dr MAREK SIKORA
dr inż. ZDZISŁAW KRZYSTANEK
dr inż. BOŻENA BOJKO
mgr inż. KAROL ŚPIECHOWICZ
Centrum Elektryfikacji i Automatyzacji Górnictwa EMAG

Moduł czyszczenia i agregacji danych jako składnik systemu predykcji stężenia gazów w kopalniach węgla kamiennego

Przedstawiono opracowywany w Centrum EMAG hybrydowy system predykcji stężenia gazów w kopalniach węgla kamiennego, który może być wykorzystany w systemie monitorowania parametrów środowiska w wyrobiskach kopalnianych. Funkcją systemu jest wypracowywanie krótkoterminowych prognoz zmian stężenia metanu na podstawie analizowanych on-line danych metanometrycznych i w razie potrzeby generowanie ostrzeżeń w celu uniknięcia przymusowych, spowodowanych automatycznym wyłączeniem energii, postojów zabezpieczanych przez system metanometryczny obiektów (ścian, przodków). Projektowany system wykorzystuje liniowe i nieliniowe metody prognozy, a w szczególności algorytm M5 zaproponowany przez R. Quinlan'a oraz algorytm predykcji liniowej. Niezbędne do zastosowania tych metod modele matematyczne są opracowywane na podstawie danych historycznych archiwizowanych w bazie danych systemu SMP-NT/A. Omówiono problemy związane z przygotowaniem danych dla procedur tworzących modele predykcyjne oraz przetestowane sposoby ich rozwiązania (eliminacja danych niewiarygodnych, wygładzanie, agregacja). Podano przykłady wyznaczania prognoz z wykorzystaniem metody ex-post.

1. WSTĘP

Artykuł przedstawia rozwijany w Centrum EMAG system predykcji stężenia gazów w kopalniach węgla kamiennego. System może być wykorzystany w celu monitorowania parametrów środowiska w wyrobiskach kopalnianych. Głównym zadaniem systemu jest generowanie ostrzeżeń w celu uniknięcia przymusowych, spowodowanych automatycznym wyłączeniem energii, postojów zabezpieczanych przez system metanometryczny obiektów (ścian, przodków).

Projektowany system predykcji wykorzystuje liniowe i nieliniowe metody prognozy. W szczególności wykorzystywany jest algorytm M5 umożliwiający indukcję reguł z liniowymi konkluzjami [7] oraz algorytm predykcji liniowej. Modele predykcyjne wykorzystywane przez powyższe metody tworzone są na podstawie danych historycznych archiwizowanych w bazie danych systemu metanometrycznego SMP-NT/A.

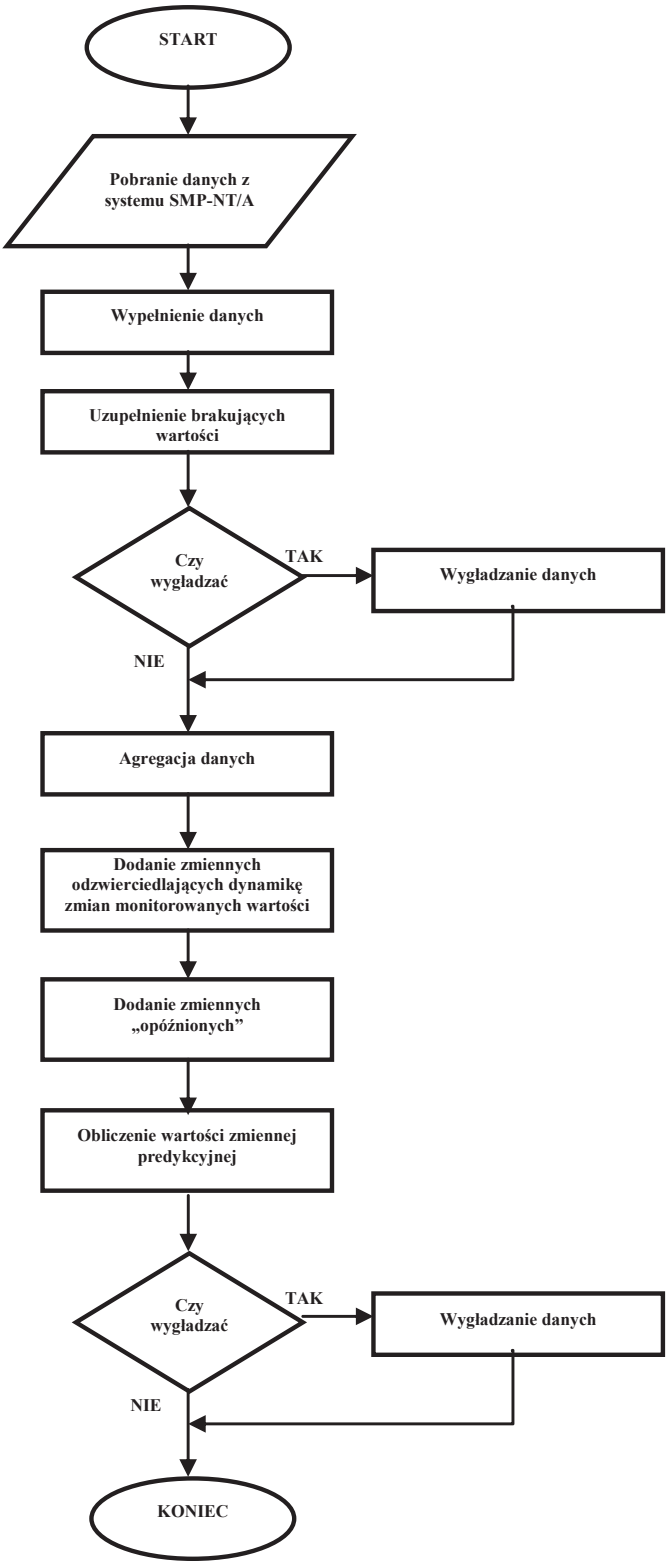
Niezwykle ważnym elementem systemu predykcji jest moduł dokonujący odpowiednich przekształceń surowych danych zawartych w przemysłowych bazach danych. Obróbka danych polega przede wszystkim na wyeliminowaniu błędnych i brakujących wartości, wygładzeniu oraz odpowiedniej agregacji danych. Zadanie agregacji wiąże się także z dodaniem do zbioru zmiennych, parametrów odzwierciedlających dynamikę zmian monitorowanych wielkości (np. tempo przyrostu stężenia metanu). Tak przygotowany zbiór danych wykorzystywany jest w procesie tworzenia modeli predykcyjnych.

Poza charakterystyką systemu predykcji w artykule przedstawiono również sposób ustalania stopnia zagrożenia na podstawie eksperckiej bazy reguł rozmytych, na którą rzutowana jest przewidywana wartość stężenia metanu.

2. PRZYGOTOWANIE DANYCH

Dane archiwizowane przez system metanometryczny, które wykorzystuje się podczas budowy modeli predykcyjnych, muszą zostać odpowiednio przygotowane. Etapy procesu przygotowania danych obrazuje rysunek pierwszy.

Aby ograniczyć przestrzeń dyskową niezbędną do przechowywania danych, system metanometryczny



Rys. 1. Schemat blokowy procesu przygotowania danych

SMP-NT/A nie przechowuje oryginalnych wskazań czujników, które do systemu przekazywane są w odstępach dwusekundowych, lecz przechowuje tylko zmiany wskazań czujników. Tak więc pierwszym etapem przygotowania danych jest takie przekształcenie danych archiwalnych, aby niejako otrzymać na powrót oryginalny zbiór surowych danych pomiarowych (krok *Wypełnienie danych* na rysunku pierwszym). Operacja ta powoduje duże zwiększenie wykorzystania zasobów komputera przetwarzającego dane. Operacja ta jednak jest konieczna ze względu na przyjęty aparat analizy danych.

Po uzyskaniu oryginalnego zbioru surowych danych pomiarowych, dane te poddawane są czyszczeniu. Czyszczenie polega głównie na uzupełnieniu brakujących wartości danych. Przyczyną braków w danych jest najczęściej awaria systemu transmisji danych lub akwizycja wartości niemożliwych (błędnych; przekroczenie zakresu pomiarowego). W realizowanym systemie predykcji gazów braki danych uzupełniane są według następujących metod:

- wpisanie ostatniej z prawidłowo zarejestrowanej wartości,
- wpisanie średniej z *k* ostatnio zarejestrowanych prawidłowych wartości,
- wpisanie wartości wynikającej z interpolacji liniowej, dokonanej na podstawie ostatniej prawidłowej (przed brakiem danych) i pierwszej prawidłowej (po ponownym pojawieniu się dopuszczalnych pomiarów) wartości pomiarowej.

Do wszystkich trzech metod wprowadzono parametr *MaxT* informujący o tym jak długo można uzupełniać brakujące wartości, według metod opisanych powyżej. W przypadku przekroczenia wartości tego parametru (upłynięcia czasu *MaxT*) system nie uzupełnia już brakujących wartości, lecz wpisuje do zbioru danych specjalny symbol, który zarezerwowany jest dla wartości brakujących (jest nim symbol „?”). Oczywiście, w systemie dostępna jest również opcja pozwalająca na to, aby nie uzupełniać brakujących wartości pomiarowych.

Tak przygotowany zbiór danych może być w następnej kolejności poddawany wygładzaniu. W etapie tym wykorzystuje się dwie metody: średnią ruchomą *n*-punktową oraz medianę ruchomą *n*-punktową. Etap wygładzania danych jest etapem opcjonalnym.

Kolejnym etapem przygotowania zbioru danych jest agregacja, której celem jest odzwierciedlenie w pojedynczym rekordzie zbioru danych stanu monitorowanego fragmentu kopalni. Stan monitorowanego fragmentu kopalni w ustalonym czasie agregacji opisywany jest przez wektor wartości zarejestrowanych przez wybrane czujniki oraz dynamikę zmian tych wartości w okresie agregacji. Agregacja danych pozwala na znaczne ograniczenie rozmiaru analizowanego zbioru danych, a w przypadku prognoz o dłuższym horyzoncie prognozy lepiej naszym zdaniem odzwierciedla sytuację w monitorowanym fragmencie kopalni. W czasie agregacji oryginalne wartości danych pomiarowych zastępowane są przez pewną wartość zagregowaną, która będzie reprezentatywną wartością danej wielkości w agregowanym czasie (stąd agregację można potraktować jako swego rodzaju metodę samplingu). W tworzonym systemie predykcyjnym, dopuszczalne są następujące funkcje agregujące:

– maksimum z okresu agregacji,
– minimum z okresu agregacji,
– średnia arytmetyczna z okresu agregacji,
– mediana z okresu agregacji,
– ostatnia wartość z okresu agregacji.

Czas agregacji definiowany jest poprzez odpowiedni parametr programu. Przykładowo – zastosowanie czasu agregacji równego 30 s spowoduje, że każde 15 rekordów (oryginalne rekordy zapisywane są co 2 s) zostanie zastąpionych jednym rekordem. Agregacja może spowodować, że utraci się pewne informacje o dynamice zmian monitorowanych zmiennych. Wprowadzono zatem możliwość dodania nowych zmiennych obrazujących zachowanie agregowanych zmiennych podczas agregacji. Każda z monitorowanych zmiennych jest źródłem nowych zmiennych, odzwierciedlających dynamikę zmian zmiennej źródłowej. Nowymi zmiennymi są:

- *rozstęp* – różnica między największą i najmniejszą wartością zmiennej źródłowej w czasie agregacji,
- *koniec-początek* – różnica pomiędzy pierwszą i ostatnią wartością zmiennej źródłowej w czasie agregacji,
- *dynamika* – dynamika jest zmienną tekstową przyjmującą dwie wartości *nierośnie* oraz *niemaleje*; moduł agregujący zlicza w czasie agregacji sytuacje, w których pomiędzy kolejnymi pomiarami następował wzrost lub spadek monitorowanej wielkości, jeśli więcej (lub tyle samo co spadków) było wzrostów wartości to *dynamika*:= *niemaleje*, jeśli więcej było spadków wartości to *dynamika*:= *nierośnie*,
- *dynamika-koniec* – wartość zmiennej wyliczana jest identycznie jak wartość zmiennej *dynamika;* różnica polega na tym, że rozpatrywanych jest jedynie 25% rekordów „znajdujących się” na końcu czasu agregacji.

Na zakończenie procesu agregacji wyliczana jest wartość zmiennej poddawanej predykcji (zmiennej zależnej). W etapie tym ustalany jest horyzont prognozy oraz sposób ustalenia wartości zmiennej poddawanej predykcji. Możliwe są dwa rozwiązania:

- wstawienie wartości zmiennej zależnej, która odpowiada czasowi rejestracji ostatniego rekordu z bieżącego okna agregacji zwiększonemu o wartość horyzontu prognozy,
- wstawienie zagregowanej wartości zmiennej zależnej (np. maksimum, minimum, średnia) obliczanej pomiędzy czasem rejestracji ostatniego rekordu z bieżącego okna agregacji a czasem tym zwiększonym o wartość horyzontu prognozy.

Ostatnim etapem przygotowania zbioru danych jest dodanie zmiennych wywiedzionych odzwierciedlających dynamikę zmian pomiędzy kolejnymi, zagregowanymi już, rekordami danych. Operacja ta realizowana jest albo poprzez odpowiednie „opóźnienie" poszczególnych zmiennych („opóźnienie" – czyli dodania do bieżącego rekordu nowych pól zawierających wartości z wcześniejszych rekordów), albo poprzez dalszą agregację zmiennych już wartości już zagregowanych. Operacje te stosowane są na danych zagregowanych i dotyczą one jedynie zmiennych niezależnych.

W tabeli pierwszej przedstawiono przykładowy uproszczony zbiór danych pobranych z systemu metanometrycznego SMP/NT. W kolumnie T[s] tabeli umieszczono czasy zmiany któregokolwiek z monitorowanych parametrów, jak już wspomniano zmiana wartości parametru powoduje wpisanie nowego rekordu do bazy danych. W tabeli drugiej przedstawiono pierwotny zbiór danych po wykonaniu operacji czyszczenia, w tabeli trzeciej zaprezentowano zbiór zagregowany, taki właśnie zbiór danych jest podstawą do wyznaczenia modelu predykcyjnego. W tabeli drugiej pogrupowano rekordy, które będą poddawane agregacji.

Poniżej podano parametry związane z procesem czyszczenia i agregacji danych:

- parametry oczyszczania
  - wygładzanie: brak,
  - uzupełnianie brakujących danych: ostatnia poprawna wartość,
  - czas uzupełniania brakujących danych w przypadku nie pojawienia się nowej, poprawnej wartości MaxT: 2 sekundy,
- parametry agregacji:
  - czas agregacji: 3 sekundy,
  - częstotliwość agregacji: co 3 sekundy,
  - funkcja agregująca zmienną A1: maksimum,
  - dodatkowe zmienne dla zmiennej A1: rozstęp,
  - funkcja agregująca zmienną A2: średnia arytmetyczna,
  - dodatkowe zmienne dla zmiennej A2: koniec-początek,
- zmienna do predykcji: A1,
  horyzont prognozy: 1 sekunda.

Tabela 1

Dane pomiarowe

A1	A2	T[s]
0,1	0,1	0
0.1	?	1
0,4	0.2	4
0,5	0,2	7
0,3	0,3	9

Tabela 2

Pierwotny zbiór danych pomiarowych, po oczyszczeniu

A1	A2	T[s]
0.1	0.1	0
0.1	0.1	1
0.1	0.1	2
0.1	?	3
0.4	0.2	4
0.4	0.2	5
0.4	0.2	6
0.5	0.2	7
0.5	0.2	8
0.3	0.3	9

Tabela 3

Dane po agregacji

A1-MAX	A1-rozs	A2-avg	A2-k-p	A1-Pred
0.1	0	0.1	0	0.1
0.4	0.3	0.2	?	0.4
0.5	0.1	0.2	0	0.3

3. PREDYKCJA STĘŻENIA GAZÓW

Ignorując aspekty przygotowania danych, utrzymania wiarygodności systemu oraz aspekty czysto techniczne, problem predykcji stężenia gazów można potraktować jako problem zadania predykcji, w którym nie istnieje model matematyczny opisujący za pomocą odpowiednich równań zależności pomiędzy zmienną zależną a zbiorem zmiennych niezależnych. Do rozwiązania tego typu problemów dobrymi metodami są: metody obliczeń miękkich i neuronowych (m.in. [Czogała, Oh, Yager]), statystyka ([Box, Hayes]) oraz metody maszynowego uczenia [Breiman, Quinlan]. Przed fazą projektowania systemu predykcji gazów przetestowano każdą z powyższych metod, a kryterium je oceniającym była wartość błędu

RMSE (6) ang. *Root Mean Squared Error* uzyskiwana na testowym zbiorze danych oraz czas potrzebny na wyznaczenie modeli predykcyjnych na podstawie zagregowanych danych historycznych).

Uwzględniając uzyskane wyniki testów, do zaimplementowania w systemie wykorzystano algorytm M5 oraz adaptacyjną liniową metodę predykcji.

Idea algorytmu M5 [15],[16] została zaczerpnięta z tzw. drzew regresji i klasyfikacji [6] (CART). Dla zbioru treningowego $Tr=(U, A\cup\{y\})$ tworzony jest zbiór reguł (1):

$$a_{i1} \in V_{a_{i1}} \wedge .. \wedge a_{ij} \in V_{a_{ij}} \rightarrow y=b_{i1}w_{i1}+\ldots+b_{il}w_{il}+w_i \quad (2)$$

gdzie:
$\{a_{i1},..,a_{ij}\}\subseteq A;\ V_{a_{il}} \subset D_{a_{il}};\ \{b_{i1},..,b_{il}\}\subseteq A;$
$w_i,w_{i1},..,w_{il}\in \boldsymbol{R}$.

W zbiorze treningowym wyróżniamy: zbiór przykładów U (u nas jest to zbiór zagregowanych rekordów), zbiór zmiennych niezależnych A (u nas jest to zbiór czujników oraz zbiór zmiennych odzwierciedlających dynamikę zmian mierzonych wielkości), zmienną zależną y (u nas jest to czujnik, którego wskazania chcemy przewidywać), dziedziny atrybutów D_a, $D_y\subset\boldsymbol{R}$. Każda zmienna zależna a jest zatem funkcją $a:U\rightarrow D_a$.

Zmienne występujące w konkluzjach reguł muszą być zmiennymi numerycznymi.

Algorytm M5 jest algorytmem budującym drzewo lokalnych modeli liniowych, które następnie zamieniane są na zbiór reguł. Na każdym etapie tworzenia drzewa (w każdym węźle nie będącym liściem), wywoływana jest procedura sprawdzania, który atrybut $a\in A$ oraz która wartość graniczna q dokona najlepszego podziału zbioru przykładów związanego z danym węzłem. Dla ustalonego atrybutu a, dla każdej wartości q, rozpatrywanej przez algorytm M5 istnieją takie wartości $v1,v2\in D_a$, że $v1<v2$ oraz $q=(v2-v1)/2$. W przypadku atrybutów symbolicznych stosowana jest procedura wyczerpująca polegająca na badaniu wszystkich możliwych podzbiorów zbioru wartości atrybutu symbolicznego.

Jeśli zbiór przykładów związanych z węzłem oznaczymy przez P, to w każdym węźle najlepszym atrybutem i punktem granicznym jest taki punkt, dla którego podział zbioru P na podzbiory $P_{<q}$ and $P_{>q}$ maksymalizuje wartość wyrażenia (3):

$$\Delta err = V(P) - \left(\frac{|P_{<q}|}{|P|}V(P_{<q}) + \frac{|P_{>q}|}{|P|}V(P_{>q})\right) \quad (3)$$

gdzie:
$V(P)$ oznacza wariancję zmiennej zależnej w zbiorze przykładów P.

Jeśli w danym węźle najlepszy z możliwych podziałów nie zmniejsza już oczekiwanej wariancji zmiennej zależnej to procedura rozbudowy drzewa zatrzymuje się (węzeł staje się liściem).

W każdym węźle (nie tylko w liściach) wyznaczana jest za pomocą regresji liniowej wielokrotnej funkcja f, której zadaniem jest predykcja zmiennej zależnej.

Po utworzeniu drzewa, uruchamiana jest procedura obcięcia. Jeśli średni błąd bezwzględny modelu liniowego przyporządkowanego do danego węzła jest mniejszy niż błąd w jego węzłach potomnych, to węzeł ten staje się liściem.

Po utworzeniu drzewa struktura ta zamieniana jest na zbiór reguł. Konkluzjami reguł stają się funkcje f znajdujące się w liściach drzewa.

Aby znacząco poprawić zdolności predykcyjne algorytm M5 stosuje procedurę smoothingu. W czasie przekształcania drzewa w zbiór reguł pamiętana jest kolejność dodawania deskryptorów warunkowych do reguły.

Wartość zmiennej zależnej propagowana jest począwszy od reguły wyjściowej r, poprzez reguły r_{-1}, r_{-2},..., aż do reguły r_{root}, która jest regułą bez jakichkolwiek deskryptorów warunkowych.

Do reguły r_{-1} przekazywana jest wartość określona przez regułę r. Załóżmy, że dane są dwie reguły częściowe r_{-i} i r_{-i-1}, wartość $PV(r_{-i-1})$ przekazywana z reguły r_{-i} do reguły r_{-i-1} określana jest jako iloraz (4):

$$\frac{n_{-i}PV(r_{-i}) + kM(r_{-i-1})}{n_{-i} + k} \quad (4)$$

gdzie:
n_{-i} jest liczbą obiektów ze zbioru U, które spełniają część warunkową reguły r_{-i}; k jest pewną ustaloną stałą, a $M(r_{-i-1})$ jest wartością przewidywaną przez regułę częściową r_{-i-1}.

Algorytm jest niezwykle szybki, a uzyskane reguły w sposób prosty opisują lokalne zależności pomiędzy zmiennymi niezależnymi i zmienną zależną.

Kolejnym składnikiem systemu predykcyjnego jest metoda prognozy liniowej [Box, Hydes]. Dla zbioru treningowego $Tr=(U, A\cup\{y\})$, model liniowy wyrazić możemy wzorem (5):

$$y(t_0 + p) = \sum_{z=-k}^{-t_1} y(z)w_z + w \quad (5)$$

gdzie:
t_0 jest obecną chwilą czasu; p jest horyzontem prognozy; $w_z,w\in\boldsymbol{R}$ są współczynnikami modelu, a $y(z)$ ($z\in\{-k, .., -t_1\}$) możemy nazwać opóźnieniami.

Należy zauważyć, że w przypadku prognozy liniowej konieczne jest, aby przykłady zawarte w zbiorze treningowym umieszczane były chronologicznie, zgodnie z kolejnością ich rejestracji. W wyrażeniu (5) założono, że do ustalenia wartości zmiennej y w chwili t_0+p, wykorzystano pewną liczbę wcześniejszych wartości tej zmiennej. W bardziej ogólnej postaci można wykorzystać również inne zmienne należące do zbiory atrybutów A lub pewne zagregowane wartości tych zmiennych, wtedy mamy do czynienia ze statystycznym podejściem do predykcji szeregów czasowych [1, 5]. Identyfikacja modelu liniowego polega na doborze opóźnień oraz ustaleniu wartości współczynników w_z oraz w.

W opisywanym systemie wybrano najprostszy model, wykorzystujący obecną $y(t_0)$ oraz wcześniejszą $y(t_0\text{-}k)$ wartość zmiennej zależnej. Na podstawie tych dwóch wartości ustalane jest równanie prostej, które następnie służy do predykcji wartości $y(t_o+p)$. W celu uproszczenia obliczeń przyjmujemy, że t_0=0. Wartość p (horyzont prognozy zależy od preferencji użytkownika), wartość k ustalana jest adaptacyjnie na podstawie treningowego zbioru danych. Dla danego zbioru treningowego ustalana jest maksymalna wartość parametru k (ozn. k_{max}). Następnie dla każdego $k=1,..,k_{max}$, obliczany jest błąd prognozy (6) na zbiorze treningowym:

$$RMSE(DB) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(i) - \bar{y}(i)\right)^2} \qquad (6)$$

gdzie:
DB jest zbiorem przykładów, na którym obliczamy błąd,
$n=|DB|$, $y(i)$ jest rzeczywista wartością zmiennej zależnej y dla przykładu i-tego,

$\bar{y}(i)$ jest wartością przewidywaną przez model liniowy.

Jako optymalną wartość k wybierana jest ta wartość, dla której osiągnięto minimalny błąd, ta wartość k będzie wykorzystywana przez system dopóki nie zaistnieje konieczność adaptacji systemu (o czym w następnym rozdziale).

W czasie przeprowadzania eksperymentów sprawdzających efektywność obu metod zauważono, że M5 uzyskuje lepsze wyniki w czasie większych (bardziej dynamicznych) zmian stężenia metanu, natomiast w czasie stabilizacji stężenia dokładniejsza jest prognoza liniowa. To spostrzeżenie było przyczynkiem do stworzenia rozwiązania wykorzystującego oba podejścia jednocześnie. W systemie do ustalenia ostatecznej wartości zmiennej predykcyjnej zastosowano model liniowy (7) przyjmujący na wejście wyniki prognoz generowanych przez algorytm M5 i model liniowy.

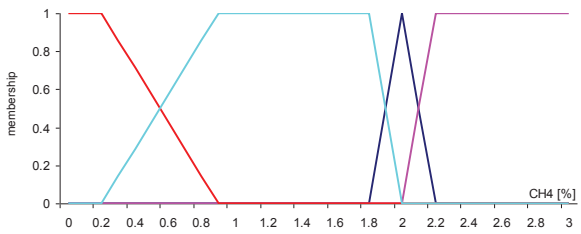
$$y(t_0 + p) = M5(t_0 + p)w + Linear(t_0 + p)(1 - w) \quad (7)$$

Wartość wagi w dobierana może być w dwojaki sposób:

- po prezentacji każdego przykładu (wtedy uwzględniane są błędy popełnione przez oba podstawowe modele, a modyfikacja wagi przebiega zgodnie z regułą Widrowa-Hoffa [14]),
- po prezentacji wszystkich przykładów ze zbioru treningowego (wtedy wartość wagi w ustalana jest za pomocą metody najmniejszych kwadratów).

4. USTALANIE STOPNIA ZAGROŻENIA

Dysponując przewidywaną wartością metanu w zadanym horyzoncie prognozy system w oparciu o ekspercką bazę reguł rozmytych dokonuje oceny stanu zagrożenia metanowego. Reguły rozmyte składają się z dwóch przesłanek: prognozowanego stężenia metanu oraz dynamiki zmian stężenia wynikającej z prognozy. Dziedziny obu tych wartości zostały podzielone na zbiory rozmyte zgodnie z wiedzą ekspercką. Stężenie metanu w atmosferze zostało podzielone na cztery zbiory rozmyte (rys. 2). Dynamikę zmian odzwierciedlono za pomocą trzech zbiorów rozmytych (brak zmian, rosnąca, szybko rosnąca). Zbiór rozmyty „brak zmian" uwzględnia również spadki stężenia metanu. Wiedza ekspercka pozwoliła także na określenie ośmiu reguł rozmytych wiążących stężenie metanu i jego dynamikę zmian z sytuacją w wyrobisku.



Rys. 2. Podział dziedziny stężenia metanu na zbiory rozmyte

Wyróżniane są trzy sytuacje: stan normalny (wartość punktowa 1), ostrzeżenie (wartość punktowa 2), zagrożenie (wartość punktowa 3). Sytuacje te opisane zostały za pomocą zbiorów rozmytych o trójkątnych funkcji przynależności, które osiągają swoje maksima odpowiednio w punktach 1, 2, 3.

Wynikowa baza wiedzy to osiem reguł postaci (8):

$$\text{IF } x_1 \text{ is } A_1 \text{ and } x_2 \text{ is } A_2 \text{ THEN } y \text{ is } B \qquad (8)$$

gdzie: x_1, x_2 są odpowiednio przewidywanym stężeniem i dynamiką zmian metanu, A_1, A_2 są zbiorami rozmytymi odzwierciedlającymi odpowiednio stężenie

i dynamikę zmian metanu w atmosferze, *B* jest zbiorem rozmytym opisującym sytuację w wyrobisku.

System wykorzystuje wnioskowanie konstruktywne typu Larsena [Yager], w którym do ustalenia poziomu zapłonu reguł wykorzystuje się operator PROD (t-norma=PROD), agregacja reguł polega na zsumowaniu zbiorów rozmytych wyprowadzanych przez każdą z reguł (suma zbiorów rozmytych – operator MAX), a jako metodę wyostrzania stosujemy standardową metodę środka ciężkości [Yager, Czogała]. Wartości wejściowe nie są poddawane rozmywaniu, traktowane są jako singeltony.

5. EKSPERYMENTY Z DANYMI

Badania eksperymentalne wykonano na zbiorze danych, w którym zawarte były informacje o stężeniach gazów zarejestrowanych na wylocie ze ściany (tam gdzie istnieje największe zagrożenie metanowe). Analizie poddano dane rejestrowane przez dwa metanomierze (M31, M32) oraz dwa anemometry (AN31, AN32). Dodatkowo dysponowaliśmy także sumarycznym wydobyciem węgla podczas całej zmiany. Akwizycja danych dokonywała się co dziesięć sekund, dla celów analizy dane zagregowano wybierając maksymalne wartości wskazań z każdej minuty. W algorytmie M5, aby dodatkowo odzwierciedlić dynamikę zmian pomiędzy kolejnymi zagregowanymi rekordami wykorzystano także zmienne (DAN31, DAN32, DM31, DM32), które zawierały sumę wskazań odpowiednich czujników z ostatnich dziesięciu minut. Celem predykcji było przewidywanie stężenia metanu zarejestrowanego przez czujnik M32 z wyprzedzeniem jedno- i dziesięciominutowym. Moduł predykcji liniowej wykorzystywał jedynie wcześniejsze wartości wskazań zarejestrowane przez czujnik M32.

Dla celów porównawczych analizy przeprowadzono także za pomocą sieci neuronowo-rozmytych o architekturach znanych jako ANNBFIS [3] i MFNN [6].

W tabelach trzeciej i czwartej zaprezentowano wyniki dla predykcji jednominutowej. Poza błędem predykcji podano również maksymalną wartość błędu oraz liczbę błędów większych niż 0.29, 0.19 i 0.09 procent stężenia metanu. Plik treningowy składał się z 6000 rekordów, plik testowy z 4000 rekordów.

Tabela 4

Wyniki predykcji dla zbioru treningowego

Metoda	RMSE	Max err	Err >0,29	Err >0,19	Err >0,09
M5	0,056	0,51	13	83	1037
Predyktor liniowy	0,058	0,51	18	113	1490
Metoda hybrydowa	0,056	0,51	13	82	1027
ANNBFIS	0,053	0,45	1	13	642
MFNN	0,052	0,45	1	24	700

Tabela 5

Wyniki predykcji dla zbioru testowego

Metoda	RMSE	Max err	Err >0,29	Err >0,19	Err >0,09
M5	0,047	0,42	7	29	358
Predyktor liniowy	0,045	0,40	9	41	560
Metoda hybrydowa	0,044	0,42	5	27	351
ANNBFIS	0,066	0,51	24	143	450
MFNN	0,063	0,50	15	132	412

6. PODSUMOWANIE I DALSZE PRACE

W artykule przedstawiono moduł przygotowania danych dla systemu predykcji stężenia gazów (w szczególności metanu) w wyrobisku górniczym. Przedstawiono także sam system umożliwiający generowanie prognoz. Moduł predykcyjny, który jest kluczowym składnikiem systemu, wykorzystuje dwie metody analityczne; jedna wykorzystująca paradygmat maszynowego uczenia, druga proste podejście predykcji liniowej. Połączenie obu metod dało rozwiązanie hybrydowe, które nieznacznie lepiej niż najlepsza z metod szczegółowych pozwala przewidywać przyszłe stężenia metanu. Badania eksperymentalne przeprowadzone dla predykcji metanu pokazują skuteczność zastosowanych metod. Porównanie do bardziej zaawansowanych metod wykorzystujących rozmyte sieci neuronowe pokazuje, że w tym konkretnym przypadku rozwiązania prostsze pozwalają uzyskać lepsze wyniki na zbiorach testowych. Niebagatelną rolę w wyborze narzędzi predykcyjnych miał także czas uczenia obu zastosowanych modeli. Dla zbioru danych złożonego z 6000 rekordów czas analizy algorytmem M5 wynosił niecałe 2 sekundy, czas adaptacyjnej metody liniowej, dla której k zmieniało się od 1 do 50, wynosił 90 sekund. Czas dostrajania każdej z metod neuronowo rozmytych wynosił ponad 180 sekund. Jednakże dla metod tych konieczne jest określenie liczby grup, na które mają zostać podzielone dane treningowe (wiąże się to z ustaleniem optymalnej liczby reguł rozmytych). Aby ustalić optymalną liczbę grup należy wykonać serię dodatkowych eksperymentów, co wielokrotnie wydłuża czas analizy.

Obecne prace nad rozwojem systemu koncentrują się na jego uruchomieniu w rzeczywistych warunkach dyspozytorni metanometrycznej. Pozwoli to na zgromadzenie sporego materiału badawczego, a tym samym na przeprowadzenie większej liczby eksperymentów. Mając nadzieję na zwiększenie trafności prognoz, w chwili obecnej opracowywana jest jeszcze jedna unikalna metoda prognozy, która łączy algorytm M5 z metodą najbliższych sąsiadów.

System może być wykorzystany również do innych celów, np. predykcji stężenia dwutlenku węgla, przepływu powietrza lub identyfikacji mieszanin wybuchowych. Oczywiście inne będą wtedy bazy reguł rozmytych wykorzystane do wnioskowania o zagrożeniu.

Literatura

1. *Box G.E., Jenkins G.M.*: Time series analysis: forecasting and control, Prentice Hall, New Jersey, 1994.
2. *Breiman L., Friedman J.H., Olshen R.A.*: C. Stone, Classification and Regresion Trees, Wadsworth, Belmont CA, 1994.
3. *Czogała E., Łęski J.*: Fuzzy and Neuro-Fuzzy Intelligent Systems, Springer-Verlag, Heidelberg, 2000.
4. *Gralewski K., Krzystanek Z.*: Nowe możliwości systemu SMP/NT. Mechanizacja i Automatyzacja Górnictwa, 2004, nr 9, s. 12-19.
5. *Hayes M.H.*: Statistical Digital Signal Processing and Modeling, J.Wiley & Sons, New York, 1996.
6. *Oh S.K., Park H.S., Pedrycz W.*: Rules based multi-FNN identification with the aid of evolutionary fuzzy granulation. Knowledge-Based Systems, No. 17, 2004, s. 1-13.
7. *Quinlan R.*: Learning with continuous classes. In Proc. of the International Conference on Artificial Intelligence (AI- 92), World Scientific, Singapore, 1992.
8. *Quinlan R.*: Combining instance-based learning and model-based learning. In Proc. of the Tenth International Conference on Machine Learning (ML-93), 1993.
9. *Salvador S., Chan P.*: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, No. 11(5), 2007, s. 561-580.
10. *Sikora M., Sikora B.*: Application of machine learning for prediction a methane concentration in a coal-mine. Archives of Mining, No. 51(4), 2006, s. 475-492.
11. *Sikora M., Kozielski M.*: Hybrid data exploration methods to prediction tasks solving. Archives of Theoretical and Applied Informatics, No. 18(1), 2006, s. 57-73.
12. *Zadeh L.A.*: Fuzzy sets. Information and Control, No. 8, 1965.
13. *Yager R.R., Filev D.P.*: Essential of Fuzzy Modeling and Control, J.Wiley & Sons, New York, 1994.
14. *Widrow B., Hoff M. E.*: Adaptive switching circuits. In IRE WESCON Convention Record, Vol. 4, 1960, s. 96-104.

Recenzent: dr inż. Zbigniew Isakow