Łączenie metody bazującej na instancjach z metodą indukcji reguł dla danych niezbalansowanych
(Autoreferat)

Grzegorz Góra
Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego

---

1. Wstęp

Jedną z głównych dziedzin badawczych sztucznej inteligencji jest uczenie maszynowe [23, 31, 36]. Najczęstszym zadaniem maszynowego uczenia jest klasyfikacja, która przypisuje dowolnemu opisowi obiektu decyzję z ustalonego skończonego zbioru decyzji. Na przykład, może wystąpić potrzeba sklasyfikowania pewnych osób poprzez przypisanie im decyzji, czy są chore czy zdrowe. Szczególnym podzadaniem klasyfikacji jest uczenie nadzorowane (w skrócie nazywane tu uczeniem). W tym podzadaniu zadany jest skończony zbiór obiektów (nazywanych również przypadkami, przykładami lub instancjami), z przypisanymi im znanymi decyzjami. Zbiór ten nazywany jest zbiorem treningowym; w rozważanym przykładzie jest to zbiór osób z przypisanym diagnozami. Celem jest przypisanie nowym obiektom, zwanym obiektami testowymi (np. nowym pacjentom do zbadania), odpowiedniej decyzji (np. chory lub zdrowy). Algorytmy maszynowego uczenia, nazywane algorytmami uczącymi, posługując się wnioskowaniem indukcyjnym, konstruują na podstawie zbiorów treningowych klasyfikatory, które dowolnym obiektom testowym przypisują decyzję. Dla jasności podkreślamy różnicę pomiędzy dwoma głównymi pojęciami związanymi z klasyfikacją danych, a mianowicie algorytmami uczącymi i algorytmami klasyfikującymi (w skrócie nazywanymi klasyfikatorami). Klasyfikator klasyfikuje dowolny przykład testowy na podstawie jego opisu, natomiast algorytm uczący stosuje się do szerokiego zakresu dziedzin, tworząc klasyfikator na podstawie zadanego zbioru treningowego. Mimo, że do tej pory opracowano wiele algorytmów uczących, to wciąż proponowane są nowe. Do najpopularniejszych algorytmów uczących należą algorytmy generujące: drzewa decyzyjne, klasyfikatory regułowe, maszyny wektorów podpierających, klasyfikatory bazujące na instancjach, proste klasyfikatory bayesowskie, sztuczne sieci neuronowe, zespoły klasyfikatorów oraz lasy losowe. W ramach tej rozprawy skupiamy się na opracowaniu nowych algorytmów uczących, które w szczególności czerpią z dwóch technik, a mianowicie indukcji reguł oraz uczenia bazującego na instancjach.

Ostatnio wiele wysiłku włożono w rozwój metod uczenia nadzorowanego, które dotyczą uczenia się z tzw. danych niezbalansowanych [12, 24, 25, 27, 29, 46]. W zadaniach klasyfikacji dla danych niezbalansowanych poprawna klasyfikacja obiektów do jednej z klas decyzyjnych jest znacznie ważniejsza niż do innych. Dla zadania klasyfikacji z decyzją binarną, na którym skupiamy się w tej rozprawie, jedna wyszczególniona klasa spośród dwóch ma szczególne znaczenie. Zazwyczaj klasa ta zawiera znacznie mniejszą liczbę obiektów niż druga klasa. Dlatego jest ona określana jako klasa mniejszościowa, a druga jako klasa większościowa. Należy zauważyć, że poprawne zdiagnozowanie pacjentów chorych na nowotwór (stanowiących klasę mniejszościową) jest znacznie ważniejsze niż poprawne zdiagnozowanie pacjentów zdrowych (stanowiących klasę większościową).

Rozpoczynając pracę nad zagadnieniem związanym z danymi niezbalansowanymi, warto zdać sobie sprawę z tego, dlaczego standardowe klasyfikatory (tzn. klasyfikatory indukowane przez algorytmy uczące zaprojektowane dla danych zbalansowanych) nie sprawdzają się w

przypadku danych niezbalansowanych. Istnieją ku temu co najmniej cztery następujące powody:

- Konstrukcja standardowych klasyfikatorów ukierunkowana jest na maksymalizację dokładności klasyfikacji (wyrażonej przez iloraz liczby poprawnych przewidywań decyzji dokonanych przez klasyfikator do całkowitej liczby dokonanych przewidywań). Jednak w przypadku danych niezbalansowanych ta miara jakości jest niezadowalająca.

- Konstrukcja standardowych klasyfikatorów powoduje, że ich zastosowanie w przypadku danych niezbalansowanych prowadzi najczęściej do uzyskania niezadowalająco niskiego współczynnika dokładności dla klasy mniejszościowej przy równoczesnym zapewnieniu wysokiego współczynnika dokładności dla klasy większościowej.

- Standardowe algorytmy identyfikując przykłady zaszumione (ang. *noisy examples*), tj. obiekty treningowe z błędnymi etykietami decyzyjnymi, dokonują tego nie uwzględniając różnicy między sytuacjami pochodzącymi z klas większościowej i mniejszościowej. Klasyfikacja obiektów z klasy mniejszościowej komplikuje się, jeśli przykład rzeczywiście należący do klasy mniejszościowej zostanie zidentyfikowany jako zaszumiony, lub prawdziwie zaszumiony przykład z klasy większościowej nie zostanie zidentyfikowany jako taki.

- Standardowe klasyfikatory zakładają jednakowe koszty błędnej klasyfikacji dla wszystkich klas. Jednakże koszt błędnej klasyfikacji może być często znacznie wyższy dla klasy mniejszościowej niż dla klasy większościowej.

W ostatnich latach zaproponowano wiele metod uczenia z danych niezbalansowanych. Zasadniczo metody te można podzielić na dwie grupy:

- rozwiązania na poziomie danych oraz

- rozwiązania na poziomie algorytmicznym.

Rozwiązania na poziomie danych (wykorzystując metody nazywane filtrami) przekształcają oryginalny zbiór danych w nowy, a następnie stosują do niego standardowy algorytm uczący. W tym podejściu można wyróżnić następujące podejścia do transformacji danych: metody próbkowania z nadmiarem (ang. *over-sampling methods*), które zwiększają liczność klasy mniejszościowej, metody próbkowania z niedomiarem (ang. *under-sampling methods*), które zmniejszają liczność klasy większościowej oraz metody hybrydowe, które łączą dwa poprzednie podejścia. Rozwiązania na poziomie algorytmicznym dotyczą opracowania algorytmów, które uwzględniają problem niezbalansowanych danych. Można tu wyróżnić następujące podejścia: adaptacja istniejących algorytmów opracowanych pierwotnie dla danych niezbalansowanych poprzez wprowadzenie przewagi na rzecz klasy mniejszościowej, uczenie jednoklasowe, uczenie wrażliwe na koszty (ang. *cost-sensitive learning*) oraz podejścia bazujące na zespołach klasyfikatorów.

2. Motywacje

W rozprawie skupiamy się na specyficznym podejściu dla danych niezbalansowanych, łączącym uczenie bazujące na instancjach i metody bazujące na regułach. W przeszłości podejmowano pewne próby łączenia metod instancyjnych i regułowych, jednak tylko dla

danych zbalansowanych (zob. np. [11, 28]). Niemniej jednak, co najmniej dwa powody przemawiają za rozwojem takich podejść nie tylko dla danych zbalansowanych, lecz również dla niezbalansowanych.

Po pierwsze, oba podejścia wykorzystują schematy rozumowania łatwo zrozumiałe dla człowieka. Do takich schematów należą reguły w postaci:

Jeśli *pewne warunki są spełnione,* ***to*** *decyzją jest X*

które są często używane przez ludzi. Analogicznie, schemat rozumowania w postaci:

Skoro *nasz nowy przykład A*
jest najbardziej podobny do *innego znanego, zbadanego przykładu B,*
to *przykład A powinien mieć taką samą decyzję jak przykład B*

stosowany w uczeniu instancyjnym, jest także łatwo zrozumiałym dla człowieka. Z tego powodu takie podejścia spełniają wymagania wyjaśnialności (ang. *explainability*) [1].

Po drugie, istnieją pewne intuicje, wynikające z rozważań matematycznych, sugerujące zastosowanie uczenia bazującego na instancjach, być może w połączeniu z indukcją regułową. Podejścia bazujące na regułach to przykład procedury składającej się z dwóch etapów. W pierwszym etapie indukujemy (estymujemy) nieznaną funkcję decyzyjną. W drugim etapie stosujemy tę indukowaną funkcję do klasyfikacji przykładów testowych. Jednakże Vapnik zauważył, że estymacja funkcji decyzyjnej jest znacznie bardziej ogólnym problemem niż ten, który zazwyczaj należy rozwiązywać w praktyce. W większości przypadków zachodzi jedynie potrzeba estymowania wartości nieznanej funkcji decyzyjnej w „kilku" nowych punktach zdefiniowanych przez obiekty testowe (zob. [43, s. 12]). Sugeruje on, że jeśli zachodzi potrzeba wnioskowania o decyzjach dla nowych przypadków na podstawie małych zbiorów treningowych, powinniśmy wziąć pod uwagę następującą zasadę:

> *Jeśli posiadasz ograniczoną ilość informacji do rozwiązania jakiegoś problemu, staraj się rozwiązać go bezpośrednio i nigdy nie rozwiązuj bardziej ogólnego problemu jako etapu pośredniego. Możliwe jest, że dostępne informacje są wystarczające do bezpośredniego rozwiązania, ale niewystarczające do rozwiązania bardziej ogólnego problemu pośredniego.* [43]

Zasada ta sugeruje, że stosowanie podejść bazujących na instancjach może mieć istotne znaczenie dla jakości konstruowanych klasyfikatorów. To samo dotyczy również metod łączących podejścia bazujące na instancjach z innymi podejściami.

Wśród metod instancyjnych dobrze znane są algorytmy kNN. Ta klasa algorytmów znalazła się na liście dziesięciu najbardziej znaczących algorytmów eksploracji danych [45]. W najprostszym przypadku zwracają one decyzję dotyczącą przykładu treningowego najbardziej podobnego do przypadku testowego. Ogólnie rzecz biorąc, algorytmy te dokonują klasyfikacji przypadku testowego na podstawie liczby wystąpień klas wśród k przykładów najbardziej podobnych do testowego – tworzących zbiór przykładów czasami nazywany sąsiedztwem przypadku testowego. Podobieństwo to mierzone jest za pomocą pewnej funkcji odległości, zwanej też metryką. Jakość metody kNN istotnie zależy od użytej funkcji odległości. W wielu pracach proponuje się różne rozwiązania w zakresie indukowania metryki z danych [40]. Ponadto jakość metody kNN zwykle istotnie zależy od wartości k. W praktyce estymacja optymalnego k jest często wykonywana za pomocą techniki walidacji krzyżowej. Generalnie, istnieje wiele podejść do automatycznego wyboru optymalnej wartości k (zob. np. [15, 47]).

Metody bazujące na regułach reprezentują wiedzę przy pomocy reguł decyzyjnych typu *jeśli-to* wiążących warunki z decyzjami. Wśród metod bazujących na regułach można wyróżnić kilka podejść (przegląd tych metod znajduje się np. w [14]). Ogólnie metody bazujące na regułach można scharakteryzować biorąc pod uwagę trzy ważne aspekty związane z następującymi pytaniami:

1. Jaki jest język opisu reguł?
2. W jaki sposób generowany jest zbiór reguł?
3. W jaki sposób uzyskany zbiór reguł jest wykorzystywany w procesie klasyfikacji? (Zwykle jest to związane z tzw. rozwiązywaniem konfliktów).

Biorąc pod uwagę pierwsze pytanie, zdecydowana większość podejść wykorzystuje koniunkcję warunków (deskryptorów) postaci $atrybut = wartość$ w poprzedniku reguły oraz $decyzja = wartość$ w następniku reguły. Istnieją jednak również inne podejścia, np. reguły monotoniczne (zob. np. [6]). W odniesieniu do drugiego pytania, można wyróżnić dwa główne podejścia: indukcja minimalnego zbioru reguł (zob. np. [13, 21, 30]) oraz indukcja nieminimalnego zbioru reguł (zob. np. [14, 38]). W odniesieniu do trzeciego pytania, można wyróżnić następujące podejścia: algorytmy generujące uporządkowany zbiór reguł (zob. np. [8]) oraz różne strategie rozwiązywania konfliktów (w [41] zawarto przegląd literatury dotyczącej tego zagadnienia).

Wśród algorytmicznych rozwiązań dla danych niezbalansowanych istnieje wiele metod bazujących na regułach (zob. np. [35] jako przykład pracy bazującej na regułach, zob. [32, Chapter 4], aby zapoznać się z ich przeglądem). Jak wspomniano wcześniej, moja praca koncentruje się na metodach klasyfikacji danych niezbalansowanych łączących podejścia instancyjne i regułowe.

3. Cel pracy i zarys wyników

Głównym celem mojej pracy jest opracowanie algorytmów uczących bazujących na połączeniu metod instancyjnych i regułowych charakteryzujących sie wysoką jakością klasyfikacji dla różnych typów zbiorów danych. Cel ten realizujemy w dwóch krokach, proponując algorytm RIONA dla danych zbalansowanych oraz algorytm RIONIDA dla danych niezbalansowanych. Na konkretnym przykładzie algorytmu RIONA pokazujemy, jak można uogólnić strukturę algorytmu opracowanego dla danych zbalansowanych, aby stał się on efektywny dla danych niezbalansowanych, co prowadzi do algorytmu RIONIDA. Najprostszym podejściem (wykorzystującym do klasyfikacji danych niezbalansowanych algorytmy opracowane dla danych zbalansowanych) byłoby zastosowanie filtru do danych niezbalansowanych przed użyciem algorytmu RIONA. W niniejszej rozprawie proponujemy jednak inne podejście, a mianowicie podejście polegające na modyfikacji algorytmu RIONA tak, aby okazał sie on skuteczny dla danych niezbalansowanych.

3.1. RIONA – algorytm dla danych zbalansowanych

W pierwszym kroku proponujemy algorytm indukcji regułowej z optymalnym sąsiedztwem (RIONA) [16]. Algorytm ten łączy podejście instancyjne i regułowe i został opracowany tak, aby jego jakość, mierzona względem miary zwanej dokładnością (zob. np. [26]), okazała się konkurencyjna w stosunku do innych rozwiązań. Algorytm ten wykorzystuje kilka pomysłów.

- RIONA generuje reguły w sposób leniwy (zob. np. [2]), to znaczy indukuje nieliczny zbiór reguł decyzyjnych istotnych tylko dla klasyfikacji rozważanego przykładu testowego. Jest to odmienna strategia od tej indukującej z góry dużą liczbę reguł decyzyjnych w celu wykorzystania ich podczas testowania.

- Klasyfikacja dokonywana przez RIONA na danym obiekcie testowym wykorzystuje reguły indukowane tylko na podstawie obiektów należących do otoczenia wyznaczonego przez sąsiedztwo danego przykładu testowego. Zauważmy, że przy zastosowaniu podejścia leniwego korzysta się jedynie ze stosunkowo niewielkiej liczby reguł.

- Używamy innego rodzaju reguł niż te powszechnie stosowane w podejściach regułowych, gdzie warunki mają postać: *atrybut równy określonej wartości*. W algorytmie RIONA stosowane są bardziej ogólne reguły z warunkami postaci: *atrybut należy do określonego zbioru wartości*. Te zbiory wartości określane są poprzez grupowanie zarówno liczbowych, jak i symbolicznych wartości atrybutów. W głosowaniu nad decyzją za pomocą reguł pokrywających klasyfikowany przykład, wykorzystywana jest agregacja zbiorów wspierających takie reguły.

- RIONA konstruuje sąsiedztwa obiektów o optymalnym rozmiarze.

- Pojęcie podobieństwa między obiektami jest dla algorytmu RIONA wykorzystywane do dwóch celów, a mianowicie do: (i) konstruowania sąsiedztwa dla danego obiektu oraz (ii) grupowania wartości atrybutów.

Przeprowadzone eksperymenty opisane przez autora rozprawy (zob. [16]) oraz w literaturze (zob. np. [4, 10, 19, 20, 37]) pokazują, że RIONA jest konkurencyjna w stosunku do wielu innych znanych systemów.

3.2. RIONIDA – algorytm dla danych niezbalansowanych

Okazuje się, że RIONA posiada pewne wady charakterystyczne dla standardowych algorytmów (tzn. zaprojektowanych dla danych zbalansowanych) w przypadku zastosowania ich do danych niezbalansowanych. W tym miejscu pojawia się drugi etap realizacji celu rozprawy. Celem jest zmodyfikowanie proponowanego podejścia łączącego metody instancyjne i regułowe (RIONA) w celu poprawy jego efektywności na danych niezbalansowanych. Mianowicie, w tym drugim etapie proponujemy algorytm RIONIDA (Rule Induction with Optimal Neighbourhood for Imbalanced Data Algorithm). Wszystkie pomysły wymienione w poprzedniej podsekcji dla algorytmu RIONA są również zrealizowane w algorytmie RIONIDA. Ponadto ten nowy algorytm realizuje kilka nowych idei w porównaniu z algorytmem RIONA.

- RIONIDA stara się maksymalizować nie dokładność, ale jedną spośród miar jakości, znacznie bardziej istotnych dla danych niezbalansowanych, jak F-miara lub G-mean (zob. np. [5, 26]).

- Rozstrzyganie konfliktów reguł w algorytmie RIONIDA jest bardziej wyrafinowane niż w algorytmie RIONA. Agregacja decyzji reguł pokrywających klasyfikowane obiekty jest definiowana z uwzględnieniem wymagania, że klasa mniejszościowa jest „ważniejsza" od klasy większościowej. Sformułowanie „ważniejsza" wyrażone jest przez stopień ważności. Stopień ważności klasy mniejszościowej (i w konsekwencji klasy większościowej) jest dostrajany w trakcie uczenia.

- W klasyfikacji dopuszcza się zastosowanie reguł sprzecznych w określonym stopniu. Poziom dozwalanej sprzeczności jest również dostrajany w trakcie uczenia.

RIONIDA prowadzi do istotnie lepszych wyników eksperymentalnych niż testowane w rozprawie znane metody opracowane dla danych niezbalansowanych. Fakt ten został zilustrowany w mojej rozprawie na kilkunastu benchmarkach (patrz Rozdział 5).

Podejście zastosowane przy tworzeniu algorytmu RIONIDA jest inne niż te prezentowane w literaturze. Według naszej wiedzy jedynym algorytmem, który łączy podejście instancyjne i regułowe, a jednocześnie należy do podejścia na poziomie algorytmicznym (modyfikującym algorytmy dla danych zbalansowanych) jest BRACID (zob. np. [32, 33]). BRACID jest modyfikacją algorytmu RISE tak, aby można go było zastosować dla danych niezbalansowanych. Istnieje kilka istotnych różnic pomiędzy algorytmami BRACID i RIONIDA. Po pierwsze, BRACID generuje reguły w fazie uczenia (z wyprzedzeniem), podczas gdy RIONIDA robi to w fazie testowania (czyli zgodnie z podejściem leniwym). Po drugie, BRACID zaczyna od reguł równoważnych instancjom i indukuje reguły quasi-optymalne dla danego zbioru danych. RIONIDA przyjmuje inną strategię i bierze pod uwagę dużą przestrzeń sparametryzowanych reguł sformułowanych w określonym języku. Warto zwrócić uwagę, że różne parametryzacje odpowiadają różnym podejściom, w tym podejściu bazującemu wyłącznie na instancjach, podejściu bazującemu wyłącznie na regułach oraz podejściu łączącemu te podejścia. Dla zadanego zbioru danych RIONIDA optymalizuje ustawienie parametrów reguł, i robi to bardzo wydajnie. Po trzecie, BRACID optymalizuje reguły dla F-miary, podczas gdy RIONIDA może optymalizować dowolną, ze zdefiniowanych przez użytkownika na podstawie macierzy konfuzji miar jakości, i robi to bardzo skutecznie.

4. Główne wyniki pracy

Głównymi wynikami pracy są opracowanie i analiza dwóch algorytmów uczących: RIONA oraz RIONIDA. Pierwszy algorytm jest dedykowany dla danych zbalansowanych, natomiast drugi dla danych niezbalansowanych. Algorytm RIONA oraz jego podstawowa analiza zostały wykonane wspólnie przez autora rozprawy oraz Arkadiusza Wojnę (zob. [16–18]). Natomiast prace związane z algorytmem RIONIDA zostały wykonane przez autora rozprawy.

W tej pracy skupiamy się bardziej na problemie uczenia niezbalansowanego. Dlatego też algorytm RIONIDA jest kluczowy dla tej rozprawy. Niemniej jednak algorytm RIONA stanowi niezbędny etap w konstrukcji algorytmu RIONIDA. Ponieważ jednak algorytm RIONA ma znaczenie tylko dla danych zbalansowanych, nie przedstawiamy jego pełnej analizy. W szczególności, nie przedstawiamy szczegółowo porównania algorytmu RIONA z innymi znanymi z literatury algorytmami stosowanymi dla danych zbalansowanych, a jedynie podajemy odnośniki do opublikowanych prac związanych z jakością algorytmu RIONA, w których takie porównania są zawarte.

Główną ideą algorytmu RIONA jest połączenie dwóch powszechnie stosowanych empirycznych podejść do uczenia z przykładów, a mianowicie uczenia instancyjnego i indukcji regułowej. Algorytm RIONA posiada kilka własności prowadzących do konstrukcji odpowiednich klasyfikatorów dla danych zbalansowanych. Skonstruowanie algorytmu, który zapewnia wszystkie te własności jest sporym wyzwaniem i stanowi istotny wynik tej pracy. Poniżej krótko przedstawiamy te własności.

1. Algorytm RIA jest szczególnym przypadkiem algorytmu RIONA z maksymalnym zbiorem wsparcia (tzn. cały zbiór treningowy jest traktowany jako sąsiedztwo przypadku testowego). RIA realizuje przytoczoną wcześniej ideę Vapnika: „staraj się rozwiązać problem bezpośrednio i nigdy nie rozwiązuj bardziej ogólnego problemu, jako kroku pośredniego” i ma bardzo ciekawą i praktyczną własność. Mianowicie, RIA jest równoważny (względem klasyfikacji) algorytmowi, który w kroku pośrednim generuje wszystkie niesprzeczne i maksymalnie ogólne reguły (patrz Twierdzenie 3.4 i

Wniosek 3.5 w rozprawie). Ten ostatni algorytm ma wykładniczą złożoność czasową, podczas gdy RIA ma znacznie mniejszą – kwadratową złożoność czasową. Ponadto, w szczególności algorytm RIA (oraz RIONA) nie wymaga dyskretyzacji (ani grupowania wartości). Grupuje on odpowiednio wartości zarówno dla atrybutów liczbowych jak i nominalnych podczas generowania reguł.

2. W ogólnym przypadku algorytmu RIONA, decyzja jest przewidywana na podstawie zbioru wsparcia ograniczonego do sąsiedztwa przypadku testowego, a nie na podstawie całego zbioru wsparcia pochodzącego ze wszystkich reguł pokrywających przypadek testowy.

3. Rozmiar optymalnego sąsiedztwa jest automatycznie indukowany w fazie uczenia. Warto zaznaczyć, że uczenie optymalnego sąsiedztwa bazuje na idei programowania dynamicznego (zob. np. [9]), co sprawia, że obliczeniowa złożoność czasowa tego etapu jest niska. Ponadto badanie empiryczne wykazało interesujący fakt, że wystarczy rozważyć niewielkie sąsiedztwo, aby uzyskać dokładność klasyfikacji porównywalną z algorytmem indukowanym z całego zbioru uczącego (zob. np. [39], gdzie jest podany algorytm obliczający kompletny zbiór niesprzecznych i maksymalnie ogólnych reguł decyzyjnych). Tak więc połączenie kNN i algorytmu bazującego na regułach prowadzi do znacznego przyspieszenia zarówno fazy uczenia jak i testowania w porównaniu z algorytmem RIA wykorzystującym wszystkie maksymalnie ogólne reguły.

4. Metoda ta jest konkurencyjna w stosunku do innych znanych z literatury podejść z punktu widzenia jakości klasyfikacji. W szczególności, prezentowany klasyfikator charakteryzuje się wysoką dokładnością dla dwóch rodzajów zbiorów danych: bardziej odpowiednich dla klasyfikatorów kNN oraz bardziej odpowiednich dla klasyfikatorów regułowych.

5. Sformułowane i udowodnione w rozprawie wyniki teoretyczne pokazują związki generowanych przez algorytm RIONA klasyfikatorów zarówno z klasyfikatorami instancyjnymi, jak i regułowymi (patrz Fakt 3.9, Twierdzenia 3.10 i 3.11, Wniosek 3.12). W szczególności, pokazujemy równoważność (względem klasyfikacji) algorytmu RIONA z algorytmem regułowym generującym wszystkie niesprzeczne i maksymalnie ogólne reguły z sąsiedztwa przypadku testowego. W konsekwencji, klasyfikator RIONA może być reprezentowany przez klasyfikator regułowy, z regułami łatwo interpretowalnymi przez człowieka. Te teoretyczne wyniki zapewniają własność wyjaśnialności wynikowych klasyfikatorów generowanych przez algorytm RIONA i mogą być wykorzystane w sytuacji, gdy wymagane jest wyjaśnienie lub uzasadnienie uzyskanej decyzji.

Ponadto, zaproponowaliśmy algorytm *Optimal Nearest Neighbour* (ONN), który jest prostą modyfikacją algorytmu RIONA. W ONN zamiast stosowania reguł do konstruowania sąsiedztwa wykorzystywana jest metoda kNN. ONN używa tej samej metryki, co algorytm RIONA i w podobny sposób uczy się optymalnego sąsiedztwa. Istnieją dwa powody, dla których warto tutaj wspomnieć o tym algorytmie: (i) dla niektórych zbiorów danych algorytm ten ma lepszą jakość klasyfikacji niż RIONA, oraz (ii) fakt ten został wykorzystany w konstrukcji algorytmu RIONIDA (patrz dalsza dyskusja na temat algorytmu RIONIDA).

Algorytm RIONA nie jest jednak odpowiedni dla niezbalansowanych danych, z powodów wymienionych we Wstępie. Poniżej odnosimy się do nich wyjaśniając, dlaczego RIONA nie sprawdza się dla takich danych.

- RIONA stara się maksymalizować dokładność. Miara ta przypisuje jednakowe koszty

błędnej klasyfikacji zarówno dla klasy mniejszościowej jak i większościowej. Jednakże, takie podejście nie jest odpowiednie dla danych niezbalansowanych.

- RIONA implicite zakłada zrównoważony rozkład klas. Oznacza to, że RIONA nie radzi sobie poprawnie z danymi, w których dla wielu obiektów z klasy mniejszościowej ich sąsiedztwo zawiera znacznie więcej obiektów z klasy większościowej. Wówczas również więcej obiektów z klasy większościowej wspiera reguły skonstruowane dla obiektów z klasy mniejszościowej. W konsekwencji, wiele przykładów testowych z klasy mniejszościowej może być błędnie zaklasyfikowanych, jako należące do klasy większościowej.

- Można uzyskać wysoki stopień dokładności przy niskiej dokładności dla klasy mniejszościowej. Fakt ten powoduje, że klasyfikator RIONA nie jest akceptowalny dla klasyfikacji danych niezbalansowanych.

Algorytm RIONIDA bazuje na modyfikacji algorytmu RIONA. Jego celem jest konstruowanie klasyfikatorów dla danych niezbalansowanych o możliwie najwyższej jakości. W celu uproszczenia zadania, liczba klas decyzyjnych w algorytmie RIONIDA jest ograniczona do dwóch, co oznacza, że algorytm ten ma bezpośrednie zastosowanie tylko dla problemów klasyfikacji binarnej. Algorytm RIONIDA, analogicznie do RIONA, bazuje na połączeniu uczenia instancyjnego i indukcji reguł. Przy konstruowaniu algorytmu RIONIDA wprowadzono jednak kilka istotnych zmian w stosunku do algorytmu RIONA. Zmiany te pozwoliły na uzyskanie algorytmu, który jest istotnym wynikiem pracy. Posiada on następujące ważne własności.

1. RIONIDA przeprowadza optymalizację w fazie uczenia nie względem dokładności, lecz względem miary bardziej adekwatnej dla danych niezbalansowanych (np. F-miary lub G-mean).

2. Ponieważ dla problemu uczenia niezbalansowanego poprawna klasyfikacja do klasy mniejszościowej jest ważniejsza niż do klasy większościowej, klasa mniejszościowa jest traktowana w szczególny sposób podczas rozwiązywania konfliktu, tj. metody wyboru ostatecznej decyzji, jeśli istnieją pewne przesłanki za klasyfikowaniem zarówno do klasy mniejszościowej, jak i do klasy większościowej. Kolejny problem związany jest z wyborem, w jakim stopniu klasa mniejszościowa jest ważniejsza od klasy większościowej.

3. Ponieważ algorytm ONN dla niektórych zbiorów danych daje lepsze wyniki niż algorytm RIONA, zdecydowaliśmy się połączyć mocne strony obu tych algorytmów. RIONIDA może wykorzystywać podejście bazujące na regułach, podejście bazujące na instancjach, lub też kombinację tych podejść. Wybór ten jest dokonywany przy użyciu parametru, który określa, w jakim stopniu użycie reguł w sąsiedztwie jest uznane za odpowiednie.

4. Wszystkie główne (wewnętrzne) parametry algorytmu RIONIDA są automatycznie indukowane w fazie uczenia. Przypomnijmy, że na te parametry składają się: rozmiar sąsiedztwa (cecha ta jest zaadaptowana z algorytmu RIONA), stopień ważności klasy mniejszościowej oraz dopuszczalny poziom (nie)sprzeczności. Ten ostatni współczynnik określa, w jakim stopniu wykorzystywane jest podejście regułowe (lub odwrotnie – instancyjne). Ponownie, należy podkreślić, że przedstawiamy efektywną w czasie metodę uczenia się wszystkich tych parametrów za pomocą techniki programowania dynamicznego (patrz Twierdzenie 4.3). Ponadto, zaprezentowaliśmy możliwość dalszego przyspieszenia algorytmu RIONIDA i zmniejszenia jego złożoności pamięciowej (patrz Twierdzenie 4.5 i Fakt 4.6).

5. Dla miar jakości G-mean i F-miary, dwa twierdzenia dostarczają oszacowań optymalnego stopnia ważności klasy mniejszościowej przy założeniu rozkładu „całkowicie losowego" (patrz Twierdzenia 4.1 i 4.2). Oszacowania te dostarczają szybszą alternatywę dla rozwiązania otrzymanego w wyniku uczenia parametrów i mogą być wykorzystane do ustawienia domyślnej wartości odpowiedniego parametru w algorytmie RIONIDA. Ponadto, z twierdzeń tych wynika ciekawy wniosek. Mianowicie, dla pewnej klasy klasyfikatorów optymalny klasyfikator może być znacząco różny (względem klasyfikacji) dla różnych miar jakości. Dodatkowo, oceny tych dwóch optymalnych klasyfikatorów mogą być znacząco różne w zależności od miary jakości użytej do oceny. Praktyczny wniosek dla rzeczywistych problemów klasyfikacji jest taki, że bez dokładnego określenia konkretnej miary jakości, którą jesteśmy zainteresowani, termin „najlepszy klasyfikator" może być niejednoznaczny lub nawet mylący.

6. RIONIDA osiąga istotnie lepsze wyniki niż renomowane podejścia znane z literatury testowane w rozprawie. Porównanie jakości algorytmu RIONIDA przeprowadziliśmy z wszystkimi głównymi renomowanymi algorytmami, których kody były dostępne dla autora rozprawy. Taki wybór gwarantował powtarzalność eksperymentów. Przewaga algorytmu RIONIDA została wykazana w eksperymentach na benchmarkach, z wykorzystaniem miar jakości istotnych dla danych niezbalansowanych. Testy porównawcze zostały starannie zaprojektowane z wykorzystaniem aktualnej wiedzy na temat ewaluacji algorytmów uczących się w kontekście danych niezbalansowanych. W szczególności wzięto pod uwagę odpowiednie miary jakości, ich właściwą metodę estymacji, co samo w sobie jest złożonym problemem, odpowiedni dobór zbiorów danych, czy wreszcie możliwość różnych ustawień algorytmów. Zamieszczono również statystyczną ocenę uzyskanych wyników. Nadmieńmy, że RIONIDA wypada znacznie lepiej, niż RIONA z rozwiązaniem na poziomie danych tzn. RIONA w kombinacji z odpowiednimi filtrami na jej wejściu, modyfikującymi dane.

7. RIONIDA posiada pożądaną własność wyjaśnialności, którą zapewniają przede wszystkim teoretyczne własności algorytmu RIONA, opisane powyżej w punkcie 5.

8. Większość z wymienionych powyżej pozytywnych własności algorytmu RIONA posiada również algorytm RIONIDA. W szczególności RIONIDA, analogicznie do algorytmu RIONA, nie wymaga uprzedniej dyskretyzacji ani grupowania wartości. Warto też dodać, że RIONIDA przy specyficznych ustawieniach parametrów staje się równoważna algorytmowi RIONA.

Podsumowując, algorytm RIONA jest metodą uczenia, która jest efektywna czasowo, z dobrą jakością klasyfikacji dla danych zbalansowanych. Algorytm RIONIDA jest połączeniem algorytmów RIONA i ONN (wraz z ich dalszym rozszerzeniem), zaprojektowanym do pracy z problemami uczenia niezbalansowanego (choć ograniczonym do klasyfikacji binarnej). Co ważne, RIONIDA osiąga znacząco lepsze wyniki niż współczesne algorytmy zaprojektowane do działania w przypadku danych niezbalansowanych, a jednocześnie ma stosunkowo niską złożoność obliczeniową. W szczególności, RIONIDA ma znacząco lepszą jakość niż algorytm RIONA poprzedzony filtrowaniem danych, co jest powszechnie znanym i bezpośrednim sposobem przystosowania standardowego algorytmu do danych niezbalansowanych.

Na koniec chcielibyśmy wspomnieć o dwóch pomniejszych rezultatach rozprawy. Pierwszym z nich jest zaproponowane podejście metodologiczne z trzema poziomami porównań algorytmów uczących, uwzględniające wiele wariantów algorytmów, w tym ich nie-domyślne

ustawienia parametrów (patrz dyskusja o eksperymentach w Rozdziale 5). Drugi to konstrukcja przykładu prowadzącego do różnych wniosków, co do przewagi jednego z porównywanych algorytmów nad drugim w zależności od sposobu agregacji cząstkowych wyników walidacji krzyżowej. Mówiąc dokładniej, w przypadku makro-uśredniania (ang. *macro-averaging*) jeden algorytm okazuje się lepszy od drugiego, a w przypadku mikro-uśredniania (ang. *micro-averaging*) jest na odwrót (patrz Dodatek B).

Wreszcie, przedstawione wyniki dotyczące algorytmów RIONA i RIONIDA mogą być postrzegane, jako przykłady kilku abstrakcyjnych i ogólnych kierunków badań nad efektywnymi i wydajnymi algorytmami uczącymi. Żywimy nadzieję, że niektóre z naszych pomysłów będą mogły być zaadaptowane w projektach, w których projektowanie algorytmów uczących się opiera się na innych niż w rozprawie pojęciach. Po pierwsze, pokazaliśmy, że w przypadku klasyfikatorów regułowych czas obliczeń miary rozwiązywania konfliktów bazującej na wszystkich niesprzecznych i maksymalnie ogólnych regułach może być znacznie przyspieszony dzięki zastosowaniu podejścia leniwego – podobne podejście może być zastosowane dla innych miar rozwiązywania konfliktów. Po drugie, pokazaliśmy, że połączenie uczenia instancyjnego z innym podejściem, np. regułowym, może być korzystne zarówno z punktu widzenia jakości, jak i efektywności – kombinacja tego rodzaju może sprawdzić się również w przypadku podejść innych niż regułowe, np. drzew decyzyjnych. Po trzecie, pokazaliśmy, że parametryzacja klasyfikatorów bazujących na podejściu leniwym może być znacznie efektywniej realizowana z wykorzystaniem programowania dynamicznego niż poprzez bezpośrednie obliczenia – podejście to może być zastosowane dla innych architektur algorytmów i/lub innych parametryzacji. Po czwarte, pokazaliśmy przykład, w jaki sposób algorytm uczący dla danych zbalansowanych może być z powodzeniem zmodyfikowany dając w wyniku algorytm uczący dla danych niezbalansowanych – analogiczne przekształcenia można byłoby zrealizować w przypadku innych algorytmów dedykowanych dla danych zbalansowanych.

Wyniki przedstawione w mojej rozprawie otwierają kilka innych możliwych kierunków dla przyszłych badań. W szczególności, opisujemy kilka użytecznych usprawnień zwiększających wydajność i użyteczność algorytmu RIONIDA, które mogłyby zostać zaimplementowane w przyszłości. Omawiamy również dalsze eksperymenty, które mogłyby zostać przeprowadzone w celu pogłębienia analizy algorytmu RIONIDA. Chociaż sprawdziliśmy już kilka kierunków możliwego ulepszenia algorytmu RIONIDA, szereg dalszych badań pozostaje jeszcze do wykonania. W szczególności, zaproponowaliśmy rozszerzenie algorytmu RIONIDA o 4-wymiarową parametryzację, która wydaje się być obiecująca dla przyszłych badań. Twierdzenia dotyczące optymalnego stopnia ważności klasy mniejszościowej mogą być rozszerzone na ogólniejsze sytuacje występujące w praktyce. Mimo uzyskanych wstępnych wyników sugerujących, że zastosowanie algorytmu RIONIDA dla dużych danych (ang. *big data*) niezbalansowanych jest wykonalne i skuteczne, konieczne są jeszcze dalsze badania w tym kierunku. Wreszcie, RIONIDA może być dalej rozwijana w celu dostosowania jej do specyficznych rzeczywistych problemów.

Literatura

[1] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] David W. Aha (ed.). *Lazy Learing*. Springer, Dordrecht, 1st edition, 1997.

[3] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[4] Ammar Almasri, Erbug Celebi, and Rami S. Alkhawaldeh. EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance. *Scientific Programming*, 2019:Article No. 3610248, 1–13, 2019.

[5] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3(10):27–38, 2013.

[6] Jerzy Błaszczyński, Salvatore Greco, and Roman Słowiński. Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence*, 25(2):284–294, 2012.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and William P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

[8] William W. Cohen. Fast Effective Rule Induction. In *Machine Learning Proceedings 1995*, pp. 115–123. Morgan Kaufmann, San Francisco, CA, 1995.

[9] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, Cambridge, MA, 3rd edition, 2009.

[10] Nilanjan Dey, Samarjeet Borah, Rosalina Babo, and Amira S. Ashour (eds.). *Social Network Analytics: Computational Research Methods and Techniques*. Academic Press, London, 1st edition, 2019.

[11] Pedro Domingos. Unifying instance-based and rule-based induction. *Machine Learning*, 24(2): 141–168, 1996.

[12] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer, Cham, 1st edition, 2018.

[13] Eibe Frank and Ian H. Witten. Generating Accurate Rule Sets Without Global Optimization. In *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*, pp. 144–151. Morgan Kaufmann, San Francisco, CA, 1998.

[14] Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrac. *Foundations of Rule Learning*. Cognitive Technologies. Springer, Heidelberg, 2012.

[15] Anil K. Ghosh. On optimum choice of k in nearest neighbor classification. *Computational Statistics & Data Analysis*, 50(11):3113–3123, 2006.

[16] Grzegorz Góra and Arkadiusz Wojna. RIONA: A New Classification System Combining Rule Induction and Instance-Based Learning. *Fundamenta Informaticae*, 51(4):369–390, 2002.

[17] Grzegorz Góra and Arkadiusz Wojna. RIONA: A Classifier Combining Rule Induction and K-nn Method with Automated Selection of Optimal Neighbourhood. In *Proceedings of the 13th European Conference on Machine Learning (ECML 2002)*, pp. 111–123. Springer-Verlag, Heidelberg, 2002.

[18] Grzegorz Góra and Arkadiusz Wojna. Local Attribute Value Grouping for Lazy Rule Induction. In *Rough Sets and Current Trends in Computing (RSCTC 2002)*, pp. 405–412. Springer, Heidelberg, 2002.

[19] Lacrimioara Grama and Corneliu Rusu. Choosing an accurate number of mel frequency cepstral coefficients for audio classification purpose. In *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis (ISPA 2017)*, pp. 225–230, 2017.

[20] Lacrimioara Grama and Corneliu Rusu. Adding audio capabilities to TIAGo service robot. In *2018 International Symposium on Electronics and Telecommunications (ISETC)*, pp. 1–4, 2018.

[21] Jerzy W. Grzymala-Busse. LERS-A System for Learning from Examples Based on Rough Sets. In Roman Słowiński (ed.), *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, pp. 3–18. Springer, Dordrecht, 1992.

[22] Jerzy W. Grzymala-Busse, Jerzy Stefanowski, and Szymon Wilk. A Comparison of Two Approaches to Data Mining from Imbalanced Data. *Journal of Intelligent Manufacturing*, 16(6): 565–573, 2005.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2nd edition, 2009.

[24] Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[25] Haibo He and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, Piscataway, NJ, 1st edition, 2013.

[26] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, Cambridge, 2011.

[27] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Computing Surveys*, 52(4):1–36, 2019.

[28] Jinyan Li, Guozhu Dong, Kotagiri Ramamohanarao, and Limsoon Wong. DeEPs: A New Instance-Based Lazy Discovery and Classification System. *Machine Learning*, 54(2):99–124, 2004.

[29] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[30] Ryszard S. Michalski, Igor Mozetic, Jiarong Hong, and Nada Lavrac. The Multi-Purpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains. In *Proceedings of the 5th AAAI National Conference on Artificial Intelligence*, pp. 1041–1045. AAAI Press, 1986.

[31] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, 1997.

[32] Krystyna Napierała. *Improving Rule Classifiers For Imbalanced Data*. PhD thesis, Poznań University of Technology, Poznań, Poland, 2012.

[33] Krystyna Napierała and Jerzy Stefanowski. BRACID: a comprehensive approach to learning rules from imbalanced data. *Journal of Intelligent Information Systems*, 39(2):335–373, 2012.

[34] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[35] Patricia Riddle, Richard Segal, and Oren Etzioni. Representation Design and Brute-force Induction in a Boeing Manufacturing Domain. *Applied Artificial Intelligence*, 8(1):125–147, 1994.

[36] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, Hoboken, NJ, 4th edition, 2021.

[37] Corneliu Rusu and Lacrimioara Grama. Recent developments in acoustical signal classification for monitoring. In *2017 5th International Symposium on Electrical and Electronics Engineering (ISEEE)*, pp. 1–10, 2017.

[38] Andrzej Skowron. Boolean reasoning for decision rules generation. In *Methodologies for Intelligent Systems (ISMIS 1993)*, pp. 295–305. Springer, Heidelberg, 1993.

[39] Andrzej Skowron and Cecylia Rauszer. The Discernibility Matrices and Functions in Information Systems. In Roman Słowiński (ed.), *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, pp. 331–362. Springer, Dordrecht, 1992.

[40] Yang Song, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles. IKNN: Informative K-Nearest Neighbor Pattern Classification. In *Knowledge Discovery in Databases (PKDD 2007)*, pp. 248–264. Springer, Heidelberg, 2007.

[41] Jerzy Stefanowski. Algorithms of rule induction for knowledge discovery (in Polish). Habilitation Thesis, 2001.

[42] Jerzy Stefanowski. On Combined Classifiers, Rule Induction and Rough Sets. In James F. Peters, Andrzej Skowron, Ivo Düntsch, Jerzy Grzymała-Busse, Ewa Orłowska, and Lech Polkowski (eds.), *Transactions on Rough Sets VI: Commemorating the Life and Work of Zdzisław Pawlak, Part I*, pp. 329–350. Springer, Heidelberg, 2007.

[43] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, New York, NY, 1st edition, 1998.

[44] Dennis L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

[45] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey McLachlan, Angus Ng, Bing Liu, Philip Yu, Zhi-Hua Zhou, Michael Steinbach, David Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[46] Qiang Yang and Xindong Wu. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*, 05(04):597–604, 2006.

[47] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient kNN Classification With Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1774–1785, 2018.