











Marek Grzyb - Wizualizacja danych - Kółko i Krzyżyk

Dane wejściowe

Plik wejściowy zawiera wszystkie możliwości zakończenia gry w kolko i krzyżyk (958 różnych zakończeń). Klasyfikacja zawiera informacje czy gracz, który rozpoczynał wygrał. Dane pochodzą z repozytorium UCI <https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame>

Każdy rekord zawiera 10 trybutów, 9 odpowiadają za wartości pól, ostatnio to informacje czy gracz 1 wygrał.

	Name	Type	Role	Values
1	top-left-square	 categorical	feature	b, o, x
2	top-middle-squ...	 categorical	feature	b, o, x
3	top-right-square	 categorical	feature	b, o, x
4	middle-left-squ...	 categorical	feature	b, o, x
5	middle-middle-...	 categorical	feature	b, o, x
6	middle-right-sq...	 categorical	feature	b, o, x
7	bottom-left-sq...	 categorical	feature	b, o, x
8	bottom-middle...	 categorical	feature	b, o, x
9	bottom-right-s...	 categorical	feature	b, o, x
10	class	 categorical	feature	negative, positive

Wartości 9 atrybutów oznaczaj:

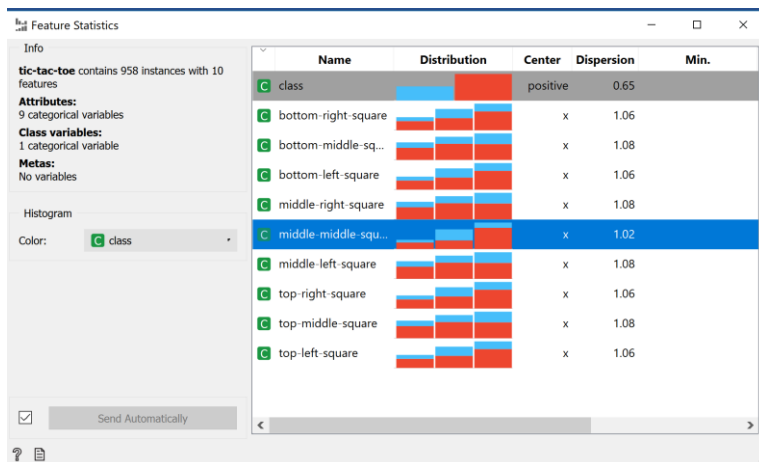
X- gracz 1 zajął to pole

O – gracz 2 zajął to pole

B – pole nie zostało zajęte

Wizualizacja

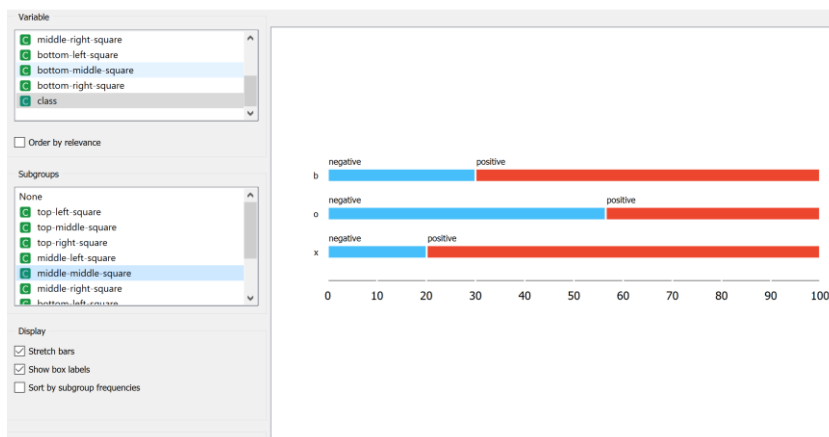
Najlepszym sposobem wizualizacji danych tego typu jest rozkład:



Z rozkładu wprost wynika:

- 1) gracz który rozpoczyna gra ma większe szanse na wygrana: 65% do 35%
- 2) rozrzut w punkcie w środku planszy jest mniejszy, gdy gracz wybierze to pole, częściej niż zwykle wygrywa, przez co drzewo gry jest krótsze
- 3) punkty w rogach mają również mniejszy rozrzut niż skrajne-środkowe (gracz 1 częściej wygrywa)

Innym sposobem prezentacji danych tego typu jest Box Plot (wykres pudełkowy):



Wykres ten również pokazuje że gracz wybierając punkt środkowy ma duże szanse na wygrana (80%).

Analiza

Do analizy takiego zbioru danych nie wszystkie metody dają zadowalające wyniki, przykładem jest tutaj metoda kNN:

Method	AUC	CA	F1	Precision	Recall
kNN	0.581	0.680	0.573	0.785	0.680

Inne metoda dają zadowalające wyniki, np.: SVM:

Method	AUC	CA	F1	Precision	Recall
SVM	0.990	0.960	0.960	0.960	0.960

Podjąłem próbę wyboru najlepszych atrybutów do klasyfikacji, sprawdziłem dwa przypadki:

1) rogi (4 atrybuty) i środek

Method	AUC	CA	F1	Precision	Recall
SVM	0.855	0.760	0.749	0.754	0.760

2) środek i skrajnie-środkowe (4 punkty)

Method	AUC	CA	F1	Precision	Recall
SVM	0.545	0.518	0.527	0.586	0.518

Pierwsza kombinacja atrybutów daje dużo lepsze wyniki.

Dlaczego zestaw atrybutów 1 sprawdza się lepiej w klasyfikacji niż zestaw 2?

Grę kółko i krzyżyk można wygrać stawiając trzy X na 8 różnych sposobów, tylko 2 sposoby nie wykorzystują punktu na rogu. 6 sposobów wymaga dwóch rogów na trzy punkty.

Z tego wynika że prognozowanie zwycięstwa jest łatwiejsze przy zastosowaniu punktów na rogu.

Analizowany schemat

