

Confusion matrix:
(a) (b) <- classifi
80 3 (a): clas
1 51 (b): clas
< > 100.75
Credit > 1087
| Statu DM
| Duration
| | Duration
| Age > good
| | | age <=
| | | Saving -accr
| | | Saving account
Status = less-2C
IF A8 < 10.75 A
AND 5.50 < A1
THEN DECISION
IF
IF
AND A9 = 1
AND A14 < 9
THEN DECISION
IF A3 > 1.79 AN
> 241.50 THEN
IF A6 = X A
THEN DE
IF A 26 8
AND A9 =
AND 2 <
T
@attrib outlo
@attribute temp
@attribute humi
@d sunny, 85
85 use, no, ca
@attrib use {your
presp ic, pres
@attribute spect
{myope, hyperm
@attribute astig

Uczenie maszynowe i sieci neuronowe

Krzysztof Krawiec
Jerzy Stefanowski

Wydawnictwo
Politechniki
Poznańskiej



Spis treści

Przedmowa	5
1. Wybrane pojęcia konstruowania systemów uczących się	7
1.1. Wprowadzenie	7
1.2. Na czym polega uczenie się maszyn?	7
1.3. Problemy rozwiązywane przez systemy uczące się	9
1.4. Reprezentacja przykładów uczących	11
1.5. Podział metod maszynowego uczenia się	16
1.6. Uczenie nadzorowane oraz zasady konstruowania systemów klasyfikujących	18
1.7. Ocena systemów klasyfikujących	21
1.7.1. Miary oceny skuteczności klasyfikowania obiektów	21
1.7.2. Eksperymentalne oszacowanie skuteczności klasyfikatora	24
1.8. Uwagi literaturowe	25
2. Indukcja drzew decyzyjnych	28
2.1. Wprowadzenie	28
2.2. Reprezentacja drzewa decyzyjnego	28
2.3. Algorytm indukcji drzew decyzyjnych	30
2.4. Konstruowanie drzew decyzyjnych dla zróżnicowanych danych	34
2.4.1. Inne miary wyboru atrybutów	35
2.4.2. Binaryzacja drzew decyzyjnych	36
2.4.3. Uwzględnianie niezdefiniowanych wartości atrybutów	37
2.5. Przebieg ćwiczenia	38
3. Tworzenie klasyfikatorów opartych na drzewach decyzyjnych	41
3.1. Wprowadzenie	41
3.2. Zjawisko przeuczenia	41
3.3. Upraszczenie drzew decyzyjnych	43
3.4. Konstruowanie drzew decyzyjnych z wykorzystaniem techniki „okien”	46
3.5. Przebieg ćwiczenia	48
4. Indukcja reguł	52
4.1. Wprowadzenie	52
4.2. Reprezentacja reguł	53
4.3. Algorytmy indukcji reguł	55
4.3.1. Strategie sekwencyjnego pokrywania	55
4.3.2. Algorytm LEM2	57
4.4. Miary oceny reguł	59
4.5. Klasyfikowanie obiektów za pomocą reguł	60
4.6. Przebieg ćwiczenia	63
5. Klasyfikacja bayesowska	65
5.1. Wprowadzenie	65
5.2. Twierdzenie Bayesa	66
5.3. Optymalny klasyfikator bayesowski	68

5.4.	Naiwny klasyfikator bayesowski	68
5.5.	Klasyfikacja dokumentów tekstowych.....	70
5.6.	Przebieg cwiczenia.....	72
6.	Wprowadzenie do sieci neuronowych	82
6.1.	Wprowadzenie	82
6.2.	Geneza sztucznych sieci neuronowych.....	82
6.3.	Reprezentacja wiedzy w sieciach neuronowych.....	82
6.4.	Architektury sieci neuronowych.....	87
6.5.	Sieci neuronowe jako narzędzia maszynowego uczenia sie.....	89
6.6.	Uczenie sieci neuronowych	92
6.7.	Uczenie pojedynczego neuronu liniowego	93
6.8.	Wlasciwosci neuronu liniowego	96
6.9.	Uczenie pojedynczego neuronu nieliniowego	99
6.10.	Przebieg cwiczenia.....	100
7.	Wsteczna propagacja błędu.....	102
7.1.	Wprowadzenie	102
7.2.	Algorytm wstecznej propagacji błędu	102
7.3.	Zjawisko przeuczenia w sieciach neuronowych.....	105
7.4.	Przebieg cwiczenia.....	110
8.	Uczenie warstwowych sieci neuronowych	111
8.1.	Parametryzacja algorytmu wstecznej propagacji błędu	111
8.2.	Odmiany algorytmu wstecznej propagacji błędu	112
8.3.	Interpretowanie decyzji podejmowanych przez siec	113
8.4.	Sieci z jednostkami o symetrii kolowej	115
8.5.	Przebieg cwiczenia.....	117
9.	Sieci neuronowe uczone przez współzawodnictwo	119
9.1.	Wprowadzenie	119
9.2.	Grupowanie przykładów – kwantyzator wektorów	119
9.3.	Mapa odwzorowania cech istotnych (SOM).....	124
9.4.	Inne modele sieci neuronowych.....	127
9.5.	Przebieg cwiczenia.....	127
Dodatek A.	Wykorzystywane oprogramowanie	129
A.1.	Oprogramowanie do indukcji drzew decyzyjnych.....	129
A.2.	Oprogramowanie do indukcji reguł decyzyjnych	133
A.3.	Srodowisko maszynowego uczenia sie WEKA.....	134
A.4.	Srodowisko STATISTICA: Neural Networks	137
Dodatek B.	Stosowane formaty plików.....	140
B.1.	Pliki dla systemu C4.5	140
B.2.	Pliki w formacie ISF	141
B.3.	Pliki w formacie ARFF.....	145
	Spis literatury	147

Przedmowa

Niniejszy skrypt jest przewodnikiem do ćwiczeń i laboratorium z przedmiotu „uczenie maszynowe i sieci neuronowe” nauczanego na studiach magisterskich na kierunku informatyka. Ma on na celu pomóc studentom w zapoznaniu się z wybranymi algorytmami uczenia się z przykładów oraz z ich zastosowaniem do analizy problemów rzeczywistych i dydaktycznych. Zakres omawianych zagadnień obejmuje dwie podstawowe grupy metod, tj.: metody uczenia nadzorowanego tworzące symboliczną reprezentację wiedzy (np. algorytmy automatycznego konstruowania drzew lub reguł decyzyjnych) oraz sztuczne sieci neuronowe. Realizacja ćwiczeń zawartych w skrypcie powinna także przygotować czytelnika do samodzielnego rozwiązywania praktycznych problemów z zakresu konstruowania i wdrażania systemów informatycznych zdolnych do automatycznego ulepszania swojego działania poprzez analizę doświadczenia w danej dziedzinie.

Część omawianych zagadnień może być także przydatna jako materiał uzupełniający do nauczania przedmiotów związanych z metodami sztucznej inteligencji, analiza i eksploracja danych (ang. *data mining*) oraz odkrywaniem wiedzy w bazach danych. Należy tutaj zauważyć, że właśnie algorytmy maszynowego uczenia się stanowią często podstawę metodyczną dla algorytmów odkrywania wiedzy z danych, stosowanych w ostatnich latach w zaawansowanych systemach informatycznych.

Każdy rozdział skryptu, z wyjątkiem pierwszego, będącego rodzajem wstępu metodologicznego do realizacji ćwiczeń, jest skonstruowany z dwóch części. Pierwsza część zawiera syntetyczne przedstawienie podstawowych pojęć związanych z wybraną metodą lub algorytmem. Druga część obejmuje propozycje realizacji ćwiczenia w postaci zestawu zadań. Na ogół pierwsze z proponowanych zadań dotyczy wykonania obliczeń na prostych przykładach. Ma to na celu ułatwienie zrozumienia omawianych pojęć i algorytmów. Dalsze zadania wymagają analizy rzeczywistych zbiorów danych reprezentujących problemy z różnych dziedzin. Dane te są dostępne w plikach zapisanych w odpowiednim formacie i powinny być przetwarzane za pomocą oprogramowania zawierającego gotowe implementacje potrzebnych algorytmów. W skrypcie krótko scharakteryzowano polecane oprogramowanie oraz podano informacje o formatach używanych plików.

Skrypt składa się z dziewięciu rozdziałów i dwóch dodatków. Pierwszy rozdział ma specjalny charakter, gdyż zawiera krótki wstęp metodologiczny do konstruowania i eksploatacji systemów uczących się. Ma on dostarczyć wspólnych pojęć dla algorytmów omawianych w kolejnych rozdziałach skryptu. W szczególności przedstawiono zasady tworzenia i eksperymentalnej oceny systemów klasyfikacyjnych, które są wykorzystywane w dalszych ćwiczeniach. Rozdział 2 jest poświęcony najpopularniejszym algorytmom indukcji drzew decyzyjnych. Następnie w roz-

dziale 3 przedstawiono problemy związane z tworzeniem klasyfikatorów opartych na drzewach decyzyjnych, w tym zjawisko przeuczenia oraz mechanizmy upraszczania drzew. W rozdziale 4 omówiono algorytmy indukcji reguł z przykładów oraz różne miary oceny wygenerowanych reguł. Rozdział 5 jest poświęcony probabilistycznym klasyfikatorom opartym na twierdzeniu Bayesa oraz ich zastosowaniom do automatycznej klasyfikacji dokumentów tekstowych.

W rozdziałach od 6 do 9 zaprezentowano możliwości wykorzystania sztucznych sieci neuronowych jako narzędzi uczenia maszynowego. W szczególności rozdział 6 ma charakter wstępny i wprowadza czytelnika w tematykę, prezentując podstawową architekturę sztucznego neuronu i sieci neuronowych oraz metody uczenia pojedynczego neuronu. W rozdziale 7 omówiono najbardziej popularny model sieci warstwowej i algorytm wstecznej propagacji błędów (ang. *backpropagation*). W rozdziale 8 zaprezentowano różne odmiany algorytmu wstecznej propagacji błędów, a w rozdziale 9 – architektury i algorytmy przeznaczone do nienadzorowanego uczenia sieci neuronowych, w tym sieć odwzorowania cech istotnych (SOM).