Matematyka, studia dzienne, mag II st. semestr II

Eksploracja Danych w Javie **2016/2017**

Prowadzący: dr. Andrzejczak poniedziałek 12:15

Data oddania: _____ Ocena: _____

Norbert Landrat 213518
Adrian Grzelak 213506

Rozpoznawanie podrabianych banknotów

1. Cel projektu

Projekt polegał na nauce rozpoznawania czy badany banknot jest prawdziwy używając dostępnej bazy danych opisaną w rozdziale 2. Baza zawiera 1372 wpisy z których każdy należy do jednego z 2 rodzajów banknotów (prawdziwy i sfałszowany). Wejściem jest pięć parametrów rzeczywistych, a jako wyjście oczekiwano odpowiedzi, czy badany banknot jest sfałszowany. Przeprowadzono badanie, które miało określić, jaka metoda najlepiej rozwiąże ten problem. Pierwsza metoda użyta do badania dostępnej bazy danych to perceptron wielowarstwowy. Następnie wykorzystano klasyfikator SVM. Obie metody zostały szczegółowo opisane w rozdziałach 3 i 4.

2. Opis danych

Rozdział ten będzie poświęcony szczegółowemu opisowi danych, które zostaną poddane klasyfikacji. Dane zostały pobrane z repozytorium UCI (https://archive.ics.uci.edu/ml/datasets/bank
Autorem powyższych danych jest Volke Lohweg (University of Applied Sciences, Ostwestfalen-Lippe, volker.lohweg '@' hs-owl.de), a donatorem Helene Doerksen (University of Applied Sciences, Ostwestfalen-Lippe, helene.doerksen '@' hs-owl.de). Pochodzą z sierpnia 2012 r.

Dane zostały wydobyte ze zdjęć, które zostały zrobione prawdziwym i sfałszowanym banknotom. W celu transformacji danych na postać cyfrową została użyta kamera przemysłowa, która jest najczęściej używana przy inspekcji wydruku banknotów. Zdjęcia mają wymiary 400 x 400 pikseli. Z powodów technicznych (obiektyw kamery, odległość od badanych przedmiotów) zdjęcia zostały robione w odcieniu szarości o rozdzielczości 660 dpi. Aby uzyskać konkretne cechy ze zdjęć została użyta transformata falkowa.

Informacje o atrybutach

1. wariancja zdjęcia po przekształceniu transformatą falkową

2. skośność zdjęcia po przekształceniu transformatą falkową
3. kurtoza zdjęcia po przekształceniu transformatą falkową
4. entropia zdjęcia
5. klasa

Liczba instancji: 1372

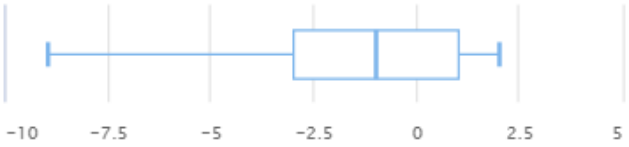
Szczegółowe informacje dotyczące poszczególnych cech

1. wariancja - wartości numeryczne

2. skośność - wartości numeryczne

3. kurtoza - wartości numeryczne

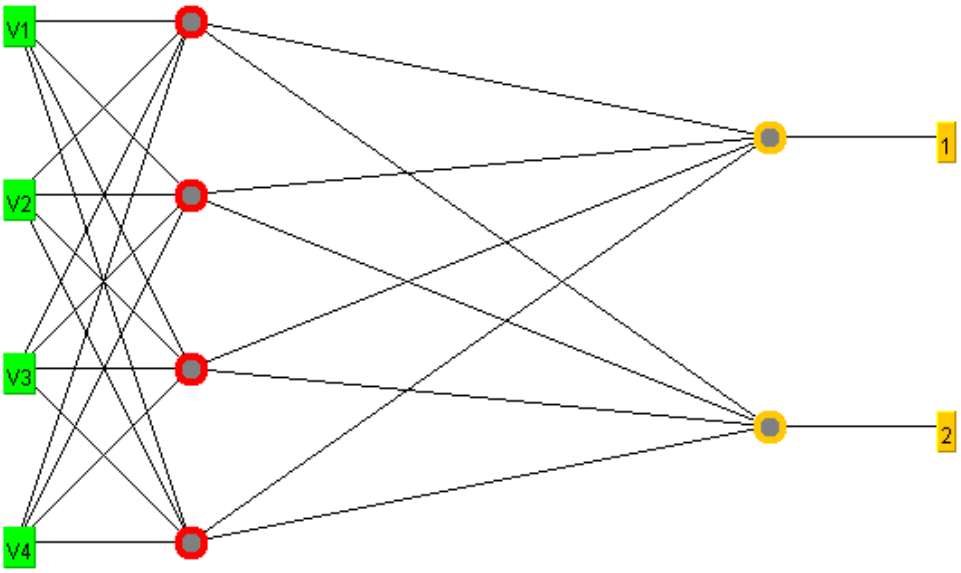
4. entropia - wartości numeryczne



5. klasa – dwie wartości (1 – banknot prawdziwy, 2 – banknot sfałszowany)

3. Perceptron wielowarstwowy

Perceptron wielowarstwowy – prosta sieć neuronowa składająca się z co najmniej dwóch neuronów McCullocha-Pittsa ułożonych warstwowo, implementująca algorytm uczenia nadzorowanego klasyfikatorów binarnych. Perceptron wielowarstwowy jest funkcją, która potrafi określić przynależność parametrów wejściowych do jednej z dwóch klas. W przeciwieństwie do perceptronu jednowarstwowego może być wykorzystywany do klasyfikowania zbiorów, które nie są liniowo separowalne.



Rysunek 1. Otrzymana sieć neuronowa

Dla celów naszego eksperymentu stowrzyliśmy sieć z 4 neuronami w warstwie ukrytej, przyjęliśmy współczynnik momentum 0.2, Współczynnik nauki 0.3. I ustawiliśmy czas uczenia się zbioru na 800 epok. W trakcie procesu uczenia 10% elementow stanowilo zbior walidacyjny.

Perceptron w tak zdefiniowanym procesie zdołał się nauczyć rozpoznawać elementy ze 100% skutecznością!

4. Klasyfikator leniwy k – najbliższych sąsiadów

4.1. Algorytm k-NN

Ustalamy wartość k (najlepiej liczbę nieparzystą, zwykle ok. 5-15). Dla każdego obiektu testowego o*:

1. wyznaczamy odległość r(o*,x) pomiędzy o* i każdym obiektem treningowym x
2. znajdujemy k obiektów treningowych najbliższych o*
3. wśród wartości decyzji odpowiadających tym obiektom wykonujemy głosowanie
4. najczęściej występującą wartość decyzji przypisujemy obiektowi o*

4.2. Uwagi techniczne

Parametr k możemy dobrać eksperymentalnie. Licząc na próbce testowej wyniki dla pewnego k, otrzymujemy przy okazji wyniki dla wszystkich wartości mniejszych. Czas uczenia (w wersji podstawowej algorytmu) jest bardzo krótki, gdyż nauka polega na zapamiętaniu całej próbki treningowej. Łatwo stosować metodę leave-one-out. Klasyfikacja nowych przypadków jest dosyć powolna. Sposoby na przyspieszenie:

1. selekcja obiektów – wybór pewnego podzbioru dającego zbliżone wyniki klasyfikacji
2. podział zbioru obiektów na podzbiory i przeszukiwanie tylko niektórych z nich.

W przypadku badania klasyfikatorem kNN kroswalidacja dzielona na 10 podzbiorów tylko w jednym przypadku daje 100

5. Istotność atrybutów

Istotność atrybutów została obliczona za pomocą miary relief. Miara Relief to wynik działania algorytmu wyznaczającego relatywną ważność atrybutów. Ocenia jak dobrze poszczególne atrybuty nadają się do przewidywania wartości jednego wybranego atrybutu binarnego, tzw. atrybutu decyzyjnego. Poniżej zaprezentowany został ranking istotności atrybutów przy użyciu domyślnych parametrów, tzn. Attribute Evaluator – ReliefFAttributeEval o liczbie sąsiadów – 10, Search Method – Ranker.

```
=== Run information ===

Evaluator:    weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     banknote-authentication
Instances:    1372
Attributes:   5
              Variance
              Skewness
              Curtosis
              Entropy
              Class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 Class):
ReliefF Ranking Filter
Instances sampled: all
Number of nearest neighbours (k): 10
Equal influence nearest neighbours

Ranked attributes:
 0.1272  1 Variance
 0.1006  2 Skewness
 0.0647  3 Curtosis
 0.0213  4 Entropy

Selected attributes: 1,2,3,4 : 4
```

6. Wnioski z badania

Analizowany zbiór jest bardzo trafnie dobrany do celów klasyfikowania. Dane w przypadku obu metod są klasyfikowane z wysoką skutecznością (dochodząc nawet do 100%), co może oznaczać, że metoda sprawdzania sfałszowanych banknotów może znaleźć odzwierciedlenie w rzeczywistości. W metodzie k – najbliższych sąsiadów zauważalny jest wpływ liczby sąsiadów na skuteczność kroswalidacji. Począwszy od 11 sąsiadów wraz ze wzrostem parametru skuteczność sklasyfikowanych instancji maleje. W procesie nauczania perceptronu z kolei duży wpływ na osiągane wyniki ma ilość neuronów w wartswie ukrytej. Zbyt mała ich ilość może doprowadzić do słabego nauczenia się wzorca, natomiast zbyt duża do zjawiska przeuczenia (Perceptron doskonale rozpoznaje elementy ze zbioru nauczającego, ale natrafia na problemy przy danych pochodzących spoza tego zbioru.