**WYŻSZA SZKOŁA INFORMATYKI**

**I ZARZĄDZANIA**

Z SIEDZIBĄ W RZESZOWIE

Sprawozdanie

Sztuczna inteligencja

**KNN**

Prowadzący: dr inż. Mariusz Wrzesień Wykonawca: Sylwia Babiarz 46741

6IID – GAK, SL04

Rzeszów 2014

# Opis problemu

Celem laboratorium jest identyfikacja nieznanych obiektów, zamieszczonych w zbiorze treningowym, przy pomocy wcześniej stworzonego zbioru uczącego. Taka metoda polega na odnalezieniu podobnych obiektów dla konkretnego przypadku. Do tego celu wykorzystane zostaną trzy metryki, przy pomocy których będą obliczone odległości pomiędzy obiektami.

Do klasyfikacji obiektów zostanie użyty algorytm k najbliższych sąsiadów dla jednego, trzech oraz pięciu sąsiadów. Następnie wyniki zostaną poddane trzem metodą głosowania: metodzie normalnej, metodzie sumy odwrotności kwadratów odległości oraz sumy odległości.

Kolejny krok ma na celu sprawdzenie, jak otrzymane wyniki nawiązują do decyzji zbioru testującego.

# Badane zbiory danych

Zbiór uczący posiada dwadzieścia dwa przypadki oraz dotyczy problemu doboru soczewek. Kolejne z kolumn są indeksowane nazwami atrybutów takimi jak ,,wiek”, ,,wada wzroku”, ,,astygmatyzm”, „łzawienie”, oraz kolumna decyzji ,,soczewki”. Wszystkie wartości atrybutów są liczbowe, ponieważ tylko wtedy można obliczać metryki.

Tabela 1 Zbiór uczący

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LP | Wiek | Wada wzroku | Astygmatyzm | Łzawienie | Soczewki |
| 1 | 1 | 2 | 1 | 1 | 3 |
| 2 | 3 | 2 | 2 | 1 | 3 |
| 3 | 3 | 4 | 1 | 1 | 2 |
| 4 | 3 | 1 | 5 | 2 | 3 |
| 5 | 2 | 2 | 2 | 2 | 1 |
| 6 | 2 | 2 | 2 | 7 | 3 |
| 7 | 1 | 2 | 2 | 2 | 3 |
| 8 | 3 | 1 | 4 | 2 | 3 |
| 9 | 2 | 3 | 6 | 1 | 3 |
| 10 | 3 | 2 | 0 | 2 | 2 |
| 11 | 2 | 2 | 1 | 8 | 3 |
| 12 | 5 | 2 | 4 | 2 | 1 |
| 13 | 1 | 6 | 1 | 1 | 3 |
| 14 | 3 | 1 | 6 | 1 | 1 |
| 15 | 2 | 1 | 1 | 6 | 3 |
| 16 | 2 | 1 | 5 | 5 | 3 |
| 17 | 3 | 2 | 5 | 2 | 2 |
| 18 | 1 | 4 | 5 | 1 | 3 |
| 19 | 1 | 3 | 2 | 1 | 3 |
| 20 | 1 | 1 | 9 | 1 | 1 |
| 21 | 1 | 1 | 2 | 3 | 3 |
| 22 | 3 | 2 | 2 | 9 | 3 |

Zbiór treningowy, dla którego zostaną wykonane obliczenia, składa się z czterech przypadków należących do różnych klas. Są to wyżej opisane, tak zwane, nieznane obiekty.

Tabela 2 Zbiór testujący

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LP | wiek | wada | astyg | łzaw. | Socz. |
| 1 | 1 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 2 | 2 | 1 |
| 3 | 2 | 1 | 1 | 2 | 2 |
| 4 | 2 | 2 | 2 | 2 | 3 |

# Obliczenia

Program, w jakim wykonano wszelkie obliczenia to Microsoft Office Excel. Po zaimportowaniu zbioru uczącego oraz testującego przystąpiono do obliczania odległości, za pomocą trzech metryk, dla każdego z przypadków.

Pierwszą z nich jest metryka City Block. Ta odległość jest sumą różnic mierzonych wzdłuż wymiarów, a jej wzór przedstawia się następująco:

Następną użytą metryką jest odległość Euklidesa, czyli odległość geometryczna w przestrzeni wielowymiarowej, dzięki której można wyliczyć najkrótszą drogę:

Ostatnią użytą jest metryka Kombinowana. Polega ona na dodaniu do siebie dwóch wyżej wymienionych metryk.

Po wyliczeniu wszystkich wartości metryk należy zastosować algorytm k najbliższych sąsiadów (1,3,5). Aby to zrobić, dane zbioru uczącego muszą być posortowane według konkretnej metryki, względem wartości, od najmniejszej do największej. Kolejny krok to wybranie metod głosowania. W powyższym sprawozdaniu, przeanalizowano:

* metodę normalna, gdzie wybór pada na tą klasę, dla której jest najwięcej przypadków pośród k najbliższych sąsiadów,
* metodę sumy odwrotności kwadratów odległości, gdzie wygrywa ta klasa, która ma największą wagę głosów,
* metodę sumy odległości, gdzie wygrywa ta klasa, która ma najmniejszą wagę głosów.

Poniżej zaprezentowano obliczenia dla przypadku pierwszego:

* metryka City Block – z kolumn indeksowanych tymi samymi nazwami liczono sumę wartości bezwzględną różnicy wartości ze zbioru testującego (t) od uczącego (u):

D = |Wiekt – Wieku| + |Wadat – Wadau| + |Astygmatyzmt – Astygmatyzmu| + |Łzawieniet – Łzawienieu| = |1-1| + |1 – 2| + |1- 1| + |1-1| = 1

Gdy wszystkie 22 przypadki zostały w ten sposób policzone, tablica została posortowana i wyłonione zostały wartości dla sąsiadów.

Tabela 3 Tabela przedstawiająca jednego, trzech i pięciu sąsiadów

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Soczewki | CityBlock | Soczewki | Euklidesa | Soczewki | kombinowana |
| 3 | 1 | 3 | 1 | 3 | 2 |
| 3 | 3 | 3 | 1,732051 | 3 | 4,732050808 |
| 3 | 3 | 1 | 2 | 3 | 5,236067977 |
| 3 | 3 | 3 | 2,236068 | 3 | 5,236067977 |
| 3 | 4 | 3 | 2,236068 | 1 | 6 |
| 1 | 4 | 3 | 2,44949 | 3 | 6,449489743 |
| 2 | 5 |
| 2 | 5 |
| 3 | 5 |
| 3 | 6 |
| 3 | 6 |

Głosowanie metodą normalną:

* dla 1 sąsiada obiekt został zaklasyfikowany do klasy 3,
* dla 3 sąsiadów (pary (3,1), (3,3), (3,3), (3,3), (3,4), (1,4)) obiekt został zaklasyfikowany do klasy 3, ponieważ klasa 1 występuje tylko raz,
* dla pięciu sąsiadów (cała powyższa tabela) także wygrywa klasa 3, ponieważ klasa 1 występuje tylko raz, a klasa 2 tylko dwa razy.

Głosowanie metodą sumy odwrotności kwadratów odległości:

* dla 1 sąsiada – klasa 3
* dla 3 sąsiadów:
  + waga głosów klasa 1 =
  + waga głosów klasa 3 =

Ponieważ waga głosów klasy 3 jest największa, to ona zostaje wybrana.

* Dla 5 sąsiadów analogicznie jak dla trzech:
  + waga głosów klasa 1 =
  + waga głosów klasa 2 =
  + waga głosów klasa 3 = 1,49139

Obiekt przydzielono do klasy 3.

Głosowanie metodą sumy odległości:

* dla 1 sąsiada – klasa 3.
* dla 3 sąsiadów:
  + klasa 1 = 4
  + klasa 3 =

Ponieważ waga głosów klasy 3 jest mniejsza, to obiekt do niej zostaje przypisany.

* dla 5 sąsiadów:
  + klasa 1 = 4
  + klasa 2 =
  + klasa 3 =

Ponieważ waga głosów klasy 3 jest mniejsza, to obiekt do niej zostaje przypisany.

* Metryka Euklidesa – polega na tym samym, co dystans City Block, z tym że podnosi się ją do kwadratu, aby przypisać większą wagę obiektom, które są bardziej oddalone. Dla przypadku pierwszego:

Głosowanie metodą normalną:

* dla 1 sąsiada: klasa 3,
* dla 3 sąsiadów: klasa 3,
* dla 5 sąsiadów: klasa 3.

Głosowanie metodą sumy odwrotności kwadratów odległości:

* dla 1 sąsiada: klasa 3,
* dla 3 sąsiadów: klasa 3,
* dla 5 sąsiadów: klasa 3.

Głosowanie metodą sumy odległości:

* dla 1 sąsiada: klasa 3,
* dla 3 sąsiadów: klasa 3,
* dla 5 sąsiadów: klasa 3.
* Metryka Kombinowana – polega na dodaniu do siebie, dwóch wcześniej wyliczonych wartości z poprzednich metryk:

Wyniki obliczeń dla każdej z metod głosowania uzyskano takie jak wcześniej. Każda jedna metoda dała odpowiedź, że obiekt powinien być przypisany do klasy 3.

# Wyniki

Tabela 4 Podsumowanie wyników



W powyższej tabeli zostały zebrane wszystkie wyniki obliczeń. Wartość 1 oznacza, że przypadek ze zbioru testowego został sklasyfikowany dobrze, wartość 0 oznacza, że obiekt został sklasyfikowany źle, natomiast wartość sprzeczność oznacza, że według metryki obiekt może należeć do dwóch klas jednocześnie.

# Wnioski

Szeroki wybór metryk, metod głosowania oraz algorytm k najbliższych sąsiadów daje wiele możliwości analizy i klasyfikacji danych.

Dla powyższego zbioru tylko pierwszy obiekt zbioru testującego został sklasyfikowany poprawnie przez wszystkie parametry.

W drugim obiekcie sprawdziła się tylko metoda głosowania sumy odległości dla pięciu sąsiadów. Inne dawały błędne odpowiedzi, a nawet sprzeczne. Sprzeczną klasyfikację uzyskano w metryce City Block dla sumy odległości dla trzech sąsiadów. Pod uwagę należy wziąć to, że w zbiorze uczącym na 22 przypadki tylko 4 z nich miały klasę 1.

Obiekt trzeci nie został sklasyfikowany poprawnie żadnym sposobem. Może to wynikać z faktu, że w zbiorze uczącym są tylko trzy przypadki z przypisaną klasą drugą. Mogły być one za „daleko” by wyszły poprawne wyniki.

Ostatni z obiektów w zbiorze testującym miał wpisaną klasę 3. Najwięcej przypadków w zbiorze uczącym posiadało taką klasę. Mimo to, w wielu obliczeniach obiekt został sklasyfikowany błędnie. Dobra klasyfikacja jest dla metody normalnej oraz dla metody sumy odwrotności kwadratów odległości dla trzech i pięciu sąsiadów.

Na dołączone do sprawozdania zbiory większy wpływ miały metody głosowania i liczba sąsiadów użytych w algorytmie k najbliższych sąsiadów niżeli stosowana metryka.