ANALIZA WYBRANYCH METOD WALIDACJI KRZYŻOWEJ W PROGRAMIE RSES

Radosław Kołpacki

Uniwersytet Kazimierza Wielkiego
Wydział Informatyki
e-mail: radekk@ukw.edu.pl

Streszczenie: *W artykule przeprowadzono analizę zbioru danych za pomocą dwóch metod walidacji krzyżowej. Wykorzystano program RSES do identyfikacji kluczowych właściwości i relacji w zbiorze. Wyniki wykazują wpływ niektórych parametrów na potencjalną dokładność wyników.*

Słowa kluczowe: *Walidacja krzyżowa, RSES, analiza danych, zależność, algorytm genetyczny*

ANALYSIS OF SELECTED CROSS-VALIDATION METHODS IN THE RSES PROGRAM

Abstarct: *This article presents an analysis of a dataset using two cross-validation methods. The RSES program was employed to identify key properties and relationships within the dataset. The results indicate the impact of certain parameters on the potential accuracy of the outcomes.*

Keywords: *Cross-validation, RSES, data analysis, dependency, genetic algorithm*

1. WSTĘP

W dzisiejszych czasach, charakteryzujących się intensywnym rozwojem technologii i dynamicznym postępem w dziedzinie analizy danych, kluczowym aspektem staje się zastosowanie skutecznych metod walidacji w celu uzyskania wyników o wysokiej wiarygodności i trafności. Jednym z narzędzi cieszących się uznaniem w obszarze eksploracyjnej analizy danych jest program RSES. Niniejszy artykuł skupia się na analizie zbioru danych, wykorzystując dwie metody walidacji krzyżowej przy użyciu wspomnianego programu.
Celem przeprowadzonego badania jest zrozumienie wpływu różnych metod walidacji krzyżowej na potencjalną dokładność wyników, a także identyfikacja kluczowych właściwości i relacji w analizowanym zbiorze danych. Prezentowane wyniki mają na celu dostarczenie wglądu w skuteczność programu RSES w kontekście analizy zbioru, a także wskazanie istotnych parametrów wpływających na jakość otrzymywanych rezultatów.
W kolejnych sekcjach artykuł przedstawi krótkie wprowadzenie do badanego problemu, ukazując kontekst badawczy. Dodatkowo, omówione zostaną teoretyczne aspekty związane z wykorzystywaną metodą walidacji krzyżowej.

2. METODY

Do analizy zbioru danych zastosowano zaawansowane podejście oparte na teorii zbiorów przybliżonych oraz korzystając z programu RSES, który jest specjalnie zaprojektowanym narzędziem do analizy zbiorów danych w kontekście tej teorii.

Walidacja krzyżowa to powszechnie stosowana technika oceny skuteczności modeli predykcyjnych w analizie danych. Procedura ta umożliwia lepsze zrozumienie ogólnej wydajności modelu poprzez podzielenie dostępnych danych na zbiór treningowy i testowy. Głównym celem jest ocena, jak dobrze model radzi sobie z danymi, które nie były wykorzystane podczas treningu. Rozpoczynamy od podziału zbioru danych na zbiór treningowy i testowy,

zazwyczaj stosując proporcje, na przykład 70% danych na trening i 30% na testowanie. Następnie przeprowadzamy trening modelu na zbiorze treningowym, używając różnych algorytmów, takich jak algorytm Exhaustive czy Genetic dostępny w programie RSES.

Po treningu modelu przechodzimy do etapu testowania, gdzie model jest oceniany na zbiorze danych, który nie był używany podczas treningu, eliminując tym samym potencjalne zjawisko nadmiernego dopasowania. Cały proces (podział danych, trening, testowanie) jest wielokrotnie powtarzany, każdorazowo inne dane są wykorzystywane jako zbiór testowy, co eliminuje losowość i przyczynia się do uzyskania bardziej wiarygodnych wyników. Na podstawie wyników z każdego etapu walidacji krzyżowej obliczamy średnią skuteczność modelu, a także analizujemy wariancję wyników, co pozwala ocenić stabilność modelu w różnych warunkach.

Zacznijmy od przybliżenia pierwszej metody. Jest to algorytm wyczerpujący (Exhaustive Algorithm), będący zaawansowaną techniką analizy danych w programie RSES, wyróżnia się pełnym przeszukiwaniem przestrzeni rozwiązań, co pozwala na identyfikację kluczowych właściwości i relacji w analizowanym zbiorze danych. Proces rozpoczyna się od inicjalizacji analizy, gdzie algorytm wyczerpujący szczegółowo badanie wszystkie możliwe kombinacje cech w zbiorze danych, co umożliwia dokładne zrozumienie struktury danych.
W trakcie analizy, algorytm wyczerpujący bada każdą kombinację cech, dokonując oceny ich wpływu na analizowany zbiór danych. To podejście umożliwia identyfikację zarówno istotnych, jak i mniej istotnych cech, co przekłada się na bardzo szczegółową analizę.
Po kompletnym przeszukaniu przestrzeni rozwiązań, uzyskane wyniki są analizowane, dostarczając szczegółowych informacji na temat każdej kombinacji cech.

Kolejną metodą jest algorytm genetyczny (Genetic Algorithm) stanowi powszechnie stosowaną technikę optymalizacyjną w analizie danych, wykorzystującą inspirację z procesu ewolucji biologicznej. W ramach programu RSES, algorytm genetyczny jest używany do identyfikacji kluczowych właściwości i relacji w analizowanym zbiorze danych.

Rozpoczynamy od inicjalizacji populacji, gdzie każdy osobnik reprezentuje potencjalne rozwiązanie problemu, a w kontekście analizy danych - kombinację różnych cech. Następnie oceniamy przystosowanie każdego osobnika do rozwiązania problemu, a więc oceniamy przydatność kombinacji cech pod kątem identyfikacji kluczowych właściwości. Kolejnym etapem jest selekcja, w której osobniki o lepszym przystosowaniu mają większą szansę na reprodukcję, odbywającą się na zasadzie "przeżywają tylko najsilniejsi". Operator krzyżowania pozwala na wymianę cech pomiędzy rodzicami, tworząc potomstwo z kombinacją ich cech. W kontekście analizy danych, krzyżowanie może prowadzić do odkrycia nowych zależności czy istotnych właściwości. Proces mutacji wprowadza niewielkie zmiany w genotypie potomstwa. Kolejne generacje populacji powstają poprzez wielokrotne powtarzanie kroków selekcji, krzyżowania i mutacji.

Poziomy "low", "normal" i "high" w kontekście algorytmu genetycznego odnoszą się do stopnia, w jakim algorytm ma działać bardziej eksploracyjnie, zrównoważenie między eksploracją a eksploatacją lub bardziej eksploatacyjnie. W przypadku ustawienia na "low", algorytm genetyczny będzie skupiał się głównie na eksploracji, czyli odkrywaniu nowych kombinacji cech w celu identyfikacji potencjalnie istotnych właściwości w analizowanym zbiorze danych. Dla opcji "normal", algorytm będzie dążył do zrównoważonego podejścia, równoważąc odkrywanie nowych rozwiązań z wykorzystywaniem już znalezionych. Jest to wartość domyślna, idealna w przypadku poszukiwania optymalnych rozwiązań przy utrzymaniu efektywności obliczeniowej. W przypadku ustawienia na "high", algorytm genetyczny będzie bardziej skoncentrowany na eksploatacji, czyli rozwijaniu i utrzymywaniu istniejących już rozwiązań, co jest korzystne, gdy mamy pewność co do znalezionych cech i chcemy zoptymalizować wyniki w oparciu o te już istniejące informacje. Dostosowanie tych parametrów pozwala na lepsze dopasowanie algorytmu do konkretnych celów analizy danych, co może wpłynąć na skuteczność identyfikacji kluczowych właściwości i relacji w analizowanym zbiorze danych przy wykorzystaniu programu RSES.

3. TEORIA ZBIORÓW PRZYBLIŻONYCH

Teoria zbiorów przybliżonych jest specjalnym podejściem matematycznym do analizy danych, które zostało zaprojektowane w celu radzenia sobie z niepewnością i niekompletnością informacji w zbiorach danych. Została ona opracowana w latach 80. XX wieku przez polskiego matematyka Zdzisława Pawlaka i od tego czasu znalazła szerokie zastosowanie w wielu dziedzinach nauki i technologii.

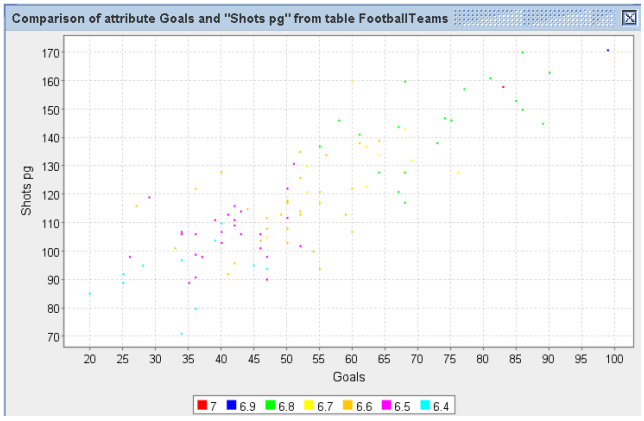
Głównym celem teorii zbiorów przybliżonych jest umożliwienie identyfikacji istotnych informacji w zbiorach danych, pomimo ich niekompletności lub niepewności. W praktyce, wiele zbiorów danych, zwłaszcza te pochodzące z rzeczywistych zastosowań, może zawierać brakujące wartości, niejednoznaczne informacje lub

niekompletne obserwacje. Teoria zbiorów przybliżonych pozwala na reprezentację takich zbiorów w formie zbiorów przybliżonych, które zawierają jedynie istotne informacje, które można w pełni potwierdzić na podstawie dostępnych danych.

Zbiór jest uważany za przybliżony, jeśli zawiera elementy, które nie można jednoznacznie zidentyfikować na podstawie dostępnych danych, ale można je przybliżyć na podstawie innych, bardziej pewnych informacji. W praktyce, teoria zbiorów przybliżonych pozwala na ekstrakcję istotnych wzorców, reguł i zależności z danych, nawet jeśli są one niekompletne lub niejednoznaczne.
Do analizy zbiorów danych za pomocą teorii zbiorów przybliżonych często wykorzystuje się różne algorytmy i techniki, które umożliwiają identyfikację kluczowych atrybutów, wzorców decyzyjnych oraz relacji między danymi.

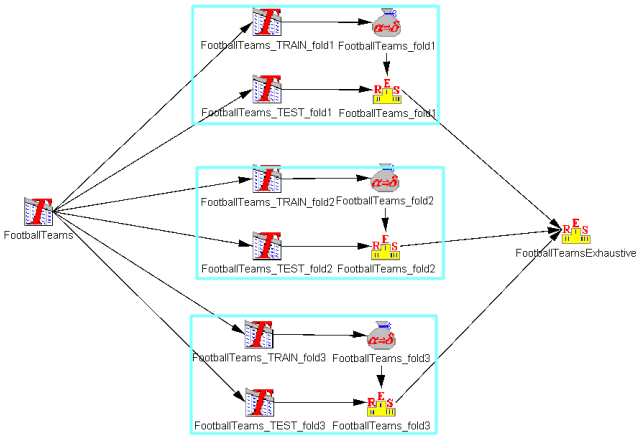
4. IMPLEMENTACJA METOD W PROGRAMIE RSES

RSES (Rough Set Exploration System) to program do analizy zbiorów przybliżonych. Pozwala automatycznie generować reguły przybliżone na podstawie danych, optymalizować je, analizować istotne cechy, oraz oceniać jakość reguł. Dzięki interfejsowi graficznemu ułatwia korzystanie z różnych przydatnych funkcji. To narzędzie jest używane w eksploracji danych, szczególnie w kontekście analizy danych z niepewnością i nieprecyzyjnością.



Rysunek.1 Wykres punktowy porównujący ilość bramek w stosunku do oddanych strzałów , źródło: opracowanie własne.

Do przeprowadzenia badania użyte zostały dane najlepszych klubów piłkarskich z 5 najlepszych lig w Europie, takie jak bramki, strzały, posiadanie piłki, skuteczność podań, liczba czerwonych i żółtych kartek.
Powyższy wykres przedstawia stosunek ilości bramek do oddanych strzałów, drużyny mają przydzielone kategorie w rankingu od 6,4 do 7. Legenda przedstawia podział każdej wartości rangi, oznaczone jest to odpowiednimi kolorami.



Rysunek.2 Schemat w programie RSES przedstawiający działanie walidacji krzyżowej , źródło: opracowanie własne.

Dane wgrane do programu zostały poddane walidacji krzyżowej. Na powyższym przykładzie widoczny jest schemat działania. W tym konkretnym przykładzie zastosowano metodę Exhaustive. Każdy przypadek miał zaprogramowane 3 powtórzenia, co jest zaznaczone na rysunku nr.2. Czynnikiem, który wyróżnia każdy osobny przypadek są inne ustawienia jeśli chodzi o użytą metodę oraz zastosowaną szybkość.

Metoda	Total accuracy	Total coverage
Exhaustive	0,272	0,958
Genetic High 2	0,285	0,656
Genetic Normal 2	0,25	0,667
Genetic Low 2	0,22	0,573
Genetic High 5	0,31	0,875
Genetic Normal 5	0,24	0,906
Genetic Low 5	0,319	0,875
Genetic High 20	0,2	0,99
Genetic Normal 20	0,277	0,948
Genetic Low 20	0,221	0,99
Genetic High 50	0,187	0,948
Genetic Normal 50	0,234	0,979
Genetic Low 50	0,245	0,979

Tabela. 1 Wyniki walidacji krzyżowej z zastosowaniem różnych metod, źródło: opracowanie własne.

Powyższa tabela przedstawia listę wszystkich przeanalizowanych wariantów metod w walidacji krzyżowej. Kolejne kolumny przedstawiają wartości Total_accuracy oraz Total_coverage.

Całkowita dokładność (Total accuracy) to miara, która informuje, jak dobrze system, model, czy algorytm radzi sobie w przewidywaniu lub klasyfikowaniu obiektów. Wartość 0.2 oznacza, że tylko 20% przewidywań lub klasyfikacji były poprawne wśród wszystkich przetestowanych obiektów. Niska wartość dokładności sugeruje, że model lub algorytm może mieć trudności z efektywnym przewidywaniem.

Całkowite pokrycie (Total coverage) to miara, która informuje, jak wiele obiektów w zbiorze danych zostało objętych testami lub analizą. Wartość 0.99 oznacza, że 99% wszystkich obiektów zostało uwzględnionych w testach. Wysoka wartość pokrycia sugeruje, że analiza objęła prawie wszystkie dostępne dane.

Wyniki zaprezentowane w Tabeli nr.1 pokazują, że skuteczność algorytmu jest niska, jednakże to nie jest celem w tym przypadku. Warto zwrócić uwagę, że stosując metodę genetyczną o najmniejszej i największej ilości reduktów., w tym przypadku 2 i 50, średnia dokładność jest niższa niż w przypadku ustawienia ilości reduktów na 5 lub 20. Powtarzającą się tendencją jest przypadek, że użycie metody genetycznej o normalnej szybkości jest najmniej skuteczne. Bez względu na to ile jest reduktów. Być może analizowane dane nie są odpowiednie, by sądzić, że dokładność przy zastosowaniu walidacji krzyżowej jest niska. O wiele wyraźniej widać różnicę w przypadku wyników zawartych w całkowitym pokryciu. Metoda wyczerpująca jest solidna, gdyż i w przypadku dokładności, której wartość mieści się w średniej z reszty wyników, tak jest bardzo przydatna w kontekście całkowitego pokrycia. Tendencja, którą widać jest to, że im większa ilość reduktów tym większa wartość pokrycia. W przypadku reduktów o wartości 2 to pokrycie wynosi w przybliżeniu do dziesiętnych 0,6. Co oznacza tylko 60% danych zostało użyte do walidacji krzyżowej. Dla porównania gdy tych reduktów jest 20 lub 50, pokrycie wynosi 0,9. To jest średnio 90% danych. Jest to bardziej wiarygodne. Warto zwrócić uwagę na przypadek, w którym reduktów jest 20, bo wartość total_coverage wyniosła 0,99. Niezależnie od ilości reduktów najbardziej solidną prędkością metody genetycznej jest normal. Wyróżnia się względem innych szybkości przy niskich ilościach reduktów.

5. PODSUMOWANIE I WNIOSKI

Badanie analizy danych z wykorzystaniem programu RSES i dwóch metod walidacji krzyżowej dostarczyło ważnych wniosków. Eksploracyjna analiza za pomocą algorytmu wyczerpującego pozwoliła na dogłębne zrozumienie struktury danych, podczas gdy algorytm genetyczny zachowywał równowagę między czasem a jakością rezultatów.

Wnioski z analizy parametrów, takich jak total_accuracy i total_coverage, ukazują kluczowe aspekty wpływające na jakość analizy danych. Optymalizacja czasu obliczeniowego stała się istotnym kryterium, a algorytm genetyczny wykazał się potencjalnie bardziej efektywny w tym kontekście.

Warto podkreślić, że to badanie stanowi jedynie wstęp do eksploracji potencjału programu RSES. Dalsze prace badawcze mogą prowadzić do jeszcze bardziej efektywnych strategii analizy danych, a uzyskane wyniki mają potencjał zastosowania w różnych dziedzinach.

Literatura

1. Tadeusiewicz R., Szaleniec M., „Leksykon sieci neuronowych”, Projekt Nauka, Wrocław 2015
2. Podręcznik użytkownika “RSES 2.1 Rough Set Exploration System”, Warszawa 2004
3. Horzyk A., “Metody Inżynierii Wiedzy- Walidacja Krzyżowa”, prezentacja multimedialna, materiały zajęciowe AGH, Kraków
4. Szczuko P., Kostek B., “Sztuczna inteligencja w medycynie”, materiały zajęciowe PG , 19-26, Gdańsk 2015