Zadanie 1 i 2 - Raport

Łukasz Dworakowski

Krzysztof Dziemiańczuk

Część teoretyczna:

**Dyskretyzacja** - zamiana ciągłych wartości zmiennej losowej na wartości dyskretne. Oznacza to podzielenie danych zebranych dla pewnego atrybutu na określoną liczbę przedziałów. Dyskretyzacja może zostać dokonana w założeniu o podaną przez użytkownika liczbę przedziałów. Wtedy uzyskuje się przedziały o równej długości. Kolejną metodą jest preferowanie najliczniejszych klas. Najliczniejsza wartość zostaje przyporządkowana do przedziału 1, kolejna pod względem liczebności do przedziału 2 itd., aż do osiągnięcia zadanej przez użytkownika maksymalnej ilości przedziałów n. Wtedy wszystkie następne liczebności zostaną przyporządkowane do przedziału n+1.

**Normalizacja** - polega na podzieleniu danej wartości pomniejszonej o średnią dla całego atrybutu do którego należy ta wartość przez odchylenie standardowe. Odchyleniem standardowym #TODO

*Metryki oceny odległości*:

* **odległość Euklidesowa** -  jest to odległość na płaszczyźnie pomiędzy dwoma punktami podanymi w argumentach funkcji jako wektory. Oblicza się ją jako pierwiastek z sumy kwadratów odległości w poszczególnych wymiarach.
* **odległość Mahalanobisa** -  potocznie nazywa się ją ważoną odległością euklidesową. Macierzą wag jest odwrócona macierz kowariancji obliczona dla całego zbiory danych.
* **metryka Manhattan** -  odległością w tej metryce jest suma bezwzględnych różnic współrzędnych dwóch podanych punktów.
* **metryka L\_nieskończoność** -  inaczej zwana odległością Czebyszewa. Odnosząc się do szachownicy jest to odległość między dwoma polami, wyrażona w ruchach jakie musi wykonać figura króla.[[1]](#footnote-1)

Część techniczna:

Graficzny interfejs użytkownika został stworzony w oparciu o framework Kivy. Jest to multiplatformowa biblioteka umożliwiająca uruchamianie aplikacji na takich systemach jak Windows, MacOS, Linux, Android, iOS,

Metody służące do dyskretyzacji danych zostały napisane natywnie, tzn. bez wykorzystania żadnych zewnętrznych framework'ów, czy bibliotek.

Do implementacji metryk oceny odległości, tj. odległość Euklidesowa, odległość Mahalanobisa, metryka Manhattan, metryka L\_nieskończoność została wykorzystana zewnętrzna biblioteka pythona o nazwie SCIPY. Oferuje ona odpowiednie metody, które po podaniu wektorów o dowolnej długości n, pozwala obliczyć odległości pomiędzy punktami o współrzędnych w przestrzeni n-wymiarowej.

Wyniki eksperymentów:

- dyskretyzacja (przedziały równej długości) - zbiór arrhythmia.txt, atrybut 1 (wiek pacjenta), 10 i 5 przedziałów

- dyskretyzacja (preferowanie najliczniejszych klas) - zbiór arrhythmia.txt, klasa decyzyjna, 6 i 3 klasy

- standaryzacja (normalizacja) - zbiór irisdat.txt, atrybut 2 (szerokość listków)

- standaryzacja (przedział min-max) - zbiór income.txt, atrybut 1 i 2, przedział 0-10

- klasyfikacja k-nn - zbiory irisdat.txt, income.txt, arrhythmia.txt, - pokazać na wykresie zależność jakości klasyfikatora mierzonej metodą leave-one-out od ilości sąsiadów dla różnych metryk oceny odległości, ilość sąsiadów 1 do n-1, gdzie n-liczność zbioru danych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| wynik | metoda | parametry | k | e |
|  | | | | |
| 0.326666666667 60 1 | MAHALANOBIS | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 10 | 40 |
| 0.326666666667 50 1 | MAHALANOBIS | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 70 | 40 |
| 0.326666666667 70 1 | MAHALANOBIS | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 50 | 40 |
| 0.326666666667 50 1 | MAHALANOBIS | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 60 | 40 |
| 0.326666666667 50 1 | MAHALANOBIS | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 80 | 40 |
| 32,66666667 | MANHATAN | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 10 | 40 |
| 0.313333333333 | MANHATAN | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 70 | 40 |
| 0.32 | MANHATAN | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 50 | 40 |
| 0.326666666667 | MANHATAN | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 60 | 60 |
| 0.32 | MANHATAN | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 80 | 60 |
| 33,33333333 | EUCLIDES | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 10 | 40 |
| 32,66666667 | EUCLIDES | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 70 | 40 |
| 0.32 | EUCLIDES | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 50 | 40 |
| 0.326666666667 | EUCLIDES | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 50 | 40 |
| 33,33333333 | CZEBYSZEW | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 10 | 40 |
| 0.313333333333 | CZEBYSZEW | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 70 | 40 |
| 0.313333333333 | CZEBYSZEW | ["LISSZE","LISDLG","PLADLG", "PLASZE"] | 50 | 40 |
| 0.340909090909 | MAHALANOBIS | ["Aktywa","Przych"] | 50 | 40 |
| 0.333333333333 | MAHALANOBIS | ["Aktywa","Przych"] | 60 | 40 |
| 0.333333333333 | MAHALANOBIS | ["Aktywa","Przych"] | 70 | 40 |
| 0.371212121212 | MAHALANOBIS | ["Aktywa","Przych"] | 2 | 70 |
| 0.371212121212 | MAHALANOBIS | ["Aktywa","Przych"] | 3 | 70 |
| 0.371212121212 | MAHALANOBIS | ["Aktywa","Przych"] | 4 | 70 |
| 0.356060606061 | MANHATAN | ["Aktywa","Przych"] | 50 | 70 |
| 0.333333333333 | MANHATAN | ["Aktywa","Przych"] | 60 | 70 |
| 0.348484848485 | MANHATAN | ["Aktywa","Przych"] | 70 | 70 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 5 | 50 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 3 | 50 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 4 | 50 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 6 | 50 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 7 | 50 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 10 | 50 |
| 0.378787878788 | MANHATAN | ["Aktywa","Przych"] | 8 | 50 |
| 0.371212121212 | MANHATAN | ["Aktywa","Przych"] | 2 | 70 |
| 0.371212121212 | MANHATAN | ["Aktywa","Przych"] | 3 | 70 |
| 0.371212121212 | MANHATAN | ["Aktywa","Przych"] | 4 | 70 |
| 0.356060606061 | EUCLIDES | ["Aktywa","Przych"] | 50 | 40 |
| 0.333333333333 | EUCLIDES | ["Aktywa","Przych"] | 60 | 40 |
| 0.333333333333 | EUCLIDES | ["Aktywa","Przych"] | 70 | 40 |
| 0.378787878788 | EUCLIDES | ["Aktywa","Przych"] | 3 | 50 |
| 0.378787878788 | EUCLIDES | ["Aktywa","Przych"] | 5 | 50 |
| 0.378787878788 | EUCLIDES | ["Aktywa","Przych"] | 2 | 50 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 2 | 70 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 3 | 70 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 4 | 70 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 6 | 70 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 8 | 70 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 9 | 70 |
| 0.371212121212 | EUCLIDES | ["Aktywa","Przych"] | 10 | 70 |
| 0.340909090909 | CZEBYSZEW | ["Aktywa","Przych"] | 50 | 40 |
| 0.325757575758 | CZEBYSZEW | ["Aktywa","Przych"] | 60 | 40 |
| 0.333333333333 | CZEBYSZEW | ["Aktywa","Przych"] | 70 | 40 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 3 | 50 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 5 | 50 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 4 | 50 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 2 | 50 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 6 | 50 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 7 | 50 |
| 0.378787878788 | CZEBYSZEW | ["Aktywa","Przych"] | 1 | 50 |
| 0.371212121212 | CZEBYSZEW | ["Aktywa","Przych"] | 8 | 70 |
| 0.371212121212 | CZEBYSZEW | ["Aktywa","Przych"] | 9 | 70 |
| 0.371212121212 | CZEBYSZEW | ["Aktywa","Przych"] | 10 | 70 |

Wnioski:

Jak wiemy odnośnie dyskretyzacji, dane mogą być kontrolowane albo niekontrolowane w zależności, od tego czy posiadają informacje o klasie. Podobnie, kontrolowane metody dyskretyzacji biorą pod uwagę klasę, natomiast niekontrolowane metody tej informacji nie wykorzystują. Dyskretyzacja niekontrolowana widoczna jest w metodach prymitywnych takich jak równa szerokość i równa częstość. W metodach niekontrolowanych, obszary ciągłe są dzielone przez użytkownika na przedziały z góry określoną szerokością albo częstością (jednakowa ilość przykładów w każdym przedziale).

Jednak są to metody, które przy niejednolitych wartościach ciągłych nie dają dobrych wyników. Ponadto narażone są na wartości izolowane, tj. wartości znacznie oddalone od pozostałych, mogące wynikać z przekłamań, które znacząco wpływają na zasięg.

Stosując metoda leave-one-out musimy się liczyć z tym, że jest to proces dosyć czasochłonny, im większa ilość obiektów wejściowych tym skokowo wzrasta nam koszt wszystkich iteracji. Wadą tego sposobu jest to, ze każda próbka będąca czy to w części testowej czy treningowej jest użyta tylko raz. Ponadto realny błąd, który możemy otrzymać w momencie estymacji będzie zakłamany, ponieważ mniejsza, a nie większa część zbioru jest używana do testów.

1. http://pl.wikipedia.org/wiki/Odleg%C5%82o%C5%9B%C4%87\_Czebyszewa, [dostęp: 12.04.2014] [↑](#footnote-ref-1)