Индивидуальная практическая задача

|  |  |  |  |
| --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **Y** |
| 5 | -3,2 | 45,3 | 0 |
| 0,1 | 99 | 78,4 | 0 |
| 98 | 4,1 | 14 | 1 |
| 3,4 | -40,4 | 8,7 | 0 |
| 3,4 | 80 | 1 | 1 |
| 2,2 | 18,1 | 12 | 0 |
| 0 | 15 | 47,1 | 0 |
| 1000 | -1000 | 1 | 0 |
| 45 | 46 | -7,1 | 1 |

Имеется выборка данных, необходимо построить дерево принятия решений с использованием критерия информационной выгоды.

Необходимо разбить выборку на 2 части, используя 1 из атрибутов, и продолжать пока не останутся однородные выборки. Для разбиения на каждом шаге нужно выбирать атрибут, по которому оно будет сделано. Результирующие подмножества должны содержать как можно меньше примесей. Для выбора атрибута будем использовать теоретико-информационный критерий.

Необходимо выбрать наибольший прирост информации (Information Gain), вычисляемый по следующей формуле:

где q – число групп после разбиения, – число элементов выборки, у которых признак Q имеет i-ое значение. -энтропия родительского множества, - энтропия подмножества, полученного в результате разбиения.

Энтропия вычисляется по формуле

Где pi –вероятности нахождения системы в i-ом состоянии.( Вероятность, что случайное Y из данной выборки = 1 или 0)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **Y** |  | **X1** | **X2** | **X3** | **Y** |  | **X1** | **X2** | **X3** | **Y** |
| 0 | 15 | 47,1 | 0 |  | 1000 | -1000 | 1 | 0 |  | 45 | 46 | -7,1 | 1 |
| 0,1 | 99 | 78,4 | 0 |  | 3,4 | -40,4 | 8,7 | 0 |  | 1000 | -1000 | 1 | 0 |
| 2,2 | 18,1 | 12 | 0 |  | 5 | -3,2 | 45,3 | 0 |  | 3,4 | 80 | 1 | 1 |
| 3,4 | -40,4 | 8,7 | 0 |  | 98 | 4,1 | 14 | 1 |  | 3,4 | -40,4 | 8,7 | 0 |
| 3,4 | 80 | 1 | 1 |  | 0 | 15 | 47,1 | 0 |  | 2,2 | 18,1 | 12 | 0 |
| 5 | -3,2 | 45,3 | 0 |  | 2,2 | 18,1 | 12 | 0 |  | 98 | 4,1 | 14 | 1 |
| 45 | 46 | -7,1 | 1 |  | 45 | 46 | -7,1 | 1 |  | 5 | -3,2 | 45,3 | 0 |
| 98 | 4,1 | 14 | 1 |  | 3,4 | 80 | 1 | 1 |  | 0 | 15 | 47,1 | 0 |
| 1000 | -1000 | 1 | 0 |  | 0,1 | 99 | 78,4 | 0 |  | 0,1 | 99 | 78,4 | 0 |

Отсортируем выборку по возрастанию каждого атрибута. (Сортируем по возрастанию)

Рассчитаем энтропию исходной выборки:

В каждом столбце рассматриваем те значения, при переходе которых целевой класс изменяет свои значения с 0 на 1 или наоборот.

При использовании атрибута X1 для разбиения: (среднее между 2.9 и 3.4)

А) Разбиваем по признаку X1 < 3.4

Вычисляем прирост информации

Б) Разбиваем по признаку X1 < 4.25

Вычисляем прирост информации

В) Разбиваем по признаку X1 < 25

Вычисляем прирост информации

Г) Разбиваем по признаку X1 < 71.50

Вычисляем прирост информации

При использовании атрибута X2 для разбиения:

А) Разбиваем по признаку X2 < 0.3

Вычисляем прирост информации

Б) Разбиваем по признаку X2 < 9.55

Вычисляем прирост информации

В) Разбиваем по признаку X2 < 32.05

Вычисляем прирост информации

Г) Разбиваем по признаку X2 < 89.50

Вычисляем прирост информации

При использовании атрибута X3:

А) Разбиваем по признаку X3 < -3.05

Вычисляем прирост информации

Б) Разбиваем по признаку X3 < 1

Вычисляем прирост информации

В) Разбиваем по признаку X3 < 4.85

Вычисляем прирост информации

Г) Разбиваем по признаку X3 <13

Вычисляем прирост информации

Д) Разбиваем по признаку X3 <29.65

Вычисляем прирост информации

Согласно информационному, критерию самым лучшим оказалось разбиение по X1 < 3.4. В результате получили 2 выборки, однородную и неоднородную. С неоднородной выборкой снова проделаем те же операции:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **Y** |  | **X1** | **X2** | **X3** | **Y** |  | **X1** | **X2** | **X3** | **Y** |
| 0 | 15 | 47,1 | 0 |  | 1000 | -1000 | 1 | 0 |  | 1000 | -1000 | 1 | 0 |
| 0,1 | 99 | 78,4 | 0 |  | 3,4 | -40,4 | 8,7 | 0 |  | 3,4 | 80 | 1 | 1 |
| 2,2 | 18,1 | 12 | 0 |  | 5 | -3,2 | 45,3 | 0 |  | 3,4 | -40,4 | 8,7 | 0 |
| 3,4 | -40,4 | 8,7 | 0 |  | 0 | 15 | 47,1 | 0 |  | 2,2 | 18,1 | 12 | 0 |
| 3,4 | 80 | 1 | 1 |  | 2,2 | 18,1 | 12 | 0 |  | 5 | -3,2 | 45,3 | 0 |
| 5 | -3,2 | 45,3 | 0 |  | 3,4 | 80 | 1 | 1 |  | 0 | 15 | 47,1 | 0 |
| 1000 | -1000 | 1 | 0 |  | 0,1 | 99 | 78,4 | 0 |  | 0,1 | 99 | 78,4 | 0 |

При использовании атрибута X1 для разбиения:

А) Разбиваем по признаку X1 < 3.4

Вычисляем прирост информации

Б) Разбиваем по признаку X1 < 4.2

Вычисляем прирост информации

При использовании атрибута X2 для разбиения:

А) Разбиваем по признаку X2 < 49.05

Вычисляем прирост информации

Б) Разбиваем по признаку X2 < 23.15

Вычисляем прирост информации

При использовании атрибута X3 для разбиения:

А) Разбиваем по признаку X3 <1

Вычисляем прирост информации

Б) Разбиваем по признаку X3 < 4.85

Вычисляем прирост информации

В результате также остается 1 неоднородная выборка, в которой с точки зрения информационного критерия не важно, какой атрибут выбирать.

|  |  |  |  |
| --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **Y** |
| 3,4 | 80 | 1 | 1 |
| 3,4 | -40,4 | 8,7 | 0 |