

Wykorzystanie technik bazujących na teorii zbiorów przybliżonych w procesie detekcji i klasyfikacji niezgodności spawalniczych

The use of rough sets theory in process of detection and classification of welding imperfections

Streszczenie

Przetwarzanie znaczących ilości informacji pochodzących z obrazów radiograficznych oraz automatyczne wykrywanie wad połączeń spawalniczych z dużą dokładnością jest możliwe dzięki zastosowaniu rozwiązań opartych na teorii zbiorów przybliżonych oraz zastosowaniu jej w systemach komputerowych, które umożliwiają szybkie przetwarzanie znaczących ilości danych. Wspomniana metoda ma solidne podstawy matematyczne, których zastosowanie umożliwia określenie istotności atrybutów mających znaczenie dla identyfikacji niedoskonałości, natomiast ostateczna ocena reguł tworzy bazę wiedzy umożliwiającą komputerowe wskazywanie określonej klasy wady spawu. Opisana technika umożliwiła klasyfikację wad w bardzo wysoką dokładnością dla rzeczywistych danych pochodzących ze zdjęć rentgenowskich połączeń spawalniczych.

Abstract

Processing of large amount of information deriving from radiographic images and automatically detecting of welding joints imperfections with high accuracy is possible by applying solutions based on rough sets theory and usage of this theory on computer systems that are capable to make fast calculations on huge number of information. The theory possesses solid and confirmed mathematical foundation that allows applying it for calculation of attribute importance that have huge significance for identification of weld imperfections, whereas final extracting rules creates knowledge base that gives possibility for computer aided pointing specific class of weld imperfection. Technique that is described in the paper was capable of classification of weld defects with very high accuracy for real data originating from radiographic images of weld joints.

Wstęp

Teoria zbiorów przybliżonych ma ugruntowaną teorię, opartą na rozszerzeniu teorii zbiorów i definiującej pojęcia dolnego oraz górnego przybliżenia. Umożliwia to tworzenie tabeli decyzyjnej na podstawie zebranych danych, atrybutów decyzyjnych oraz zdefiniowanych przybliżeń, które przypisują każdy z przypadków do określonej klasy wraz z wyznaczeniem siły przynależności oraz innych parametrów jakościowych dla uzyskanych reguł. Taka tabela może

zostać przekształcona w łatwo interpretowalne reguły [1] w postaci jeżeli ... to, jak również zostać zapisana w formie lingwistycznej, która umożliwia łatwą interpretację reguł przez ekspertów. Rozwiązania zaproponowane w tej teorii dają również możliwość na określenia poziomu spójności danych, które mają zostać poddane klasyfikacji, co z kolei umożliwia stwierdzenie przydatności informacji do wyznaczenia reguł [2]. Możliwe jest również redukcowanie wymiaru na podstawie badania współczynników określających poziom istotności poszczególnych atrybutów ze względu na wpływ na docelową zmienną decyzyjną.

Wymienione cechy opisanej teorii oraz wykorzystanie komputerowego przetwarzania danych umożliwiają automatyczną detekcję wad połączeń spawalniczych

Dr inż. Leszek Misztal – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie.

z dużą dokładnością. Jest to możliwe dzięki przetworzeniu informacji pochodzących z obrazów spawów oraz pozyskaniu wiedzy i identyfikacji poszczególnych niedokładności połączeń za pomocą wspomnianej teorii z wykorzystaniem mocy obliczeniowej współczesnych komputerów. Ma to szczególne znaczenie w branżach, gdzie połączenia spawalnicze decydują o ludzkim zdrowiu i życiu, czyli np w przemyśle konstrukcyjnym, stoczniowym, lotniczym oraz in.. Przedstawiona teoria oraz jej komputerowa implementacja umożliwiła wykrywanie wad spawalniczych w zrealizowanym projekcie inteligentnego systemu analizy radiogramów, który był oparty na rzeczywistych obrazach radiograficznych spawów pochodzących z przemysłu stoczniowego.

Metoda badań

Do opisanego sposobu funkcjonowania przedstawionej teorii zbiorów przybliżonych niezbędny jest opis podstawowych pojęć związanych z definicją systemu informacyjnego, relacji nierozróżnialności, dolnego i górnego przybliżenia oraz regionu granicznego [3]. Zostały one przedstawione w kolejnych akapitach.

W zbiorach przybliżonych definiuje się system *informacyjny* S , który składa się z następujących elementów:

$$S = \langle U, Q, V, f \rangle,$$

gdzie U jest skończonym uniwersum N obiektów $\{x_1, x_2, \dots, x_N\}$, Q jest niepustym zbiorem n atrybutów $\{q_1, q_2, \dots, q_n\}$, które charakteryzują obiekty, $V = \bigcup_{q \in Q} V_q$, gdzie V_q jest wartością atrybutu q , natomiast $f: U \times Q \rightarrow V$ jest funkcją informacyjną, dla której $f(x, q) \in V_q$ dla każdego $q \in Q, x \in U$. Ważnym pojęciem jest nierozróżnialność zbiorów. Przyjmując, że $A \subseteq Q$ jest podzbiorem atrybutów, a jest atrybutem w zbiorze A oraz obiekt $y \in U$, oznacza, że relacje nierozróżnialności przedstawia się w następujący sposób:

$$I(A) = \{(x, y) \in U \times U: \text{dla każdego } a \in A, f(x, a) = f(y, a)\}$$

Oznacza to, że para elementów (x, y) należy do $I(A)$, jeżeli x, y są obiektami nierozróżnialnymi w zakresie zbioru atrybutów A . Nie można wówczas rozróżnić pomiędzy x i y , bazując na zawartości informacyjnej zawartej w zbiorze A .

Kolejnymi bardzo ważnymi definicjami, na których jest oparty algorytm, są dolne oraz górne przybliżenia, które przedstawia się w następujący sposób:

$$A^*(X) = \{x \in U: A(x) \subseteq X\}$$

$$A^*(X) = \{x \in U: A(x) \cap X \neq \langle \text{ZBIÓR PUSTY} \rangle\}$$

gdzie $A(x)$ oznacza rodzinę wszystkich klas równoważności $I(A)$ dla partycji $UII(A)$ zawierającej x .

Dolne przybliżenie $A^*(X)$ zbioru X zawiera wszystkie elementy, bazując na zawartości informacyjnej atrybutów A , które są sklasyfikowane jako należące do X .

Natomiast górne przybliżenie $A^*(X)$ zawiera wszystkie elementy zawarte w A , które mogą zostać sklasyfikowane jako możliwa przynależność do X . Występuje również pojęcie *regionu granicznego*, jest różnicą wyżej zdefiniowanych pojęć i przedstawia się następująco:

$$GR_A(X) = A^*(X) - A^*(X)$$

Jeżeli region graniczny jest zbiorem pustym, wówczas zbiór X jest dokładny w odniesieniu do zbioru atrybutów A , natomiast gdy nie jest pusty, oznacza to przybliżenie do zbioru A .

W celu przystąpienia do właściwej klasyfikacji, czyli utworzenia tablicy decyzyjnej, konieczna jest ponowna definicja systemu informacyjnego [4], w którym atrybut Q jest podzielony na dwa rozdzielne zbiory, gdzie C jest zbiorem atrybutów warunkowych, natomiast D jest atrybutem decyzyjnym, takim że $C \cup D = Q$ i $C \cap D = \langle \text{ZBIÓR PUSTY} \rangle$. Oznacza to, że C reprezentuje atrybuty związane z cechami i właściwościami połączeń spawalniczych, a D jest atrybutem oznaczającym decyzję o klasie wady połączenia spawanego. Wówczas tablica decyzyjna przyjmuje następującą postać:

$$DT = \langle U, C \cup D, V, f \rangle,$$

gdzie U – skończony zbiór wszystkich obiektów zwany *uniwersum*, C i D zostały opisane powyżej, $V = \bigcup_{q \in Q} V_q$, gdzie V_q jest zbiorem dyskretnych wartości atrybutów $q \in Q$, $f: U \times (C \cup D) \rightarrow V$ jest opisem funkcji definiowanej jako $f(x, q) \in V_q$ dla każdego $q \in Q$ i $x \in U$. Tabele decyzyjne dzieli się na deterministyczne, czyli takie, gdzie tryb decyzyjny jest wyznaczony przez kombinację atrybutów warunkowych oraz tablice niedeterministyczne, gdzie atrybut decyzyjny nie jest wyznaczony. Ocenę rodzaju tablicy można wykonać na podstawie wartości współczynnika zwanego *jakością przybliżenia conceptów decyzyjnych*. Jego wartości $\gamma(C, D) = 1$ oznacza występowanie tablicy deterministycznej, czyli pełnej spójności, natomiast $\gamma(C, D) \neq 1$ oznacza tablicę niedeterministyczną, co w praktyce występuje najczęściej. Powstałą w wyniku operacji tablicę decyzyjną można w łatwy sposób przekształcić w formę lingwistyczną, która może przyjmować następującą postać: PARKSZTAŁT 1 = WARTOŚĆ, PARKSZTAŁT 2 = WARTOŚĆ, PARKSZTAŁT 3 = WARTOŚĆ \rightarrow WYNIK_KON = PEŁCZERZ.

Dla oceny wytworzonych reguł decyzyjnych definiuje się współczynniki jakościowe [5], których celem jest pomiar różnych parametrów związanych z jakością reguł. Przedstawiają się następująco:

– *współczynnik zaufania (pewności)*, który oznacza częstość występowania obiektów w systemie informacyjnym S mających konkluzję Ψ w zbiorze obiektów mających przesłankę Φ . Przyjmuje następującą definicję:

$$\pi_s(\Psi | \Phi) = \frac{\text{card}(\|\Phi \wedge \Psi\|_s)}{\text{card}(\|\Phi\|_s)}$$

gdzie Ψ oznacza konkluzję reguły, Φ to przesłanka reguły, $\|\Psi\|$ to zbiór znaczeniowy konkluzji Φ w systemie informacyjnym S , $\|\Psi\|$ to zbiór znaczeniowy przesłanki Φ w systemie informacyjnym

S, natomiast $\|\Phi \wedge \Psi\|$ jest zbiorem znaczeniowy reguły decyzyjnej $\Phi \rightarrow \Psi$ w systemie informacyjnym S,
 – *współczynnik pokrycia* oznacza częstość występowania obiektów z przesłanką Φ w zbiorze obiektów posiadających konkluzję Ψ :

$$\pi_s(\Phi | \Psi) = \frac{\text{card}(\|\Phi \wedge \Psi\|_s)}{\text{card}(\|\Psi\|_s)}$$

– *wsparcie reguły* jest to liczność danej reguły decyzyjnej w systemie informacyjnym S, w przypadku, gdy ta reguła jest dopuszczalna. Definiowane jest następująco:

$$\text{sup } p_s(\Phi, \Psi) = \text{card}(\|\Phi \wedge \Psi\|_s)$$

– *siła reguły* – oznacza stosunek liczności danej reguły decyzyjnej w systemie informacyjnym S do liczności całego uniwersum U:

$$\sigma_s(\Phi, \Psi) = \frac{\text{sup } p_s(\Phi, \Psi)}{\text{card}(U)} = \frac{\text{card}(\|\Phi \wedge \Psi\|_s)}{\text{card}(U)}$$

Do określenia poziomu zaufania dla atrybutów decyzyjnych [6], koniecznym jest zdefiniowanie współczynnika jakości przybliżenia konceptów decyzyjnych $0 \leq k \leq 1$, gdzie zbiór atrybutów decyzyjnych D zależy od zbioru atrybutów warunkowych C w stopniu zdefiniowanym przez k :

$$k = \gamma(C, D) = \frac{\text{card}(POS_c(D))}{\text{card}(U)}$$

gdzie $POS_c(D)$ jest pozytywnym regionem dla partycji U/D w zakresie atrybutów C , card jest kardynalnością, czyli ilością elementów w danym zbiorze.

W celu określenia poziomu istotności atrybutu i podjęcia decyzji o usunięciu wymiaru, definiowana jest względna istotność atrybutu warunkowego $\sigma(a)$ dla atrybutu a , gdzie $a \in C$, poprzez następujące równanie:

$$\sigma_{(C,D)}(a) = \frac{\gamma(C, D) - \gamma(C - \{a\}, D)}{\gamma(C, D)} = 1 - \frac{\gamma(C - \{a\}, D)}{\gamma(C, D)}$$

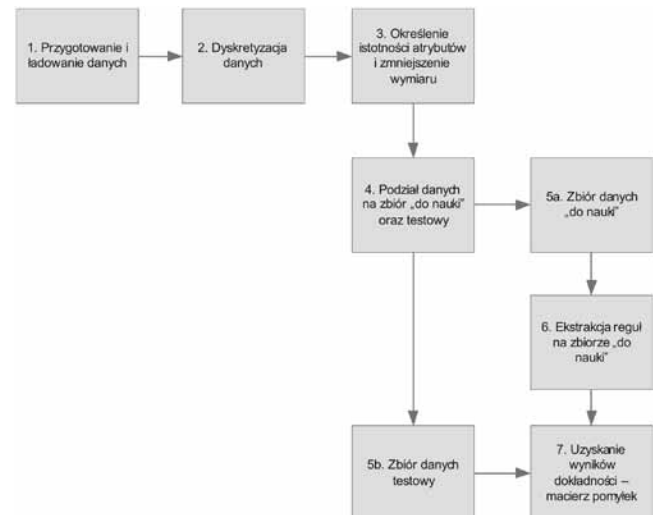
Wartość współczynnika mieści się w zakresie $0 \leq \sigma(a) \leq 1$. Jeżeli jest równa zero lub też ma wartość poniżej zadanego niskiego progu określonego dla badania, wówczas atrybut a powinien zostać wyeliminowany z obliczeń tablicy decyzyjnej, ponieważ nie ma wpływu lub też jego wpływ jest nieistotny na atrybut decyzyjny.

Sposób przeprowadzenia badań

Wykorzystanie potencjału oferowanego przez zbiory przybliżone jest możliwe poprzez zaproponowanie procedury badawczej oraz wykorzystanie jej w procesie ekstrakcji wiedzy dotyczącej wykrywania wad połączeń spawalniczych. Zgodnie z [7] odmienna specyfika problemów występujących w różnych branżach powoduje, że zaproponowany model powinien być znacząco

dopasowany do konkretnego zadania, w opisywanym przypadku do identyfikacji niedoskonałości spawów. W związku z tym w zaproponowanej procedurze (rys. 1) przewidziano siedem etapów, prowadzących do ekstrakcji wiedzy umożliwiającej automatyczne wykrywanie poszczególnych klas wad spawów.

Procedura jest złożona z następujących kroków: przygotowanie i ładowanie danych, dyskretyzacja danych, określenie istotności atrybutów i zmniejszenie wymiaru, podział danych na zbiór „do nauki” oraz testowy, 5a. Zbiór danych „do nauki”, 6. Ekstrakcja reguł na zbiorze „do nauki”, 5b. Zbiór danych testowy, 7. Uzyskanie wyników dokładności – macierz pomyłek



Rys. 1. Procedura identyfikacji wad spawów z wykorzystaniem zbiorów przybliżonych

Fig. 1. Procedure of identification of weld imperfections with the usage of rough sets theory

testowy, utworzenie dwóch tablic informacyjnych dla wymienionych zbiorów, ekstrakcja reguł czyli niezbędnej wiedzy do klasyfikacji oraz uzyskanie wyników dokładności. Sama precyzja metody określona została na podstawie tak zwanej macierzy pomyłek oraz współczynników jakościowych określanych na jej podstawie. Jest to powszechnie uznany sposób oceny dokładności metod klasyfikacji [8]. Do wspomnianych współczynników należą m.in.: pozytywny współczynnik predykcji, negatywny współczynnik predykcji, współczynnik czułości, współczynnik specyficzności, współczynnik całkowitej dokładności, współczynnik całkowitego poziomu błędów. W celu dodatkowego potwierdzenia dokładności zaproponowanego systemu do wykrywania rodzaju wad została zaproponowana dodatkowa weryfikacja uzyskanych wyników z wykorzystaniem tak zwanej *crossvalidacji* [9]. Polega ona na losowym podziale zbioru danych na n podzbiorów testowych i treningowych, na których oddzielnie zostaną wyznaczone reguły i zweryfikowana dokładność klasyfikacji. Umożliwia to zdecydowanie dokładniejszą weryfikację zaproponowanego modelu, gdyż w ten sposób eliminuje się możliwość przypadkowego, jednorazowego podziału danych zwiększającego dokładność (tzw. *peak* dokładności dla jednego z podziałów).

Uzyskane wyniki

Wyniki automatycznej detekcji poszczególnych wad spawów zostały zrealizowane na podstawie rzeczywistych radiogramów spawów wykonanych na błonach rentgenowskich zebranych podczas projektu ISAR (Inteligentny System Analizy Radiogramów). Pochodziły one z procesu kontroli jakości złączy spawanych płaskich zebranych w branży stoczniowej. W sumie badaniom poddano ponad 1500 próbek. W wyniku identyfikacji zostały rozpoznane następujące rodzaje wad zebrane w klasy oznaczone za pomocą liczb: 1 – pęknięcie, 2 – pustki, 3 – wtrącenia stałe, 4 – przyklejenia i braki przetopu, 5 – niezgodności kształtu. W tabelicy I została umieszczona dokładność (pozytywny współczynnik predykcji) dla poszczególnych pięciu

klas, jak również współczynnik całkowitej dokładności uzyskanej przez modelowanie dla wszystkich klas.

Przedstawione wyniki reprezentują średnią dokładność z uwzględnieniem tzw. *crosswalidacji* dla dziesięciu losowych podziałów całego zbioru informacyjnego dotyczącego wad spawalniczych. Potwierdza to wysoką skuteczność zaproponowanej metody opartej na teorii zbiorów przybliżonych, gdyż zastosowanie wielokrotnej walidacji zazwyczaj obniża wyniki dokładności predykcji, natomiast w przypadku zaproponowanej procedury różnice pomiędzy dokładnością dla poszczególnych kroków nie przekraczały 1%, biorąc pod uwagę pozytywny współczynnik predykcji. Dokładność wyników uzyskanych dla wykrywania poszczególnych klas, jak również całkowita precyzja jest bardzo wysoka, gdyż przekracza 90%.

Tablica I. Wyniki dokładności klasyfikacji wad spawów z wykorzystaniem zbiorów przybliżonych

Table I. Results of classification accuracy of weld imperfections with the usage of rough sets theory

Klasa (class) → współczynnik (coefficient) ↓	1 %	2 %	3 %	4 %	5 %	Całkowita dokładność (overall accuracy) %
pozytywny współczynnik predykcji (positive prediction coefficient)	> 90	> 90	> 90	> 90	> 90	>90

Podsumowanie

W artykule przedstawiono nowatorskie wykorzystanie teorii zbiorów przybliżonych do wykrywania wad połączeń spawalniczych. Opisana procedura umożliwiła automatyzację detekcji niedoskonałości z wysoką dokładnością, potwierdzoną przez dane z rzeczywistych obrazów radiograficznych spawów. Uzyskane wyniki przyjmują formę reguł, które mogą zostać interpretowane i rozumiane przez człowieka, jak również mogą zostać przedstawione w formie

lingwistycznej. Wykorzystanie wspomnianej metody w postaci oprogramowania dla systemów komputerowych umożliwia szybką oraz akceptowalną wysoką dokładność wskazywania klas wad spawalniczych. Dzięki temu realna staje się automatyzacja procesu kontroli jakości w różnych dziedzinach przemysłu, która umożliwi wspomaganie procesu sprawdzania produktów wykorzystujących techniki spawalnicze.

Literatura

- [1] Inuiguchi M., Generalizations of Rough Sets and Rule Extraction, Springer, 2005.
- [2] Bazan J., Synak P., Wroblewski J., Rough Set Algorithms in Classification Problem, Springer, 2000.
- [3] Pawlak Z., Rough Sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.
- [4] Olson D., Delen D., Advanced Data Mining Techniques, Springer, 2008.
- [5] Duntsch I., Gediga G., Rough set data analysis, Methodos Publisher, 2000.
- [6] Pawlak Z.: Some issues on rough sets, Springer Science, 2005
- [7] S.Bandyopadhyay, U.Maulik, L.B.Holder, D.J.Cook Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005.
- [8] Han J., Kamber M.: Data mining concepts and techniques, Morgan Kaufman, 2006.
- [9] Michie D., Spiegelhalter D.J., Taylor C.C.: Machine Learning, Neural and Statistical Classification, Ellis Horwood, New York, 1994.