



SPRAWOZDANIE

Porównanie wybranych metod uczenia maszynowego

Przedmiot: **Uczenie maszynowe**

Prowadząca: dr Barbara Marszał-Paszek

Autor: Klaudia Dyjas

Kognitywistyka, II stopnia, 2 rok
Luty 2022

1. Wprowadzenie

Należało wygenerować macierze błędów przy użyciu wybranych metod uczenia maszynowego i porównać je ze sobą. Zbiór danych trzeba podzielić na część treningową i testową, a także przeprowadzić na nich sześciokrotne testowanie.

2. Materiały i metody

Testowanie train-and-test przeprowadzono w programie Rough Set Exploration System (RSES). Wybrano zbiór danych pt. „diabetes.tab” dotyczącego cukrzycy. Dostępne dane mają 768 obserwacji i 9 atrybutów (liczba ciąży, stężenie glukozy w osoczu, ciśnienie rozkurczowe, grubość fałdu skórno tricepsa, insulina w surowicy w ciągu 2 godzin, wskaźnik masy ciała, funkcja rodowodu cukrzycy, wiek, klasa).

Do testowania wykorzystano następujące metody:

- a. Klasyfikatory regułowe:
 - algorytm wyczerpujący (*exhaustive algorithm*)
 - algorytm genetyczny (*genetic algorithm*)
 - algorytm pokryciowy (*covering algorithm*)
- b. Klasyfikator neuronowy LTF-C (sieci neuronowe)
- c. Metoda cross-validation

Wyliczono macierze błędów klasyfikacji przy sposobie rozwiązywania konfliktów typu standardowy system głosowania, w którym każda reguła głosuje z siłą równą liczbie obiektów ją wspierających („inteligentniejsze” rozwiązanie).

3. Wyniki

Poniżej zebrano wyniki klasyfikacji wybranych metod.

3.1. Scenariusz testowania z klasyfikatorami regułowymi

Tabela 1. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody algorytmu wyczerpującego.

Results of experiments by train&test method: pima-indians-diabetes_0.3_ex						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	32	48	81	0.4	0.988
	0	26	124	150	0.827	1
	True positive rate	0.55	0.72			
Total number of tested objects: 231						
Total accuracy: 0.678						
Total coverage: 0.996						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 2. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody algorytmu genetycznego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	33	47	81	0.412	0.988
	0	29	121	150	0.807	1
	True positive rate	0.53	0.72			
Total number of tested objects: 231						
Total accuracy: 0.67						
Total coverage: 0.996						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 3. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody algorytmu pokryciowego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	28	41	81	0.406	0.852
	0	27	99	150	0.786	0.84
	True positive rate	0.51	0.71			
Total number of tested objects: 231						
Total accuracy: 0.651						
Total coverage: 0.844						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 4. Wyniki klasyfikacji w iteracji 2 – macierz błędów dla metody algorytmu wyczerpującego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	28	55	83	0.337	1
	0	33	115	148	0.777	1
	True positive rate	0.46	0.68			
Total number of tested objects: 231						
Total accuracy: 0.619						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 5. Wyniki klasyfikacji w iteracji 2 – macierz błędów dla metody algorytmu genetycznego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	29	54	83	0.349	1
	0	33	115	148	0.777	1
	True positive rate	0.47	0.68			
Total number of tested objects: 231						
Total accuracy: 0.623						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 6. Wyniki klasyfikacji w iteracji 2 – macierz błędów dla metody algorytmu pokryciowego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	22	45	83	0.328	0.807
	0	33	93	148	0.738	0.851
	True positive rate	0.4	0.67			
Total number of tested objects: 231						
Total accuracy: 0.596						
Total coverage: 0.835						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 7. Wyniki klasyfikacji w iteracji 3 – macierz błędów dla metody algorytmu wyczerpującego.

Results of experiments by train&test method: pima-indians-diabetes_...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	29	53	84	0.354	0.976
	0	25	122	147	0.83	1
	True positive rate	0.54	0.7			
Total number of tested objects: 231						
Total accuracy: 0.659						
Total coverage: 0.991						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 8. Wyniki klasyfikacji w iteracji 3 – macierz błędów dla metody algorytmu genetycznego.

Results of experiments by train&test method: pima-indians-diabetes_...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	32	49	84	0.395	0.964
	0	24	123	147	0.837	1
	True positive rate	0.57	0.72			
Total number of tested objects: 231						
Total accuracy: 0.68						
Total coverage: 0.987						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 9. Wyniki klasyfikacji w iteracji 3 – macierz błędów dla metody algorytmu pokryciowego.

Results of experiments by train&test method: pima-indians-diabetes_...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	24	41	84	0.369	0.774
	0	26	107	147	0.805	0.905
	True positive rate	0.48	0.72			
Total number of tested objects: 231						
Total accuracy: 0.662						
Total coverage: 0.857						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 10. Wyniki klasyfikacji w iteracji 4 – macierz błędów dla metody algorytmu wyczerpującego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	36	45	82	0.444	0.988
	0	28	121	149	0.812	1
True positive rate		0.56	0.73			
Total number of tested objects: 231						
Total accuracy: 0.683						
Total coverage: 0.996						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 11. Wyniki klasyfikacji w iteracji 4 – macierz błędów dla metody algorytmu genetycznego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	33	47	82	0.412	0.976
	0	29	120	149	0.805	1
True positive rate		0.53	0.72			
Total number of tested objects: 231						
Total accuracy: 0.668						
Total coverage: 0.991						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 12. Wyniki klasyfikacji w iteracji 4 – macierz błędów dla metody algorytmu pokryciowego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	32	33	82	0.492	0.793
	0	31	93	149	0.75	0.832
True positive rate		0.51	0.74			
Total number of tested objects: 231						
Total accuracy: 0.661						
Total coverage: 0.818						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 13. Wyniki klasyfikacji w iteracji 5 – macierz błędów dla metody algorytmu wyczerpującego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	37	48	85	0.435	1
	0	30	116	146	0.795	1
	True positive rate	0.55	0.71			
Total number of tested objects: 231						
Total accuracy: 0.662						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 14. Wyniki klasyfikacji w iteracji 5 – macierz błędów dla metody algorytmu genetycznego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	35	49	85	0.417	0.988
	0	33	113	146	0.774	1
	True positive rate	0.51	0.7			
Total number of tested objects: 231						
Total accuracy: 0.643						
Total coverage: 0.996						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 15. Wyniki klasyfikacji w iteracji 5 – macierz błędów dla metody algorytmu pokryciowego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	27	39	85	0.409	0.776
	0	24	102	146	0.81	0.863
	True positive rate	0.53	0.72			
Total number of tested objects: 231						
Total accuracy: 0.672						
Total coverage: 0.831						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 16. Wyniki klasyfikacji w iteracji 6 – macierz błędów dla metody algorytmu wyczerpującego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	34	44	78	0.436	1
	0	38	115	153	0.752	1
True positive rate		0.47	0.72			
Total number of tested objects: 231						
Total accuracy: 0.645						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 17. Wyniki klasyfikacji w iteracji 6 – macierz błędów dla metody algorytmu genetycznego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	38	40	78	0.487	1
	0	38	115	153	0.752	1
True positive rate		0.5	0.74			
Total number of tested objects: 231						
Total accuracy: 0.662						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 18. Wyniki klasyfikacji w iteracji 6 – macierz błędów dla metody algorytmu pokryciowego.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	24	35	78	0.407	0.756
	0	27	102	153	0.791	0.843
True positive rate		0.47	0.74			
Total number of tested objects: 231						
Total accuracy: 0.67						
Total coverage: 0.814						

Źródło: opracowanie własne przy użyciu programu RSES.

3.2. Scenariusz testowania z klasyfikatorem regułowym i skalowaniem

Tabela 19. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody algorytmu wyczerpującego wraz ze skalowaniem.

Results of experiments by train&test method: pima-indians-diabetes_...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	47	34	81	0.58	1
	0	27	123	150	0.82	1
	True positive rate	0.64	0.78			
Total number of tested objects: 231						
Total accuracy: 0.736						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 20. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody algorytmu genetycznego wraz ze skalowaniem.

Results of experiments by train&test method: pima-indians-diabetes_...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	48	33	81	0.593	1
	0	27	123	150	0.82	1
	True positive rate	0.64	0.79			
Total number of tested objects: 231						
Total accuracy: 0.74						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 21. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody algorytmu pokryciowego wraz ze skalowaniem.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	18	15	81	0.545	0.407
	0	21	64	150	0.753	0.567
	True positive rate	0.46	0.81			
Total number of tested objects: 231						
Total accuracy: 0.695						
Total coverage: 0.511						

Źródło: opracowanie własne przy użyciu programu RSES.

3.3. Scenariusz testowania z klasyfikatorem neuronowym LTF-C

Tabela 22. Wyniki klasyfikacji w iteracji 1 – macierz błędów dla metody sieci neuronowych.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	48	33	81	0.593	1
	0	21	129	150	0.86	1
	True positive rate	0.7	0.8			
Total number of tested objects: 231						
Total accuracy: 0.766						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 23. Wyniki klasyfikacji w iteracji 2 – macierz błędów dla metody sieci neuronowych.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	49	34	83	0.59	1
	0	31	117	148	0.791	1
True positive rate		0.61	0.77			
Total number of tested objects: 231						
Total accuracy: 0.719						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 24. Wyniki klasyfikacji w iteracji 3 – macierz błędów dla metody sieci neuronowych.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	51	33	84	0.607	1
	0	19	128	147	0.871	1
True positive rate		0.73	0.8			
Total number of tested objects: 231						
Total accuracy: 0.775						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 25. Wyniki klasyfikacji w iteracji 4 – macierz błędów dla metody sieci neuronowych.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	45	37	82	0.549	1
	0	14	135	149	0.906	1
True positive rate		0.76	0.78			
Total number of tested objects: 231						
Total accuracy: 0.779						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 26. Wyniki klasyfikacji w iteracji 5 – macierz błędów dla metody sieci neuronowych.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	58	27	85	0.682	1
	0	26	120	146	0.822	1
	True positive rate	0.69	0.82			
Total number of tested objects: 231						
Total accuracy: 0.771						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 27. Wyniki klasyfikacji w iteracji 6 – macierz błędów dla metody sieci neuronowych.

Results of experiments by train&test method: pima-indians-diabetes_...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	48	30	78	0.615	1
	0	28	125	153	0.817	1
	True positive rate	0.63	0.81			
Total number of tested objects: 231						
Total accuracy: 0.749						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 28. Zebranie wyników z sześciokrotnego testowania wraz ze średnią dla sześciu iteracji.

	Dokładność (Total accuracy)	Pokrycie (Total coverage)
Iteracja 1		
Algorytm wyczerpujący	0.678	0.996
Algorytmu genetyczny	0.670	0.996
Algorytmu pokryciowy	0.651	0.844
Sieci neuronowe	0.766	1.000
Średnia	0.691	0.959

Iteracja 2

Algorytm wyczerpujący	0.619	1.000
Algorytmu genetyczny	0.623	1.000
Algorytmu pokryciowy	0.596	0.835
Sieci neuronowe	0.719	1.000
Średnia	0.639	0.959

<u>Iteracja 3</u>		
Algorytm wyczerpujący	0.659	0.991
Algorytmu genetyczny	0.680	0.987
Algorytmu pokryciowy	0.662	0.857
Sieci neuronowe	0.775	1.000
Średnia	0.694	0.959

<u>Iteracja 4</u>		
Algorytm wyczerpujący	0.683	0.996
Algorytmu genetyczny	0.668	0.991
Algorytmu pokryciowy	0.661	0.818
Sieci neuronowe	0.779	1.000
Średnia	0.698	0.951

<u>Iteracja 5</u>		
Algorytm wyczerpujący	0.662	1.000
Algorytmu genetyczny	0.643	0.996
Algorytmu pokryciowy	0.672	0.831
Sieci neuronowe	0.771	1.000
Średnia	0.687	0.957

<u>Iteracja 6</u>		
Algorytm wyczerpujący	0.645	1.000
Algorytmu genetyczny	0.662	1.000
Algorytmu pokryciowy	0.670	0.814
Sieci neuronowe	0.749	1.000
Średnia	0.682	0.954

Źródło: opracowanie własne.

3.4. Scenariusz testowania metodą cross-validation

Tabela 29. Wyniki klasyfikacji metodą cross-validation z użyciem algorytmu wyczerpującego.

Results of experiments by cross-validation method: pima-indians-diab...						
Predicted						
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	19.67	24.83	44.67	0.434	0.996
	0	15.83	67.5	83.33	0.809	1
	True positive rate	0.54	0.73			
Total number of tested objects: 128						
Total accuracy: 0.682						
Total coverage: 0.999						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 30. Wyniki klasyfikacji metodą cross-validation z użyciem algorytmu genetycznego.

Results of experiments by cross-validation method: pima-indians-diab...						
Predicted						
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	21.67	22.33	44.67	0.497	0.984
	0	17.17	66.17	83.33	0.795	1
	True positive rate	0.57	0.75			
Total number of tested objects: 128						
Total accuracy: 0.69						
Total coverage: 0.995						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 31. Wyniki klasyfikacji metodą cross-validation z użyciem algorytmu pokryciowego.

Results of experiments by cross-validation method: pima-indians-diab...						
Predicted						
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	14.5	18.83	44.67	0.438	0.749
	0	16.83	52.33	83.33	0.756	0.831
	True positive rate	0.47	0.74			
Total number of tested objects: 128						
Total accuracy: 0.653						
Total coverage: 0.801						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 32. Wyniki klasyfikacji metodą cross-validation przy pomocy sieci neuronowych.

Results of experiments by cross-validation method: pima-indians-diab...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	29.33	15.33	44.67	0.656	1
	0	12.33	71	83.33	0.852	1
	True positive rate	0.7	0.82			
Total number of tested objects: 128						
Total accuracy: 0.784						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

3.5. Scenariusz testowania metodą cross-validation i skalowaniem

Tabela 33. Wyniki klasyfikacji metodą cross-validation z użyciem algorytmu wyczerpującego i skalowania.

Results of experiments by cross-validation method: pima-indians-diab...						
	Predicted					
Actual		1	0	No. of obj.	Accuracy	Coverage
	1	27.67	17	44.67	0.622	1
	0	18.5	64.83	83.33	0.777	1
	True positive rate	0.6	0.79			
Total number of tested objects: 128						
Total accuracy: 0.723						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 34. Wyniki klasyfikacji metodą cross-validation z użyciem algorytmu genetycznego i skalowania.

Results of experiments by cross-validation method: pima-indians-diab...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	27.33	17.33	44.67	0.611	1
	0	20.67	62.67	83.33	0.751	1
	True positive rate	0.56	0.78			
Total number of tested objects: 128						
Total accuracy: 0.703						
Total coverage: 1						

Źródło: opracowanie własne przy użyciu programu RSES.

Tabela 35. Wyniki klasyfikacji metodą cross-validation z użyciem algorytmu pokryciowego i skalowania.

Results of experiments by cross-validation method: pima-indians-diab...						
Actual	Predicted					
		1	0	No. of obj.	Accuracy	Coverage
	1	12.33	9.33	44.67	0.571	0.483
	0	11.17	31.17	83.33	0.73	0.509
	True positive rate	0.52	0.77			
Total number of tested objects: 128						
Total accuracy: 0.678						
Total coverage: 0.5						

Źródło: opracowanie własne przy użyciu programu RSES.

4. Wnioski

Jak pokazują wyniki sześciokrotnego testowania, klasyfikatory LTF-C (Local Transfer Function Classifier) bazujące na sieci neuronowej, pozwalają uzyskać najlepsze wyniki w każdej iteracji dla wybranego zbioru danych zarówno pod względem dokładności jak i sklasyfikowania wszystkich obiektów (w każdej iteracji „Total coverage” ma wartość 1).

Dla porównania wyników, w pierwszej iteracji wzbogacono scenariusz o skalowanie (dyskretyzację) tablicy treningowej. Operacja skalowania znacząco polepszyła wyniki dokładności oraz sklasyfikowania wszystkich obiektów przy klasyfikacji metody algorytmu wyczerpującego (Tabela 19) oraz algorytmu genetycznego (Tabela 20). Co ciekawe, w przypadku algorytmu pokryciowego (Tabela 21) polepszyła się dokładność, ale przy równoczesnym zmniejszeniu się

liczby obiektów przydzielonych do klas decyzyjnych („Total coverage” wynosi 0.511). Nie zastosowano skalowania przy sieciach neuronowych.

Dodatkowo, wybrane metody porównano poprzez sześciokrotną cross-validation. Również i w tym przypadku największą dokładność uzyskały sieci neuronowe wraz ze sklasyfikowaniem wszystkich obiektów. Drugi pod względem dokładności okazał się algorytm genetyczny. Na trzecim miejscu znajduje się algorytm wyczerpujący, a najmniej dokładny okazał się algorytm pokryciowy.

Na koniec wzbogacono cross-validation o dyskretyzację. Również i w tym przypadku polepszyła się dokładność, a system decyzyjny sklasyfikował wszystkie obiekty dla algorytmu wyczerpującego (Tablica 33) oraz algorytmu genetycznego (Tablica 34). Ponownie dla algorytmu pokryciowego (Tablica 35) poprawiła się dokładność, ale kosztem liczby przydzielonych obiektów do klas decyzyjnych (wartość tej statystyki wynosi zaledwie 0.5). Opcja dyskretyzacji danych jest niedostępna dla klasyfikatora LTF-C (sieci neuronowych).