub modyfikacji) kolejnego deskryptora warunkowego oceniana jest bieżącą postać reguły. Jako reguła wyjściowa pamiętana jest ta, która uzyskała najlepszą ocenę. Proces tworzenia reguły kończy się, kiedy dodanie nowego deskryptora do części warunkowej powoduje spadek jakości tworzonej reguły. Zmodyfikowaną wersję algorytmu MODLEM można skrótowo zapisać w następującej postaci:

$RUL:=\varnothing$; $P:=U$; $q:=-1$;
Dla każdej wartości v atrybutu decyzyjnego d
  Utwórz regułę r bez przesłanek, której konkluzja jest postaci (d,v)
  $G:=\{x: d(x)=v\}$
  Dopóki $G\neq\varnothing$ **powtarzaj**
    Dla każdego atrybutu warunkowego a
    Znajdź najlepszy deskryptor (a, Va)
    Dodaj deskryptor do części przesłankowej reguły r
    Ogranicz zbiór U do obiektów rozpoznawanych przez r
    Oblicz wartość miary oceniającej $q(r)$
    Jeśli $q(r)<q$ **to**
      Usuń ostatnią modyfikację reguły r
      $RUL:=RUL\cup\{r\}$
      $G:=G-supp(r)$ ($supp(r)$ to zbiór obiektów wspierających regułę r)
      Rozszerz zbiór ($U:=P-G$)
      Utwórz nową regułę r bez przesłanek, której konkluzja jest postaci (d,v)
    w przeciwnym przypadku $q:=q(r)$
  Koniec // Dopóki
Koniec

Niezależnie od sposobu wyznaczania, wyjściowy zbiór reguł można poddać redukcji za pomocą algorytmu filtracji [21]. Filtracja polega na usuwaniu ze zbioru reguł, tych reguł które są nieistotne zarówno dla czytelności opisu jak również zdolności uogólniania klasyfikatora. Regułowym opisem klasy decyzyjnej X_v będziemy nazywali zbiór reguł o konkluzjach postaci $d=v$.

Prosty, aczkolwiek skuteczny, algorytm filtracji „W przód" [21] wykorzystuje ranking reguł utworzony przez dowolną z miar oceniających jakość reguł. Początkowy opis każdej klasy decyzyjnej składa się z jednej – najlepszej reguły, następnie dodaje się po jednej regule do opisu każdej klasy decyzyjnej, jeżeli dokładność klasy się zwiększy, regułę pozostawiamy w opisie, jeżeli nie, rozpatrywana jest kolejna reguła. O kolejności rozpatrywania reguł decyduje ranking reguł ustalony przez miarę oceniającą (reguły sortowane są malejąco względem wartości wykorzystywanej przez algorytm miary oceniającej). Dodawanie reguł do opisu klasy decyzyjnej kończy się z chwilą osiągnięcia identycznej dokładności klasyfikacji, jaką uzyskuje niefiltrowany zbiór reguł lub kiedy zbiór reguły się skończy. Ze względu na kryterium decydujące o dodaniu reguły do nowego opisu klasy decyzyjnej algorytm nie gwarantuje, iż odfiltrowane reguły uzyskają identyczną dokładność klasyfikacji, jak zbiór wszystkich reguł.

3.3. Klasyfikator i ocena jakości klasyfikatora

Wyznaczony zbiór reguł chcemy wykorzystać do wnioskowania o wartości atrybutu decyzyjnego obiektów należących zarówno do tablicy treningowej

jak również tablicy testowej, do której należą obiekty (rekordy) nie występujące w tablicy treningowej.

Przyporządkowując obiektowi odpowiadającą mu wartość atrybutu decyzyjnego możemy spotkać się z trzema przypadkami: żadna z wyznaczonych reguł nie rozpoznaje obiektu testowego, wszystkie reguły rozpoznające obiekt testowy mają identyczne konkluzje, reguły rozpoznające obiekt testowy mają różne konkluzje. W pierwszym przypadku klasyfikator nie jest w stanie podjąć żadnej decyzji i chociaż istnieją sposoby rozwiązania tej niedogodności, to w niniejszym artykule nie będą one stosowane. Kiedy obiekt testowy rozpoznawany jest przez reguły należące do opisu jednej klasy decyzyjnej, wybór decyzji, jaką należy przyznać obiektowi testowemu jest oczywisty. Problem pojawia się wtedy, kiedy obiekt rozpoznawany jest przez reguły należące do opisów różnych klas decyzyjnych. Jaką wartość decyzji przypisać wtedy obiektowi testowemu? Znane są metody rozwiązania tego problemu. Jeśli do każdej z reguł przyporządkowano jej jakość obliczaną zgodnie z jednym ze wzorów (2), (3), (4) to, wzorując się na pracach [8],[11],[16], można m.in.:

– klasyfikować obiekt do tej klasy decyzyjnej, na którą wskazuje reguła o maksymalnej jakości,
– sumować jakość reguł rozpoznających obiekt testowy i klasyfikować obiekt do tej klasy, dla której suma ta jest maksymalna.

Podstawową cechą charakteryzującą efektywność algorytmu decyzyjnego jest jego dokładność klasyfikacji. W celu zbadania dokładności klasyfikacji algorytmu decyzyjnego, dostępną do analizy tablicę decyzyjną dzieli się na dwie części [24]. Pierwszą, na podstawie której dokonujemy indukcji reguł, jest tablica treningowa, drugą – za pomocą której bada się efektywność stworzonego algorytmu decyzyjnego, jest tablica testowa. Efektywność algorytmu rozumiana jest jako stopień poprawności, z jakim klasyfikuje on obiekty z tablicy testowej.

Jeżeli $\boldsymbol{DT_{Ts}}=(U,A\cup\{d\})$ jest pewną testową tablicą decyzyjną oraz RUL jest zbiorem reguł, to współczynnikiem dokładności klasyfikacji algorytmu decyzyjnego wykorzystującego zbiór reguł RUL i klasyfikującego obiekty ze zbioru $\boldsymbol{DT_{Ts}}$ nazywamy liczbę (6).

$$accuracy(RUL,DT_{Ts})=\frac{|\{u\in U: f(u)=d(u)\}|}{|U|} \qquad (6)$$

W wyrażeniu (6) *f* jest funkcją przyporządkowującą obiektowi testowemu wartości atrybutu decyzyjnego wykorzystującą do tego celu zbiór reguł RUL.

Liczba w liczniku wyrażenia (6) to, liczba obiektów poprawnie sklasyfikowanych, tzn. takich, których wartość funkcji decyzyjnej obliczona za pomocą zbioru reguł RUL jest taka sama, jak wartość decyzji wynikająca z tablicy testowej.

Gdy liczebność przykładów reprezentujących poszczególne klasy decyzyjne jest znacząco różna, celowe jest obliczanie współczynnika dokładności klasyfikacji dla każdej klasy decyzyjnej $X_v\subseteq U$ osobno (7).

$$accuracy_{x_v}(RUL,DT_{Ts})=\frac{|\{u\in X_v: f(u)=v\}|}{|X_v|} \qquad (7)$$

W zależności od sposobu podziału wejściowej tablicy na tablicę treningową i testową można otrzymać różne wartości dokładności klasyfikacji. Istnieje kilka metod postępowania pozwalających na wyznaczenie wiarygodnego estymatora współczynnika dokładności klasyfikacji. W zależności od rozmiaru badanych danych najbardziej popularne są metody znane pod nazwami: *train-and-test* i *cross-validation* [16],[18].

Metoda *train-and-test* polega na tym, iż poddawaną analizie tablicę decyzyjną dzieli się, w sposób losowy, na dwie podtablice: treningową i testową. Zazwyczaj podtablice te są rozłączne. W tablicy testowej znajduje się zazwyczaj od 20-50% wszystkich dostępnych obiektów. Metodę *train-and-test* stosuje się, kiedy analizowany zbiór danych zawiera więcej niż 1000 obiektów. Metoda *cross-validation* stosowana jest wtedy, gdy liczba obiektów w poddawanej analizie tablicy decyzyjnej jest mniejsza niż 1000. Metoda polega na tym, że dane dzieli się w sposób losowy na *r* równolicznych i rozłącznych podzbiorów, a następnie wykonuje się *r* eksperymentów. W każdym z eksperymentów, jeden z *r* podzbiorów jest zbiorem testowym, a suma pozostałych *r-1* jest zbiorem treningowym. Po każdym eksperymencie oblicza się wartość współczynnika dokładności klasyfikacji. Po wykonaniu *r* eksperymentów, ostateczną wartość współczynnika dokładności klasyfikacji oblicza się jako średnią arytmetyczną współczynników ze wszystkich *r* eksperymentów. Zazwyczaj liczba *r* jest liczbą całkowitą z przedziału 5 do 15. W niniejszym artykule jako estymator dokładności klasyfikacji algorytmu decyzyjnego wykorzystano metodą 5-*fold cross-validation*.

4. WYNIKI BADAŃ EKSPERYMENTALNYCH

W celu weryfikacji jakość klasyfikatorów uzyskanych metodami przedstawionymi w poprzednim rozdziale, przeanalizowano dane zmianowe i godzinowe pochodzące z dwóch wyrobisk. W tablicy

pierwszej zamieszczono wyniki uzyskane przez algorytm decyzyjny, w którym indukcji reguł dokonano wykorzystując miarę Cohena. Po indukcji reguł zastosowano algorytm filtracji, który jako kryterium jakości reguły również wykorzystywał miarę Cohena. Miara ta stosowana również była w czasie klasyfikacji obiektów testowych. W tablicy pierwszej podano wyniki dokładności dla każdej klasy decyzyjnej oddzielnie, w przypadku SC508 podano oddzielnie wyniki dla tablicy decyzyjnej, w której wykorzystano uśrednione i maksymalne wartości pomiarowe z geofonów. Ściana SC503 monitorowana była jedynie przez jeden geofon (takie dane dostępne były w systemie Hestia). Poza średnią dokładnością klasyfikacji uzyskaną na zbiorze danych testowych (jak już wspominano wykorzystano 5-*fold cross-validation* jako metodologie testowania) podano także odchylenie standardowe. Wartości w tablicy pierwszej podane są w procentach

Tablica 1
Wyniki klasyfikacji stanów zagrożenia za pomocą algorytmu indukcji reguł

Zbiór danych	Dokładność stan „zagrożony”	Dokładność stan „niezagrożony”
Zmianowe SC503	74.4 ± 0.00	94.6 ± 0.00
Zmianowe SC508-max	84.6 ± 0.02	73.7 ± 0.00
Zmianowe SC508-śr	68.3 ± 0.05	79.5 ± 0.00
Godzinowe SC503	0.0 ± 0.00	97.3 ± 0.00
Godzinowe SC508-max	73.1 ± 0.00	67.5 ± 0.00
Godzinowe Sc508-śr	79.7 ± 0.00	62.6 ± 0.00

Analiza wyników zamieszczonych w tablicy pierwszej prowadzi do konkluzji, że lepsze wyniki predykcji stanu zagrożenia uzyskujemy dla agregacji zmianowej (i zmianowego horyzontu predykcji). Oczywiście z punktu widzenia użytkownika (operatora stacji geofizycznej) bardziej użyteczne byłoby wykorzystywanie prognozy godzinowej. W skrajnym przypadku (Godzinowe SC503) dokładność predykcji stanu „zagrożony” wynosi 0%. Rozpatrując jedynie ten wynik można by powiedzieć, że proponowana metoda nie jest dobra dla rozwiązywania zadania prognozy stanu zagrożenia sejsmicznego. Należy jedna pamiętać, że w przypadku zbioru (Godzinowe SC503) dysponowaliśmy jedynie trzema rekordami wskazującymi na stan „zagrożone”, a zatem w czasie testowania klasyfikator uczył się jedynie na podstawie informacji zawartej w dwóch rekordach i testował jakość nauki na jednym rekordzie. Widać, że w przypadku ściany SC508 jakość prognoz godzinowych jest już dobra.

Warto zauważyć, że pomimo nierównomiernego rozkładu przykładów pomiędzy klasy decyzyjne dokładności klasyfikacji poszczególnych klas nie różnią się zbytnio od siebie. Efekt ten uzyskano stosując odpowiednią miarę oceniającą oraz filtrując większość reguł wskazujących na klasę decyzyjną „niezagrożony”. Nie wgłębiając się zbytnio w szczegóły techniczne zauważmy o jak ważnej cesze algorytmu decyzyjnego mówimy. Pracownikowi stacji geofizycznej zależy na dokładnym przewidywaniu stanu „zagrożony”, ważne jest jednak również to, że system nie powinien w zbyt wielu przypadkach przewidywać stanu „zagrożonego” w przypadku kiedy będzie on „niezagrożony”. Zbyt częste i niedokładne przewidywanie stanu „zagrożony” spowoduje utratę zaufania i „znieczulenie” operatora na ostrzeżenia generowane przez system. Jednym słowem zależy nam na tym, aby dokładność obu klas decyzyjnych była wysoka i w zasadzie podobna. Zauważmy, że w przypadkach prezentowanych w tablicy pierwszej mamy do czynienia z taką właśnie sytuacją.

Dla celów porównania z innymi metodami, w tablicy drugiej podano wyniki klasyfikacji uzyskane na tych samych zbiorach danych za pomocą programów RSES [3] (algorytmy LEM, ze skracaniem reguł [9]; lub *exhaustive algorithm* [22] – w zależności od tego, który z algorytmów da lepsze wyniki) oraz CART [4] (*gini* jako kryterium podziału węzła), które są znanymi programami umożliwiającymi indukcję reguł decyzyjnych.

Tablica 2
Wyniki klasyfikacji stanów zagrożenia uzyskane za pomocą programów RSES i CART

Zbiór danych		Dokładność stan „zagrożony”	Dokładność stan „niezagrożony”
RSES	Zmianowe SC503	49.4	97.1
	Zmianowe SC508-max	46.4	93.0
	Zmianowe SC508-śr	50.0	89.3
	Godzinowe SC503	0.0	97.4
	Godzinowe SC508-max	33.4	92.6
	Godzinowe Sc508-śr	29.1	91.5
CART	Zmianowe SC503	87.7	86.8
	Zmianowe SC508-max	81.4	73.9
	Zmianowe SC508-śr	88.6	65.7
	Godzinowe SC503	0	97.3
	Godzinowe SC508-max	87.8	50.8
	Godzinowe Sc508-śr	82.9	65.4

Porównując wyniki zamieszczone w tablicach pierwszej i drugiej widać, że w przypadku danych sejsmicznych wyniki uzyskane za pomocą metody przedstawionej w rozdziale 3.2 oraz za pomocą programu CART są porównywalne. Widać także, że w przypadku predykcji zmianowej lepiej wykorzystywać maksymalne wartości rejestrowane przez geofon. Z inną sytuacją mamy do czynienia w przypadku predykcji godzinowej, tutaj lepszą dokładność klasyfikacji uzyskuje się dla zbioru z uśrednionymi wynikami pomiarów z wszystkich przyporządkowanych do wyrobiska geofonów.

Na zakończenie części eksperymentalnej opisano wyniki klasyfikacji polegające na zastosowaniu klasyfikatora uzyskanego na podstawie analizy danych pochodzących z jednego wyrobiska do klasyfikacji danych pochodzących z drugiego z rozważanych w artykule wyrobisk. W tym przypadku klasyfikator trenowano na całym dostępnym zbiorze danych i stosowano do całego dostępnego zbioru danych pochodzącego z drugiego wyrobiska. Wyniki dla danych zmianowych umieszczono w tablicy trzeciej, dla danych godzinowych podobnej analizy nie przeprowadzono, ze względu na niewielką liczbę przykładów stanów „zagrożony" dla ściany SC503.

Tablica 3
Wyniki stosowania reguł uzyskanych w jednym wyrobisku do danych pochodzących z innego wyrobiska

Zbiór danych – uczący Zbiór danych - testujący	Dokładność stan „zagrożony"	Dokładność stan „niezagrożony"
Zmianowe SC503 – uczący Zmianowe SC508-max - testujący	53.6	91.3
Zmianowe SC508-max – uczący Zmianowe SC503 – testujący	76.0	89.5

Jak widać stosowanie klasyfikatora trenowanego na danych pochodzących z innego wyrobiska daje bardzo dobre rezultaty w przypadku przewidywania stanu „niezagrożony" i zadowalające lub dobre w przypadku stanu zagrożony. Lepszy wynik dla stanu zagrożony uzyskano trenując klasyfikator na danych pochodzących z wyrobiska SC508. Uzyskane wyniki mają duże znaczenie dla praktycznej implementacji klasyfikatorów regułowych w systemie Hestia. W przypadku nowego wyrobiska, stan zagrożenia może być oceniany według modelu uzyskanego na podstawie analizy danych pochodzących z innego wyrobiska. Widać, że jakość tak generowanych prognoz jest jednak nieznacznie gorsza niż prognoz generowanych na podstawie analizy danych pochodzących z wyrobiska poddawanego ocenie; wynika stąd wniosek, że w miarę napływu danych pomiarowych z nowego wyrobiska należy dążyć do ponownego wyznaczenia regułowego modelu danych.

5. PODSUMOWANIE

W artykule przedstawiono możliwości zastosowania metody maszynowego uczenia, jaką jest indukcja reguł do rozwiązania problemu predykcji stanu zagrożenia sejsmicznego w wyrobisku górniczym. Przedstawiono algorytm indukcji reguł, a także wyniki eksperymentów przeprowadzanych na danych pochodzących z dwóch ścian KWK Wesoła. Uzyskane wyniki porównano z innymi algorytmami umożliwiającymi indukcję reguł. Źródłem danych był system wspomagania stacji geofizyki górniczej Hestia, dane przed analizą musiałyby zostać zagregowane do godzinowych i zmianowych przedziałów czasu.

Uzyskane wyniki wykazują, że przedstawiona metoda jest interesującą alternatywą dla metod predykcji liniowej i funkcji wskaźnikowych, dobra jakość predykcji (zweryfikowana na niezależnych zbiorach danych) pokazuje, że możliwe jest stosunkowo dokładne przewidywanie w pewnym przedziale czasu nadchodzącego zagrożenia. Zastosowanie metody w praktyce górniczej wymaga jednak dalszych badań. Obecnie trwają prace nad poprawą dokładności klasyfikacji poprzez zastosowanie również innych miar dokładności oraz zastosowaniu reguł rozmytych do predykcji stanu zagrożenia. Wykonana zostanie większa liczba eksperymentów porównujących różne metody eksploracyjnej analizy danych oraz metod obliczeń miękkich [15] (ang. *soft computing*). Przeanalizowane zostanie także, które z monitorowanych zmiennych mają największy wpływ na podejmowane decyzje oraz jakie są wzajemne powiązania pomiędzy zmiennymi. Autorzy chcieliby także zbadać, czy stosując wszystkie rozwijane metody predykcji zagrożenia uzyska się efekt synergii polegający na poprawienie dokładności predykcji, w tym kontekście szczególnie interesujące byłoby wykorzystanie wartości funkcji wskaźnikowych i wyników predykcji liniowej jako zmiennych warunkowych w algorytmie indukcji reguł dokonujących oceny stanu zagrożenia.

Literatura

1) *Ajdukiewicz K.*: Logika pragmatyczna. Warszawa: PWN, 1974.
2) *An, A., Cercone, N.*: Rule quality measures for rule induction systems – description and evaluation. Computational Intelligence 17 (2001) 409-424.

3) *Bazan, J., Szczuka, M., Wróblewski, J.*: A new version of rough set exploration system. Lecture Notes in Computer Sciences 2475, Springer (2002) 14-16.
4) *Breiman, L., Friedman J., Olshen R., Stone R.*: Classification and Regression Trees. Pacific Grove: Wadsworth (1984).
5) *Bruha, I.*: Quality of Decision Rules: Definitions and Classification Schemes for Multiple Rules. In: Nakhaeizadeh, G., Taylor, C.C. (Eds.) Machine Learning and Statistics, The Interface. Wiley, NY, USA (1997) 107-131.
6) *Cianciara A., Cianciara B.*: The meaning of seismoacustic emission for estimation of time of the mining tremors occurrence. Archives of Mining Sciences. Vol. 51(4) 2006, pp.563-575.
7) *Cichosz P.*: Systemu uczące się. WNT Warszawa 2000.
8) *Grzymała-Busse J., Wang C. P.*: Classification Methods in Rule Induction. Intelligent Information Systems. Proceedings of the Workshop held in Dęblin, Poland 2-5 June, 1996, pp. 120-126.
9) *Grzymała-Busse, J.W.*: LERS - a system for learning from examples based on rough sets. In: Słowiński, R. (Ed.) Intelligent Decision Support. Handbook of applications and advances of the rough set theory. Kluwer Academic Publishers, Dordrecht, Boston, London (1992) 3-18.
10) *Guillet F., Hamilton H.J. (Eds.)*: Quality Measures in Data Mining. Computational Intelligence Series, Springer 2007.
11) *Kubat M., Bratko I., Michalski R.*: Machine learning and data mining. Methods and applications. John Wiley and Sons 1998.
12) *Dubiński J., Lurka A., Mutke I.*: Zastosowanie metody tomografii pasywnej do oceny zagrożenia sejsmicznego w kopalniach. Przegląd Górniczy Nr. 3, 1998.
13) *Kornowski J.*: Linear prediction of aggregated seismic and seismoacustic energy emitted from a mining longwall. Acta Montana Ser. A, No 22 (129), 2003, str.4-14.
14) *Kornowski J.*: Linear prediction of hourly aggregated AE and tremors energy emitted form a longwall and its performance in practice. Archives of Mining Sciences, Vol. 48, No. 3, 2003, str. 315-337.
15) *Łęski J.*: Systemu neuronowo-rozmyte. WNT Warszawa 2008.
16) *Michie D., Spiegelhalter D. J., Taylor C. C.*: Machine Learning, neural and statistical classification. England: Ellis Horwood Limited, 1994.
17) *Pawlak Z.*: Rough sets: Theoretical aspects of reasoning about data. Dordrecht Kluwer 1991.
18) *Salzberg S. L.*: On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Mining and Knowledge Discovery 1, 1997, pp. 317-328.
19) *Sikora M.*: System wspomagania pracy stacji geofizycznej – Hestia. Mechanizacja i Automatyzacja Górnictwa, 12/395, Katowice 2003.
20) *Sikora M.*: Rule quality measures in creation and reduction of data rule models. Lecture Notes in Artificial Intelligence Vol. 4259, Springer-Verlag, Berlin Heidelberg, 2006, pp. 716-725.
21) *Sikora M.*: Decision rules-based data models using TRS and NetTRS - methods and algorithms. Transaction on Rough Sets X. LNCS 5300 (w druku, ukaże się w 2009).
22) *Skowron A., Rauszer C.*: The Discernibility Matrices and Functions in Information systems. Słowiński R. (ed.): Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer, 1992, pp. 331-362.
23) *Stefanowski J.*: Rough set based rule induction techniques for classification problems. In: Proceedings of the 6th European Congress of Intelligent Techniques and Soft Computing, Achen, Germany (1998) 107-119.
24) *Weiss S. M. Kulikowski C. A.*: Computer Systems That Learn. San Mateo: Morgan Kaufmann, 1991.
25) *Zhong, N., Skowron, A.*: A rough set-based knowledge discovery process. International Journal of Applied Mathematics and Computer Sciences. No. 11 (2001) 603-619.
26) Zasady stosowania metody kompleksowej i metod szczegółowych oceny stanu zagrożenia tąpaniami w kopalniach węgla kamiennego. Główny Instytut Górnictwa, Seria Instrukcje Nr 20, Katowice 2007.

Recenzent: prof. dr hab. inż. Jerzy Kornowski

KOMUNIKAT

Centrum Badań i Certyfikacji Centrum Elektryfikacji i Automatyzacji Górnictwa EMAG
– Jednostki Certyfikującej Wyroby: (Certyfikat akredytacji nr AC 053)
o wydanych i cofniętych certyfikatach

Wydano:

1. Certyfikat zgodności nr 1/09 uzyskany w certyfikacji dobrowolnej, system 1b ISO (kwiecień 2009 r.)
   Dostawca: **Fabryka Kabli ELPAR Szczygielski Spółka z o.o., Al. Jana Pawła II, 21-200 Parczew**
   Wyrób: **Przewód wielożyłowy o izolacji i powłoce gumowej, do odbiorników ruchomych i przenośnych**
   Typ (odmiany): **HO5RR-F**

2. Certyfikat zgodności nr 2/09 uzyskany w certyfikacji dobrowolnej, system 1b ISO (kwiecień 2009 r.)
   Dostawca: **Fabryka Kabli ELPAR Szczygielski Spółka z o.o., Al. Jana Pawła II, 21-200 Parczew**
   Wyrób: **Przewód w izolacji i powłoce gumowej, do odbiorników ruchomych i przenośnych**
   Typ (odmiany): **HO7RN-F**

3. Certyfikat zgodności nr 3/09 uzyskany w certyfikacji dobrowolnej, system 1b ISO (kwiecień 2009 r.)
   Dostawca: **Fabryka Kabli ELPAR Szczygielski Spółka z o.o., Al. Jana Pawła II, 21-200 Parczew**
   Wyrób: **Przewód jednożyłowy w izolacji polwinitowej do układania na stałe**
   Typ (odmiany): **HO7V-K**