

POLITECHNIKA BIAŁOSTOCKA
ROZPRAWY NAUKOWE NR 158

Marek Krętowski

OBLICZENIA EWOLUCYJNE
W EKSPŁORACJI DANYCH

GLOBALNA INDUKCJA DRZEW DECYZYJNYCH



WYDAWNICTWO POLITECHNIKI BIAŁOSTOCKIEJ
BIAŁYSTOK 2008

Spis treści

Wstęp	7
1 Wprowadzenie do obliczeń ewolucyjnych	13
1.1 Struktura algorytmu ewolucyjnego	15
1.2 Reprezentowanie dopuszczalnych rozwiązań	17
1.3 Różnicowanie osobników	20
1.3.1 Operator mutacji	21
1.3.2 Operator krzyżowania	22
1.4 Selekcja	23
2 Eksploracja danych	27
2.1 Eksploracja danych w procesie pozyskiwania wiedzy	28
2.1.1 Zadania eksploracji danych	31
2.1.2 Komponenty eksploracji wiedzy	32
2.1.3 Klasyfikacja	33
2.2 Drzewa decyzyjne	35
2.2.1 Rodzaje testów i drzew	37
2.2.2 Metody indukcji drzew decyzyjnych	39
2.2.3 Poszukiwanie optymalnego testu w węźle	43
2.2.4 Przycinanie drzew decyzyjnych	46
2.3 Reguły decyzyjne	48
2.3.1 Algorytm sekwencyjnego pokrywania	49
2.3.2 Zbiór reguł jako klasyfikator	49
2.3.3 Przegląd systemów indukcji reguł	50
2.4 Obliczenia ewolucyjne jako narzędzie eksploracji danych	51
2.4.1 Obliczenia ewolucyjne w indukcji drzew decyzyjnych	51
2.4.2 Obliczenia ewolucyjne w generowaniu reguł decyzyjnych	54

3	Globalna indukcja drzew jednowymiarowych	57
3.1	Reprezentacja	58
3.2	Tworzenie populacji początkowej	60
3.3	Operatory różnicowania	61
3.3.1	Mutowanie drzewa decyzyjnego	62
3.3.2	Krzyżowanie drzew decyzyjnych	64
3.3.3	Dodatkowe przetwarzanie	67
3.4	Selekcja i warunek zatrzymania	69
3.5	Funkcja dopasowania	69
3.6	Weryfikacja eksperymentalna globalnej indukcji	71
3.6.1	Poznanie własności systemu	73
3.6.2	Wyniki uzyskane na zbiorach sztucznych i rzeczywistych	78
3.6.3	Skalowalność indukcji na dużych zbiorach	80
4	Algorytm memetyczny w indukcji drzew jednowymiarowych	83
4.1	Hybrydowa indukcja drzew jednowymiarowych	84
4.1.1	Tworzenie populacji początkowej	84
4.1.2	Modyfikacja operatora mutacji	85
4.2	Weryfikacja eksperymentalna	86
4.2.1	Wpływ częstości lokalnej optymalizacji na indukcję	86
4.2.2	Wyniki na zbiorach sztucznych i rzeczywistych	87
5	Generowanie drzew skośnych	91
5.1	Ewolucyjna indukcja drzew skośnych	92
5.1.1	Tworzenie populacji początkowej	93
5.1.2	Różnicowanie drzew skośnych	94
5.1.3	Funkcja dopasowania	96
5.2	Weryfikacja eksperymentalna	97
5.2.1	Rola operatora dipolowego	97
5.2.2	Eksperymenty ze zbiorami zawierającymi szum	98
5.2.3	Wyniki uzyskane na sztucznych i rzeczywistych zbiorach	101
5.2.4	Skalowalność indukcji na dużych zbiorach	102
6	Konstruowanie drzew mieszanych	105
6.1	Ewolucyjna indukcja drzew mieszanych	107
6.1.1	Różnicowanie drzew mieszanych	107

6.1.2	Funkcja dopasowania	108
6.2	Wyniki eksperymentów obliczeniowych	109
6.2.1	Zbiory sztuczne	110
6.2.2	Zbiory rzeczywiste	112
6.2.3	Szybkość i skalowalność ewolucji drzew mieszanych	115
7	Klasyfikacja uwzględniająca koszty	117
7.1	Ewolucyjna indukcja drzew jednowymiarowych uwzględniająca koszty	119
7.1.1	Funkcja dopasowania czuła na koszty	119
7.1.2	Modyfikacje operatorów genetycznych	121
7.1.3	Przypisanie klas w liściach drzewa	122
7.2	Weryfikacja eksperymentalna	122
7.2.1	Minimalizacja kosztów błędnych decyzji	123
7.2.2	Eksperymenty z dwoma rodzajami kosztów	125
8	Obliczenia równoległe w globalnej indukcji drzew	129
8.1	Implementacja równoległa globalnej indukcji drzew	131
8.1.1	Zaproponowane rozwiązanie	132
8.1.2	Realizacja na klastrze obliczeniowym	134
8.2	Rozwiązanie hybrydowe	136
8.2.1	Weryfikacja eksperymentalna	137
8.3	Rozproszona globalna indukcja drzew decyzyjnych	138
8.3.1	Wyniki eksperymentów obliczeniowych	139
9	Podsumowanie	143
9.1	Możliwe kierunki badań	145