Politechnika Warszawska

WYDZIAŁ ELEKTRYCZNY

Instytut Elektrotechniki Teoretycznej
i Systemów Informacyjno-Pomiarowych
Zakład Elektrotechniki Teoretycznej
i Informatyki Stosowanej

Praca dyplomowa magisterska

na kierunku Informatyka
w specjalności Inżynieria oprogramowania

Meta-metody służące do poprawy jakości klasyfikacji

Konrad Ziaja
nr albumu 272170

promotor
dr inż. Łukasz Skonieczny

konsultant
brak

WARSZAWA 2017

BŁĄD: BRAK POLECENIA \streszczenia !

WARSZAWA, **BŁĄD: BRAK POLECENIA \terminwykonania !**

POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRYCZNY

OŚWIADCZENIE

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa magisterska pt. Meta-metody służące do poprawy jakości klasyfikacji:

- została napisana przeze mnie samodzielnie,
- nie narusza niczyich praw autorskich,
- nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam, że przedłożona do obrony praca dyplomowa nie była wcześniej podstawą postępowania związanego z uzyskaniem dyplomu lub tytułu zawodowego w uczelni wyższej. Jestem świadom, że praca zawiera również rezultaty stanowiące własności intelektualne Politechniki Warszawskiej, które nie mogą być udostępniane innym osobom i instytucjom bez zgody Władz Wydziału Elektrycznego.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Konrad Ziaja.....

Spis treści

Rozdział 1

Wstęp

Trzeba napisać jakiś wstęp :(

Cel pracy

I jakiś cel pracy :((:()))

Rozdział 2

Wstęp teoretyczny

2.1 Klasyfikacja danych

może coś pisać o uczeniu maszynowym i że klasyfikacja jest nadzorowana? inline

Klasyfikacja jest to proces przyporządkowania danych do jednej z predefiniowanych klas na podstawie atrybutów tych danych. Algorytm klasyfikacji na podstawie analizy danych trenujących, zawierających atrybuty oraz klasę, tworzy model klasyfikacyjny. Stworzony model klasyfikacyjny wykorzystywany jest do predykcji klasy (kategorii) nowych danych bez określonej klasy. Celem algorytmu budującego model, jest odnalezienie wzorców, w jaki sposób atrybuty obiektu wpływają na przynależność do danej klasy, starając się o to, aby wiedza na temat analizowanych danych była możliwie ogólna oraz niezależna od próby.

Klasyfikacja danych jest procesem dwuetapowym:

- budowa modelu – proces ten polega na analizie obiektów z przyporządkowaną klasą oraz na budowie modelu opisującego predefiniowany zbiór klas danych,
- właściwa klasyfikacja – otrzymany model stosuje się do przydzielania klasy nowym obiektom.

Budowa modelu jest także procesem dwu-etapowym. Dzieli się ona na:

- uczenie – klasyfikator budowany jest w oparciu o dane treningowe,
- ocena jakości klasyfikacji – jakoś klasyfikacji badana jest w oparciu o dane testowe.

W zależności od liczebności klas w zbiorze danych, możemy wyróżnić:

- klasyfikację binarną – klasyfikator decyduje o przypisaniu obiektu do jednej z dwóch klas (np. czy człowiek jest zdrowy lub nie)

- klasyfikację wieloklasową – obiektowi przypisuje się jedną z wielu predefiniowanych klas.

Do reprezentacji danych uczących, testowych oraz do klasyfikacji najczęściej stosuje się system informacyjny.

Zachmurzenie	**Temp.**	**Temp. wody**	**Opady**	**Wiatr**	**Pływać**	**h(x)**
słonecznie	32	25	brak	słaby	tak	tak
słonecznie	31	26	brak	umiarkowany	tak	nie
pochmurnie	22	15	brak	b. mocny	nie	nie
pochmurnie	20	18	brak	słaby	tak	tak
całkowite zachmurzenie	12	6	brak	umiarkowany	nie	nie
całkowite zachmurzenie	10	8	duże	słaby	nie	nie
pochmurnie	21	10	brak	mocny	nie	tak
słonecznie	25	17	brak	umiarkowany	tak	nie
pochmurnie	23	17	przelotne	umiarkowany	nie	tak

Tablica 2.1: Przykład danych treningowych składających się z 5 atrybutów oraz klasy decyzyjnej. W ostatniej kolumnie znajduje się wynik klasyfikacji. W pięciu przypadkach, klasyfikator poprawnie wskazał klasę.

2.2 Przegląd algorytmów klasyfikacji danych

2.2.1 Drzewo decyzyjne

2.2.2 Naiwny klasyfikator bayesowski

2.2.3 Klasyfikator k najbliższych sąsiadów (kNN)

2.2.4 Las losowy

2.2.5 Bagging

2.2.6 Boosting

2.2.7 Stacking

2.3 Ocena poprawności klasyfikacji

2.3.1 Miary jakości klasyfikacji danych

Jakość klasyfikacji można ocenić na podstawie kilku współczynników. Do ich obliczenia wykorzystuje się macierz pomyłek. Tworzona jest ona w oparciu o wynik klasyfikacji. Dla klasyfikacji binarnej macierz składa się z dwóch kolumn oraz dwóch wierszy. W wierszach znajdują się poprawne klasy decyzyjne, natomiast w kolumnach przewidziane przez klasyfikator. Zaklasyfikowane obiekty, umieszcza się w odpowiedniej grupie.

		Klasa predykowana	
		pozytywna	negatywna
Klasa rzeczywista	pozytywna	prawdziwie pozytywna (TP)	fałszywie negatywna (FN)
	negatywna	fałszywie pozytywna (FP)	prawdziwie negatywna (TN)

Tablica 2.2: Macierz pomyłek