**WYŻSZA SZKOŁA INFORMATYKI**

**I ZARZĄDZANIA**

Z SIEDZIBĄ W RZESZOWIE

Sprawozdanie

Sztuczna inteligencja

**Drzewa decyzji C4.5**

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

6IID – GAK, SL04

Rzeszów 2014

# Opis problemu

Celem laboratorium jest przygotowanie drzewa decyzji na podstawie danych dotyczących problemu doboru soczewek. Drzewa jest to alternatywny sposób reprezentacji wiedzy zrozumiałej zarówno dla człowieka jak i komputera. Drzewo składa się z korzenia, gałęzi, liści i węzłów. Korzeń to wybrany atrybut, z którego wychodzą gałęzie reprezentujące jego wartości. Gałęzie prowadzą do węzłów bądź liści. Węzłami nazywa się wierzchołki, z których wychodzi co najmniej jedna krawędź, pozostałe to liście. W każdym węźle sprawdzany jest warunek, na podstawie którego wybierana jest jedna z gałęzi prowadząca do kolejnego wierzchołka. Klasyfikacja polega na przejściu od korzenia do liścia, w którym znajduje się informacja do jakiej klasy obiekt jest zaliczony.

Ponadto należy obliczyć średnią liczbę pytań dla drzewa przedstawionego w badanych zbiorach danych.

# Badane zbiory danych

Dane, jakie należy przetworzyć na drzewo decyzji, dotyczą problemu doboru soczewek. Są one przedstawione za pomocą tablicy decyzji, której skład to pięć kolumn i dwadzieścia dwa wiersze. Wszystkie wiersze to przypadki, jakie są rozpatrywane. Kolumna pierwsza jest indeksowana atrybutem porządkowym o nazwie „wiek”. Zawiera trzy wartości atrybutów, tj. młody, starczy i prestarczy. Kolumna druga indeksowana jest atrybutem nominalnym „wada wzroku”. Posiada dwie wartości, dalekowidz oraz krótkowidz. Kolejna kolumna ma nazwę atrybutu nominalnego „astygmatyzm” i dysponuje także dwoma wartościami, tak i nie. Przedostatnia z kolumn indeksowana jest atrybutem porządkowym „łzawienie” wraz z dwoma wartościami, normalne i zmniejszone. Ostatnia kolumna ma nazwę atrybutu „soczewki” i jest to kolumna decyzji. Składa się z trzech klas: brak, miękkie, twarde.

Tabela 1 Tablica decyzji dotycząca problemu doboru soczewek

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LP | Wiek | Wada\_wzroku | Astygmatyzm | Łzawienie | Soczewki |
| 1 | mlody | dalekowidz | nie | zmniejszone | brak |
| 2 | mlody | krotkowidz | tak | zmniejszone | brak |
| 3 | prestarczy | krotkowidz | tak | zmniejszone | brak |
| 4 | prestarczy | krotkowidz | nie | normalne | miekkie |
| 5 | mlody | krotkowidz | tak | normalne | twarde |
| 6 | starczy | dalekowidz | tak | zmniejszone | brak |
| 7 | prestarczy | dalekowidz | nie | zmniejszone | brak |
| 8 | prestarczy | dalekowidz | tak | zmniejszone | brak |
| 9 | prestarczy | krotkowidz | nie | zmniejszone | brak |
| 10 | mlody | dalekowidz | tak | zmniejszone | brak |
| 11 | starczy | krotkowidz | tak | normalne | twarde |
| 12 | prestarczy | dalekowidz | nie | normalne | miekkie |
| 13 | starczy | dalekowidz | tak | normalne | brak |
| 14 | starczy | krotkowidz | tak | zmniejszone | brak |
| 15 | starczy | krotkowidz | nie | normalne | brak |
| 16 | prestarczy | krotkowidz | tak | normalne | twarde |
| 17 | mlody | krotkowidz | nie | zmniejszone | brak |
| 18 | starczy | krotkowidz | nie | zmniejszone | brak |
| 19 | starczy | dalekowidz | nie | zmniejszone | brak |
| 20 | mlody | dalekowidz | nie | normalne | miekkie |
| 21 | starczy | dalekowidz | nie | normalne | miekkie |
| 22 | prestarczy | dalekowidz | tak | normalne | brak |

Zdjęcie 1 Drzewo decyzji



Powyższe drzewo decyzji dotyczy kwiatów irysa. Zdjęcie przedstawia tekstowy zapis drzew.

# Obliczenia

Na przykładzie bazy informacyjnej z problemem doboru soczewek można przedstawić kroki tworzenia drzew decyzji. Poniżej zostanie przedstawiony sposób odnalezienia korzenia. Wszystkie pozostałe operacje są umieszczone w pliku Microsoft Office Excel o nazwie LAB05.

Pierwszym krokiem jest podział zbioru uczącego na podzbiory. Ich ilość jest zgodna z liczebnością atrybutów występujących w tablicy decyzji. Przedstawiają się one następująco:

krótkowidz

dalekowidz

prestarczy

starczy

młody

normalne

zmniejszone

nie

tak

Zgodnie z podziałem należy policzyć entropię informacji dla poszczególnych wartości wszystkich atrybutów. Entropia to średnia liczba bitów potrzebna do zakodowania decyzji d dla losowo wybranego obiektu ze zbioru S, a jej wzór wygląda następująco:

Kolejno dla atrybutów wykonano obliczenia:

Dla Wiek = młody

Info([1,1,4])= entropy(1/6, 1/6, 4/6) = -1/6 log2(1/6) -1/6 log2(1/6) -4/6 log2(4/6)= 1,25 bitów

Dla Wiek = prestarczy

Info([2,1,5]) =entropy(2/8, 1/8, 5/8)= -2/8 log2(2/8) -1/8 log2(1/8) -5/8 log2(5/8)= 1,30 bitów

Dla Wiek = starczy

Info([1,1,6])=entropy(1/8, 1/8, 6/8) = -1/8 log2(1/8) -1/8 log2(1/8) -6/8 log2(6/8)= 1,06 bitów

Wartość dla atrybutu Wiek

Info([1,1,4],[2,1,5],[1,1,6])=(6/22)\*1,25 + (8/22)\*1,30 + (8/22)\*1,06= 1,20 bitu

Wartość dla całej tablicy decyzji

Info([4,3,15])=entropy(4/22,3/22,15/22)=-4/22 log2(4/22) -3/22 log2(3/22) -15/22 log2(15/22) =1,22bitu

Później miara entropii jest wykorzystywana w obliczaniu zysku informacji (Information Gain), który stanowi kryterium wyboru atrybutu do korzenia. Jest on definiowany:

Information Gain = informacja przed podziałem – informacja po podziale

Gain(„Wiek”)= Info([4,3,15])- Info([1,1,4], [2,1,5], [1,1,6])=1,22 – 1,20 = 0,02 bitu

Analogiczne obliczenia wykonujemy dla pozostałych atrybutów. Poniższe tabele prezentują wyniki obliczeń:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Wiek** | Młody | Prestarczy | Starczy | Klasa |
| ilość |  | 6 | 8 | 8 | 22 |
| dla atrybutu | **1,20** |  |  |  |  |
| GAIN | **0,02** | **1,25** | **1,30** | **1,06** | **1,22** |
|  | miękkie | 1 | 2 | 1 | 4 |
|  | twarde | 1 | 1 | 1 | 3 |
|  | brak | 4 | 5 | 6 | 15 |
|  |  |  |  |  |  |
|  | **Wada** | dalekowidz | krótkowidz |  | Klasa |
| ilość |  | 11 | 11 |  | 22 |
| dla atrybutu | **0,79** |  |  |  |  |
| Gain | **0,43** | **0,85** | **0,73** |  | **1,22** |
|  | miękkie | 3 | 1 |  | 4 |
|  | twarde | 0 | 3 |  | 3 |
|  | brak | 8 | 7 |  | 15 |
|  |  |  |  |  |  |
|  | **Astyg.** | tak | nie |  | Klasa |
| ilość |  | 11 | 11 |  | 22 |
| dla atrybutu | **0,90** |  |  |  |  |
| Gain | **0,32** | **0,85** | **0,95** |  | **1,22** |
|  | miękkie | 0 | 4 |  | 4 |
|  | twarde | 3 | 0 |  | 3 |
|  | brak | 8 | 7 |  | 15 |
|  |  |  |  |  |  |
|  | **Łzawienie** | zmniejszone | normalne |  | Klasa |
| ilość |  | 12 | 10 |  | 22 |
| dla atrybutu | **0,71** |  |  |  |  |
| Gain | **0,50** | **0,00** | **1,57** |  | **1,22** |
|  | miękkie | 0 | 4 |  | 4 |
|  | twarde | 0 | 3 |  | 3 |
|  | brak | 12 | 3 |  | 15 |

Na podstawie powyższych tabel, można zauważyć iż wartość Gain przy Łzawieniu ma największy zysk informacji, czyli jest to najistotniejszy z atrybutów opisujących. W związku z powyższym będzie to początek drzewa, a dokładnie korzeń pokazany poniżej.

Wybranie atrybutu Łzawienie jako korzenia spowodowało utworzenie dwóch podzbiorów z czego jeden jest liściem, który opisuje aż 12 z 15 obiektów klasy Brak.

Postępując analogicznie wykonujemy takie same obliczenia, lecz na zbiorze uczącym pomniejszonym o wyznaczone obiekty.

Do sprawdzenia rozłożystości drzewa stosuje się parametr zwany średnią liczbą pytań, którego wzór wygląda następująco:

gdzie:

Sk(xi) - liczba pytań prowadzących do i-tego węzła terminalnego (decyzji);

(pi) - prawdopodobieństwo rozpoznania przypadku przez i-tą ścieżkę w drzewie.

Poniższe obliczenia dotyczą drzewa przedstawionego na zdjęciu 1:

Ilość wszystkich obiektów = 43+2+1+2+2+2+1+47+50= 150

E(Sk)=3\*(43/150) +4\*(2/150) +4\*(1/150) +3\*(2/150) +5\*(2/150) +5\*(2/150) +5\*(1/150) +5\*(47/150) +1\*(50/150) ≈ 3,05

Średnia ilość pytań dla drzewa przedstawionego na zdjęciu 1 wynosi w przybliżeniu 3,05. Oznacza to, że aby odnaleźć przypadek należy zadać średnio 4 pytania.

# Wyniki

W wyniku opisanego powyżej algorytmu powstało następujące drzewo decyzji:

# Wnioski

Drzewo decyzji to graficzny sposób przedstawienia wiedzy. Dzięki takiemu rozwiązaniu, zbudowanemu na podstawie zbioru uczącego, można sklasyfikować nowe obiekty, które nie brały udziału w procesie tworzenia drzewa. Charakteryzują się one strukturą hierarchiczną.

Dla danych z powyższego zbioru uczącego otrzymano drzewo składające się z pięciu węzłów decyzyjnych oraz siedmiu liści. Średnia liczba pytań dla tego drzewa wynosi 2. Oznacza to, że aby sklasyfikować pewien obiekt, teoretycznie, należy średnio zadać 2 pytania.

Drzewo decyzji ma szeroki wachlarz zastosowań, ponieważ można je z powodzeniem stosować w wielu dziedzinach takich jak medycyna, botanika, ekonomia i tym podobne. Proces klasyfikacji z wykorzystaniem drzew decyzyjnych jest efektywny obliczeniowo, ponieważ wyznaczenie kategorii obiektu wymaga w najgorszym razie przetestowania raz wszystkich jego atrybutów. Drzewo odzwierciedla, w jaki sposób na podstawie atrybutów były podejmowane decyzje klasyfikujące. Dodatkową ich zaletą jest czytelność dla ludzi oraz to, że w prosty sposób można przekształcić je do reprezentacji regułowej.

Drzewa mają jednak swoje wady, bardzo rozgałęzione, złożone drzewa, z wieloma przypadkami nie są już tak czytelne dla człowieka, aby mógł odnaleźć w nich rozwiązanie. Dzieje się tak dlatego, że algorytmy generowania drzew wykorzystują tylko pojedynczy atrybut. Ponadto, raz utworzone drzewo bardzo trudno jest zaktualizować, gdy zostaną do zbioru dodane nowe przypadki.