Porównanie metod wstępnego przetwarzania i klasyfikacji danych biomedycznych

Seminarium Dyplomowe 2

Piotr Tąkiel

Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

21 kwietnia 2015

Plan prezentacji

Cel pracy magisterskiej

Cel pracy magisterskiej

Porównanie wyników wstępnego przetwarzania i klasyfikacji danych medycznych dotyczących zachorowań na raka piersi.

Motywacja

Powody:

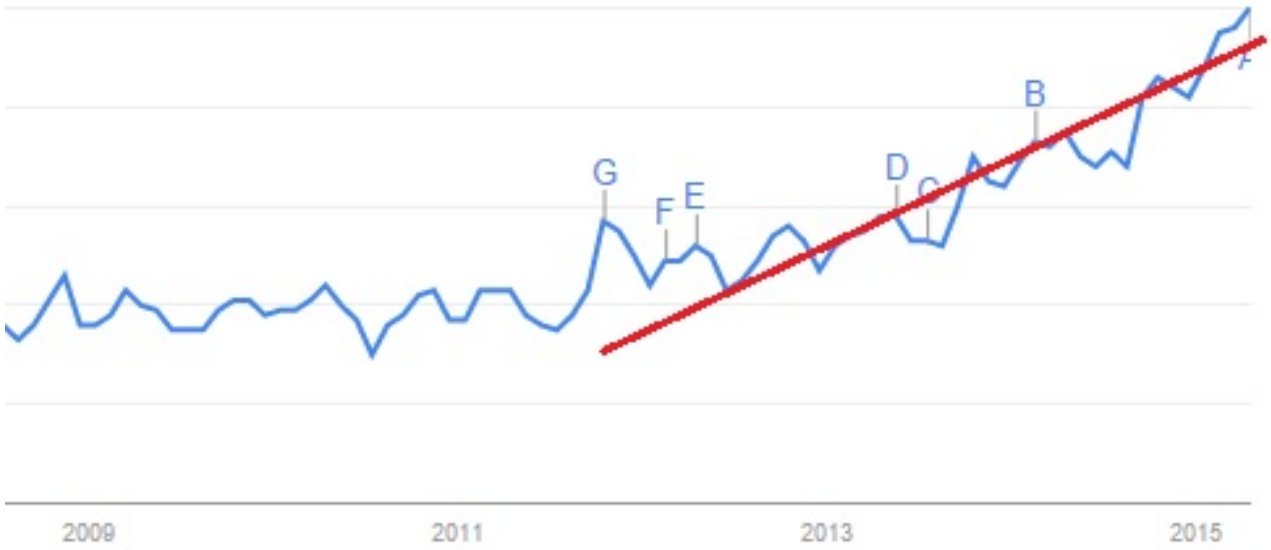
1. Aktualność problemu raka piersi
2. Zainteresowanie uczeniem maszynowym

Problem raka piersi

1. Wzrostowy wskaźnik zachorowalności w krajach wysoko rozwiniętych
2. Zwiększa się liczba zgonów pomimo postępów w diagnozowaniu i leczeniu
3. Co 14 żyjąca Polka zachoruje na raka piersi
4. Obecnie co druga chora Polka umiera na tę chorobę

Uczenie maszynowe

Uczenie maszynowe jest bardzo popularną i bujnie rozwijaną dziedziną informatyki.



Google Trends dla hasła *Machine learning*

Uczenie maszynowe kont.

Uczenie maszynowe znajduje bardzo wiele zastosowań praktycznych:

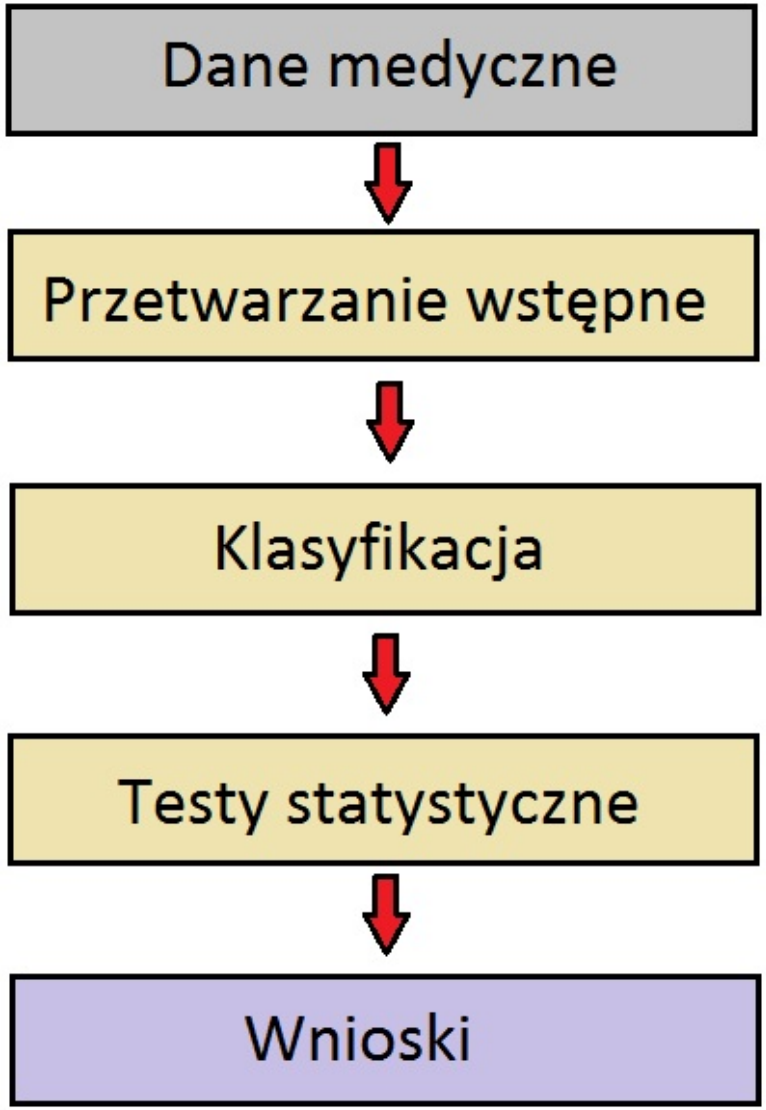
1. Rozpoznawanie obrazów
2. Wyszukiwarki
3. Detekcja spamu
4. Wykrywanie oszustw finansowych
5. *Rozpoznawanie chorób na podstawie symptomów*

Badania

Charakterystyka dostępnych danych medycznych:

1. Sześć atrybutów o wartościach ciągłych
2. Kategoria dwuwartościowa: choruje lub nie choruje
3. Ok. 1000 przykładów
4. Brakujące wartości atrybutów (ok. 8%)

Badania: uproszczony plan



Klasyfikacja

Czym jest klasyfikacja?

Klasyfikacja statystyczna – rodzaj algorytmu statystycznego, który przydziela obserwacje statystyczne do klas, bazując na atrybutach (cechach) tych obserwacji.

— Wikipedia

Klasyfikacja cd.

Algorytmy klasyfikacji tworzą *model klasyfikacji* w oparciu o dane trenujące zawierające znane klasy (kategorie). Model można następnie wykorzystywać do *predykcji* kategorii nowych danych.

Klasyfikacja jest przykładem *uczenia się z nadzorem*. Nadzór polega na obecności informacji o kategoriach danych trenujących.

Klasyfikacja: przykład

x1	x2	c(x)	*h(x)*
1	0,56	1	*1*
1	2,5	1	*1*
2	2,13	1	*0*
1	1,14	1	*0*
3	0,75	1	*0*
2	1,15	0	*0*
1	2,53	0	*0*
1	1,98	0	*0*
3	4,3	0	*1*

Przykład użycia pewnego modelu klasyfikacji

Metody oceny modelu

Jedną z metod oceny modelu klasyfikacji jest użycie miar liczonych na podstawie macierzy pomyłek:

		Stan faktyczny	
		0	1
Odpowiedź klasyfikatora	0	TN	FN
	1	FP	TP

Macierz pomyłek

Miary jakości klasyfikacji

Precyzja - jaka część zdiagnozowanych przypadków była prawdziwa

$$precision = \frac{TP}{TP + FP}$$

Wrażliwość - jaką część chorych została wykryta

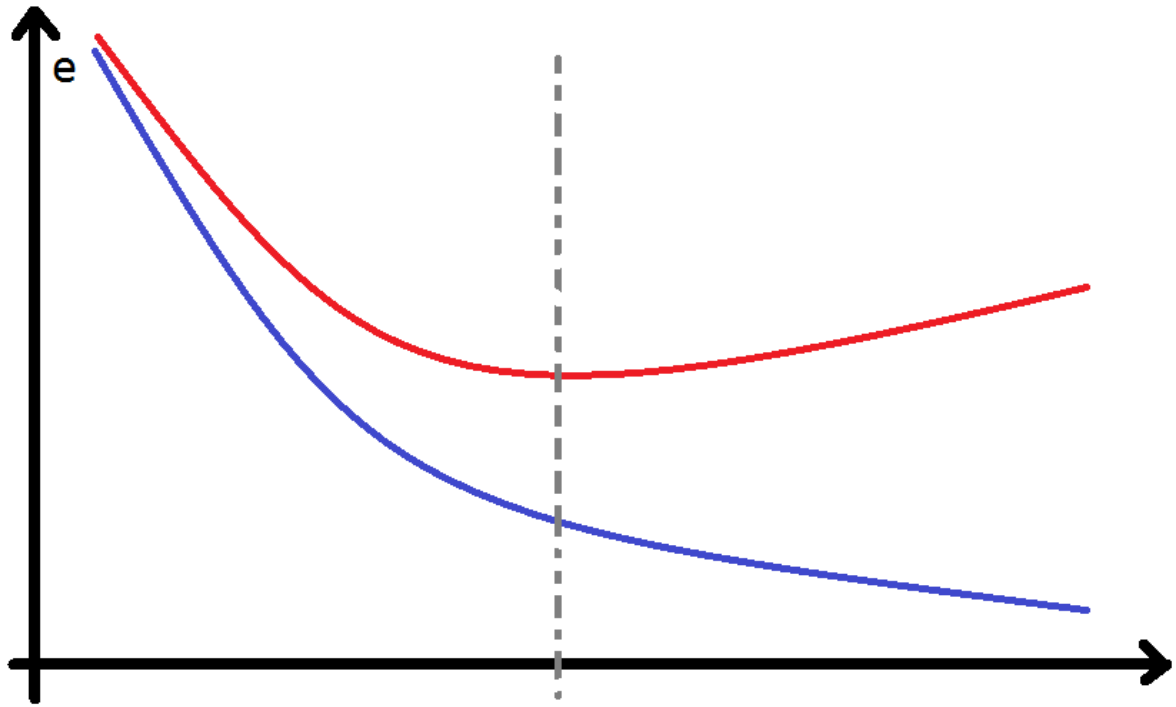
$$sensitivity = \frac{TP}{TP + FN}$$

Miara F1 - średnia harmoniczna precyzji i wrażliwości

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Nadmierne dopasowanie

Nadmierne dopasowanie to zjawisko zachodzące, gdy model klasyfikacji daje gorsze wyniki dla nowych danych niż dla danych trenujących.



Wykres funkcji błędu na próbie losowej i na całej populacji

Nadmierne dopasowanie a ocena

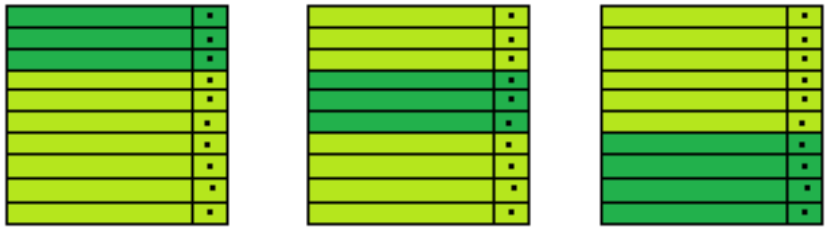
Aby ocena klasyfikacji odnosiła się do całej populacji (a nie tylko znanej próby):

- Model klasyfikacji tworzymy na *zbiorze trenującym*
- Poprawność modelu sprawdzamy na *zbiorze testowym*
- Zbiór trenujący i zbiór testowy powinny być rozłączne

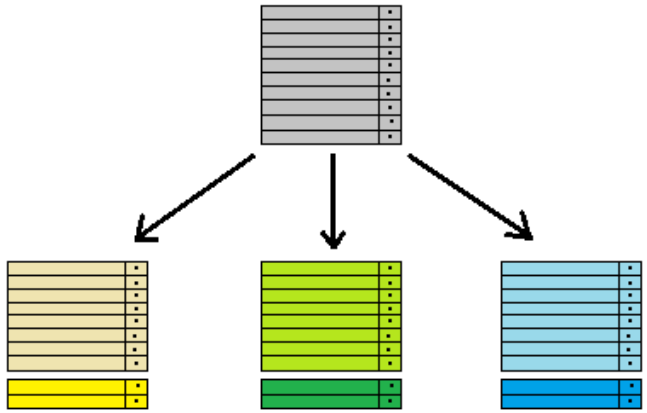
Problem, gdy dysponujemy ograniczonym zbiorem danych:

- Dostępne dane dzielimy na zbiór trenujący i testowy
- Mały zbiór trenujący - nauka modelu na niepełnych danych
- Mały zbiór testowy - duża wariancja oceny
- Rozwiązanie: wiele testów

Testowanie wielokrotne



Sprawdzian krzyżowy



Metoda bootstrap

Przykłady modeli klasyfikacji

Popularne modele klasyfikacji:

- Naiwny Klasyfikator Bayesowski
- Drzewo Decyzyjne
- Losowy las
- Maszyna Wektorów Nośnych

Przetwarzanie wstępne

Cel przetwarzania wstępnego

W kontekście uczenia maszynowego i klasyfikacji, celem wstępnego przetwarzania jest zamiana oryginalnej postaci danych wejściowych na taką, która umożliwia uzyskiwanie lepszych wyników klasyfikacji.

Metody wstępnego przetwarzania danych

- Standaryzacja atrybutów
- Binaryzacja atrybutów
- Normalizacja przykładów
- Skalowanie atrybutów
- Imputacja
- Redukcja wymiarów
- Grupowanie

Brakujące wartości atrybutów

Bardzo często rzeczywiste zbiory danych zawierają brakujące wartości. Algorytmy klasyfikacji często wymagają, aby dane każdego przykładu były kompletne. Możliwe postępowanie:

- Usuwanie przykładów lub
- Imputacja brakujących wartości atrybutów

Brakujące wartości cd.

Problemy z usuwaniem przykładów:

1. Utrata informacji
2. Predykcja niekompletnych przykładów

Brakujące wartości cd.

Imputacja może polegać na zastępowaniu braków:

- Wartością średniej atrybutu
- Wartością mediany atrybutu
- Dominantą atrybutu
- Wynikiem metod sztucznej inteligencji, np. klasyfikacji, regresji

Redukcja wymiarów

Redukcja wymiarów polega na zmniejszeniu liczby atrybutów w danych. Przyczyny:

- Skrócenie czasu obliczeń
- Redukcja szumu

Redukcja wymiarów cd.

Metody:

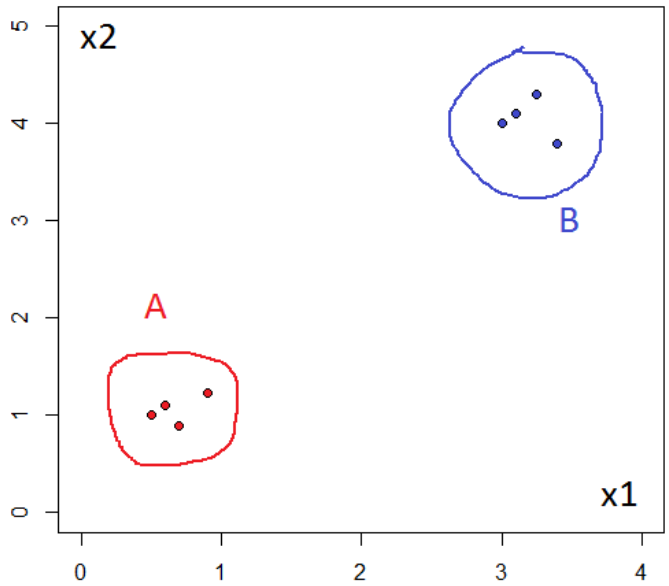
- Arbitralne usuwanie wybranych atrybutów
- PCA (ang. *Principal Component Analysis*) - poszukiwanie kombinacji atrybutów, które wiernie oddają wariancję oryginalnych danych
- Grupowanie atrybutów

Grupowanie

Grupowanie (ang. *data clustering*) jest jednym z przykładów uczenia maszynowego bez nadzoru. Algorytmy grupowania łączą elementy we względnie jednorodne klasy. Grupowanie wykorzystuje ustalone metryki podobieństwa elementów.

Grupowanie: przykład

x1	x2
0.5	1.0
0.9	1.23
3.0	4.0
3.1	4.1
0.7	0.89
3.25	4.3
0.6	1.1
3.4	3.8



x
A
A
B
B
A
B
A
B

Dane przed grupowaniem, ich reprezentacja na płaszczyźnie oraz dane po grupowaniu

Grupowanie cd.

Popularne algorytmy grupowania danych:

- Metoda K-Średnich
- Grupowanie hierarchiczne
- Algorytm DBSCAN

Testy statystyczne

Dlaczego testy statystyczne?

1. Celem pracy jest porównanie metod
2. Porównanie powinno być oparte na ocenie na populacji
3. Dysponujemy próbą, a nie populacją
4. Do porównania wykorzystamy oszacowanie - test statystyczny

Test statystyczny

> Test statystyczny - formuła matematyczna pozwalająca oszacować prawdopodobieństwo spełnienia pewnej hipotezy statystycznej w populacji na podstawie próby losowej z tej populacji.
>
> — Wikipedia

Hipotezą, którą pragniemy obalić, jest jednakowa skuteczność klasyfikacji dla dwóch wybranych metod.

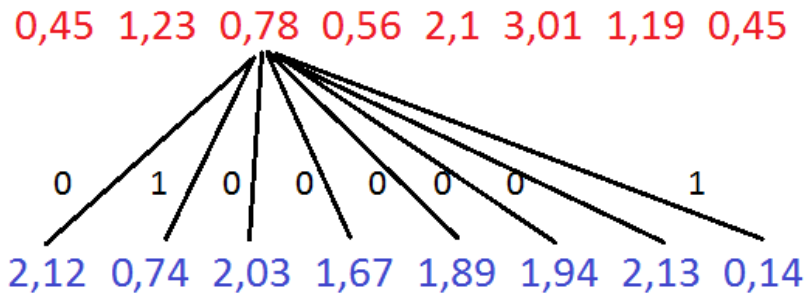
Rodzaje testów

Główny podział:

- Testy parametryczne - zakładają rozkład danych oraz pewne parametry tych rozkładów
- Testy nieparametryczne - nie wymagają założeń odnośnie do rozkładu populacji z której losowana jest próba

Test Manna-Whitneya-Wilcoxona

- Test nieparametryczny
- Obliczanie statystyki *U*
- Dla małych danych używamy tablic
- Dla dużych danych *U* ma rozkład normalny i dlatego możemy obliczyć *p-wartość*



Ilustracja przebiegu testu Manna Whitneya Wilcoxona

Zarys projektowy

Plan badań

1. Zdefiniowanie grup algorytmów A
2. Dla każdej grupy $A_i \in A$:
   1. Dla każdego algorytmu $x_j \in A_i$:
      1. Wyliczenie miar klasyfikacji M_{x_j}
   2. Zbiór ocen $Y_i = \bigcup M_{x_j}$
   3. Porównanie ocen Y_i z ocenami poprzednich grup $Y_k, k \in [1, i)$
3. Analiza wyników porównań i wyciągnięcie wniosków

Etapy w pojedynczym algorytmie

Każdy algorytm składa się z następujących etapów:

1. Pozbywanie się brakujących wartości
2. Usuwanie atrybutów
3. Skalowanie atrybutów
4. Grupowanie atrybutów
5. Klasyfikacja

Każdy etap przyjmuje jakieś parametry, np. zbiór atrybutów do usunięcia, rodzaj klasyfikatora i parametry klasyfikatora.

Definiowanie grup algorytmów

Definiując grupę algorytmów chcemy osiągnąć złoty środek pomiędzy:

1. Skrótową parametryzacją grupy
2. Możliwością zdefiniowania dowolnego zbioru algorytmów w grupie

Grupy algorytmów cd.

Moja propozycja dla:

1. Eliminowania braków: zbiór metod, np. $\{Mean, Median\}$
2. Usuwania i skalowania:
   1. Zbiór atrybutów, np. $\{X_1, X_4, X_5\}$
   2. Zbiór liczności podzbiorów, np. $\{0, 1, 3\}$
3. Grupowania:
   1. Zbiór algorytmów, np. $\{DBSCAN\}$
   2. Zbiór atrybutów, np. $\{X_1, ..., X_6\}$
   3. Zbiór liczb algorytmów używanych jednocześnie, np. $\{2\}$
   4. Maksymalna liczba atrybutów podlegających grupowaniu przez algorytm, np. 2
   5. Granularność, np. 5
4. Klasyfikacji
   1. Zbiór algorytmów, np. $\{RandomForest\}$
   2. Granularność, np. 6

Granularność

Granularnością to liczba elementów wybranych z danego przedziału wartości zgodnie z pewną formułą. Granularność:

- Liniowa
- Logarytmiczna

Czasochłonność obliczeń

1. Bardzo duża liczba kombinacji wewnątrz grup
2. Bardzo czasochłonne obliczenia
3. Jak skrócić czas obliczeń?
    1. Praca na niewielkich grupach algorytmów
    2. Optymalizacje
    3. Rozproszenie obliczeń

Przykład optymalizacji

Niektóre złożoności czasowe poszczególnych etapów:

1. Prosta imputacja: $O(nlogn)$
2. Skalowanie, usuwanie: $O(n)$
3. Grupowanie: $O(n^3)$
4. Klasyfikacja: $O(n^3)$

Z powyższego oraz z algorytmu badań wynika, że warto zapamiętywać dane (*cache*) po etapie grupowania.

Rozproszenie obliczeń

Główną korzyścią płynącą z rozproszenia obliczeń jest skrócenie czasu (proporcjonalne do ilości dodanych zasobów). Należy wziąć pod uwagę:

1. Opóźnienia sieci komputerowej
2. Zawodność maszyn
3. Używanie danych z wcześniejszych etapów algorytmów
4. Skomplikowanie implementacji

Rozproszenie obliczeń: projekt

1. Implementacja:
   1. Leniwe generowanie obiektów reprezentujących konkretny algorytm
   2. Wysyłanie obiektów do maszyn z aplikacją liczącą
2. Wysyłanie obiektów blokami posiadającymi zbliżone wczesne etapy algorytmu
3. Ponowne wysłanie bloku jeżeli nie ma odpowiedzi w ustalonym czasie
4. Architektura klient-serwer oparta na systemie kolejkowym (*Celery, RabbitMQ*)

Wstępne wyniki badań

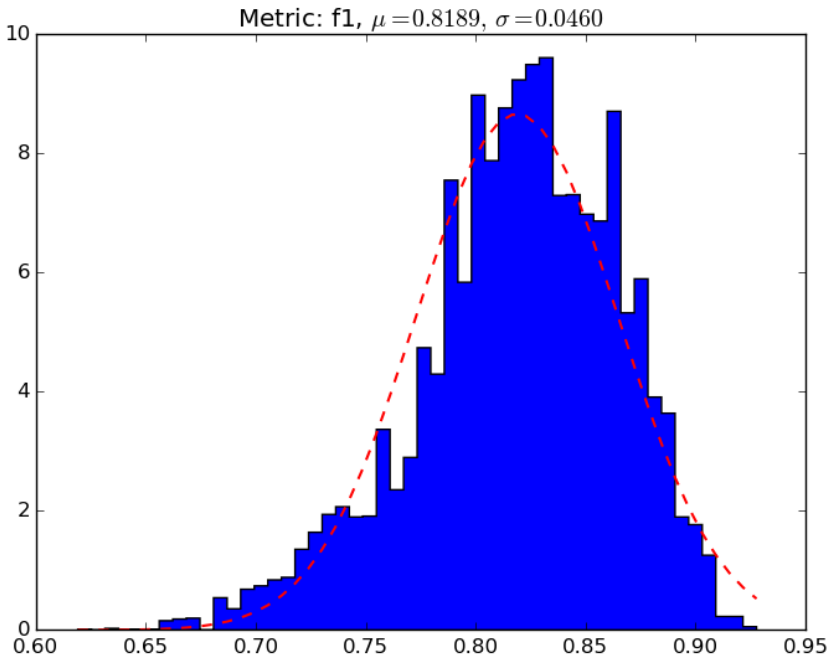
Badane grupy algorytmów

Dla różnych rodzajów klasyfikacji oraz metod usuwania brakujących danych:

1. Bez przetwarzania wstępnego
2. Usuwanie wszystkich możliwych kombinacji atrybutów
3. Skalowanie wszystkich możliwych kombinacji atrybutów
4. Usuwanie i skalowanie atrybutów

Wyniki (miara F1)

Pod kątem miary F1 najlepsza okazała się grupa imputująca wartością średnią, używająca klasyfikatora losowy las oraz skalująca kombinacje atrybutów. Grupa ta okazała się również najlepsza pod kątem precyzji. W rankingu wrażliwości zajęła jednak 10 miejsce.



Histogram miar F1 metody

Wnioski z analizy wyników

1. Do klasyfikacji dostępnych danych najlepszy jest klasyfikator losowy las.
2. Możliwe jest jednoczesne zapewnienie precyzji powyżej 84% oraz wrażliwości powyżej 80%.
3. Grupy używające maszyny wektorów nośnych mają bardzo wysoką wrażliwość (powyżej 93%) ale niską precyzję (ok. 51-60%)
4. Wartość średnia miary nie decyduje o pozycji w rankingu miar, po uwzględnieniu testu statystycznego

Bibliografia


Paweł Cichosz
Systemy Uczące Się
Wydawnictwa Naukowo-Techniczne, Warszawa, 2000


Podręcznik użytkownika biblioteki *scikit-learn*


Portal abczdrowie.pl


Wikipedia

Dziękuję