Autorzy:

Tomasz Szewczuk, 226163
Adam Węglowski, 226175

Projekt z przedmiotu: Obrazowanie Biomedyczne
Prowadzący: dr Paweł Ksieniewicz

Temat: Porównanie wpływu działania SMOTE na wybrane algorytmy

1. Opis problemu badawczego

Badanie miało za zadanie pokazać wpływ działania klasy SMOTE na dokładność klasyfikacji z zastosowaniem wybranych algorytmów tj. SVM (maszyna wektorów nośnych), KNN (K-najbliższych sąsiadów), LR (regresja logistyczna) oraz GNB (Gaussian Naive Bayes). Eksperyment został przeprowadzony na wybranej bazie danych `Yeast` składającej się z 8 klas oraz 1484 instancji.
Pierwszy etap badań polegał na zaimplementowaniu oraz przetestowaniu 4 algorytmów (SVM, KNN, LR oraz GNB) do przeprowadzenia klasyfikacji drożdży. W trakcie drugiego etapu badań została zaimplementowana klasa SMOTE, który miała za zadanie wyrównać ilości instancji w poszczególnych klasach. Następnie po zastosowaniu SMOTE na bazie danych została przeprowadzona ponowna klasyfikacja z zastosowaniem tych samych algorytmów. Ostatni etap polegał na porównaniu otrzymanych wyników oraz podsumowaniu.

2. Zastosowane algorytmy

SVM (maszyna wektorów nośnych) – konstruuje hiperpłaszczyznę w wielowymiarowej przestrzeni w celu oddzielenia różnych klas. SVM generuje optymalną hiperpłaszczyznę w sposób powtarzalny, który jest wykorzystywany do minimalizowania błędu. Podstawową ideą SVM jest znalezienie maksymalnej hiperpłaszczyzny brzegowej (MMH), która najlepiej dzieli zestaw danych na klasy.

KNN – wyznacza k-sąsiadów, do których badany element ma najbliżej dla wybranej metryki (np. euklidesowej), a następnie wyznacza wynik w oparciu o głosy większości sąsiadów.

Regresja logistyczna to algorytm klasyfikacji uczenia maszynowego, który służy do przewidywania prawdopodobieństwa zmiennej zależnej. W regresji logistycznej zmienna zależna jest zmienną binarną, która zawiera dane zakodowane jako 1 (tak, sukces itp.) lub 0 (nie, niepowodzenie itp.). Innymi słowy, model regresji logistycznej przewiduje P (Y = 1) jako funkcję X.

Naiwny klasyfikator bayesowski jest prostym probabilistycznym klasyfikatorem, bazującym na twierdzeniu Bayesa, nadaje się szczególnie do problemów o bardzo wielu wymiarach na wejściu. Mimo prostoty metody, często działa ona lepiej od innych, bardzo skomplikowanych metod klasyfikujących. Naiwne klasyfikatory bayesowskie są oparte na założeniu o wzajemnej niezależności predyktorów. Często nie mają one żadnego związku z rzeczywistością i właśnie z tego powodu nazywa się je naiwnymi.

3. Badania eksperymentalne

Pierwszy etap badań - bez zastosowania klasy SMOTE.

- SVM

	precision	recall	f1-score	support
CYT	0.33	0.84	0.47	44
ERL	0.00	0.00	0.00	1
EXC	0.00	0.00	0.00	2
ME1	0.00	0.00	0.00	5
ME2	0.00	0.00	0.00	7
ME3	1.00	0.04	0.07	26
MIT	0.64	0.58	0.61	24
NUC	0.64	0.23	0.34	39
VAC	0.00	0.00	0.00	1
accuracy			0.41	149
macro avg	0.29	0.19	0.17	149
weighted avg	0.54	0.41	0.34	149

Rysunek 1. Wyniki dla algorytmu SVM (bez klasy SMOTE)

- KNN

	precision	recall	f1-score	support
CYT	0.45	0.59	0.51	44
ERL	1.00	1.00	1.00	1
EXC	0.25	0.50	0.33	2
ME1	0.67	0.80	0.73	5
ME2	0.00	0.00	0.00	7
ME3	0.90	0.73	0.81	26
MIT	0.56	0.58	0.57	24
NUC	0.50	0.41	0.45	39
VAC	0.00	0.00	0.00	1
accuracy			0.54	149
macro avg	0.48	0.51	0.49	149
weighted avg	0.54	0.54	0.54	149

Rysunek 2. Wyniki dla algorytmu KNN (bez klasy SMOTE)

- LR

	precision	recall	f1-score	support
CYT	0.46	0.75	0.57	44
ERL	0.00	0.00	0.00	1
EXC	0.00	0.00	0.00	2
ME1	0.50	0.20	0.29	5
ME2	0.00	0.00	0.00	7
ME3	0.88	0.58	0.70	26
MIT	0.68	0.71	0.69	24
NUC	0.52	0.44	0.47	39
VAC	0.00	0.00	0.00	1
accuracy			0.56	149
macro avg	0.34	0.30	0.30	149
weighted avg	0.55	0.56	0.53	149

Rysunek 3. Wyniki dla algorytmu LR (bez klasy SMOTE)

- GNB

	precision	recall	f1-score	support
CYT	0.00	0.00	0.00	44
ERL	0.50	1.00	0.67	1
EXC	0.15	1.00	0.27	2
ME1	0.29	0.40	0.33	5
ME2	0.00	0.00	0.00	7
ME3	0.00	0.00	0.00	26
MIT	0.80	0.17	0.28	24
NUC	0.58	0.18	0.27	39
VAC	0.01	1.00	0.02	1
accuracy			0.11	149
macro avg	0.26	0.42	0.20	149
weighted avg	0.30	0.11	0.14	149

Rysunek 4. Wyniki dla algorytmu GNB (bez klasy SMOTE)

Drugi etap badań - z zastosowaniem klasy SMOTE.

- SVM

	precision	recall	f1-score	support
CYT	0.52	0.57	0.54	44
ERL	0.50	1.00	0.67	1
EXC	0.20	0.50	0.29	2
ME1	0.50	0.40	0.44	5
ME2	0.38	0.43	0.40	7
ME3	0.75	0.81	0.78	26
MIT	0.71	0.62	0.67	24
NUC	0.73	0.28	0.41	39
VAC	0.00	0.00	0.00	1
accuracy			0.53	149
macro avg	0.48	0.51	0.47	149
weighted avg	0.63	0.53	0.55	149

Rysunek 5. Wyniki dla algorytmu SVM (z klasą SMOTE)

- KNN

	precision	recall	f1-score	support
CYT	0.54	0.43	0.48	44
ERL	1.00	1.00	1.00	1
EXC	0.11	0.50	0.18	2
ME1	0.75	0.60	0.67	5
ME2	0.00	0.00	0.00	7
ME3	0.71	0.77	0.74	26
MIT	0.52	0.54	0.53	24
NUC	0.52	0.36	0.42	39
POX	0.00	0.00	0.00	0
VAC	0.00	0.00	0.00	1
accuracy			0.48	149
macro avg	0.42	0.42	0.40	149
weighted avg	0.54	0.48	0.50	149

Rysunek 6. Wyniki dla algorytmu KNN (z klasą SMOTE)

- LR

	precision	recall	f1-score	support
CYT	0.62	0.34	0.44	44
ERL	0.50	1.00	0.67	1
EXC	0.14	0.50	0.22	2
ME1	0.50	0.60	0.55	5
ME2	0.22	0.29	0.25	7
ME3	0.72	0.81	0.76	26
MIT	0.65	0.62	0.64	24
NUC	0.55	0.44	0.49	39
VAC	0.00	0.00	0.00	1
accuracy			0.50	149
macro avg	0.43	0.51	0.45	149
weighted avg	0.59	0.50	0.53	149

Rysunek 7. Wyniki dla algorytmu LR (z klasą SMOTE)

- GNB

	precision	recall	f1-score	support
CYT	0.00	0.00	0.00	44
ERL	0.50	1.00	0.67	1
EXC	0.05	0.50	0.09	2
ME1	0.29	0.40	0.33	5
ME2	0.00	0.00	0.00	7
ME3	0.00	0.00	0.00	26
MIT	0.75	0.25	0.38	24
NUC	0.50	0.18	0.26	39
VAC	0.01	1.00	0.02	1
accuracy			0.12	149
macro avg	0.23	0.37	0.19	149
weighted avg	0.27	0.12	0.15	149

Rysunek 8. Wyniki dla algorytmu GNB (z klasą SMOTE)

4. Podsumowanie i wnioski

Z przeprowadzonych badań wynika, że klasa SMOTE, poprzez wyrównanie ilości instancji w poszczególnych klasach, istotnie poprawia wynik dokładności tylko dla algorytmu SVM. Dla pozostałych algorytmów dokładność klasyfikacji oscylowała na tym samym poziomie bądź gorszym od poprzedniego (bez SMOTE).

Trzy z czterech zastosowanych algorytmów wykazuje dokładność klasyfikacji na poziomie około 45% - 55% poprawności, co jest wynikiem średnio zadowalającym.

Najgorzej z algorytmów radził sobie naiwny klasyfikator bayesowski, który poprawnie klasyfikował tylko około 10% próbek.

Zauważyć można, że zaimplementowane metody poradziły sobie średnio zadowalająco z klasyfikacją drożdży. Na uzyskane wyniki mogła mieć jednak wpływ zastosowana baza „Yeast”. Przy zastosowaniu innej bazy danych wynik mógłby być inny. Być może dalsze badania z różnymi bazami danych mogłyby poprawić wyniki dokładności klasyfikacji.