Dokumentacja wstępna UM

Temat: Algorytm maszynowego uczenia się w zastosowaniu do predykcji wartości szeregu czasowego. Zadanie polega na przewidywaniu temperatury w chłodziarce w zależności od temperatury zewnętrznej i stanu agregatu.

Opis Algorytmu

W projekcie planuje wykorzystać algorytm lasu losowego drzew regresyjnych.

Dane

Najpierw należy rozdzielić dane na dane trenujące i dane testowe. Ostatnie 10% danych przeznaczę na testy. Reszta danych będzie podzielona na mniejsze fragmenty na podstawie których będą tworzone poszczególne drzewa regresyjne. Dane zawierają temperaturę wewnątrz chłodziarki, temperaturę zewnętrzną i stan pracy agregatu z poprzedniego kwantu czasu, oraz temperaturę wewnętrzną chłodziarki z bieżącego czasu.

Tworzenie drzew regresyjnych

Drzewo składa się z węzłów i liści oraz korzenia (pierwszego węzła). Każdy węzeł zawiera warunek decydującym czy wybieramy prawą czy lewą gałąź. Gałęzie zakończone są liśćmi, które reprezentują wynik algorytmu.

Tworzenie drzewa zaczynamy od korzenia. Przebieg tworzenia każdego węzła jest jednakowy. Najpierw sprawdzamy rozmiar zbioru, jeśli jest mniejszy niż dany parametr to węzeł staje się liściem. A jego wartość to średnia wartości szukanej w wybranym zbiorze. Jeśli zbiór jest większy szukamy najlepszego miejsca podziału zbioru.

Losujemy atrybut według którego podzielimy zbiór. Następnie sortujemy według tego atrybutu. Liczymy średnią temperaturę wewnątrz urządzenia w podzbiorze przed wartością podziału. Od tej średniej odejmujemy wartość w poszczególnych wierszach sumę tych różnic podniesionych do kwadratu nazywamy SSR. Taką samą operacje wykonujemy na podzbiorze po wartości podziału. Warunkiem podziału zostanie wartość w której suma SSR obu zbiorów będzie najmniejsza. W węźle zostaje zapisany warunek podziału. Operacje powtarza się na obu zbiorach uzyskanych takim podziałem.

Przykład

czas(min)	alpha	temp	Ts	Ts+1
0	1	6.50	2.00	2.18
15	0	6.35	2.18	2.19
30	0	6.20	2.19	1.68
45	1	6.05	1.68	1.50

czas(min)	alpha	temp	Ts	Ts+1
60	1	5.90	1.50	1.64

Czas w tym nie będzie potrzebny więc usunę go dla przejrzystości. Załóżmy że ostatni wiersz to dane testowe. I rozmiar zbioru niepodlegającego dzieleniu jest równy 2. Powiedzmy że będziemy dzielić na podstawie temperatury zewnętrznej '*temp*' sortujemy więc dane.

alpha	temp	Ts	Ts+1
1	6.05	1.68	1.50
0	6.20	2.19	1.68
0	6.35	2.18	2.19
1	6.50	2.00	2.18
1	5.90	1.50	1.64

podział na >= 6.2
zbiór pierwszy [1.5]
średnia = 1.5
zbiór drugi [1.68 2.19 2.18]
średnia = $\frac{1.68+2.19+2.18}{3} = 2.02$
SSR = $(1.5-1.5)^2 + (1.68-2.02)^2 + (2.19-2.02)^2 + (2.18-2.02)^2 = 0.17$

analogicznie postępujemy dla kolejnych

podział na >= 6.35
zbiór pierwszy [1.5 1.68]
średnia = 1.59
zbiór drugi [2.19 2.18]
średnia = 2.185
SSR = 0.016

podział na >= 6.5
zbiór pierwszy [1.5 1.68 2.19]
średnia = 1.79
zbiór drugi [2.18]
średnia = 2.18
SSR = 0.26

Podziału zbioru dokonujemy na '*temp*' >= 6.35 ponieważ SSR jest wtedy najmniejsze. Powstałe zbiory nie są większe od założonego rozmiaru liścia więc kończymy budowę drzewa.

'*temp*' z danych testowych 5.9 < 6.35 więc przewidziana wartość to 1.59

Tworzenie lasu losowego

Las losowy to zbiór drzew regresyjnych utworzonych na podstawie różnych danych. Dane dla każdego drzewa są losowane z powtórzeniami.

Wynik przewidziany przez las jest średnią wyników ze wszystkich drzew.

Eksperymenty

W ramach eksperymentu można porównać działanie lasu losowego z metodą naiwną (wartość z poprzedniego kwantu czasu). Można również sprawdzić wpływ parametrów takich jaki ilość drzew, rozmiar liści, ilość danych trenujących na działanie algorytmu.

Zbiory Danych

Zbiór testujących będzie to ostatnie 10% danych które nie będą użyte do tworzenia drzew. Reszta danych to dane trenujące.

Jacek Dobrowolski