Lab3 Data Preprocessing in Weka

**แบบฝึกหัด   
ให้ฝึกปฏิบัติและตอบคำถามต่อดังต่อไปนี้**

**Missing Value**

1. ให้ **Download ไฟล์ bank-data-missvalue.csv** มาไว้ที่เครื่อง
2. กด Open file และกดดู attribute ของ income ถามว่า **มี missing value กี่ %** ถ้ามองหน้าจอให้หาคำว่า missing value บริเวณด้านขวา

**คำตอบ missing : 11 หรือ 2%รูปภาพประกอบด้วย ข้อความ, โต๊ะ

คำอธิบายที่สร้างโดยอัตโนมัติ**

1. จากนั้นให้กด choose ด้านบน Filter และกด +filter และกด + unsupervised และ + attribute และกดเลือก ReplaceMissingValues กด OK และ กด Apply
2. กดปุ่ม Save..และบันทึกเป็นชื่อ bank-data-missvalue-R1.csv และกดดูว่า**โปรแกรมแทนค่าที่ขาดหายไปเติมค่าเท่ากับเท่าใด และคุณคิดว่าคือวิธีการใด**

**ตอบคือ 27470.335195**

**รูปภาพประกอบด้วย โต๊ะ

คำอธิบายที่สร้างโดยอัตโนมัติ**

**ด้วยวิธีการUse the attribute mean to fill in the missing value ของค่าincome**

จากที่เรามีไฟล์ต้นฉบับการไม่เติมค่าที่ขาดหายไป bank-data-missvalue.csv ให้สังเกตเทคนิควิธีต่อไปนี้ใช้ได้หรือไม่ แลคุณคิดว่าเช่นไร

1. ให้ทำการทดสอบโดยใช้ค่าข้อมูลเข้าไปใช้เทคนิค classify -> tree -> ID3/J48

และ classify -> function -> MultilayerPerceptron: MLP

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ไฟล์** | **ใช้กับเทคนิค ID3 (ได้หรือไม่ได้)** | **ใช้กับ J48 (ได้หรือไม่ได้)** | **ใช้กับ MLP**  **(ได้หรือไม่ได้)** | **วิเคราะห์เหตุผล** |
| bank-data-missvalue.csv (ค่าที่ขาดหายไป ยังไม่ได้เติม) | ไม่ได้ | ได้ | ได้ | ID3 ที่ใช้ไม่ได้เพราะ ไม่อนุญาตให้ missing value ใช้งาน |
| bank-data-missvalue-R1.csv (เติมค่าที่ขาดหายไป+ Remove id) | ไม่ได้ | ได้ | ได้ | ID3 ใช้ไม่ได้เพราะ ไม่ได้ทำ Binning |
| bank-data-missvalue-R1-B3 (เติมค่าที่ขาดหาย + Remove id+ binining=3) | ได้ | ได้ | ได้ | ID3 ใช้ได้เพราะการทำ Binning หรือการจัดกลุ่ม |

1. ให้ลองจัดกลุ่มข้อมูลโดยวิธี Binning Method โดยลองแบ่ง N เท่ากับ 3,4,5 แล้ววิเคราะห์ว่ามีผลต่อความถูกต้องโดยใช้เทคนิควิธีดังต่อไปนี้

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **N of Binning Method** | **%ความถูกต้องเทคนิค ID3** | **%ความถูกต้องเทคนิค J48** | **%ความถูกต้องของเทคนิค MLP** | **%ความถูกต้องของเทคนิค Naïve Bayes** | **%ความถูกต้องของเทคนิค IBk** |
| 3 | 64.5 | 73.1667 | 68.6667 | 62.5 | 64 |
| 4 | 79.5 | 86.8333 | 79.6667 | 71.3333 | 74.5 |
| 5 | 76.5 | 87.8333 | 73.1667 | 72.8333 | 70.6667 |

1. จากข้อที่ 2 คุณคิดว่าให้เลือก N ของวิธีการ Binning Method ที่ดีที่สุด

ตอบ N เท่ากับ 5 วิธี J48?

1. จากข้อที่ 3 ให้นำข้อมูลดังกล่าวที่ทำการแบ่งกลุ่ม N ให้ทดลองวิจัย Reduction Attribute data ดังต่อไปนี้โดยวิธีการของ Best First และ Ranker จากโปรแกรม

* เข้าไปที่ tab Select Attribute -> search method -> best first

**Best first เลือก attribute ที่ดีที่สุดคือ Income married children**

* เข้าไปที่ tab Select Attribute -> search method -> Ranker โดยที่ให้จำนวน attribute เท่ากับผลลัพธ์ของ Best first ซึ่งผลลัพธ์ **คือ age sex region income married children**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Select Attribute Method** | | |
| **No select attribute** | **Best-first** | **Ranker** |
| J48 | 87.8333% | 81.3333 % | 81.5 % |
| MLP | 73.1667% | 81.8333 % | 75.8333 % |
| Naïve bayes | 72.8333% | 71.6667 % | 72.6667 % |
| IBk | 70.6667% | 81.8333 % | 69.6667 % |