DOI: 10.24427/978-83-66391-96-3_1

Rozdział 1
Obliczenia granularne w przetwarzaniu danych niezbalansowanych

Katarzyna Borowska
Wydział Informatyki, Politechnika Białostocka

Streszczenie: Dane niezbalansowane od ponad dwóch dekad stanowią jeden z najbardziej interesujących problemów eksploracji danych. Najnowsze badania wykazały, że zagadnienie to nie powinno być rozpatrywane jedynie w kontekście deficytu informacji o jednej z klas. W rzeczywistości spadek jakości standardowych klasyfikatorów w zetknięciu z tego typu zbiorami danych wiąże się przede wszystkim z wysoką złożonością rozkładu danych. Złożoność ta wzrasta, gdy w charakterystyce danych pojawiają się dodatkowe trudności, takie jak dekompozycja klas, niejednoznaczność strefy brzegowej oraz występowanie szumu. Wymaga to zastosowania dedykowanych rozwiązań, w szczególności uwzględniających lokalną charakterystykę rozkładu. Niniejsza praca opisuje algorytm RGA, który wykorzystuje granule informacyjne i zbiory przybliżone do ukierunkowanego modyfikowania zbioru uczącego oraz proponuje automatyzację doboru parametrów. Przeprowadzone eksperymenty dowiodły wysokiej skuteczności zaproponowanej metody w porównaniu z podobnymi technikami.

Słowa kluczowe: dane niezbalansowane, przetwarzanie wstępne danych, obliczenia granularne, granule informacyjne, zbiory przybliżone, SMOTE

Wprowadzenie

Rozwój technologii jest obecnie szybki jak nigdy dotąd. Spektakularne wynalazki, o których dawniej nikt nawet nie marzył, stają się częścią naszej codzienności. Roboty, autonomiczne pojazdy, asystenci głosowi, zaawansowana diagnostyka medyczna, telechirurgia, biodruk 3D, inteligentne domy, rozszerzona rzeczywistość to tylko nieliczne przykłady dziedzin, które w bardzo spektakularny sposób rozwinęły się w ostatnim czasie. Każda z nich korzysta z szeroko pojmowanej sztucznej inteligencji. Ta zaś opiera się przede wszystkim na zgromadzonych danych. Wydobywanie cennej wiedzy z danych jest złożonym procesem, wymagającym eksperckiej wiedzy, doświadczenia oraz odpowiednio przygotowanych i wdrożonych technik. Niestety, nawet najbardziej kompetentni badacze napotykają w trakcie eksploracji danych na problemy,

które niejednokrotnie wynikają z samej natury danych [22]. Wśród takich problemów szczególnie interesujący i dominujący pod względem częstości występowania jest problem danych niezbalansowanych. Dane tego typu charakteryzują się znaczną dysproporcją liczebności próbek reprezentujących poszczególne klasy. W praktyce oznacza to, że nie udało się pozyskać dostatecznych informacji na temat jednej z klas – liczba odnotowanych obserwacji należących do tej klasy (nazywanej mniejszościową lub pozytywną) jest diametralnie mniejsza niż liczba pozostałych obserwacji (z klasy nazywanej większościową lub negatywną) [9]. Co więcej, to właśnie ta klasa, która występuje w zbiorze danych w niedomiarze, jest zazwyczaj przedmiotem szczególnego zainteresowania. Sam fakt niewielkiej liczby próbek reprezentujących jedną z klas nie stanowiłby tak znaczącego problemu, gdyby nie tendencja narzędzi używanych standardowo w klasyfikacji do uogólniania, czyli podejmowania decyzji na korzyść klas, których reprezentacja jest największa [9, 26]. Ponadto, na poziom złożoności problemu wpływa również rozkład danych. Okazuje się, że nie tylko deficyt danych z jednej z klas może skutkować negatywnym wpływem na proces uczenia maszynowego, ale również sama charakterystyka danych, w szczególności jej poziom skomplikowania, może doprowadzić do degradacji wyników [9, 17, 26]. Te dwa czynniki: niezbalansowanie danych oraz ich złożoność stanowią główne przyczyny komplikacji w procesie uczenia maszynowego. Wśród proponowanych rozwiązań problemu można wyróżnić metody modyfikujące zbiór uczący, zmodyfikowane klasyfikatory oraz techniki hybrydowe, łączące obydwa podejścia [9, 10, 15, 26].

W ramach niniejszej pracy opisana zostanie metoda, która proponuje zastosowanie obliczeń granulanych [21] we wstępnym przetwarzaniu danych niezbalansowanych, aby uzyskać zbalansowany pod względem liczebności zbiór próbek oraz doprowadzić do zmniejszenia złożoności dystrybucji danych. Biorąc pod uwagę indywidualny charakter rozkładów danych z różnych dziedzin oraz ich wewnętrzną dystrybucję, opracowany algorytm pozwala na zachowanie elastyczności przez możliwość automatycznego doboru parametrów.

1.1. Istota problemu danych niezbalansowanych

1.1.1. Źródła problemu

W zagadnieniu danych niezbalansowanych głównym wyznacznikiem powagi problemu jest liczba oraz rodzaj dziedzin, które są reprezentowane przez zbiory danych dotknięte tym zjawiskiem. Okazuje się, że dane tego typu charakteryzują zagadnienia z bardzo wielu obszarów otaczającej nas rzeczywistości, również tych, które uznawane są za wyjątkowo istotne. W szczególności są to: diagnostyka medyczna, wykrywanie wszelkich anomalii, takich jak wycieki ropy naftowej, detekcja wad produkcyjnych, wykrywanie przestępstw finansowych oraz włamań, zarządzanie ryzykiem,

kategoryzacja tekstu oraz analiza sentymentu [9, 12, 15, 16, 28]. W większości tych dziedzin to poprawne rozpoznawanie obiektów z klasy, która jest mniej liczna, stanowi główny cel analizy [28]. Spadek jakości algorytmu względem mniej licznej klasy w wielu dziedzinach może prowadzić do dramatycznych skutków, takich jak zbyt późne wdrożenie (lub całkowity brak wdrożenia) odpowiednich do danego problemu rozwiązań (np. niepodjęcie leczenia we wczesnym stadium choroby nowotworowej [12]). Potwierdza to, jak ważna jest świadomość istnienia niniejszego problemu oraz odpowiednie reagowanie we wczesnych etapach eksploracji danych.

Podstawowym krokiem, który pozwala wykrywać potencjalny problem danych niezbalansowanych, jest zbadanie wskaźnika *imbalanced ratio* (IR), [24]:

$$IR = \frac{N^-}{N^+}, \tag{1.1}$$

gdzie N^- stanowi liczbę obiektów z klasy większościowej, zaś N^+ oznacza liczbę obiektów z klasy mniejszościowej.

Nie zdefiniowano jednoznacznego progu określającego sytuację, w której można nazwać analizowaną dystrybucję danych niezbalansowaną [26]. Zarówno proporcje między klasami takie, jak 100:1, 1000:1, czy 1000000:1 mogą wskazywać na konieczność szczególnego traktowania danych oraz uwzględnienia w dalszych pracach tej dysproporcji [28].

Oprócz wspomnianej weryfikacji wskaźnika IR, niezwykle ważne jest uwzględnienie dodatkowych trudności, które mają bezpośredni wpływ na złożoność rozkładu danych. Ostatnie szczegółowe badania potwierdzają, że to właśnie te komplikacje, w połączeniu z dysproporcją liczebności próbek z poszczególnych klas, stanowią główne źródło problemu [9, 12, 14, 17, 18, 24, 26, 28, 29]. Zaburzona charakterystyka danych oraz ich niezbalansowanie ma ogromne znaczenie w procesie uczenia klasyfikatorów. Większość badań pokazuje w takich przypadkach ukierunkowanie modelu do podejmowania decyzji na korzyść klasy większościowej. Próbki z klasy mniejszościowej, ich cechy szczególne, wyróżniające właściwości, nie mogą zostać prawidłowo rozpoznane i w rezultacie zostają zdominowane przez bardziej liczne obserwacje z klasy większościowej. Wśród wspomnianych dodatkowych komplikacji można wyróżnić:

- Dekompozycję klas – zbiór danych składa się z rozdrobnionych klastrów, skupiających bardzo niewielkie zbiory próbek z poszczególnych klas [12, 14, 18, 24, 28, 29]. Zjawisko to jest szczególnie niekorzystne w przypadku klasy mniejszościowej, której liczebność nawet bez podziału na klastry jest zbyt mała, aby standardowe klasyfikatory miały możliwość wyznaczenia charakterystyki wyróżniającej obiekty reprezentujące daną klasę [14]. Podział na rozproszone podzbiory dodatkowo utrudnia rozpoznawanie klasy mniejszościowej.
- Niejednoznaczność strefy brzegowej – występuje w sytuacji, gdy istnieją w rozkładzie danych obszary, które koncentrują bardzo podobne do siebie obiekty z różnych klas. Stąd klasyfikatory nie mają możliwości wyznaczenia reguł odpowiednio

rozróżniających próbki między klasami, gdyż odbywałoby się to kosztem zdolności uogólniania [9]. Przeprowadzone badania wskazują, że im większy jest poziom nakładania się klas w obszarze brzegowym, tym większa podatność algorytmu uczenia na problem niezbalansowania danych [28].
- Występowanie szumu – istnienie specyficznych przykładów, które mimo, że reprezentują klasę mniejszościową, posiadają cechy specyficzne dla klasy większościowej. Obiekty tego typu w przestrzeni cech zajmują miejsca oddalone od próbek z tej samej klasy, jednocześnie znajdując się w otoczeniu przykładów z klasy przeciwnej [24]. Próbki tego rodzaju mogą stanowić błędne dane, jednak nie można wykluczyć, że są to bardzo rzadkie przypadki [29].

Opisane trudności, zwiększające złożoność danych, powinny zawsze być rozpatrywane równolegle z problemem danych niezbalansowanych.

1.1.2. Parametry jako zagadnienie zwiększające złożoność problemu

Oprócz wspomnianych istotnych komplikacji, które można nazwać wewnętrznymi, gdyż wynikają z samej natury danych niezbalansowanych oraz złożoności ich rozkładów, występują również trudności niejako zewnętrzne, związane z narzędziami, które potencjalnie mogłyby stanowić skuteczne rozwiązanie problemu danych niezbalansowanych. Otóż, nawet te narzędzia, których wyniki w zetknięciu z niezbalansowanymi zbiorami danych są obiecujące, niezwykle często nie spełniają kryterium oczekiwanej elastyczności. W istocie pozwalają na uzyskiwanie rezultatów o wysokiej jakości, jednak zazwyczaj jest to poprzedzone żmudną pracą nad odpowiednim doborem parametrów, które są adekwatne jedynie do danego, konkretnego przypadku. Jednocześnie te same parametry najprawdopodobniej nie znajdą zastosowania przy pracy nad innym zbiorem danych (w szczególności takim, który będzie miał zupełnie inną charakterystykę rozkładu). W ten sposób można wyszczególnić kolejne, niekiedy pomijane oraz poniekąd ukryte, źródło problemów w przetwarzaniu danych niezbalansowanych: trudności w doborze parametrów.

Wiele technik przedstawianych jako środki zaradcze na niezbalansowanie danych wymaga konfiguracji różnych parametrów. Wśród nich można wyróżnić między innymi wylistowane poniżej parametry:
- liczba najbliższych sąsiadów,
- miary odległości,
- proporcje i wartości progowe pozwalające określić złożoność rozkładu danych,
- liczba klastrów,
- docelowa liczebność klasy mniejszościowej,
- wagi cech,
- wagi obiektów reprezentujących poszczególne klasy,
- próg określający podobieństwo obiektów,
- tryby przetwarzania dobrane w zależności od poziomu złożoności rozkładu danych.

Występowanie wymienionych parametrów jest uzależnione od wybranej metody, jednakże dobór nawet pojedynczych z nich może wpływać znacząco na wyniki. Brak doświadczenia lub wiedzy dziedzinowej może skutkować zaprzepaszczeniem skuteczności danego algorytmu przy nieodpowiednio dobranych parametrach. Ponadto, nawet przy dysponowaniu tymi pożądanymi cechami również istnieje ryzyko błędnej oceny odnośnie zbiorów danych o wysokiej złożoności rozkładu. Nie należy więc lekceważyć tych dodatkowych trudności i warto zautomatyzować proces dobierania parametrów celem uzyskania większej pewności co do tego, czy możliwości algorytmów są wykorzystywane w pełni w ramach prac nad danymi o konkretnej charakterystyce.

1.2. Przegląd dotychczasowych rozwiązań

Niniejsza praca obejmuje propozycję rozwiązania problemu danych niezbalansowanych należącą do technik modyfikujących dane (przetwarzania wstępnego). Stąd również w tym rozdziale zostaną przedstawione wyłącznie techniki pokrewne do przedstawionego algorytmu. Nie sposób jednak nie wspomnieć, że oprócz tej kategorii metod, istnieje wiele innych rozwiązań ukierunkowanych na dostosowanie algorytmów klasyfikacji na poprawne wykrywanie próbek z klasy mniejszościowej lub narzędzia hybrydowe, łączące modyfikacje klasyfikatorów z modyfikacjami danych [9]. Nie umniejszając skuteczności wymienionych alternatywnych metod, w niektórych zastosowaniach mogą okazać się niewystarczająco elastyczne ze względu na konieczność ingerencji w mechanizmy działania algorytmów. Analiza ta będzie więc dotyczyła wybranych metod niezależnych od klasyfikatorów.

1.2.1. Synthetic Minority Oversampling Technique (SMOTE)

Rozpatrując dostępne techniki niwelowania problemu danych niezbalansowanych, oparte na modyfikacjach w zbiorze uczącym, niewątpliwie należy wyróżnić algorytm SMOTE [6]. Metoda ta zawdzięcza swój sukces idei generowania nowych próbek danych z klasy mniejszościowej bazując na podobieństwie między istniejącymi przykładami reprezentującymi tę klasę [9]. W porównaniu z losowym tworzeniem obiektów z klasy mniejszościowej, stanowiących wierne kopie istniejących już obserwacji, algorytm ten można nazwać przełomowym. Niestety, biorąc pod uwagę dodatkowe trudności, wynikające ze złożoności rozkładów danych, nawet tak skuteczna metoda jak SMOTE okazuje się niepozbawiona wad [26, 28]. Generowanie nowych próbek z jednej z klas, bez uwzględniania charakterystyki konkretnego zbioru danych, może prowadzić do wprowadzenia niejasności oraz niespójności, a tym samym do zwiększenia złożoności rozkładu danych. W szczególności umiejscowienie pojedynczych próbek z klasy mniejszościowej w obszarach zajmowanych głównie przez obserwacje z klasy większościowej, stanowi jedynie wprowadzanie niepotrzebnego szumu, który

nic nie wniesie w procesie klasyfikacji oraz nie poprawi wyników (a może wręcz przeciwnie – zaburzyć je). Stąd w wielu pracach postanowiono skupić się na udoskonaleniu tak obiecującej metody [8].

1.2.2. Modyfikacje SMOTE

Pierwsza grupa rozwiązań opartych na SMOTE obejmuje techniki, które zakładają generowanie nowych, syntetycznych próbek z klasy mniejszościowej jedynie w określonych obszarach przestrzeni cech. Metoda Borderline-SMOTE [11] ogranicza tworzenie nowych obiektów jedynie do granicy między przykładami z obydwu klas. Autorzy sugerują, że to właśnie obszar graniczny jest punktem newralgicznym w procesie klasyfikacji i wzmocnienie reprezentacji próbek z klasy mniejszościowej w tym rejonie polepszy zdolność algorytmów do poprawnego klasyfikowania. MSMOTE [13] z kolei przyjmuje odmienne podejście. Na początku przydziela próbki do trzech grup w zależności od ich sąsiedztwa w przestrzeni cech. Podział ten decyduje o sposobie tworzenia nowych obiektów z klasy mniejszościowej: w przypadku obiektów leżących na granicy między klasami syntetyczna próbka tworzona jest na podstawie połączenia cech analizowanej aktualnie obserwacji oraz jedynie jej najbliższego sąsiada. Pozostałe obiekty podlegają identycznemu procesowaniu jak w przypadku klasycznej metody SMOTE. Przykłady uznane za szum nie biorą udziału w przetwarzaniu. Podobny mechanizm działania przyjmuje metoda Safe-Level SMOTE [5]. Próbki z klasy mniejszościowej są kategoryzowane pod względem poziomów bezpieczeństwa, które reprezentują. Poziomy te odzwierciedlają sąsiedztwo, w jakim znajduje się dana próbka – jeśli będzie to otoczenie złożone głównie z obiektów z klasy przeciwnej (większościowej), wtedy obserwacja uznawana jest za należącą do zmniejszonego poziomu bezpieczeństwa. Nowe próbki generowane są jedynie z pobliżu obszarów złożonych z obiektów o wysokich poziomach bezpieczeństwa (czyli obszarach składających się głównie z obserwacji z klasy mniejszościowej).

1.2.3. Metody filtrujące

Druga grupa rozwiązań nawiązujących do techniki SMOTE wzbogaca przetwarzanie oparte na ukierunkowanym generowaniu syntetycznych próbek o dodatkowy krok czyszczenia danych. Metoda SPIDER [18] w trybie „Weak" opiera się na wzmacnianiu liczebności przykładów z klasy mniejszościowej, które zostały uznane za bezpieczne (bazując na ich położeniu w przestrzeni cech). Tryb „Relabel" wprowadza fazę zmian etykiet obiektów z klasy większościowej, które zostały uznane za szum. W trybie „Strong" zarówno bezpieczne, jak i uznane za szum próbki z klasy mniejszościowej są wzmacniane (biorą udział w przetwarzaniu wstępnym analogicznym do metody SMOTE). *SMOTE-RSB*$_{\star}$ [23] zakłada generowanie syntetycznych przykładów algorytmem SMOTE. Dodatkowo, opierając się na teorii zbiorów przybliżonych, wprowadza

etap czyszczenia nowoutworzonych danych, które nie należą do dolnej aproksymacji. Metoda SMOTE + Tomek-links [1] po kroku tworzenia nowych obiektów, usuwa z obydwu klas przykłady, które wprowadzają niejasności do rozkładu danych. Wykorzystuje do tego pary najbliżej sąsiadujących obserwacji, tzw. Tomek-links, które zawierają próbki z dwóch różnych klas. Wszystkie tak zidentyfikowane pary są usuwane. SMOTE-IPF [24] do usuwania wprowadzających szum danych wykorzystuje filtr IPF. Filtr ten bazuje na modelach klasyfikatora C4.5 utworzonych na partycjach z podzielonego źródłowego zbioru danych. W przypadku, gdy ponad połowa z tak zbudowanych klasyfikatorów nie jest w stanie poprawnie sklasyfikować danej próbki, taka obserwacja jest usuwana. Filtrowaniu podlegają wszystkie przykłady: zarówno te pierwotnie występujące w zbiorze danych, jak i wygenerowane w procesie przetwarzania wstępnego metodą SMOTE. Algorytm VISROT [4] łączy podejście ukierunkowanego tworzenia nowych próbek (przez generowanie ich jedynie w wybranych obszarach przestrzeni cech – tych, które można uznać za bezpieczne, czyli oddalone od przykładów z klasy większościowej) z etapem czyszczenia danych. Wszystkie nowe, syntetyczne przykłady, które są uznane za szum, zostają usuwane.

1.3. Algorytm RGA

Głównym celem powstania algorytmu RGA (Rough-Granular Computing Algorithm) [2] jest zapewnienie elastyczności w połączeniu z efektywnością działania przy wydobywaniu wiedzy z danych niezbalansowanych. Opisywane podejście wykorzystuje wyłącznie proces przetwarzania wstępnego danych, nie ingerując w żaden sposób w mechanizmy działania algorytmów klasyfikacji. Stąd zachowana jest niezależność metody od wybranego klasyfikatora. Algorytm opiera się na wybiórczym i ukierunkowanym generowaniu nowych próbek z klasy mniejszościowej oraz dodatkowym kroku filtrowania. Etap filtrowania został ograniczony jedynie do syntetycznych próbek, aby nie stracić żadnych potencjalnie istotnych informacji z oryginalnego zbioru danych. Dodatkowo istotną cechą wyróżniającą metodę RGA jest możliwość automatycznego doboru parametrów do konkretnego problemu poprzez iteracyjne powtarzanie kroku przetwarzania wstępnego z różnymi kombinacjami wartości parametrów.

1.3.1. Wyznaczenie granul

Algorytm rozpoczyna się od identyfikacji granul informacyjnych [21, 27]. Identyfikacja ta opiera się na założeniu, że obserwacje mogą być traktowane jako podobne, jeśli znajdują się w bliskiej odległości w przestrzeni cech. Za pomocą metody kNN wyznaczane są granule skupiające instancje z klasy mniejszościowej wraz z sąsiadującymi próbkami z obydwu klas. Kluczowe jest w tym kroku określenie miary podobieństwa obiektów. Jako że rzeczywiste dane przeważnie wymagają możliwości przetwarzania

zarówno jakościowych, jak i ilościowych atrybutów, stosowana jest dedykowana metryka: Heterogeneous Value Distance Metric (HVDM), [30]. Pozwala ona skutecznie określić podobieństwo obserwacji bazując na ich sąsiedztwie, czyli bliskim położeniu w przestrzeni cech. Bliskie położenie wskazuje na zbliżone wartości poszczególnych atrybutów rozpatrywanych obserwacji.

1.3.2. Badanie stopnia zawierania granul

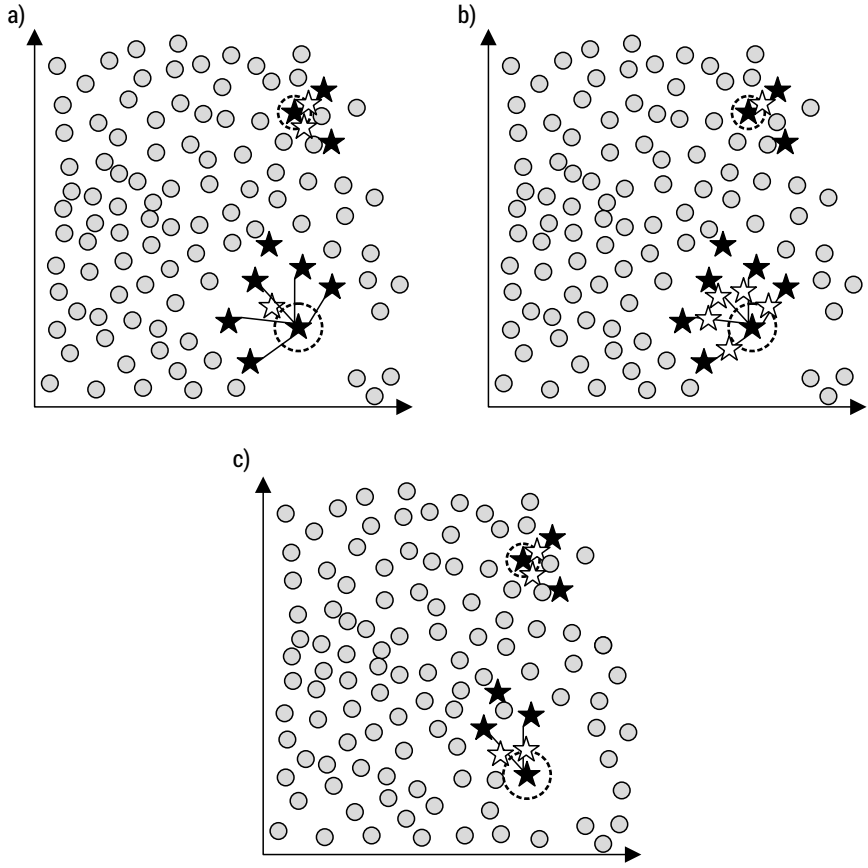
Wykonanie kroku formowania granul informacyjnych, skupiających próbki z klasy mniejszościowej otoczone podobnymi obiektami z obydwu klas, pozwala na podział problemu na mniejsze zadania [21]. Dzięki temu granule mogą zostać przetworzone w sposób adekwatny do specyfiki danych, które zawierają. Proces ten stanowi kluczowy etap działania algorytmu, identyfikujący złożoność rozkładu danych i determinujący mechanizm postępowania, który powinien zostać zastosowany wobec danych określonego typu. Tym samym uzyskiwana jest elastyczność, pozwalająca na skuteczną pracę ze zbiorami danych o różnej charakterystyce.

Proces dobierania odpowiedniej metody przetwarzania do typu granuli obejmuje ustalenie, w jakim stopniu określona granula zawiera się w granuli definiującej całą klasę mniejszościową. W rezultacie każdej z granul przydzielana jest etykieta wyznaczająca dalszy proces postępowania. Poniższe punkty zawierają opis poszczególnych etykiet:

- SAFE – wysoki stopień zawierania granuli informacyjnej oznacza, że znajduje się ona w obszarze jednolicie zajmowanym przez obiekty z klasy mniejszościowej. Obszar taki można uznać za bezpieczny – algorytm klasyfikacji nie powinien napotkać trudności w rozpoznawaniu próbek tego typu. Wskaźnikiem umożliwiającym przydzielenie granuli tej etykiety jest proporcja zawierających się w niej obserwacji z dwóch klas: jeśli większość stanowią przykłady z klasy mniejszościowej i jednocześnie próbka położona najbliżej centralnego punktu granuli należy do klasy mniejszościowej, granula otrzymuje etykietę SAFE.
- BOUNDARY – niski stopień zawierania jest wynikiem znacznej reprezentacji próbek z klasy większościowej w danej granuli. Tak określona granula znajduje się w mniej jednorodnym obszarze, stanowiącym granicę między klasami w przestrzeni cech. Występowanie granul tego rodzaju oznacza, że próbki z obydwu klas są wymieszane. Doprecyzowując, granula jest oznaczana jako BOUNDARY, gdy przynajmniej połowa jej elementów reprezentuje klasę większościową lub próbka położona najbliżej elementu centralnego granuli należy do klasy większościowej.
- NOISE – zerowy stopień zawierania granuli zachodzi w sytuacji, gdy żaden z obiektów tworzących granulę, z wykluczeniem centralnego przykładu, nie należy do klasy mniejszościowej. Jako że jedynie centralna próbka reprezentuje klasę mniejszościową, należy ją rozpatrywać w kategoriach bardzo rzadkiego przypadku. Otoczenie sąsiadującymi obserwacjami z klasy większościowej będzie

przysparzało znacznych trudności w procesie uczenia standardowymi klasyfikatorami. Granula oznaczona tą etykietą niesie informacje, które charakteryzują przede wszystkim klasę większościową.

Wymienione etykiety odzwierciedlają, w jakim stopniu dana granula definiuje specyficzne cechy klasy mniejszościowej, jak bardzo są one charakterystyczne dla tej klasy i tym samym, czy będzie powodowała trudności w procesie uczenia. Na podstawie tak dokonanego podziału granul podejmowana jest decyzja odnośnie trybu dalszego przetwarzania danych. W zależności od przyjętego trybu działania, algorytm generuje w określonych obszarach przestrzeni cech odpowiednią liczbę syntetycznych próbek (rysunek 1.1).



RYSUNEK 1.1. Centralne punkty granul zostały oznaczone linią przerywaną. Wygenerowane przykłady z klasy mniejszościowej oznaczono znakiem gwiazdy bez wypełnienia. Koła z szarym wypełnieniem reprezentują próbki z klasy większościowej. Proces generowania syntetycznych próbek w zależności od przyjętego trybu działania: (a) tryb *LowComplexity*, (b) tryb *HighComplexity*, (c) tryb *noSAFE*

Wybór sposobu przetwarzania wstępnego danych niezbalansowanych odbywa się poprzez weryfikację wartości parametrycznej definiującej procentowy udział granul oznaczonych etykietą BOUNDARY w zbiorze wszystkich granul stworzonych wokół obiektów z klasy mniejszościowej. Wartość progowa określa, z jakim typem problemu będzie trzeba się zmierzyć w procesie przygotowania danych do uczenia maszynowego. Eksperymenty przeprowadzone w [12] udowadniają, że co najmniej 30% przykładów z klasy mniejszościowej będących na granicy klas (gdzie próbki są wymieszane) może doprowadzić do znacznego pogorszenia skuteczności klasyfikacji. Algorytm RGA umożliwia określanie wartości progowej jako parametru: *complexity_th*.

Detekcja niewielu granul typu BOUNDARY w zbiorze danych umożliwia zastosowanie trybu *LowComplexity*. Dane tego typu niosą dużo informacji charakterystycznych dla danej klasy, pozwalając odróżniać od siebie obiekty z poszczególnych klas. Przykłady z klasy mniejszościowej są przeważnie otoczone próbkami reprezentującymi tę samą klasę. Stąd można wnioskować, że granule te położone są w relatywnie jednorodnych obszarach. Poniższy wzór definiuje kryterium uznania zbioru danych za odpowiedni do przetwarzania w tym trybie:

$$\frac{card\left(\left\{x \in X_{d=+} : Label(x) = BOUNDARY\right\}\right)}{card\left(X_{d=+}\right)} < complexity\_th, \tag{1.2}$$

gdzie: $card(U)$ to liczba elementów zbioru U, x stanowi pojedynczy obiekt, $X_{d=+}$ oznacza przykłady z klasy mniejszościowej, zaś $Label(x)$ jest etykietą przypisaną obiektowi x.

Po ustaleniu metody działania algorytmu jako *LowComplexity*, przetwarzanie wstępne granul poszczególnych rodzajów odbywa się w dostosowany do ich właściwości sposób. Dokładny opis tego mechanizmu znajduje się poniżej:

- $Label(x) = SAFE$: w większości występujące granule oznaczone jako bezpieczne nie wymagają znacznego wsparcia nowymi syntetycznymi próbkami – tworzone jest wyłącznie po jednej próbce na granulę posiadającą tę etykietę. Nowy obiekt opisany jest przez zbiór cech stanowiących kombinację właściwości rozpatrywanego centralnego przykładu danej granuli oraz najbliższego sąsiada.
- $Label(x) = BOUNDARY$: większość syntetycznych próbek generowana jest w tych granicznych obszarach. Dzięki temu można uzyskać wzmocnienie roli obserwacji z klasy mniejszościowej w tym trudnym rejonie przestrzeni cech. Jako że nie dysponujemy wieloma przypadkami tego typu, utworzenie wielu egzemplarzy instancji z klasy mniejszościowej w obszarze nakładania się klas pozwoli klasyfikatorowi na wygenerowanie poprawnych reguł. Nowe próbki tworzone są w bliższej odległości od centralnego przykładu danej granuli, aby zapobiegać zbytniemu zwiększaniu złożoności rozkładu przez nadmierne wymieszanie klas.
- $Label(x) = NOISE$: nowe przykłady nie są generowane.

Przewaga granul oznaczonych jako BOUNDARY, czyli:

$$\frac{card\left(\left\{x \in X_{d=+} : Label(x) = BOUNDARY\right\}\right)}{card\left(X_{d=+}\right)} \geq complexity\_th, \tag{1.3}$$

implikuje konieczność przyjęcia bardziej restrykcyjnych reguł generowania syntetycznych próbek. Wiąże się to z potencjalnymi znacznymi komplikacjami w procesie uczenia w wyniku wysokiej złożoności problemu. Przyjmowany jest więc tryb *HighComplexity*, który zakłada dalsze przetwarzanie danych w następujący sposób:

- $Label(x) = SAFE$: większość nowych próbek generowana jest w tym uznanym za bezpieczny obszarze. Zakładając, że granule tego typu skupiają przykłady posiadające specyficzne dla klasy mniejszościowej cechy i stanowiące proste do wyuczenia wzorce, utworzenie głównej puli syntetycznych przykładów właśnie w tych rejonach nie spowoduje wzrostu złożoności problemu, a jednocześnie umożliwi klasyfikatorom lepsze rozeznanie odnośnie właściwości charakteryzujących klasę mniejszościową.
- $Label(x) = BOUNDARY$: liczba obserwacji tego typu jest podwajana poprzez tworzenie pojedynczych nowych przykładów, wykorzystując kombinację cech centralnej próbki granuli oraz jej najbliżej położonego sąsiada – przy czym nowy przykład będzie wykazywał większe podobieństwo do centralnej próbki granuli. Tworzenie wielu nowych obiektów w obszarze granicznym w trybie *HighComplexity* mogłoby doprowadzić do nadmiernego efektu nakładania się danych różnych klas. Między dwoma obiektami z klasy mniejszościowej może występować obiekt z klasy większościowej. Stąd nowa próbka mogłaby być zbyt podobna do reprezentanta klasy negatywnej. Aby zapobiec takiej sytuacji, w tym niejednoznacznym obszarze generowane są wyłącznie pojedyncze nowe przykłady z klasy mniejszościowej.
- $Label(x) = NOISE$: nowe przykłady nie są generowane.

Ostatni przypadek dotyczy szczególnej sytuacji, w której żadne granule typu bezpiecznego (SAFE) nie zostały rozpoznane:

$$\left\{x \in X_{d=+} : Label(x) = SAFE\right\} = \varnothing, \tag{1.4}$$

Wybierany jest wtedy tryb *noSAFE*. Określa on dane o szczególnie dużej złożoności. Brak bezpiecznych obszarów oznacza bardzo wysoki poziom wymieszania danych pochodzących z różnych klas. Stosowana jest w takim wypadku dedykowana metoda przetwarzania, zgodna z opisem poniżej:

- $Label(x) = BOUNDARY$: wszystkie syntetyczne obiekty z klasy mniejszościowej generowane są w obszarach granicznych. Nowe przykłady nie mogą być tworzone w bezpiecznych rejonach, gdyż takie nie istnieją. W tej szczególnej sytuacji jedynie wzmocnienie reprezentacji próbek z klasy mniejszościowej w obszarach wymieszania klas może poprawić skuteczność klasyfikacji.
- $Label(x) = NOISE$: nowe przykłady nie są generowane.

Granule typu NOISE zostały wykluczone z przetwarzania, gdyż nie ma pewności odnośnie przyczyny ich występowania. Mogą stanowić zarówno błędnie zapisane dane, jak i niespotykane przypadki. Zakładając możliwość prawdziwości każdej z tych potencjalnych przyczyn, tego typu próbki nie powinny być usuwane, ale jednocześnie nie warto poddawać ich przetwarzaniu, aby nie doprowadziły do nadmiernego zwiększenia złożoności rozkładu, wprowadzając szum.

1.3.3. Filtrowanie danych

W celu zapewnienia jak najwyższej jakości tworzonych obiektów klasy mniejszościowej, wprowadzony został etap usuwania niejednoznaczności. Po wygenerowaniu nowych próbek weryfikowana jest ich przynależność do dolnej aproksymacji zbioru zawierającego elementy z klasy mniejszościowej. Jeśli nowy obiekt nie należy do dolnej aproksymacji, jest usuwany ze zbioru syntetycznych próbek.

Przyjmując U jako uniwersum, relacja nierozróżnialności $IND \subseteq U \times U$ w teorii zbiorów przybliżonych identyfikuje przykłady, które są opisywane identycznymi informacjami [20, 27]. Jednym z głównych pojęć opartych na tym twierdzeniu jest dolna aproksymacja zbioru X (stanowiącego dowolny podzbiór uniwersum U):

$$\left\{x \in U : [x]_{IND} \subseteq X\right\}, \tag{1.5}$$

Jest to zbiór wszystkich próbek, które mogą być jednoznacznie sklasyfikowane jako należące do X w odniesieniu do relacji nierozróżnialności IND [19].

Możliwość przetwarzania danych o wartościach typu ciągłego zapewniona została poprzez zastosowanie pojęcia podobieństwa obiektów jako odległości między próbkami w przestrzeni cech w ramach relacji tolerancji uogólnionej teorii zbiorów przybliżonych [25, 27]. Modyfikacja ta wymagała ustanowienia określonej wartości progowej, która jest swoistą granicą między podobnymi przykładami a pozostałymi obserwacjami. Próg podobieństwa (*distance_th*) jest parametrem aplikacji. Jest to wartość procentowa średniej odległości HVDM między obiektami z różnych klas w przetwarzanym zbiorze danych. Po wykonaniu kroku filtrowania, dane mogą zostać poddane procesowi klasyfikacji.

Przewidzenie jak wiele z wygenerowanych próbek zostanie usuniętych nie jest możliwe. Zaistnienie pesymistycznego scenariusza mogłoby oznaczać, że nadmiernie dużo syntetycznych obiektów będzie wykluczonych z dalszego przetwarzania w etapie filtrowania i w związku z tym całkowita liczebność klasy mniejszościowej nie zostanie odpowiednio zwiększona. Mając na uwadze istotny cel wstępnego przetwarzania danych niezbalansowanych – a jest nim próba zrównania liczebności przykładów z obydwu klas – wprowadzono dodatkowe zabezpieczenie w postaci określanej jako parametr nadmiarowości w tworzeniu nowych próbek (*card_redundancy*). Parametr ten jest wartością procentową, wykorzystywaną do wyliczania całkowitej liczby wygenerowanych obiektów zgodnie z następującym wzorem:

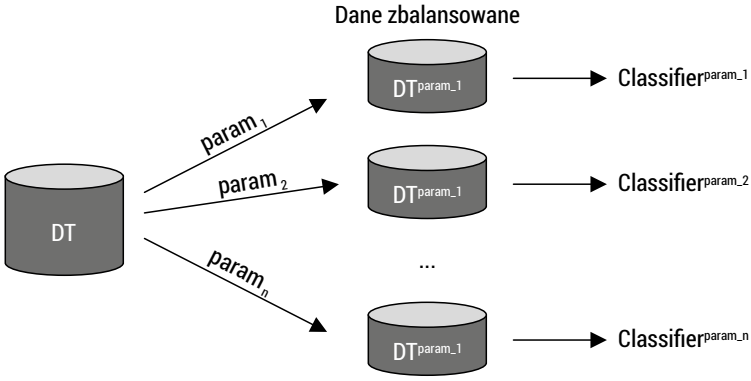
$$generated\_samples = card\_redundancy * (N^- - N^+), \tag{1.6}$$

gdzie N^- oznacza liczbę obiektów z klasy większościowej, zaś N^+ oznacza liczbę obiektów z klasy mniejszościowej.

1.3.4. Dobieranie parametrów

Opisany algorytm posiada kilka szczególnie istotnych parametrów, których modyfikacja może skutkować lepszym dopasowaniem mechanizmu działania do konkretnego rozkładu danych. Parametry te są następujące:

- *k* – liczba najbliższych sąsiadów,
- *complexity_th* – próg złożoności,
- *distance_th* – próg podobieństwa,
- *card_reundancy* – liczba dodatkowych obiektów.



RYSUNEK 1.2.: **Metoda automatycznej ewaluacji parametrów**

Aby uprościć proces doboru wymienionych parametrów, została opracowana metoda automatycznej ewaluacji (rysunek 1.2). Na początku przygotowywane są kombinacje możliwych wartości, jakie może przyjąć każdy z parametrów. W wyniku uzyskiwane są skończone zbiory potencjalnych wartości. Następnie algorytm przetwarzania danych rozpoczyna działanie wykorzystując pierwszą kombinację przygotowanych parametrów. Przetwarzanie powtarzane jest wielokrotnie, za każdym razem z użyciem innej puli wartości parametrów. W rezultacie budowane są modele klasyfikatorów korzystające z oddzielnych zbiorów danych, stanowiących oryginalny zbiór danych wzbogacony o wygenerowane według odpowiednio sparametryzowanego algorytmu RGA syntetyczne próbki z klasy mniejszościowej. Następnie uzyskane wyniki są poddawane weryfikacji tak, aby wybrać najlepszy z nich.

1.4. Eksperymenty

Przedstawiona metoda została przetestowana na piętnastu rzeczywistych zbiorach danych. Charakterystyka poszczególnych zbiorów zaprezentowana jest w tabeli 1.1. Wszystkie dane zostały pobrane z repozytorium UCI [7]. Cechują się różnymi poziomami niezbalansowania liczebności próbek z poszczególnych klas (na co wskazuje parametr IR) oraz w większości wysoką złożonością (kolumna „Obszar graniczny").

TABELA 1.1. Charakterystyka wykorzystanych zbiorów danych

Nazwa	Liczba próbek	Liczba atrybutów	IR	Obszar graniczny
ecoli-0-1 vs 5	240	6	11,00	niepusty
ecoli-0-1 vs 2-3-5	244	7	9,17	niepusty
ecoli-0-1-4-6 vs 5	280	6	13,00	niepusty
ecoli-0-1-4-7 vs 5-6	332	6	12,28	niepusty
ecoli-0-3-4-7 vs 5-6	257	7	9,28	pusty
ecoli-0-4-6 vs 5	203	6	9,15	niepusty
ecoli-0-6-7 vs 3-5	222	7	9,09	niepusty
ecoli-0-2-6-7 vs 3-5	224	7	9,18	niepusty
glass-0-1-6 vs 5	184	9	19,44	pusty
glass-0-4 vs 5	92	9	9,22	pusty
glass-0-6 vs 5	108	9	11,00	pusty
glass5	214	9	22,78	pusty
led7digit-0-2-4-5-6-7-8-9 vs 1	443	7	10,97	niepusty
page-blocks-1-3 vs 4	472	10	15,86	niepusty
yeast-0-3-5-9 vs 7-8	506	8	9,12	niepusty

Każdy ze zbiorów został podzielony na mniejsze partycje, aby umożliwić pięciokrotną walidację krzyżową. Wykorzystane w eksperymentach zbiory często występują w publikacjach o tematyce danych niezbalansowanych, co usprawnia proces porównywania wyników między różnymi metodami.

W tabeli 1.2 zaprezentowano otrzymane w trakcie eksperymentów wyniki. Algorytm RGA porównano z sześcioma podobnymi technikami przetwarzania wstępnego. Dodatkowo przedstawione zostały wyniki klasyfikacji oryginalnych zbiorów danych, bez etapu przetwarzania (kolumna noPRE). Konfiguracja algorytmu RGA pozwoliła na przetestowanie 81 kombinacji wartości parametrów. Liczba najbliższych sąsiadów *k* przyjmowała wartości 3, 5 oraz 7. Wartości *complexity_th* to 0,2, 0,3 i 0,4. Parametr *distance_th* przyjmował wartości 10, 15, 25. Z kolei *card_redundancy*

miało przypisane wartości 20, 30 oraz 40. Szczegółowe wyniki klasyfikacji w zależności od wybranych dwóch spośród wszystkich opisanych parametrów, a mianowicie parametru *k* i *complexity_th*, można znaleźć w [3].

Analiza rezultatów ujawnia, że metoda RGA osiągała największe wartości miary AUC w większości przypadków. Oznacza to, że algorytm spełnia swoje zadanie: sprawdza się w różnych okolicznościach, dostosowując swoje działanie do złożoności problemu. Opisywana metoda RGA okazała się skuteczniejsza od innych w przypadku siedmiu zbiorów danych. Eksperymenty na czterech spośród wszystkich zbiorów danych skutkowały dzieleniem najlepszego wyniku przez algorytm RGA z innymi metodami o tych samych wartościach AUC. Rezultaty osiągnięte w ramach badań na trzech zbiorach danych były nieznacznie lepsze w przypadku dwóch innych algorytmów (Safe-Level SMOTE oraz *SMOTE RSB*$_{\star}$). W tabeli 1.3 przedstawiono podsumowanie porównania. Zawiera sumaryczne liczby przypadków, w których każdy z algorytmów osiągnął najwyższy wynik. Pierwsza wartość oznacza liczbę zwycięstw w sensie absolutnym (gdy żadna inna metoda nie osiągnęła takiego samego wyniku), druga zaś wskazuje na liczbę sytuacji, w których więcej niż jedna metoda miała taki sam najlepszy wynik.

TABELA 1.2. **Wyniki klasyfikacji – porównanie metody RGA z sześcioma innymi algorytmami oraz klasyfikacją z pominięciem kroku przetwarzania wstępnego**

Zbiór danych	noPRE	SMOTE	S-ENN	Border-S	SafeL-S	S-RSB	VISROT	RGA
ecoli-0-1 vs 5	0,8159	0,7977	0,8250	0,8318	0,8568	0,7818	0,8636	**0,8886**
ecoli-0-1 vs 2-3-5	0,7136	0,8377	0,8332	0,7377	0,7550	0,7777	0,7314	**0,8673**
ecoli-0-1-4-6 vs 5	0,7885	**0,8981**	**0,8981**	0,7558	0,8519	0,8231	0,8366	0,8654
ecoli-0-1-4-7 vs 5-6	0,8318	0,8592	0,8424	0,8420	0,8197	0,8670	0,8220	**0,8853**
ecoli-0-3-4-7 vs 5-6	0,7757	0,8568	0,8546	0,8427	0,7995	0,8984	0,8471	**0,9113**
ecoli-0-4-6 vs 5	0,8168	0,8701	0,8869	0,8615	0,8923	**0,9476**	0,8060	0,8891
ecoli-0-6-7 vs 3-5	0,8250	0,8500	0,8125	0,8550	0,7950	0,8525	0,8500	**0,8700**
ecoli-0-2-6-7 vs 3-5	0,7752	0,8155	0,8179	0,8352	**0,8380**	0,8227	0,7977	0,8327
glass-0-1-6 vs 5	**0,8943**	0,8129	0,8743	0,8386	0,8429	0,8800	**0,8943**	**0,8943**
glass-0-4 vs 5	**0,9941**	0,9816	0,9754	**0,9941**	0,9261	**0,9941**	**0,9941**	**0,9941**
glass-0-6 vs 5	**0,9950**	0,9147	0,9647	**0,9950**	0,9137	0,9650	**0,9950**	**0,9950**
glass5	0,8976	0,8829	0,7756	0,8854	0,8939	0,9232	0,9951	**0,9976**
led7digit-0-2-4-5-6-7-8-9 vs 1	0,8788	0,8908	0,8379	0,8908	0,9023	0,9019	0,8918	**0,9056**
page-blocks-1-3 vs 4	**0,9978**	0,9955	0,9888	0,9978	0,9831	**0,9978**	0,9944	**0,9978**
yeast-0-3-5-9 vs 7-8	0,5868	0,7047	0,7024	0,6228	0,7296	**0,7400**	0,6868	0,7105

TABELA 1.3. Podsumowanie wyników klasyfikacji dla poszczególnych metod: liczba pełnych/współdzielonych zwycięstw

	noPRE	SMOTE	S-ENN	Border-S	SafeL-S	S-RSB	VISROT	RGA	ŁĄCZNIE
Test	0/4	0/1	0/1	0/1	1/0	2/2	0/3	7/4	10/5

W tabeli 1.4 przedstawiono wyniki testu Friedmana, wskazujące na skuteczność algorytmów we wszystkich zbiorach danych, opierając się na wartościach AUC. Metoda RGA, wykazując się bardzo stabilnym poziomem skuteczności, osiągnęła w rankingu pierwszą pozycję. Drugie miejsce uzyskała metoda *SMOTE RSB*$_*$ (również stosująca dodatkowy etap filtrowania danych). Na trzeciej pozycji pod względem całkowitej skuteczności uplasowała się metoda VISROT. Najgorsze wyniki spośród technik przetwarzania wstępnego osiągnął algorytm SMOTE-ENN.

TABELA 1.4. Wyniki testu Friedmana

Algorytm	Średnia ranga
RGA	1,93
S_RSB	3,43
VISROT	4,73
Border_S	4,83
SMOTE	4,97
SafeL_S	5,13
S_ENN	5,3
noPRE	5,67

Zgodnie z przewidywaniem oraz kolejny raz potwierdzając wcześniejsze analizy, klasyfikacja danych niezbalansowanych standardowymi klasyfikatorami z pominięciem kroku wstępnego przetwarzania skutkuje najniższą efektywnością. Dokładniejsza weryfikacja pokazuje, że brak czynności podjętych w celu zrównania liczebności próbek z poszczególnych klas może sprawdzać się jedynie w przypadku zbiorów o niedużej złożoności. Widać to szczególnie w przypadku zbiorów danych glass-0-1-6 vs 5, glass-0-4 vs 5, glass-0-6 vs 5, które posiadają pusty obszar graniczny (wskazujący na niższą złożoność rozkładu), gdy wiele algorytmów (w tym brak użycia jakiejkolwiek metody przetwarzania wstępnego) uzyskało dobre wyniki. Niemniej jednak pozostałe zbiory danych, z których większość posiada skomplikowaną charakterystykę, pozwoliły wykazać skuteczność stosowania przetwarzania wstępnego, w szczególności metodą RGA.

Podsumowanie

W niniejszej pracy opisano metodę bazującą na obliczeniach granularnych oraz teorii zbiorów przybliżonych wzbogacając klasyczne podejście znane z technik rozszerzających algorytm SMOTE o możliwość adaptacji do złożoności problemu. Główne cele powstania metody RGA można podsumować następująco:

- zwiększenie liczebności próbek z klasy mniejszościowej ukierunkowane na umożliwienie jak najwyższego poziomu rozpoznawalności przez standardowe klasyfikatory;
- uwzględnienie dodatkowych trudności, pojawiających się w trakcie eksploracji niezbalansowanych zbiorów danych poprzez zastosowanie granul informacyjnych;
- wprowadzenie kroku filtrowania bazującego na teorii zbiorów przybliżonych w celu usunięcia wszelkich niejednoznaczności, które mogłyby przyczynić się do zwiększenia złożoności rozkładu danych;
- usprawnienie procesu doboru parametrów do konkretnej charakterystyki danych.

Rezultaty przeprowadzonych eksperymentów potwierdziły poprawność przyjętych założeń. Połączenie techniki ukierunkowanego generowania przykładów z klasy mniejszościowej z dodatkowym krokiem filtrowania danych okazało się skutecznym rozwiązaniem. Badania zdecydowanie ukazały wartościowość podejścia opartego na uwzględnianiu lokalnej charakterystyki zbiorów danych w przetwarzaniu wstępnym. W szczególności obiecujące jest automatyczne dobieranie istotnych parametrów algorytmu, pozwalające na lepsze dostosowanie podejmowanych kroków przetwarzania do konkretnego problemu.

W ramach kolejnych eksperymentów może zostać przeprowadzona weryfikacja wpływu liczby atrybutów na skuteczność proponowanego rozwiązania. Równie istotna będzie analiza złożoności metody RGA w zależności od rozmiaru danych, uwzględniając także w porównaniu złożoność alternatywnych algorytmów.

Bibliografia

[1] Batista G.E.A.P.A., Prati R.C., Monard M.C., *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*, „Association for Computing Machinery" 2004, 6 (1)

[2] Borowska K., Stepaniuk J., A rough-granular approach to the imbalanced data classification problem, „Applied Soft Computing" 2019

[3] Borowska K., Stepaniuk J., Granular Computing and Parameters Tuning in Imbalanced Data Preprocessing [w:] K. Saeed, W. Homenda (red.), Computer Information Systems and Industrial Management. CISIM 2018, Lecture Notes in Computer Science 11127, Springer, 2018, s. 233–245

[4] Borowska K., Stepaniuk J., Rough Sets in Imbalanced Data Problem: Improving Re–sampling Process [w:] K. Saeed, W. Homenda, R. Chaki (red.), Computer Information Systems and Industrial Management. CISIM 2017, Lecture Notes in Computer Science 10244, Springer, 2017, s. 459–469

[5] Bunkhumpornpat C., Sinapiromsaran K., Lursinsap C., Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem [w:] T. Theeramunkon, B. Kijsirikul, N. Cercone, TB. Ho (red.), Advances in Knowledge Discovery and Data Mining. PAKDD 2009, Lecture Notes in Computer Science 5476, Springer, 2009, s. 475–482

[6] Chawla N., Bowyer W.K., Hall L.O., Kegelmeyer W.P., *SMOTE: Synthetic Minority Oversampling Technique*, „AI Access Foundation" 2002, 16 (1)

[7] *Data Sets – UCI Machine Learning Repository,* strona internetowa, https://archive.ics.uci.edu/ml/datasets.php [dostęp: 27.02.2021]

[8] Fernadez A., Herrera F., Chawla N.V., *SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary*, „Journal of Artificial Intelligence Research" 2018, 61

[9] Galar M., Fernandez A., Barrenechea E., Bustince H., Herrera F., *A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches*, „IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)" 2012, 42 (4)

[10] Garcia S., Luengo J., Herrera F., *Data Preprocessing in Data Mining*, „Springer International Publishing" 2015, 72

[11] Han H., Wang WY., Mao BH., *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning* [w:] DS. Huang, XP. Zhang, GB. Huang (red.), *Advances in Intelligent Computing.* ICIC 2005. Lecture Notes in Computer Science 3644, Springer, 2005, s. 878–887

[12] He H., Garcia E.A., *Learning from Imbalanced Data*, „IEEE Transactions on Knowledge and Data Engineering" 2009, 21 (9)

[13] Hu S., Liang Y., Ma L., He Y., *MSMOTE: Improving Classification Performance When Training Data is Imbalanced*, Second International Workshop on Computer Science and Engineering, 2009, s. 13–17

[14] Jo T., Japkowicz N., *Class Imbalances Versus Small Disjuncts*, „Association for Computing Machinery" 2004, 6 (1)

[15] Krawczyk B., *Learning from imbalanced data: open challenges and future directions*, „Progress in Artificial Intelligence" 2016, 5 (4)

[16] Lopez V., Fernandez A., Garcia S., Palade V., Herrera F., An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, „Information Sciences" 2013, 250

[17] Napierala K., Stefanowski J., Types of minority class examples and their influence on learning classifiers from imbalanced data, „Journal of Intelligent Information Systems" 2016, 46 (3)

[18] Napierała K., Stefanowski J., Wilk S., *Learning from Imbalanced Data in Presence of Noisy and Borderline Examples* [w:] M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen, Q. Hu (red.), *Rough Sets and Current Trends in Computing.* RSCTC 2010. Lecture Notes in Computer Science 6086, Springer, 2010, s. 158–167

[19] Pawlak Z., *Rough sets*, „International Journal of Computer & Information Sciences" 1982, 11 (5)

[20] Pawlak Z., Skowron A., *Rudiments of Rough Sets*, „Information Sciences" 2007, 177 (1)

[21] Pedrycz W., *Granular Computing: An Introduction* [w:] N. Kasabov (red.), *Future Directions for Intelligent Systems and Information Sciences*. Studies in Fuzziness and Soft Computing 45, Physica, s. 309–328

[22] Pietruszkiewicz W., *Practical Evaluation, Issues and Enhancements of Applied Data Mining* [w:] A. Abd Manaf, A. Zeki, M. Zamani, S. Chuprat, E. El-Qawasmeh (red.), *Informatics Engineering and Information Science*. ICIEIS 2011, Communications in Computer and Information Science 252, Springer, 2011, s. 717–731

[23] Ramentol E., Caballero Y., Bello R., Herrera F., SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, „Knowledge and Information Systems" 2012, 33 (2)

[24] Saez J.A., Luengo J., Stefanowski J., Herrera F., SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, „Information Sciences" 2015, 291

[25] Skowron A., Stepaniuk J., *Tolerance Approximation Spaces*, „IOS Press" 1996, 27 (2–3)

[26] Stefanowski J., *Dealing with Data Difficulty Factors While Learning from Imbalanced Data* [w:] S. Matwin, J. Mielniczuk (red.), *Challenges in Computational Statistics and Data Mining*. Studies in Computational Intelligence 605, Springer, 2015, s. 333–363

[27] Stepaniuk J., *Rough–Granular Computing in Knowledge Discovery and Data Mining*, „Studies in Computational Intelligence" 152, Springer, 2009

[28] Sun Y., Kamel M.S., Wong A.K.C., Wang Y., *Cost-sensitive boosting for classification of imbalanced data*, „Pattern Recognition" 2007, 40 (12)

[29] Weiss G.M., *Mining with Rarity: A Unifying Framework*, „Association for Computing Machinery" 2004, 6 (1)

[30] Wilson D, Martinez T.R., *Improved heterogeneous distance functions*, „Journal of Artificial Intelligence Research" 1997, 6 (1)

Granular computing in imbalanced data preprocessing

Abstract: Imbalanced data problem occurs when the numer of examples representing a class of major interest is significantly lower than the number of instances from another class. It has been one of the most interesting and important data mining problems for more than two decades. Recent studies have shown that the issue of imbalanced data should not be considered only in the context of information deficit of one of the classes. In fact, the decrease in the quality of standard classifiers when confronted with this type of datasets is mainly related to the high complexity of the data distribution. The complexity increases when additional difficulties such as minority class decomposition, class overlapping and the presence of noise occur in the data characteristics. Therefore dedicated solutions are required, in particular taking into account the local characteristics of the distribution. This paper describes an algorithm that applies information granules and rough sets theory for targeted modification of the learning set and enables to automate parameter selection.

Keywords: class imbalance, data preprocessing, granular computing, information granules, rough sets, SMOTE