Szymon Wojciechowski, Szymon Wilk

Generator sztucznych danych wielowymiarowych: weryfikacja eksperymentalna (Raport Badawczy RB-2/15)

Instytut Informatyki, Politechnika Poznańska, 2015

1. Wprowadzenie

W raporcie opisano weryfikację eksperymentalną generatora sztucznych danych opisanego w pracy [5]. Cele tej weryfikacji były następujące:

1. ocena poprawności struktury i układów zbiorów stworzonych za pomocą generatora,
2. ocena poprawności wyników eksperymentu obliczeniowego przeprowadzonego na stworzonych zbiorach danych.

Aby zrealizować pierwszy cel, wybrano cztery zbiory danych z pracy [1], stworzono dla nich pliki konfiguracyjne, wygenerowano na ich podstawie zbiory w formacie ARFF i wreszcie porównano je ze zbiorami uzyskanymi przy użyciu poprzedniej wersji generatora. W porównaniu wzięto pod uwagę zarówno strukturę uzyskanych zbiorów (np. bezwzględny rozkład obiektów w skupieniach tworzących poszczególne klasy), jak i układy i kształty poszczególnych zbiorów. Wyniki tej fazy weryfikacji opisane są w rozdziale 2.

W celu realizacji drugiego celu wykonano eksperyment obliczeniowy, w którym do wcześniej rozważanych zbiorów stopniowo wprowadzano zaburzenia w klasie mniejszościowej (zwiększając udział obiektów typu borderline, rare oraz outlier), na następnie na zaburzonych zbiorach przetestowano różne zestawy metod wstępnego przetwarzania oraz klasyfikatorów. Dobór tych metod oraz klasyfikatorów został zainspirowany wcześniej przeprowadzonymi pracami (m.in., [2, 3, 4]). Eksperyment oraz jego wyniki przedstawiono w rozdziale 3.

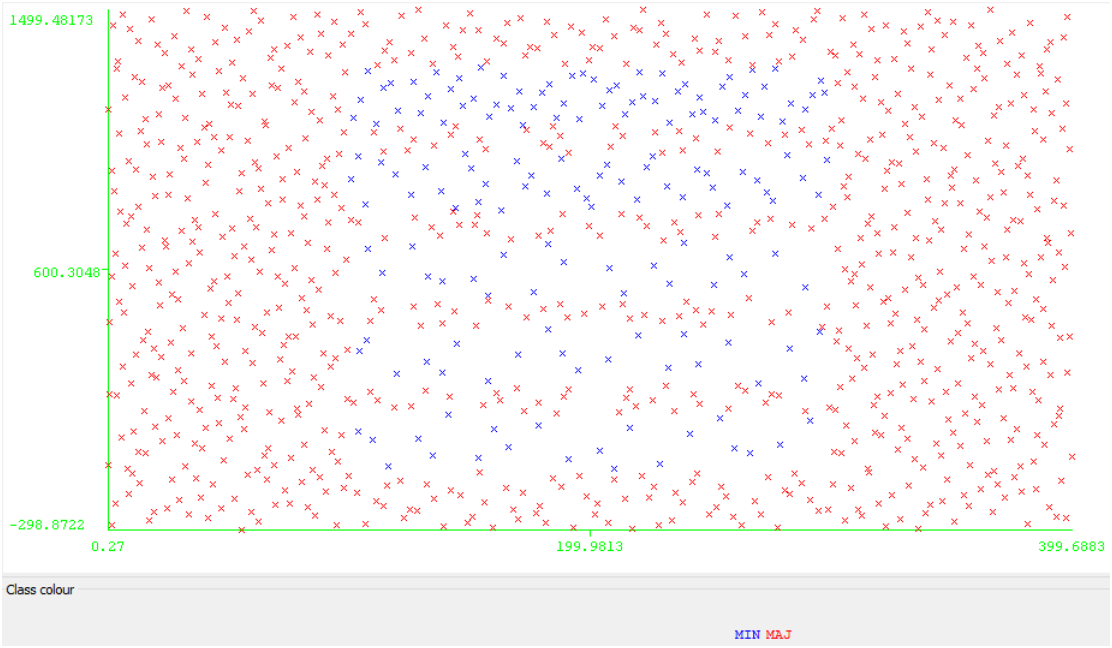
2. Rozważane zbiory danych

Na podstawie podstawie analizy wyników z pracy [1] wybrano cztery reprezentatywne zbiory danych (*02a*, *02b*, *03subcls5* oraz *04clover5*). Z uwagi na kształty i układy skupień tworzących klasę mniejszościową prezentują one różne poziomy trudności dla metod uczących. We wszystkich tych zbiorach obiekty opisane są za pomocą dwóch atrybutów warunkowych (X i Y, odpowiadających współrzędnym na płaszczyźnie) i należą do jednej z dwóch klas (MIN lub MAJ, gdzie pierwsza klasa to klasa mniejszościowa, a druga – większościowa) Krótka charakterystyka poszczególnych zbiorów (od najłatwiejszego do najtrudniejszego) przedstawiona jest poniżej.

Zbiór *03subcls5* przedstawiony jest na rysunku 1, a wykorzystana konfiguracja dla generatora danych znajduje się na listingu 1. W celu bardziej wyraźniej wizualizacji kształtów skupień z klasy mniejszościowej ten zbiór i kolejne zawierają 1000 obiektów i wygenerowano je dla stopnia niezrównoważenia równego 1:5. Poza tym na tym etapie do zbiorów nie wprowadzano zakłóceń (wszystkie obiekty w klasie mniejszościowej są typu safe). W zbiorze *03subcls5* klasa mniejszościowa składa się z 5 prostokątnych (łatwo separowalnych) skupień, przez co zbiór ten powinien sprawiać najmniejszy problem klasyfikatorom, zwłaszcza tak, jak drzewa i reguły decyzyjne.

Listing 1. Plik konfiguracyjny dla zbioru *03subcls5*

```
# 03subcl5
attributes = 2
classes = 2
names.attributes = X, Y
names.decision = CLASS
names.classes = MIN, MAJ
```



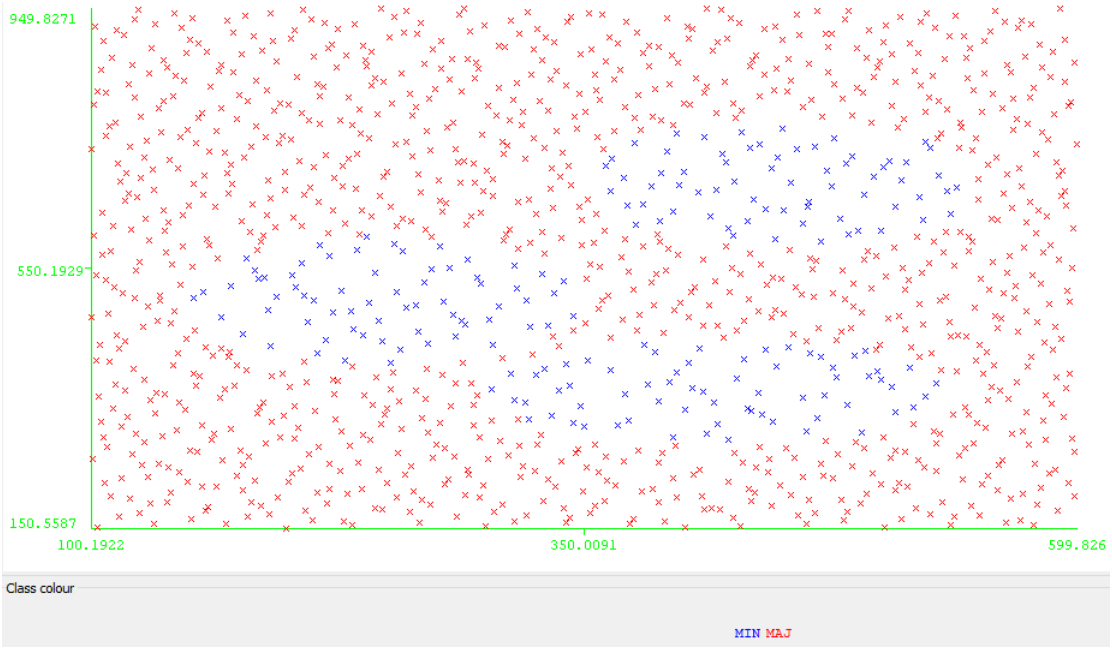
Rysunek 1. Wizualizacja zbioru *03subcl5*

```
classRatio = 1:5
examples=1000
minOutlierDistance = 40
defaultRegion.weight = 1
defaultRegion.distribution = U
defaultRegion.borderZone = 40
defaultRegion.noOutlierZone = 40
defaultRegion.shape = R
defaultRegion.radius = 100, 100
class.1.regions = 5
class.1.exampleTypeRatio = 50:30:20:10
class.1.region.1.center = 200, 0
class.1.region.2.center = 200, 300
class.1.region.3.center = 200, 600
class.1.region.4.weight = 2
class.1.region.4.center = 200, 900
class.1.region.5.weight = 3
class.1.region.5.center = 200, 1200
class.2.regions = 1
class.2.region.1.shape = I
class.2.region.1.center = 200, 600
class.2.region.1.radius = 200, 900
fileName=03subcl5.arff
```

Zbiór *02a* przedstawiony jest na rysunku 2, a użyta do jego wygenerowania konfiguracja na listingu 2. Zbiór ten stanowi nieco większą trudność dla klasyfikatorów niż *03subcl5* – klasa mniejszościowa składa się z 3 eliptycznych skupień (żadne z nich nie jest obrócone). O ile taki zbiór może stanowić trudność dla reguł i drzew, o tyle klasyfikator KNN powinien sobie z nim poradzić.

Listing 2. Plik konfiguracyjny dla zbioru *02a*

```
# 02a
attributes = 2
classes = 2
names.attributes = X, Y
names.decision = CLASS
names.classes = MIN, MAJ
classRatio = 1:5
```



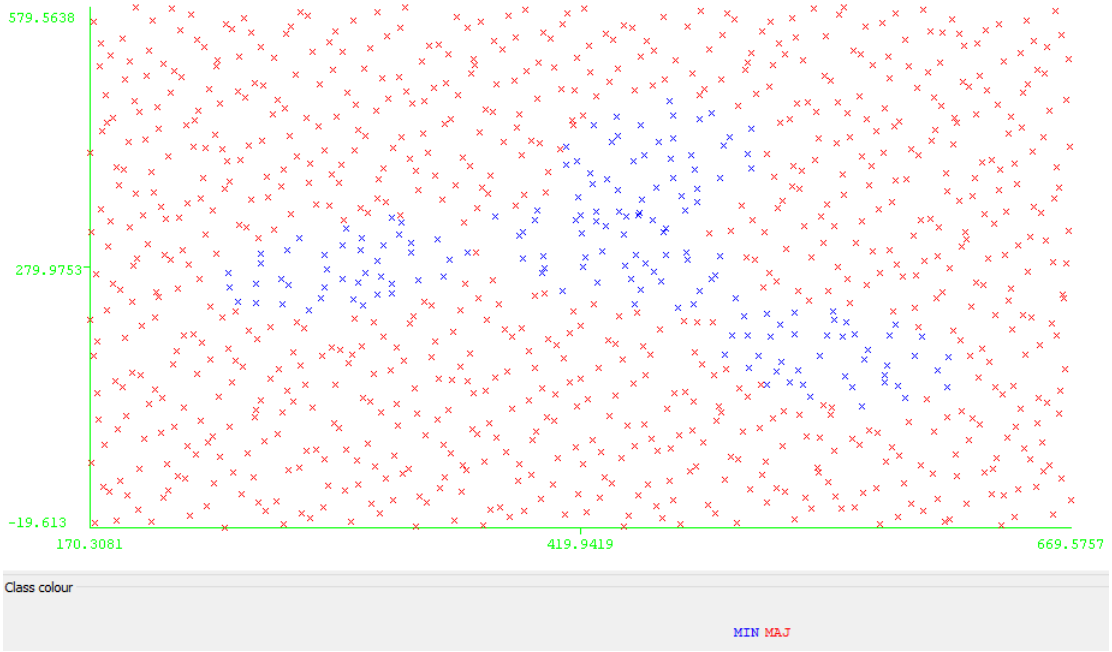
Rysunek 2. Wizualizacja zbioru *02a*

```
examples=1200
minOutlierDistance = 40
defaultRegion.weight = 1
defaultRegion.distribution = U
defaultRegion.borderZone = 40
defaultRegion.noOutlierZone = 40
defaultRegion.shape = C
defaultRegion.radius = 100, 100
class.1.regions = 3
class.1.region.weight = 4
class.1.region.1.center = 420, 360
class.1.region.1.radius = 120, 80
class.1.region.weight = 3
class.1.region.2.center = 450, 680
class.1.region.weight = 1
class.1.region.3.center = 250, 500
class.2.regions = 1
class.2.region.1.shape = I
class.2.region.1.center = 350,550
class.2.region.1.radius = 250, 400
fileName=02a.arff
```

Zbiór *02b* przedstawiony jest na rysunku 3, a użyta do jego wygenerowania konfiguracja na listingu 2. Tutaj klasa mniejszościowa składa się z 4 poobracanych eliptycznych skupień, z których część nachodzi na siebie – przez to poziom trudności rośnie.

Listing 3. Plik konfiguracyjny dla zbioru *02b*

```
# 02b
attributes = 2
classes = 2
names.attributes = X, Y
names.decision = CLASS
names.classes = MIN, MAJ
classRatio = 1:5
examples=1000
minOutlierDistance = 40
defaultRegion.weight = 1
```



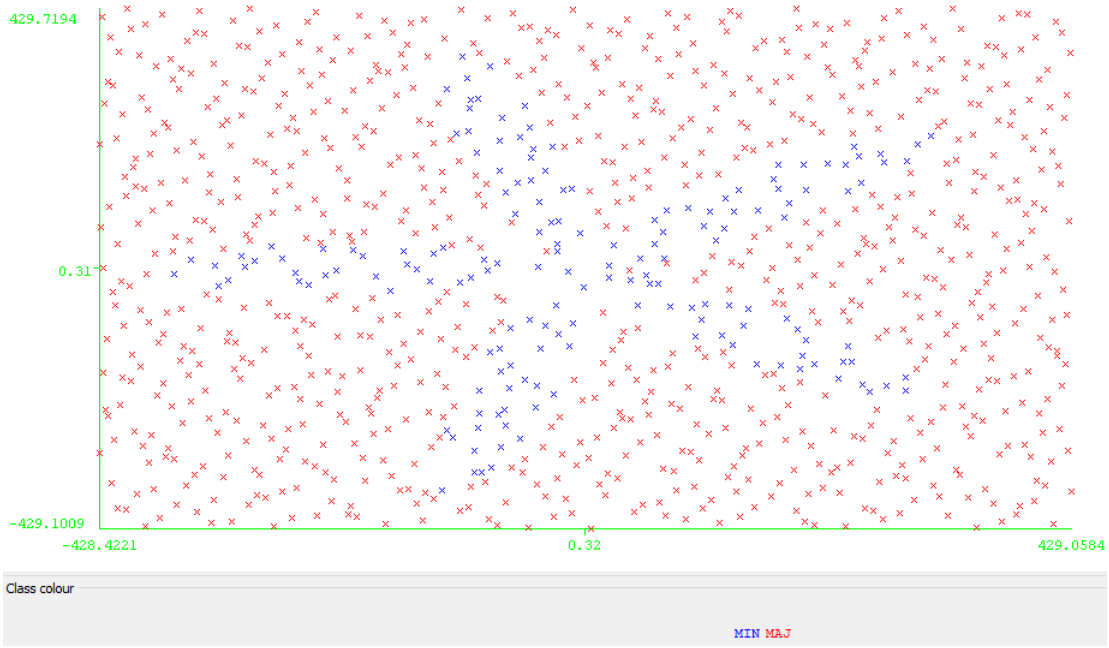
Rysunek 3. Wizualizacja zbioru *02b*

```
defaultRegion.distribution = U
defaultRegion.borderZone = 40
defaultRegion.noOutlierZone = 40
defaultRegion.shape = C
defaultRegion.radius = 100, 100
class.1.regions = 4
class.1.region.1.center = 440, 290
class.1.region.1.radius = 50, 80
class.1.region.1.rotation = 1, 2, 50
class.1.region.2.center = 300, 280
class.1.region.2.radius = 70, 50
class.1.region.2.rotation = 1, 2, 35
class.1.region.3.center = 460, 400
class.1.region.3.radius = 50, 80
class.1.region.3.rotation = 1, 2, -10
class.1.region.4.center = 550, 180
class.1.region.4.radius = 70, 50
class.1.region.4.rotation = 1, 2, -45
class.2.regions = 1
class.2.region.1.shape = I
class.2.region.1.center = 420, 280
class.2.region.1.radius = 250, 300
fileName=02b.arff
```

Wreszcie zbiór *04clover5* pokazany jest na rysunku 4, a jego konfiguracja znajduje się na listingu 4. W tym zbiorze klasa mniejszościowa przypomina „kwiat” (dlatego też zbiór ten w części prac, np. [3], oznaczony jest jako *flower*) z pięcioma wąskimi płatkami, nachylonymi pod różnym kątem. Taki układ i kształty stanowią największe wyzwanie dla metod uczących spośród rozważanych w tym opracowaniu zbiorów.

Listing 4. Plik konfiguracyjny dla zbioru *02b*

```
# 04clover5
attributes = 2
classes = 2
names.attributes = X, Y
names.decision = CLASS
names.classes = MIN, MAJ
```



Rysunek 4. Wizualizacja zbioru *04clover5*

```
classRatio = 1:5
examples=1000
minOutlierDistance = 40
defaultRegion.weight = 1
defaultRegion.distribution = U
defaultRegion.borderZone = 10
defaultRegion.noOutlierZone = 10
defaultRegion.shape = C
defaultRegion.radius = 190, 40
class.1.regions = 5
class.1.region.1.center = 162, 118
class.1.region.1.rotation = 1, 2, 36
class.1.region.2.center = -62, 190
class.1.region.2.rotation = 1, 2, 108
class.1.region.3.center = -200, 0
class.1.region.3.rotation = 1, 2, -180
class.1.region.4.center = -62, -190
class.1.region.4.rotation = 1, 2, -108
class.1.region.5.center = 162, -118
class.1.region.5.rotation = 1, 2, -36
class.2.regions = 1
class.2.region.1.shape = I
class.2.region.1.center = 0, 0
class.2.region.1.radius = 430, 430
fileName=04clover5.arff
```

Wszystkie opisane w tym rozdziale zbiory porównano ze zbiorami uzyskanymi za pomocą wcześniejszej wersji generatora. Stwierdzono zgodność pomiędzy nimi, zarówno na poziomie struktury zbiorów, jak i układu skupień, potwierdzając tym samym poprawne działanie nowej wersji generatora.

3. Eksperyment obliczeniowy

3.1. Przebieg eksperymentu

W eksperymencie wykorzystano zbiory danych opisane w poprzednim rozdziale, przy czym zwiększono poziom niezrównoważenia z 1:5 do 1:9 oraz zwiększono liczbę wygenerowanych obiektów z 1000

Tablica 1. Zakłócenia w zbiorach danych

Safe [%]	Boderline [%]	Rare [%]	Outlier [%]
90	10	0	0
70	30	0	0
60	30	0	10
60	20	10	10
50	30	10	10
40	40	10	10
30	50	10	10
20	60	10	10
10	60	20	10
10	50	20	20
10	40	20	30
10	30	30	30
0	40	30	30
0	30	30	40
0	20	30	50
0	10	30	60
0	0	30	70
0	0	10	90

do 1200. Poza tym w systematyczny sposób do poszczególnych zbiorów danych wprowadzano zakłócenia polegające na zwiększaniu udziału obiektów borderline, rare oraz outlier w klasie mniejszościowej – kolejne poziomy zakłóceń przedstawione są w tabeli 1. W ten sposób stworzono 18 wariantów każdego zbioru danych odpowiadających poszczególnym poziomom zakłóceń, przy czym wyniki są przedstawione tylko dla wybranych (reprezentatywnych) poziomów, zaznaczonych w tabeli 1 dodatkowym obramowaniem.

Do oceny metod wstępnego przetwarzania oraz klasyfikatorów wykorzystano schemat warstwowej walidacji skrośnej (*stratified cross validation*) z liczbą podziałów równą 10, przy czym obliczenia zostały powtórzone 5 razy w celu ograniczenia zmienności wyników. W opisywanym eksperymencie wykorzystano następujące metody wstępnego przetwarzania (w nawiasach podane są oznaczenia wykorzystywane podczas prezentacji wyników):

— brak wstępnego przetwarzania (none) – wyniki uzyskane w tym wypadku stanowią punkt odniesienia (*baseline*) dla pozostałych metod wstępnego przetwarzania,
— *random undersampling* (RU),
— *random oversampling* (RO),
— *neighborhood cleaning rule* (NCR),
— SMOTE (SM),
— SPIDER2 (SP2).

Metody RU, RO oraz SMOTE zostały sparametryzowane w taki sposób, aby równoważyć klasy oraz dla ostatniej metody przyjęto $k = 5$. Natomiast w przypadku metody SP2 zastosowano silne wzmocnienie oraz zmianę etykiet dla obiektów z klasy większościowej.

Wymienione powyżej metody wstępnego przetwarzania zostały połączone z następującymi klasyfikatorami (podobnie jak poprzednio w nawiasach podane są oznaczenia wykorzystywane podczas prezentacji wyników) zaimplementowanymi w systemie WEKA:

— klasyfikator KNN z liczbą najbliższych sąsiadów równą 1 i 3 (odpowiednio 1NN i 3NN),
— reguły decyzyjne wygenerowane za pomocą algorytmu PART (PART),
— drzewa decyzyjne wygenerowane za pomocą algorytmu C4.5/J48 (J48),
— naiwny klasyfikator Bayesa (NB).

W przypadku klasyfikatorów PART i J48 zrezygnowano z przycinania (*pruning*) z uwagi na wcześniejsze wyniki [4] pokazujące, że w przypadku zastosowania metod wstępnego przetwarzania staje się ono zbędne. Poza tym, w porównaniu z wcześniej cytowaną pracą [4] zrezygnowano z sieci neuronowych RBF oraz z maszyn wektorów wspierających SVM z uwagi na konieczność strojenia ich parametrów w celu uzyskania zadowalających wyników.

Do oceny poszczególnych kombinacji metod wstępnego przetwarzania i klasyfikatorów wykorzystano miary czułości (*sensitivity*), swoistości (*specificity*) oraz ich średniej geometrycznej (*geometric mean*) oznaczonej dalej jako GM. Wartości poszczególnych miar zostały uzyskane poprzez uśrednienie wyników uzyskanych dla poszczególnych podziałów w walidacji.

3.2. Wyniki eksperymentu

Szczegółowe wyniki eksperymentu zamieszczono w tabelach od 2 do 9. W szczególności dla każdego zbioru zaprezentowano wartości miary GM, swoistości, aby zapewnić szczegółowy wgląd w działanie różnych kombinacji metod wstępnego przetwarzania i klasyfikatorów na zbiorach danych o różnych poziomach zakłócenia (zrezygnowano z prezentacji czułości, ponieważ zachowanie tej miary pokrywa się w dużej mierze z miarą GM). We wszystkich tych tabelach w kolumnie „zakłócenie" przedstawiono udziały procentowe poszczególnych typów obiektów w klasie mniejszościowej (safe:borderline:rare:outlier) – odpowiada on zaznaczonym poziomom w tabeli 1.

Obserwacja wyników dla miary GM (tabele 2–6) pozwala na sformułowanie następujących wniosków:

— zwiększanie poziomy zakłóceń w klasie mniejszościowej powoduje obniżenie zaobserwowanej wartości GM. Jest to szczególnie widoczne w przypadku braku zastosowania jakichkolwiek metod wstępnego przetwarzania, gdzie przy dużym rozproszeniu klasy mniejszościowej tylko klasyfikator 1NN (i w niektórych przypadkach 3NN) jest w stanie uzyskać niezerową wartość GM,

— wyniki potwierdziły wstępne przypuszczenia co do trudności poszczególnych zbiorów danych – zbiór *03subcl5* okazał się najłatwiejszy ze względu na separowalność liniową skupień i najlepsze wyniki uzyskano dla klasyfikatorów PART i J48. Najtrudniejszym okazał się zbiór *04clover5*, gdzie najlepiej zachowywały się klasyfikatory 1NN oraz 3NN. Wartości miary GM pokazały też, że pomimo prostszej struktury (brak obrotów skupień) zbiór *02a* okazał się trudniejszy niż *02b* – w przypadku tego ostatniego dobre wyniki uzyskały klasyfikatory 1NN, 3NN, PART oraz J48, co wskazuje na możliwość stosunkowo łatwego odseparowania poszczególnych skupień klasy mniejszościowej od klasy większościowej,

— wyniki jednoznacznie wskazały na sensowność zastosowania metod wstępnego przetwarzania – poprawa wartości miary GM jeszcze szczególnie widoczna dla mocniejszych zakłóceń. W szczególności, zastosowanie metod RO, SM oraz SP2 wiązało się z bardzo dużą poprawa wyników (w porównaniu do wyników referencyjnych),

— nieco zaskakujące wydaje się poprawna wyników na ostatnim poziomie zakłóceń (0:0:10:90) przy zastosowaniu metod wstępnego przetwarzania. Dla tego wariantu zbiorów większość obiektów z klasy mniejszościowej jest rozproszona i wrzucona do klasy większościowej. Metody wstępnego przetwarzania wzmocniły te obiekty (lub ich sąsiedztwo) oraz oczyściły ich otoczenie, co mogło doprowadzić do powstania wielu małych skupień dobrze „wyłapujących" obiekty testowe.

Wyniki uzyskane dla miary swoistości (przedstawione w tabelach 6– 9) są uzupełnieniem wyników dla miary GM. Na ich podstawie można sformułować następujące wnioski:

— wartości swoistości w przypadku braku wstępnego przetwarzania były stabilne, a zwiększanie stopnia powodowało ich ograniczoną poprawę (i dojście do poziomu 1.0),

— wyniki potwierdziły wcześniejsze wnioski dotyczące trudności poszczególnych zbiorów oraz dopasowania rozważanych klasyfikatorów do charakterystyki do poszczególnych zbiorów danych,

— zastosowanie metod wstępnego przetwarzania wiązało się z obniżenie swoistości – najmniejsza zmiana występowała dla tych metod, dla których zaobserwowano małą poprawę miary GM lub brak takiej poprawy (np. dla NCR). Metody, które poprawiały GM, prowadziły także do spadków wartości swoistości – w szczególności warto podkreślić dobre zachowanie metod SM, SP2 oraz RO (dla wszystkich zbiorów poza *04clover5*).

Zaprezentowane wyniki rozszerzają rezultaty zaprezentowane w pracach [3, 1]. W części wspólnej (początkowe poziomy zakłóceń) są one spójne, co potwierdza poprawność działania generatora danych. Natomiast nowe wyniki dla silniejszych zakłóceń dają lepszy wgląd w zachowanie przetestowanych metod wstępnego przetwarzania i klasyfikatorów.

4. Podsumowanie i dyskusja

W niniejszym opracowaniu opisaną eksperymentalną weryfikację nowego generatora danych. Uzyskane w obu etapach wyniki potwierdzają jego poprawne działanie. W ramach dalszych prac planowane jest jego wykorzystanie do wygenerowania złożonych kształtów wielowymiarowych (5-10 wymiarów) oraz wykorzystanie tych danych w rozszerzonym eksperymencie obliczeniowym, stanowiącym znaczące rozwinięcie pracy [3].

Planowany jest również dalszy rozwój generatora, a w szczególności jego dostosowanie do pracy w środowisku MOA i generowanie niezrównoważonych strumieni danych, w których poziom niezrównoważenia oraz poziom zakłóceń będą zmieniały się w czasie.

Tablica 2. Wartości miary GM uzyskane na zbiorze *03subcl5*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.6636	0.7914	0.6636	0.7174	0.8123	0.6626
	3NN	0.5689	0.8174	0.7761	0.6711	0.8349	0.7215
	PART	0.9069	0.8613	0.8797	0.9038	0.9023	0.8772
	J48	0.9084	0.8751	0.8989	0.9039	0.9100	0.8966
	NB	0.0000	0.7676	0.7712	0.0000	0.7645	0.1851
50:30:10:10							
	1NN	0.4454	0.5713	0.4454	0.5150	0.5820	0.4451
	3NN	0.3238	0.6127	0.5663	0.4474	0.5749	0.5470
	PART	0.4133	0.6340	0.5627	0.4758	0.6093	0.6111
	J48	0.4083	0.6389	0.6041	0.4828	0.6554	0.6198
	NB	0.0000	0.6416	0.6475	0.0000	0.6651	0.3576
10:50:20:20							
	1NN	0.2536	0.4795	0.2536	0.3622	0.3980	0.2529
	3NN	0.0000	0.5045	0.3957	0.1414	0.4335	0.3959
	PART	0.0000	0.1753	0.4901	0.0000	0.5202	0.4978
	J48	0.0000	0.1816	0.4505	0.0000	0.5238	0.4937
	NB	0.0000	0.5475	0.5611	0.0000	0.5722	0.4710
0:30:30:40							
	1NN	0.4139	0.5008	0.4139	0.4832	0.4744	0.4119
	3NN	0.0284	0.5054	0.4706	0.2331	0.4495	0.4744
	PART	0.0000	0.2295	0.5457	0.0057	0.5508	0.5548
	J48	0.0000	0.2318	0.5063	0.0057	0.5543	0.5386
	NB	0.0000	0.5394	0.5268	0.0000	0.5204	0.3709
0:0:10:90							
	1NN	0.1194	0.6366	0.1194	0.3532	0.5265	0.1291
	3NN	0.0562	0.6813	0.4110	0.3489	0.5655	0.4034
	PART	0.0000	0.8140	0.8077	0.6714	0.7929	0.8169
	J48	0.0000	0.8053	0.7630	0.6293	0.7880	0.8035
	NB	0.0000	0.6137	0.6587	0.0000	0.6540	0.1048

Tablica 3. Wartości miary GM uzyskane na zbiorze *02a*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.7942	0.8543	0.7942	0.8160	0.8606	0.7920
	3NN	0.7746	0.8768	0.8569	0.8308	0.8718	0.8248
	PART	0.2889	0.8038	0.8263	0.4379	0.8446	0.8128
	J48	0.5165	0.8142	0.8331	0.6206	0.8664	0.8660
	NB	0.0000	0.7640	0.7683	0.0000	0.7696	0.0000
50:30:10:10							
	1NN	0.5775	0.6492	0.5775	0.6393	0.6229	0.5781
	3NN	0.5030	0.6551	0.6698	0.5949	0.6297	0.6313
	PART	0.0000	0.5609	0.6319	0.0484	0.6815	0.5198
	J48	0.0000	0.5615	0.6097	0.1349	0.6574	0.6249
	NB	0.0000	0.6772	0.6835	0.0000	0.6846	0.0000
10:50:20:20							
	1NN	0.3146	0.5081	0.3146	0.4316	0.4147	0.3189
	3NN	0.0362	0.4775	0.4512	0.2367	0.4353	0.4515
	PART	0.0000	0.0645	0.4681	0.0000	0.5124	0.4077
	J48	0.0000	0.0645	0.3550	0.0000	0.4951	0.4656
	NB	0.0000	0.5286	0.5421	0.0000	0.5550	0.2289
0:30:30:40							
	1NN	0.3404	0.4529	0.3404	0.4285	0.3705	0.3436
	3NN	0.0000	0.4743	0.4157	0.1325	0.3703	0.4161
	PART	0.0000	0.0142	0.3198	0.0000	0.3700	0.3346
	J48	0.0000	0.0142	0.3581	0.0000	0.3546	0.3745
	NB	0.0000	0.4932	0.4952	0.0000	0.4953	0.2888
0:0:10:90							
	1NN	0.1851	0.5539	0.1851	0.2112	0.3929	0.1844
	3NN	0.0000	0.5953	0.2428	0.1858	0.3839	0.2429
	PART	0.0000	0.3763	0.6138	0.0000	0.6296	0.6044
	J48	0.0000	0.3898	0.5575	0.0000	0.6225	0.6066
	NB	0.0000	0.6086	0.6105	0.0000	0.6027	0.4317

Tablica 4. Wartości miary GM uzyskane na zbiorze *02b*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.8496	0.8987	0.8496	0.8726	0.8785	0.8486
	3NN	0.8568	0.9071	0.8999	0.8785	0.9095	0.8712
	PART	0.8344	0.8082	0.8918	0.8575	0.8868	0.8923
	J48	0.8693	0.8195	0.8885	0.8745	0.9000	0.8849
	NB	0.0000	0.7825	0.7820	0.0000	0.7775	0.0000
50:30:10:10							
	1NN	0.6607	0.7022	0.6607	0.7194	0.6849	0.6669
	3NN	0.5971	0.7000	0.7436	0.6697	0.6953	0.7198
	PART	0.1494	0.6131	0.6762	0.5532	0.6959	0.5649
	J48	0.1649	0.6214	0.6594	0.5598	0.6946	0.7041
	NB	0.0000	0.6740	0.6817	0.0000	0.6815	0.0000
10:50:20:20							
	1NN	0.4465	0.5324	0.4465	0.4993	0.5095	0.4486
	3NN	0.2314	0.5332	0.5042	0.3718	0.5018	0.5038
	PART	0.0000	0.1149	0.4726	0.0264	0.4945	0.3614
	J48	0.0000	0.1147	0.4198	0.0356	0.5017	0.4924
	NB	0.0000	0.5551	0.5769	0.0000	0.5822	0.0000
0:30:30:40							
	1NN	0.4015	0.4853	0.4015	0.4927	0.4291	0.4235
	3NN	0.0846	0.4736	0.4806	0.2283	0.4528	0.4809
	PART	0.0000	0.0000	0.2196	0.0000	0.1880	0.0977
	J48	0.0000	0.0000	0.3369	0.0000	0.2692	0.1253
	NB	0.0000	0.4637	0.4658	0.0000	0.4537	0.0791
0:0:10:90							
	1NN	0.0971	0.4995	0.0971	0.1927	0.2473	0.0970
	3NN	0.0000	0.5422	0.2041	0.0637	0.3633	0.2042
	PART	0.0000	0.2195	0.5538	0.0000	0.5724	0.5618
	J48	0.0000	0.2262	0.5403	0.0000	0.5803	0.5553
	NB	0.0000	0.5868	0.5929	0.0000	0.5940	0.5003

Tablica 5. Wartości miary GM uzyskane na zbiorze *04clover5*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.7453	0.8582	0.7453	0.8156	0.8401	0.7610
	3NN	0.7015	0.8554	0.8856	0.8275	0.8765	0.8130
	PART	0.3323	0.7878	0.7968	0.6157	0.8267	0.6808
	J48	0.3078	0.8026	0.7909	0.6408	0.8460	0.7286
	NB	0.0000	0.7428	0.7401	0.0000	0.7403	0.0000
50:30:10:10							
	1NN	0.5948	0.7129	0.5948	0.7057	0.7065	0.6197
	3NN	0.4604	0.7440	0.7414	0.6475	0.7319	0.7122
	PART	0.0728	0.6215	0.6794	0.3001	0.7313	0.6527
	J48	0.1075	0.6355	0.6912	0.3687	0.7400	0.6925
	NB	0.0000	0.6619	0.6686	0.0000	0.6638	0.0000
10:50:20:20							
	1NN	0.3564	0.5708	0.3564	0.5860	0.4728	0.3773
	3NN	0.1314	0.5504	0.5948	0.4398	0.5477	0.5889
	PART	0.0000	0.1285	0.5229	0.0869	0.5705	0.4672
	J48	0.0000	0.1285	0.4903	0.0468	0.5889	0.4993
	NB	0.0000	0.5839	0.6043	0.0000	0.6183	0.0000
0:30:30:40							
	1NN	0.3905	0.4694	0.3905	0.4771	0.4057	0.3886
	3NN	0.0000	0.4370	0.4785	0.0575	0.4057	0.4785
	PART	0.0000	0.0075	0.2348	0.0000	0.2909	0.0150
	J48	0.0000	0.0075	0.2845	0.0000	0.3027	0.0287
	NB	0.0000	0.4904	0.5047	0.0000	0.4980	0.0057
0:0:10:90							
	1NN	0.0802	0.3262	0.0802	0.1502	0.2002	0.0802
	3NN	0.0000	0.3991	0.1703	0.0166	0.1610	0.1703
	PART	0.0000	0.0000	0.0480	0.0000	0.0493	0.0033
	J48	0.0000	0.0000	0.0682	0.0000	0.0505	0.0033
	NB	0.0000	0.4629	0.4545	0.0000	0.4256	0.2940

Tablica 6. Wartości swoistości uzyskane na zbiorze *03subcl5*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.9706	0.7811	0.9706	0.9233	0.9120	0.9676
	3NN	0.9907	0.7787	0.9104	0.9457	0.9067	0.9187
	PART	0.9965	0.8176	0.9787	0.9852	0.9739	0.9683
	J48	0.9974	0.8448	0.9809	0.9854	0.9694	0.9759
	NB	1.0000	0.7689	0.7720	1.0000	0.7948	0.9846
50:30:10:10							
	1NN	0.9233	0.5963	0.9233	0.8389	0.6800	0.9165
	3NN	0.9867	0.6467	0.7983	0.9191	0.7048	0.7989
	PART	0.9937	0.6561	0.7600	0.9705	0.8878	0.7294
	J48	0.9918	0.6550	0.9017	0.9589	0.8611	0.8926
	NB	1.0000	0.6156	0.6309	1.0000	0.6756	0.9291
10:50:20:20							
	1NN	0.8756	0.5111	0.8756	0.7796	0.5391	0.8700
	3NN	0.9719	0.5372	0.7150	0.8837	0.5833	0.7157
	PART	1.0000	0.1569	0.4667	1.0000	0.4637	0.7311
	J48	1.0000	0.1654	0.7357	1.0000	0.5426	0.6913
	NB	1.0000	0.5541	0.5935	1.0000	0.6163	0.7989
0:30:30:40							
	1NN	0.8872	0.4941	0.8872	0.7902	0.5383	0.8778
	3NN	0.9587	0.5165	0.7139	0.8856	0.5433	0.7211
	PART	1.0000	0.3931	0.6083	0.9967	0.7543	0.6663
	J48	1.0000	0.3909	0.7106	0.9956	0.7444	0.6720
	NB	1.0000	0.6087	0.6567	1.0000	0.6609	0.8930
0:0:10:90							
	1NN	0.8707	0.6422	0.8707	0.7706	0.7081	0.8535
	3NN	0.9539	0.6082	0.7387	0.8046	0.6891	0.7467
	PART	1.0000	0.6856	0.6943	0.8070	0.6850	0.6930
	J48	1.0000	0.6870	0.7137	0.8189	0.7048	0.7031
	NB	1.0000	0.6341	0.6574	1.0000	0.6537	0.9946

Tablica 7. Wartości swoistości uzyskane na zbiorze *02a*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.9743	0.8363	0.9743	0.9439	0.9365	0.9691
	3NN	0.9950	0.8272	0.9274	0.9700	0.9318	0.9322
	PART	0.9943	0.6880	0.8785	0.9755	0.8544	0.9035
	J48	0.9861	0.7002	0.9502	0.9672	0.9424	0.9624
	NB	1.0000	0.7326	0.7365	1.0000	0.7446	1.0000
50:30:10:10							
	1NN	0.9246	0.6341	0.9246	0.8576	0.7211	0.9218
	3NN	0.9841	0.6904	0.8170	0.9415	0.7441	0.8178
	PART	1.0000	0.5537	0.6817	0.9952	0.7359	0.6235
	J48	1.0000	0.5509	0.8502	0.9868	0.7920	0.8119
	NB	1.0000	0.6424	0.6541	1.0000	0.6826	1.0000
10:50:20:20							
	1NN	0.8894	0.5124	0.8894	0.8043	0.5587	0.8872
	3NN	0.9706	0.5367	0.7380	0.9094	0.5902	0.7393
	PART	1.0000	0.0511	0.4763	1.0000	0.5650	0.5709
	J48	1.0000	0.0511	0.7737	1.0000	0.6002	0.6693
	NB	1.0000	0.5033	0.5322	1.0000	0.5626	0.9365
0:30:30:40							
	1NN	0.8561	0.4580	0.8561	0.7463	0.5213	0.8506
	3NN	0.9624	0.5013	0.6852	0.8878	0.5561	0.6865
	PART	1.0000	0.0254	0.5783	1.0000	0.5648	0.7902
	J48	1.0000	0.0254	0.6507	1.0000	0.6187	0.7726
	NB	1.0000	0.5204	0.4652	1.0000	0.4304	0.8987
0:0:10:90							
	1NN	0.8595	0.5565	0.8595	0.7256	0.6204	0.8513
	3NN	0.9676	0.5493	0.6755	0.8289	0.6109	0.6768
	PART	1.0000	0.3022	0.5124	0.9985	0.4731	0.4491
	J48	1.0000	0.3228	0.6030	0.9985	0.5022	0.5298
	NB	1.0000	0.5672	0.5804	1.0000	0.5720	0.8565

Tablica 8. Wartości swoistości GM uzyskane na zbiorze *02b*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.9813	0.8733	0.9813	0.9607	0.9517	0.9791
	3NN	0.9902	0.8569	0.9411	0.9720	0.9511	0.9504
	PART	0.9750	0.7259	0.9585	0.9585	0.9368	0.9302
	J48	0.9722	0.7372	0.9646	0.9613	0.9489	0.9505
	NB	1.0000	0.7707	0.7744	1.0000	0.7807	1.0000
50:30:10:10							
	1NN	0.9411	0.6656	0.9411	0.8835	0.7613	0.9326
	3NN	0.9735	0.7098	0.8443	0.9385	0.7915	0.8478
	PART	0.9957	0.6163	0.7811	0.9274	0.8189	0.7491
	J48	0.9963	0.6046	0.9037	0.9405	0.8531	0.8574
	NB	1.0000	0.6359	0.6517	1.0000	0.6852	1.0000
10:50:20:20							
	1NN	0.8965	0.5404	0.8965	0.8159	0.6272	0.8931
	3NN	0.9693	0.5630	0.7363	0.8963	0.6445	0.7383
	PART	1.0000	0.1004	0.6098	0.9922	0.7641	0.6437
	J48	1.0000	0.0998	0.8394	0.9896	0.7837	0.7183
	NB	1.0000	0.5150	0.5663	1.0000	0.5843	0.9993
0:30:30:40							
	1NN	0.8822	0.4902	0.8822	0.7880	0.5969	0.8765
	3NN	0.9678	0.5113	0.7037	0.9098	0.5817	0.7048
	PART	1.0000	0.0359	0.3993	1.0000	0.3206	0.9328
	J48	1.0000	0.0359	0.5746	1.0000	0.3185	0.9046
	NB	1.0000	0.4787	0.4841	1.0000	0.4554	0.9652
0:0:10:90							
	1NN	0.8343	0.5369	0.8343	0.7381	0.6007	0.8315
	3NN	0.9791	0.5369	0.6724	0.8419	0.5989	0.6731
	PART	1.0000	0.1994	0.4870	1.0000	0.4582	0.4424
	J48	1.0000	0.1911	0.6002	1.0000	0.4991	0.5691
	NB	1.0000	0.5843	0.5776	1.0000	0.5652	0.8093

Tablica 9. Wartości swoistości uzyskane na zbiorze *04clover5*

Zakłócenie		None	RU	RO	NCR	SM	SP2
90:10:0:0							
	1NN	0.9593	0.8194	0.9593	0.9341	0.9376	0.9480
	3NN	0.9765	0.7711	0.9078	0.9506	0.9248	0.9228
	PART	0.9709	0.6776	0.8576	0.9365	0.8267	0.9269
	J48	0.9965	0.7007	0.9030	0.9317	0.8554	0.8989
	NB	1.0000	0.7370	0.7404	1.0000	0.7437	1.0000
50:30:10:10							
	1NN	0.9420	0.6791	0.9420	0.8722	0.8059	0.9265
	3NN	0.9730	0.7020	0.8328	0.9130	0.8337	0.8363
	PART	0.9931	0.6119	0.7520	0.9639	0.7159	0.6707
	J48	0.9952	0.5980	0.8481	0.9578	0.7489	0.8306
	NB	1.0000	0.6422	0.6596	1.0000	0.6768	1.0000
10:50:20:20							
	1NN	0.8889	0.5657	0.8889	0.8156	0.6491	0.8839
	3NN	0.9589	0.5981	0.7739	0.9039	0.6861	0.7791
	PART	1.0000	0.0944	0.6293	0.9735	0.6665	0.5363
	J48	1.0000	0.0944	0.8100	0.9876	0.6706	0.7609
	NB	1.0000	0.5535	0.5756	1.0000	0.6050	1.0000
0:30:30:40							
	1NN	0.8798	0.4654	0.8798	0.7913	0.5406	0.8700
	3NN	0.9594	0.4919	0.7015	0.9043	0.5513	0.7011
	PART	1.0000	0.0067	0.1894	1.0000	0.1711	0.9756
	J48	1.0000	0.0067	0.4393	1.0000	0.2107	0.9604
	NB	1.0000	0.4620	0.4724	1.0000	0.4654	0.9917
0:0:10:90							
	1NN	0.8537	0.3880	0.8537	0.7076	0.5087	0.8532
	3NN	0.9855	0.4311	0.6200	0.9218	0.4894	0.6202
	PART	1.0000	0.0000	0.1235	1.0000	0.0831	0.9741
	J48	1.0000	0.0000	0.1420	1.0000	0.0894	0.9739
	NB	1.0000	0.5030	0.5359	1.0000	0.5100	0.7411

Podziękowania

Autorzy dziękują za wsparcie udzielone przez Narodowe Centrum Nauki w ramach grantu DEC-2013/-11/B/ST6/00963.

Literatura

[1] K. Kałużny. Metody dekompozycji w analizie niezrównoważonych liczebnie danych. praca magisterska, 2009.
[2] K. Napierała and J. Stefanowski. Types of minority class examples and their influence on learning classifiers from imbalanced data. *J. Intell. Inform. Syst.*, 2015 to appear.
[3] K. Napierała, J. Stefanowski, and Sz. Wilk. Learning from imbalanced data in presence of noisy and borderline examples. In *Proceedings of the 7th International Conference RSCTC 2010*, volume 6086 of *LNAI*, pages 158–167. Springer, 2010.
[4] Sz. Wilk, J. Stefanowski, Sz. Wojciechowski, K.J. Farion, and W. Michalowski. Application of preprocessing techniques to imbalanced clinical data: an experimental study. In *Proceedings of the 5th International Confernece on Information Technologies in Biomedicine, ITIB 2016*, 2016 (accepted).
[5] Sz. Wojciechowski and Sz. Wilk. Generator sztucznych danych wielowymiarowych: projekt i implementacja, 2014.