***Podstawy Sztucznej Inteligencji***

***Projekt 2 – Sprawozdanie***

**Zespół:**

Jarosław Zabuski - 300288

Jakub Strawa - 300266

**Temat:**

Przewidywanie czy grzyb jest jadalny przy użyciu zmodyfikowanej implementacji algorytmu ID3. Do wyboru testu w drzewie stosujemy zasadę koła ruletki - większe prawdopodobieństwo wyboru mają lepsze testy. Porównanie wyników z wynikami wersji klasycznej algorytmu.

**Doprecyzowanie tematu i założenia:**

Założyliśmy, że w podstawowej wersji algorytmu ID3 dodajemy nowe gałęzie do drzewa decyzyjnego wybierając atrybut z największym zyskiem informacyjnym aż cały zbiór testowy nie zostanie w pełni opisany. Dla zmodyfikowanej wersji ID3 nowe gałęzie są dodawane do momentu aż żaden z pozostałych atrybutów nie niesie ze sobą żadnego zysku informacyjnego. Prawdopodobieństwo wyboru atrybutu jest równe zyskowi informacyjnemu podzielonego przez sumę zysków informacyjnych wszystkich atrybutów które są możliwe do wyboru na tym etapie.

**Podział prac:**

* Jarosław Zabuski:
  + implementacja klasycznej wersji ID3
  + opisanie eksperymentów
* Jakub Strawa:
  + implementacja zmodyfikowanej wersji ID3
  + przeprowadzenie eksperymentów

**Wykorzystane narzędzia i biblioteki:**

Program został napisany w języku Python. Korzystaliśmy wyłącznie z modułów biblioteki standardowej: math, random, sys, gc, time(w celach testowych) oraz klasy Counter z collections. W finalnym programie realizującym założenia obu implementacji algorytmu ID3 nie skorzystaliśmy z żadnych bibliotek zewnętrznych.

Do przeprowadzenia testów i eksperymentów wykorzystaliśmy plik z danymi ze strony:

<https://archive.ics.uci.edu/ml/datasets/mushroom>

,używany w formie rozszerzonej (z pełnymi nazwami atrybutów oraz podpisami kolumn) jako plik tekstowy mushroom.txt.

Utworzony przez nas program implementuje klasyczną oraz zmodyfikowaną implementację algorytmu ID3. Klasyczna implementacja została przez nas wprowadzona na podstawie artykułu „An Application of Decision Tree Based on ID3” autorstwa Wang Xiaohu, Wang Lele, oraz Li Nianfeng. <https://www.sciencedirect.com/science/article/pii/S1875389212006098>

**Przeprowadzone eksperymenty:**

Naszą główną tezą, sformułowaną przez nas na początku badań, było przekonanie o tym, że modyfikacja ID3 polegająca na losowości doboru testów wpłynie negatywnie na czas tworzenia rozwiązania oraz pogorszenie jego jakości, względem klasycznej implementacji ID3. Chcieliśmy sprawdzić jej słuszność, przeprowadzając poniższe eksperymenty:

1. Zmiana skuteczności wraz ze wzrostem zbioru uczącego.
   1. Założenia: Zależało nam na przekonaniu się, jaki wpływ na skuteczność działania zmodyfikowanego algorytmu ID3 ma rozmiar danych testowych.
   2. Eksperyment: Aby przeprowadzić eksperyment, wprowadziliśmy możliwość podzielenia początkowego zbioru danych, pochodzącego z repozytorium danych uczących . Przy początkowym wywołaniu instancji programu, użytkownik ma możliwość wybrania liczby od 1 do 100, pozwalającej określić, jaki procent danych ma być przeznaczone na dane uczące, a jaki – na dane testów. (liczba 1 odpowiada temu, że jeden procent danych jest poświęconych na uczenie się, liczba 2 – dwa procent itp.) Pozwala to na testowanie działania algorytmu ID3 w zależności od podziału zbioru, i porównaniu go z działaniem klasycznej implementacji ID3 w tej sytuacji. W poniższej tabeli znajdują się średnie zgodności wyprodukowanych przez obie implementacje drzew decyzyjnych z kilku uruchomień instancji obu algorytmów dla każdego rozmiaru. Owe średnie zgodności bowiem służyły jako główny współczynnik badany przez nas w tym eksperymencie.

|  |  |  |
| --- | --- | --- |
| *Procent ilości danych testowych w stosunku do danych sprawdzających.* | *Zgodność drzewa wyprodukowanego przez zmodyfikowany ID3 (w procentach)* | *Zgodność drzewa wyprodukowanego przez oryginalny ID3 (w procentach)* |
| 1 | 89,88 | 96,79 |
| 2 | 93,39 | 98,18 |
| 3 | 94,74 | 98,71 |
| 4 | 96,09 | 99,08 |
| 5 | 95,81 | 99,22 |
| 6 | 95,48 | 99,50 |
| 7 | 96,24 | 99,51 |
| 8 | 95,35 | 99,65 |
| 9 | 95,93 | 99,78 |
| 10 | 96,13 | 99,77 |
| 11 | 96,99 | 99,81 |
| 12 | 95,77 | 99,82 |
| 13 | 96,06 | 99,81 |
| 14 | 96,93 | 99,86 |
| 15 | 96,14 | 99,88 |
| 16 | 95,82 | 99,88 |
| 17 | 96,55 | 99,90 |
| 18 | 95,99 | 99,93 |
| 19 | 96,48 | 99,93 |
| 20 | 96,78 | 99,93 |
| 21 | 95,25 | 99,93 |
| 22 | 96,30 | 99,95 |
| 23 | 96,34 | 99,94 |
| 24 | 96,51 | 99,95 |
| 25 | 95,52 | 99,95 |
| 26 | 95,56 | 99,97 |
| 27 | 95,65 | 99,97 |
| 28 | 95,34 | 99,96 |
| 29 | 96,49 | 99,97 |
| 30 | 95,14 | 99,97 |
| 31 | 96,49 | 99,98 |
| 32 | 95,93 | 99,98 |
| 33 | 95,82 | 99,98 |
| 34 | 96,59 | 99,98 |
| 35 | 95,52 | 99,98 |
| 36 | 94,48 | 99,98 |
| 37 | 94,66 | 99,98 |
| 38 | 95,86 | 99,99 |
| 39 | 94,01 | 99,98 |
| 40 | 96,23 | 99,98 |
| 41 | 96,43 | 99,98 |
| 42 | 97,66 | 99,99 |
| 43 | 95,73 | 99,99 |
| 44 | 95,79 | 99,99 |
| 45 | 96,79 | 99,99 |
| 46 | 96,45 | 99,99 |
| 47 | 95,31 | 99,99 |
| 48 | 95,87 | 99,99 |
| 49 | 97,02 | 99,99 |
| 50 | 94,93 | 100,00 |
| 51 | 95,51 | 99,99 |
| 52 | 95,23 | 100,00 |
| 53 | 96,17 | 100,00 |
| 54 | 95,79 | 100,00 |
| 55 | 96,39 | 99,99 |
| 56 | 96,04 | 99,99 |
| 57 | 96,61 | 99,99 |
| 58 | 97,08 | 100,00 |
| 59 | 95,41 | 100,00 |
| 60 | 95,25 | 100,00 |
| 61 | 95,91 | 100,00 |
| 62 | 96,92 | 100,00 |
| 63 | 96,08 | 100,00 |
| 64 | 94,03 | 100,00 |
| 65 | 96,62 | 99,99 |
| 66 | 95,41 | 100,00 |
| 67 | 96,02 | 100,00 |
| 68 | 95,35 | 100,00 |
| 69 | 97,29 | 100,00 |
| 70 | 95,29 | 100,00 |
| 71 | 96,69 | 100,00 |
| 72 | 96,18 | 100,00 |
| 73 | 95,96 | 100,00 |
| 74 | 96,95 | 100,00 |
| 75 | 95,63 | 100,00 |
| 76 | 96,37 | 100,00 |
| 77 | 94,61 | 100,00 |
| 78 | 94,49 | 100,00 |
| 79 | 94,80 | 100,00 |
| 80 | 95,57 | 100,00 |
| 81 | 96,24 | 100,00 |
| 82 | 95,84 | 100,00 |
| 83 | 93,91 | 100,00 |
| 84 | 94,27 | 100,00 |
| 85 | 96,11 | 100,00 |
| 86 | 96,27 | 100,00 |
| 87 | 96,77 | 100,00 |
| 88 | 94,80 | 100,00 |
| 89 | 95,97 | 100,00 |
| 90 | 96,48 | 100,00 |
| 91 | 97,18 | 100,00 |
| 92 | 95,48 | 100,00 |
| 93 | 95,66 | 100,00 |
| 94 | 97,26 | 100,00 |
| 95 | 95,39 | 100,00 |
| 96 | 95,91 | 100,00 |
| 97 | 95,25 | 100,00 |
| 98 | 96,50 | 100,00 |
| 99 | 96,60 | 100,00 |
| 100 | 96,87 | 100,00 |

* 1. Wniosek: Z danych skompilowanych w powyższej tabeli wynika, że zmodyfikowana implementacja ID3 jest średnio o 5% mniej dokładna od klasycznej implementacji w stosunku do zbioru danych, na którym się ona uczy. Zgadza się to z naszymi pierwotnymi podejrzeniami, albowiem mniejsza dokładność jest powiązana z wprowadzeniem elementu losowości do zmodyfikowanego ID3, jakim jest ruletkowe wybieranie testu jako nowego elementu drzewa decyzyjnego. Klasyczna implementacja bowiem osiąga lepsze wyniki, ponieważ zawsze sugeruje się ona metodą największego przyrostu informacyjnego, a nie tak jak zmodyfikowana – sugeruje się nią, ale tylko „dość często”.

1. Średni czas wykonywania.
   1. Założenia: Zależało nam na przekonaniu się, w jaki sposób wprowadzenie semi-losowego wyboru testów do metody ID3 wpływa na czas jej wykonania.
   2. Eksperyment: Tak jak dla pierwszego eksperymentu zbudowaliśmy tabelę, zawierającą wyniki wykonania kilku instancji obu algorytmów ID3 dla każdego możliwego do zadeklarowania procentu ilości danych testowych w ogólnej bazie danych. Tym razem jednak badaliśmy przede wszystkim czas wykonania obu algorytmów z podziałem na dwie kategorie: na średni czas poświęcony przez dany algorytm na budowę drzewa, i średnia czas poświęcony na przeprowadzenie na nim testów. Tabela ta znajduje się poniżej.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Procent ilości danych testowych w stosunku do danych sprawdzających.* | Średni czas poświęcony na budowę drzewa przez zmodyfikowany ID3 (w sekundach) | Średni czas poświęcony na testowanie drzewa przez zmodyfikowany ID3 (w sekundach) | Średni czas poświęcony na budowę drzewa przez klasyczny ID3 (w sekundach) | Średni czas poświęcony na budowę drzewa przez klasyczny ID3 (w sekundach) |
| 1 | 0,07 | 0,01 | 0,06 | 0,01 |
| 2 | 0,07 | 0,01 | 0,07 | 0,01 |
| 3 | 0,07 | 0,01 | 0,08 | 0,01 |
| 4 | 0,08 | 0,02 | 0,08 | 0,01 |
| 5 | 0,08 | 0,01 | 0,07 | 0,01 |
| 6 | 0,08 | 0,01 | 0,08 | 0,01 |
| 7 | 0,09 | 0,02 | 0,08 | 0,01 |
| 8 | 0,10 | 0,02 | 0,08 | 0,01 |
| 9 | 0,10 | 0,02 | 0,09 | 0,01 |
| 10 | 0,10 | 0,02 | 0,09 | 0,01 |
| 11 | 0,11 | 0,02 | 0,10 | 0,01 |
| 12 | 0,11 | 0,02 | 0,10 | 0,01 |
| 13 | 0,12 | 0,02 | 0,10 | 0,01 |
| 14 | 0,12 | 0,02 | 0,11 | 0,01 |
| 15 | 0,12 | 0,01 | 0,11 | 0,01 |
| 16 | 0,13 | 0,02 | 0,12 | 0,01 |
| 17 | 0,13 | 0,01 | 0,12 | 0,01 |
| 18 | 0,13 | 0,01 | 0,12 | 0,01 |
| 19 | 0,14 | 0,01 | 0,12 | 0,01 |
| 20 | 0,14 | 0,01 | 0,13 | 0,01 |
| 21 | 0,15 | 0,01 | 0,13 | 0,01 |
| 22 | 0,15 | 0,01 | 0,13 | 0,01 |
| 23 | 0,17 | 0,01 | 0,13 | 0,01 |
| 24 | 0,16 | 0,01 | 0,14 | 0,01 |
| 25 | 0,19 | 0,01 | 0,17 | 0,01 |
| 26 | 0,18 | 0,01 | 0,14 | 0,01 |
| 27 | 0,18 | 0,01 | 0,16 | 0,01 |
| 28 | 0,18 | 0,01 | 0,16 | 0,01 |
| 29 | 0,18 | 0,01 | 0,16 | 0,01 |
| 30 | 0,18 | 0,01 | 0,17 | 0,01 |
| 31 | 0,19 | 0,01 | 0,16 | 0,01 |
| 32 | 0,19 | 0,01 | 0,17 | 0,01 |
| 33 | 0,20 | 0,01 | 0,18 | 0,01 |
| 34 | 0,21 | 0,01 | 0,18 | 0,01 |
| 35 | 0,21 | 0,01 | 0,18 | 0,01 |
| 36 | 0,22 | 0,01 | 0,19 | 0,01 |
| 37 | 0,21 | 0,01 | 0,19 | 0,01 |
| 38 | 0,22 | 0,01 | 0,19 | 0,01 |
| 39 | 0,23 | 0,01 | 0,20 | 0,01 |
| 40 | 0,24 | 0,01 | 0,21 | 0,01 |
| 41 | 0,24 | 0,01 | 0,20 | 0,01 |
| 42 | 0,25 | 0,01 | 0,22 | 0,01 |
| 43 | 0,26 | 0,01 | 0,22 | 0,01 |
| 44 | 0,27 | 0,01 | 0,23 | 0,01 |
| 45 | 0,27 | 0,01 | 0,24 | 0,01 |
| 46 | 0,28 | 0,01 | 0,23 | 0,01 |
| 47 | 0,29 | 0,01 | 0,24 | 0,01 |
| 48 | 0,29 | 0,01 | 0,24 | 0,01 |
| 49 | 0,30 | 0,01 | 0,24 | 0,01 |
| 50 | 0,30 | 0,01 | 0,24 | 0,01 |
| 51 | 0,30 | 0,01 | 0,25 | 0,01 |
| 52 | 0,31 | 0,01 | 0,25 | 0,01 |
| 53 | 0,31 | 0,01 | 0,26 | 0,01 |
| 54 | 0,31 | 0,01 | 0,28 | 0,01 |
| 55 | 0,32 | 0,01 | 0,28 | 0,01 |
| 56 | 0,32 | 0,01 | 0,26 | 0,00 |
| 57 | 0,34 | 0,01 | 0,28 | 0,01 |
| 58 | 0,35 | 0,01 | 0,28 | 0,01 |
| 59 | 0,33 | 0,01 | 0,29 | 0,01 |
| 60 | 0,35 | 0,01 | 0,29 | 0,01 |
| 61 | 0,35 | 0,01 | 0,29 | 0,01 |
| 62 | 0,37 | 0,01 | 0,29 | 0,01 |
| 63 | 0,37 | 0,01 | 0,30 | 0,005 |
| 64 | 0,39 | 0,01 | 0,36 | 0,005 |
| 65 | 0,39 | 0,01 | 0,34 | 0,004 |
| 66 | 0,37 | 0,01 | 0,32 | 0,005 |
| 67 | 0,40 | 0,01 | 0,34 | 0,006 |
| 68 | 0,38 | 0,01 | 0,34 | 0,005 |
| 69 | 0,42 | 0,01 | 0,36 | 0,005 |
| 70 | 0,41 | 0,01 | 0,36 | 0,005 |
| 71 | 0,43 | 0,01 | 0,34 | 0,005 |
| 72 | 0,43 | 0,01 | 0,34 | 0,004 |
| 73 | 0,45 | 0,01 | 0,34 | 0,004 |
| 74 | 0,45 | 0,01 | 0,35 | 0,003 |
| 75 | 0,46 | 0,01 | 0,44 | 0,004 |
| 76 | 0,46 | 0,005 | 0,38 | 0,004 |
| 77 | 0,45 | 0,004 | 0,39 | 0,003 |
| 78 | 0,46 | 0,005 | 0,40 | 0,004 |
| 79 | 0,48 | 0,004 | 0,40 | 0,004 |
| 80 | 0,47 | 0,003 | 0,42 | 0,003 |
| 81 | 0,50 | 0,004 | 0,39 | 0,003 |
| 82 | 0,52 | 0,004 | 0,41 | 0,002 |
| 83 | 0,55 | 0,004 | 0,40 | 0,002 |
| 84 | 0,52 | 0,004 | 0,49 | 0,002 |
| 85 | 0,53 | 0,003 | 0,45 | 0,002 |
| 86 | 0,53 | 0,003 | 0,45 | 0,003 |
| 87 | 0,55 | 0,003 | 0,42 | 0,001 |
| 88 | 0,55 | 0,003 | 0,48 | 0,002 |
| 89 | 0,54 | 0,003 | 0,45 | 0,002 |
| 90 | 0,58 | 0,002 | 0,46 | 0,002 |
| 91 | 0,58 | 0,002 | 0,46 | 0,001 |
| 92 | 0,54 | 0,001 | 0,46 | 0,001 |
| 93 | 0,58 | 0,002 | 0,50 | 0,001 |
| 94 | 0,61 | 0,002 | 0,51 | 0,001 |
| 95 | 0,63 | 0,001 | 0,50 | 0,001 |
| 96 | 0,66 | 0,001 | 0,51 | 0,001 |
| 97 | 0,61 | 0,001 | 0,52 | 0,001 |
| 98 | 0,64 | 0,001 | 0,56 | 0,0004 |
| 99 | 0,59 | 0,0002 | 0,50 | 0,0001 |
| 100 | 0,60 | 0,02 | 0,50 | 0,01 |

* 1. Wniosek: Zmodyfikowany ID3 w każdym przypadku okazywał się być bardziej czasochłonny aniżeli zwykły ID3. Związane jest to ze specyfiką wprowadzonej modyfikacji – jeżeli wybrany na podstawie przyrostu informacji test nie zda testu losowości, wybierany jest inny test. Sama procedura przeprowadzenia drugiego testu może być na tyle czasochłonna, by trwale zwiększyć czas wykonywania się zmodyfikowanego ID3 w każdym wypadku, nawet gdy czasy testowania drzewa przy wysokich wartościach ilości danych uczących w stosunku do danych testowych zbiegają do zera. (ponieważ do zera zmniejsza się ilość rekordów w bazie danych, jakie poświęcamy na testowanie drzewa.)

1. Średnia głębokość drzewa.
   1. Założenia: Zależało nam na sprawdzeniu, w jaki sposób modyfikacja algorytmu ID3 wpływa na średnią głębokość drzew, generowanych za jego pomocą.
   2. Eksperyment: Tak jak i w powyższych eksperymentach, tak i tutaj zbadaliśmy parametry związane z obiema implementacjami algorytmu ID3, zależne od stosunku ilości danych uczących do danych testowych. Tym razem sprawdziliśmy, jak ma się średnia głębokość generowanego drzewa decyzyjnego do ilości danych uczących, liczona dla obu implementacji algorytmu ID3.

|  |  |  |
| --- | --- | --- |
| *Procent ilości danych testowych w stosunku do danych sprawdzających.* | Średnia głębokość drzewa decyzyjnego, wygenerowanego przez zmodyfikowany ID3 | Średnia głębokość drzewa decyzyjnego, wygenerowanego przez klasyczny ID3 |
| 1 | 3,85 | 1,87 |
| 2 | 4,80 | 2,38 |
| 3 | 5,09 | 2,78 |
| 4 | 5,40 | 3,18 |
| 5 | 5,56 | 3,26 |
| 6 | 5,59 | 3,42 |
| 7 | 5,91 | 3,58 |
| 8 | 5,81 | 3,69 |
| 9 | 5,80 | 3,80 |
| 10 | 5,92 | 3,82 |
| 11 | 5,79 | 3,77 |
| 12 | 5,98 | 3,87 |
| 13 | 6,19 | 3,90 |
| 14 | 6,08 | 3,98 |
| 15 | 6,02 | 3,90 |
| 16 | 6,28 | 3,97 |
| 17 | 6,18 | 3,98 |
| 18 | 6,11 | 4,09 |
| 19 | 6,22 | 4,04 |
| 20 | 6,06 | 4,06 |
| 21 | 6,71 | 4,06 |
| 22 | 6,28 | 4,06 |
| 23 | 6,38 | 4,06 |
| 24 | 6,35 | 4,04 |
| 25 | 6,37 | 4,04 |
| 26 | 6,40 | 4,02 |
| 27 | 6,41 | 4,08 |
| 28 | 6,44 | 4,03 |
| 29 | 6,38 | 4,10 |
| 30 | 6,28 | 4,03 |
| 31 | 6,47 | 4,08 |
| 32 | 6,40 | 4,06 |
| 33 | 6,39 | 4,08 |
| 34 | 6,35 | 4,07 |
| 35 | 6,75 | 4,07 |
| 36 | 6,41 | 4,11 |
| 37 | 6,64 | 4,11 |
| 38 | 6,22 | 4,09 |
| 39 | 6,57 | 4,07 |
| 40 | 6,44 | 4,08 |
| 41 | 6,70 | 4,05 |
| 42 | 6,71 | 4,06 |
| 43 | 6,57 | 4,08 |
| 44 | 6,62 | 4,07 |
| 45 | 6,55 | 4,10 |
| 46 | 6,46 | 4,09 |
| 47 | 6,62 | 4,03 |
| 48 | 6,65 | 4,06 |
| 49 | 6,50 | 4,08 |
| 50 | 6,88 | 4,04 |
| 51 | 6,83 | 4,09 |
| 52 | 6,87 | 4,02 |
| 53 | 6,87 | 4,03 |
| 54 | 6,66 | 4,04 |
| 55 | 6,67 | 4,04 |
| 56 | 6,47 | 4,08 |
| 57 | 6,66 | 4,04 |
| 58 | 6,78 | 4,06 |
| 59 | 6,81 | 4,06 |
| 60 | 6,60 | 4,03 |
| 61 | 6,56 | 4,02 |
| 62 | 6,63 | 4,04 |
| 63 | 6,96 | 4,09 |
| 64 | 6,44 | 4,07 |
| 65 | 6,60 | 4,01 |
| 66 | 7,10 | 4,06 |
| 67 | 6,57 | 4,04 |
| 68 | 6,80 | 4,04 |
| 69 | 6,67 | 4,04 |
| 70 | 6,61 | 4,04 |
| 71 | 6,51 | 4,04 |
| 72 | 6,47 | 4,02 |
| 73 | 6,84 | 4,06 |
| 74 | 6,65 | 4,02 |
| 75 | 6,60 | 4,01 |
| 76 | 6,46 | 4,01 |
| 77 | 6,50 | 4,00 |
| 78 | 6,76 | 4,03 |
| 79 | 6,63 | 4,03 |
| 80 | 6,43 | 4,00 |
| 81 | 6,55 | 4,01 |
| 82 | 6,62 | 4,00 |
| 83 | 6,60 | 4,01 |
| 84 | 6,63 | 4,00 |
| 85 | 6,62 | 4,00 |
| 86 | 6,54 | 4,00 |
| 87 | 6,68 | 4,00 |
| 88 | 6,52 | 4,00 |
| 89 | 6,71 | 4,00 |
| 90 | 6,71 | 4,00 |
| 91 | 6,81 | 4,00 |
| 92 | 6,47 | 4,00 |
| 93 | 6,53 | 4,00 |
| 94 | 6,58 | 4,00 |
| 95 | 6,56 | 4,00 |
| 96 | 6,82 | 4,00 |
| 97 | 6,88 | 4,00 |
| 98 | 6,49 | 4,00 |
| 99 | 6,42 | 4,00 |
| 100 | 6,65 | 4,00 |

* 1. Wniosek: Podczas gdy średnia głębokość drzewa decyzyjnego, wygenerowanego przez klasyczny ID3, znalazła optymalny podział zbioru testowego na klasy przy użyciu czterech atrybutów, średnia głębokość drzewa decyzyjnego dla modyfikacji ID3 oscylowała pomiędzy sześcioma a siedmioma atrybutami. Oznacza to, że zmodyfikowany ID3 przez fakt semi-losowego wyboru atrybutów do drzewa tworzy znacznie bardziej rozbudowane drzewo decyzyjne, próbując w ten sposób zniwelować wpływ losowości przy decydowaniu o optymalnych atrybutach dzielących i dojść do maksymalnego pokrycia danych testowych.

1. Średnia ilość wykorzystanych atrybutów.
   1. Założenia: Zależało nam na sprawdzeniu, ile atrybutów jest branych pod uwagę w obu rozpatrywanych przez nas implementacjach algorytmu ID3 przy generacji drzew decyzyjnych, i co można za pomocą tych danych powiedzieć o modyfikacji algorytmu.
   2. Eksperyment: Jak w powyższych eksperymentach, tak i tutaj stwierdziliśmy, że najlepiej będzie zbadać sposób, w jaki modyfikacja algorytmu ID3 wpływa na średnie użycie atrybutów, za pomocą dokładnego prześledzenia średniej ilości atrybutów wykorzystanych przez obie implementacje, z zależnością co do procentu wykorzystania danych do uczenia się. Tabela, kompilująca zebrane w trakcie badań współczynniki, znajduje się poniżej.

|  |  |  |
| --- | --- | --- |
| *Procent ilości danych testowych w stosunku do danych sprawdzających.* | Średnia wykorzystanych atrybutów w implementacji zmodyfikowanej | Średnia wykorzystanych atrybutów w implementacji klasycznej |
| 1 | 8,03 | 1,87 |
| 2 | 11,84 | 2,42 |
| 3 | 12,26 | 2,88 |
| 4 | 13,04 | 3,27 |
| 5 | 13,68 | 3,41 |
| 6 | 14,69 | 3,57 |
| 7 | 14,91 | 3,77 |
| 8 | 15,00 | 3,88 |
| 9 | 14,41 | 4,12 |
| 10 | 15,52 | 4,05 |
| 11 | 14,85 | 4,04 |
| 12 | 15,30 | 4,17 |
| 13 | 15,30 | 4,23 |
| 14 | 15,60 | 4,48 |
| 15 | 15,69 | 4,29 |
| 16 | 15,96 | 4,48 |
| 17 | 15,68 | 4,56 |
| 18 | 15,74 | 4,72 |
| 19 | 16,11 | 4,65 |
| 20 | 15,17 | 4,67 |
| 21 | 16,82 | 4,74 |
| 22 | 15,70 | 4,80 |
| 23 | 16,24 | 4,77 |
| 24 | 16,04 | 4,80 |
| 25 | 16,52 | 4,85 |
| 26 | 16,70 | 4,90 |
| 27 | 16,05 | 4,99 |
| 28 | 16,88 | 4,90 |
| 29 | 15,60 | 4,93 |
| 30 | 16,18 | 4,97 |
| 31 | 16,23 | 5,00 |
| 32 | 15,93 | 4,93 |
| 33 | 16,03 | 4,96 |
| 34 | 15,83 | 4,99 |
| 35 | 17,00 | 4,98 |
| 36 | 16,30 | 5,05 |
| 37 | 17,31 | 5,06 |
| 38 | 15,97 | 5,05 |
| 39 | 17,11 | 5,03 |
| 40 | 15,85 | 5,02 |
| 41 | 16,50 | 5,01 |
| 42 | 15,97 | 5,04 |
| 43 | 16,25 | 5,08 |
| 44 | 16,97 | 5,05 |
| 45 | 16,25 | 5,07 |
| 46 | 15,71 | 5,07 |
| 47 | 16,22 | 5,02 |
| 48 | 16,54 | 5,05 |
| 49 | 15,81 | 5,06 |
| 50 | 16,87 | 5,04 |
| 51 | 16,40 | 5,07 |
| 52 | 17,19 | 5,01 |
| 53 | 16,92 | 5,03 |
| 54 | 16,55 | 5,04 |
| 55 | 16,44 | 5,04 |
| 56 | 15,92 | 5,07 |
| 57 | 16,28 | 5,04 |
| 58 | 16,17 | 5,06 |
| 59 | 16,84 | 5,06 |
| 60 | 16,72 | 5,03 |
| 61 | 16,27 | 5,02 |
| 62 | 16,90 | 5,04 |
| 63 | 16,80 | 5,09 |
| 64 | 17,29 | 5,07 |
| 65 | 15,96 | 5,00 |
| 66 | 17,32 | 5,06 |
| 67 | 16,14 | 5,04 |
| 68 | 16,24 | 5,04 |
| 69 | 16,53 | 5,05 |
| 70 | 16,59 | 5,04 |
| 71 | 16,04 | 5,05 |
| 72 | 15,89 | 5,02 |
| 73 | 17,00 | 5,07 |
| 74 | 17,27 | 5,02 |
| 75 | 15,98 | 5,01 |
| 76 | 16,16 | 5,01 |
| 77 | 16,11 | 5,00 |
| 78 | 17,16 | 5,03 |
| 79 | 16,72 | 5,03 |
| 80 | 16,83 | 5,00 |
| 81 | 16,74 | 5,01 |
| 82 | 16,65 | 5,00 |
| 83 | 16,86 | 5,01 |
| 84 | 16,67 | 5,00 |
| 85 | 16,47 | 5,00 |
| 86 | 16,13 | 5,00 |
| 87 | 16,18 | 5,00 |
| 88 | 17,40 | 5,00 |
| 89 | 16,45 | 5,00 |
| 90 | 16,08 | 5,00 |
| 91 | 16,15 | 5,00 |
| 92 | 15,86 | 5,00 |
| 93 | 16,57 | 5,00 |
| 94 | 16,14 | 5,00 |
| 95 | 16,85 | 5,00 |
| 96 | 17,19 | 5,00 |
| 97 | 17,05 | 5,00 |
| 98 | 17,08 | 5,00 |
| 99 | 16,61 | 5,00 |
| 100 | 16,60 | 5,00 |

* 1. Wniosek: Od samego początku jest bardzo mocno widoczna ilość wykorzystanych przez zmodyfikowany ID3 atrybutów. Gdy klasyczna implementacja potrzebuje 5 atrybutów by całkowicie pokryć i sklasyfikować zbiór danych, zmodyfikowana (z powodu semi-losowości przy wybieraniu atrybutów do oceny) potrzebuje znacznie większej liczby zasobów. Tak duża liczba nadmiarowych atrybutów wykorzystanych przez zmodyfikowane ID3 jest również konsekwencją tego, że przy nieuważnym doborze atrybutu dzielącego daną klasę rozpatrywanych przez nas obiektów na dwie klasy o tej samej złożoności.

**Wnioski:**

* W każdym rozpatrywanym przez nas przypadku, zmodyfikowany algorytm ID3 sprawdzał się gorzej w zadaniu przewidywania, czy grzyb jest jadalny, od klasycznego ID3.
* Głównym powodem niższej niż algorytm klasyczny ID3 skuteczności jest modyfikacja, polegająca na losowym doborze nagłówków służących jako separatory klas, co można zauważyć w każdym z przeprowadzonych przez nas eksperymentów.
* Zasada koła ruletki jest gorszą metodą doboru testów w drzewie od stałej zasady wyboru atrybutu o największym przyroście informacyjnym, ponieważ do wystarczająco dobrej zasady o minimalizowaniu entropii wprowadza element losowości.
* W najlepszym wypadku, zmodyfikowany w ten sposób algorytm ID3 może działać tak samo dobrze jak klasyczny algorytm. W przypadku uwzględniania zagregowanych średnich czasów działania algorytmu czy parametrów produkowanego drzewa decyzyjnego, nie ma jednak szansy na to, by modyfikacja pozwoliła na poprawienie działania algorytmu ID3.
* Dzięki przeprowadzonym badaniom nauczyliśmy się, że w ten sposób sformułowana modyfikacja algorytmu ID3 nie nadałaby się do optymalnego konstruowania drzew decyzyjnych.

**Instrukcja obsługi i sposób odtworzenia eksperymentów:**

Program uruchamiany jest poleceniem:

*python main.py iterations percentage if\_print\_tree*

Gdzie za *iterations* wpisujemy liczbę iteracji algorytmów, za *percentage* wpisujemy liczbę określającą jaki procent zbioru danych będzie zbiorem przykładów uczących, a za *if\_print\_tree* wpisujemy 1 jeśli chcemy wyświetlić drzewo decyzyjne lub 0 jeśli nie chcemy (1 zalecana jest tylko dla pojedynczego uruchomienia algorytmów).

Przykład:

*python main.py 100 5 0*

Program wykona 100 iteracji dla obu algorytmów, gdzie zbiorem uczącym będzie 5% zbioru danych, drzewo decyzyjne nie zostanie wypisane.

Program nie jest związany stale z wykorzystywaną przez nas w eksperymentach bazą danych. Można wykorzystać dowolną bazę danych uczących w formacie csv, po zmianie jej nazwy na „mushroom.txt” i upewnieniu się, że pierwszy wiersz używanej bazy danych zawiera nagłówki kolumn.

**Aby odtworzyć eksperyment:**

Należy wykorzystać skrypt, wykorzystywany przez nas do badania poszczególnych współczynników obu implementacji algorytmu ID3: script1.sh.

Script1.sh jest skryptem bashowym, który po wywołaniu pozwala na uruchomienie działalności wytworzonego przez nas programu wprowadzającego i testującego obie implementacje ID3 wiele razy, celem sprawdzenia poprawności agregacji wyników poszczególnych instancji problemu. Zmienna skryptowa max pozwala na ograniczenie ilości uruchamianych instancji programu testowego do zadanej liczby n, jeżeli przykładowo nie chcemy zagregować oddzielnych danych i je oddzielnie uśrednić dla lepszej dokładności, lecz tylko jednokrotnie uruchomić program testujący.