

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

PROJECT 2 – Classification and Clustering

Lớp: 19KHMT – Nhóm: 17

December 26, 2021

PROGRESS: 100%

GIẢNG VIÊN

Lê Hoài Bắc

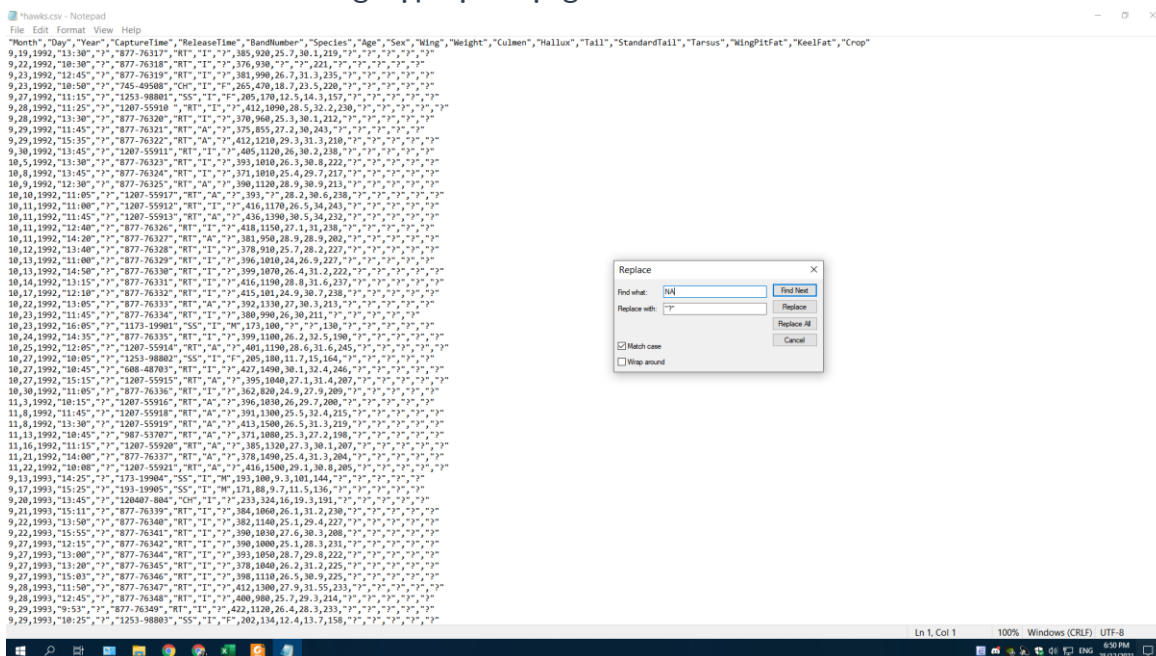
Nguyễn Khánh Toàn

Project 2	Classification			Clustering	Report	Progress
	Explorer	Experimenter	Observation			
Ngô Văn Anh Kiệt 19127191	x	x	x		x	100%
Triệu Nguyên Phát 19127505				x	x	100%

Preprocessing hawks dataset with Weka

- **Bước 1:** Bởi vì các dữ liệu bị thiếu trong file “hawks.csv” được ghi bằng chuỗi “NA”, “” hoặc “ ” nên ta cần phải chuyển về cùng một định dạng đánh dấu dữ liệu bị thiếu mà Weka có thể đọc được là “?”.

Do file data này tương đối nhẹ, ta chỉ cần bật file csv bằng Notepad (Windows 10) hoặc các trình soạn thảo văn bản tương tự để dùng chức năng Replace All thay đổi tất cả chuỗi không hợp định dạng thành “?”.



Hình 1. Thay đổi các giá trị bị thiếu từ NA thành “?”

- **Bước 2:** Ta mở file csv bằng Weka Explorer, vào giao diện Edit của tab Preprocessing để kiểm tra lại các cột có đúng kiểu dữ liệu được mô tả trong trang web nguồn của tập dữ liệu hawks.

Viewer

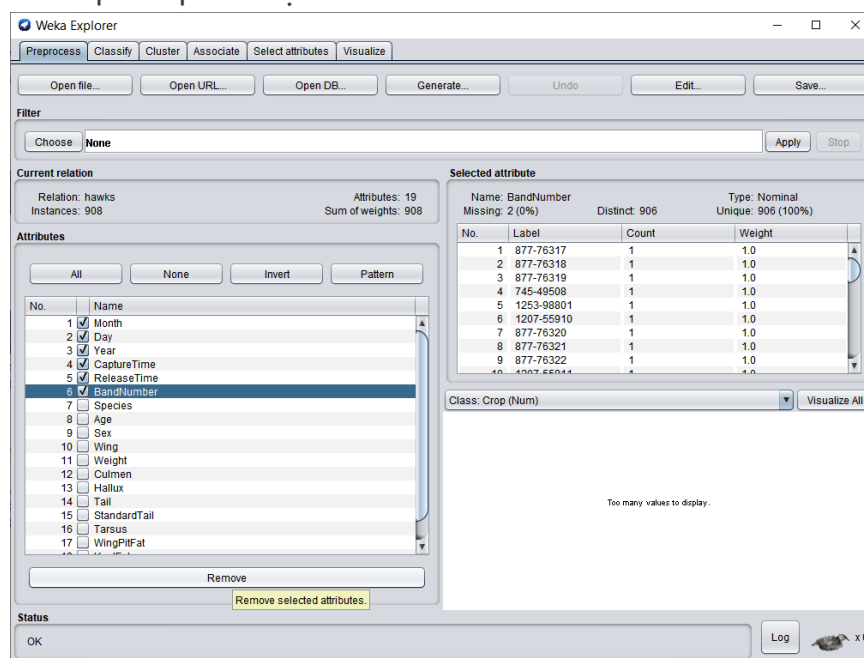
Relation: hawks

No.	1: Month	2: Day	3: Year	4: CaptureTime	5: ReleaseTime	6: BandNumber	7: Species	8: Age	9: Sex	10: Wing	11: Weight	12: Culmen	13: Hallux	14: Tail	15: StandardTail	16: Tarsus	17: WingPIFat	18: KeelFat	19: Crop
...	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
...	10.0	13.0	199...	15.00		1387-79794	RT	I		410.0	1135.0	26.4	32.4	267.0					
...	10.0	14.0	199...	10.05		1387-79195	RT	I		384.0	940.0	26.2	29.8	220.0					
...	10.0	14.0	199...	13.25		1207-64634	RT	I		385.0	920.0	25.0	32.2	238.0					
...	10.0	14.0	199...	13.40		1387-79196	RT	I		398.0	1280.0	28.0	32.4	248.0					
...	10.0	15.0	199...	14.20			CH	A			480.0		17.7	32.1	198.0				
...	10.0	15.0	199...	14.30		1387-79197	RT	I		425.0	1220.0	27.3	33.0	241.0					
...	10.0	15.0	199...	15.00		1397-79198	RT	I		401.0	1000.0	26.7	28.0	212.0					
...	10.0	16.0	199...	12.35		1387-79172	RT	I		387.0	1120.0	26.8	50.2	221.0					
...	10.0	16.0	199...	13.33		1807-53105	RT	I		376.0	925.0	26.0	30.8	233.0					
...	10.0	17.0	199...	11.10		1423-16223	SS	A	M	171.0	90.0	9.9	11.9	138.0					
...	10.0	17.0	199...	11.45		1207-64635	RT	A		420.0	1280.0	27.5	31.8	230.0					
...	10.0	18.0	199...	12.20		877-76379	RT	I		385.0	985.0	27.5	30.8	223.0					
...	10.0	18.0	199...	13.05		2107-64631	RT	I		405.0	1350.0	28.3	32.6	226.0					
...	10.0	18.0	199...	13.15		1807-53107	RT	I		350.0	730.0	24.6	25.7	208.0					
...	10.0	18.0	199...	12.00		1387-79200	RT	I		388.0	890.0	27.8	31.2	231.0					
...	10.0	18.0	199...	12.02		877-76380	RT	I		398.0	1020.0	26.5	31.1	217.0					
...	10.0	20.0	199...	10.10		1207-64637	RT	A		410.0	1000.0	27.1	30.6	230.0					
...	10.0	23.0	199...	10.30		2003-58489	SS	I	F	202.0	150.0	11.7	14.1	153.0					
...	10.0	23.0	199...	12.20		1343-78405	SS	I	F	204.0	180.0	11.5	12.4	186.0					
...	11.0	2.0	199...	12.05		8777-63481	RT	A		382.0	1020.0	26.5	29.4	225.0					
...	11.0	2.0	199...	13.00		1207-64638	RT	I		111.0	1340.0	26.85	31.9	226.0					
...	11.0	20.0	199...	14.10		1207-64639	RT	I		396.0	1300.0	27.3	30.5	214.0					
...	11.0	21.0	199...	11.45		8777-6382	RT	I		363.0	1015.0	25.5	30.1	242.0					
...	10.0	2.0	199...	9.47		1807-53108	RT	I		360.0	900.0	30.5	28.8	220.0					
...	10.0	10.0	199...	10.45		1207-64640	RT	I		390.0	1000.0	26.1	29.6	250.0					
...	10.0	10.0	199...	11.07		1204-45808	SS	I	M	195.0	150.0	12.3	14.6	178.0					
...	10.0	10.0	199...	12.15		1207-64641	RT	A		390.0	1050.0	24.8	32.5	220.0					
...	10.0	10.0	199...	13.20		1207-64642	RT	I		380.0	950.0	24.9	29.0	240.0					
...	10.0	14.0	199...	9.40		1705-24609	CH	I	M	225.0	350.0	12.6	28.0	215.0					
...	10.0	17.0	199...	13.05		1705-24610	CH	A	M	247.0	375.0	16.9	18.2	160.0					
...	10.0	18.0	199...	10.58		1207-64643	RT	A		415.0	1175.0	28.3	33.2	230.0					
...	10.0	23.0	199...	9.25		1387-92101	RT	I		354.0	980.0	25.8	29.3	220.0					
...	10.0	24.0	199...	10.35		1387-92102	RT	A		417.0	1260.0	29.0	32.8	234.0					
...	10.0	24.0	199...	11.18		1387-92103	RT	I		379.0	1050.0	25.9	31.3	226.0					
...	10.0	25.0	199...	10.40		1207-64644	RT	A		412.0	1330.0	29.2	32.2	218.0					
...	10.0	27.0	199...	16.16		1387-92104	RT	A		377.0	980.0	28.0	29.1	209.0					
...	10.0	28.0	199...	12.40		1387-92105	RT	I		372.0	920.0	25.3	30.0	228.0					

Add instance Undo OK Cancel

Hình 2. Các thuộc tính dùng kiểu dữ liệu được mô tả và thấy rõ các cột thiếu dữ liệu

- **Bước 3:** Nhận thấy trong các thuộc tính của tập dữ liệu hawks, có một vài thuộc tính không mang nhiều ý nghĩa cho việc phân loại chủng điều hâu, ví dụ như: ngày, tháng, năm ghi dữ liệu, thời điểm bắt, thả chim, mã số của chim. Ta sẽ loại bỏ các thuộc tính đó ra khỏi tập dữ liệu cần dùng, đồng thời đặt thuộc tính "Species" là thuộc tính lớp cần phân loại.



Hình 3. Loại bỏ một số thuộc tính khỏi tập dữ liệu gốc

Viewer

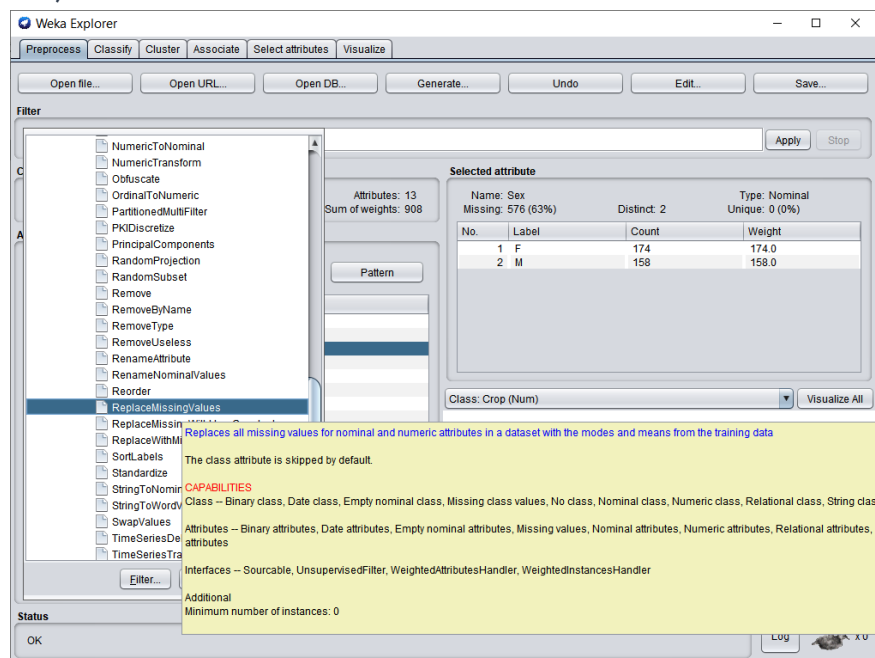
Relation: hawks-weka.filters.unsupervised.attribute.Remove-R1-6-weka.filters.unsupervised.attribute.Reorder-R2,3,4,5,6,7,8,9,10,11,12,13,1

No.	1: Age	2: Sex	3: Wing	4: Weight	5: Culmen	6: Hallux	7: Tail	8: StandardTail	9: Tarsus	10: WingPitFat	11: KeelFat	12: Crop	13: Species
	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
1	I		385.0	920.0	25.7	30.1	219.0						RT
2	I		376.0	930.0			221.0						RT
3	I		381.0	990.0	26.7	31.3	235.0						RT
4	I	F	265.0	470.0	18.7	23.5	220.0						CH
5	I	F	205.0	170.0	12.5	14.3	157.0						SS
6	I		412.0	1090.0	28.5	32.2	230.0						RT
7	I		370.0	960.0	25.3	30.1	212.0						RT
8	A		375.0	855.0	27.2	30.0	243.0						RT
9	A		412.0	1210.0	29.3	31.3	210.0						RT
10	I		405.0	1120.0	26.0	30.2	238.0						RT
11	I		393.0	1010.0	26.3	30.8	222.0						RT
12	I		371.0	1010.0	25.4	29.7	217.0						RT
13	A		390.0	1120.0	28.9	30.9	213.0						RT
14	A		393.0		28.2	30.6	238.0						RT
15	I		416.0	1170.0	26.5	34.0	243.0						RT
16	A		436.0	1390.0	30.5	34.0	232.0						RT
17	I		418.0	1150.0	27.1	31.0	238.0						RT
18	A		381.0	950.0	28.9	28.9	202.0						RT
19	I		378.0	910.0	25.7	28.2	227.0						RT
20	I		396.0	1010.0	24.0	26.9	227.0						RT
21	I		399.0	1070.0	26.4	31.2	222.0						RT
22	I		416.0	1190.0	28.8	31.6	237.0						RT
23	I		415.0	101.0	24.9	30.7	238.0						RT
24	A		392.0	1330.0	27.0	30.3	213.0						RT
25	I		380.0	990.0	26.0	30.0	211.0						RT
26	I	M	173.0	100.0			130.0						SS
27	I		399.0	1100.0	26.2	32.5	190.0						RT
28	A		401.0	1190.0	28.6	31.6	245.0						RT
29	I	F	205.0	180.0	11.7	15.0	164.0						SS
30	I		163.0	1400.0	26.4	30.1	216.0						RT

Add Instance Undo OK Cancel

Hình 4. Đặt Species làm class

- **Bước 4:** Bây giờ, ta bắt đầu tiền xử lý các dữ liệu bị thiếu. Sử dụng bộ lọc weka.filters.unsupervised.attribute.ReplaceMissingValues để thay thế các dữ liệu bị thiếu bằng giá trị mean của thuộc tính (nếu là kiểu số) hoặc mode (nếu là kiểu định danh).



Hình 5. Sử dụng filter ReplaceMissingValues để điền dữ liệu thiếu

Như vậy, ta có bảng dữ liệu sau. Lưu vào file "preprocessed_hawks.arff"

Viewer

Relation: hawks-weka.filters.unsupervised.attribute.Remove-R1-6-weka.filters.unsupervised.attribute.Reorder-R2,3,4,5,6,7,8,9,10,11,12,13,1-weka.filters.unsupervised.attribute.R...

No.	1: Age	2: Sex	3: Wing	4: Weight	5: Culmen	6: Hallux	7: Tail	8: StandardTail	9: Tarsus	10: WingPitFat	11: KeelFat	12: Crop	13: Species
Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	I	F	385.0	920.0	25.7	30.1	219.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
2	I	F	376.0	930.0	21.801...	26.41...	221.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
3	I	F	381.0	990.0	26.7	31.3	235.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
4	I	F	265.0	470.0	18.7	23.5	220.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	CH
5	I	F	205.0	170.0	12.5	14.3	157.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	SS
6	I	F	412.0	1090.0	28.5	32.2	230.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
7	I	F	370.0	960.0	25.3	30.1	212.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
8	A	F	375.0	855.0	27.2	30.0	243.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
9	A	F	412.0	1210.0	29.3	31.3	210.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
10	I	F	405.0	1120.0	26.0	30.2	238.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
11	I	F	393.0	1010.0	26.3	30.8	222.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
12	I	F	371.0	1010.0	25.4	29.7	217.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
13	A	F	390.0	1120.0	28.9	30.9	213.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
14	A	F	393.0	772.0...	28.2	30.6	238.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
15	I	F	416.0	1170.0	26.5	34.0	243.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
16	A	F	436.0	1390.0	30.5	34.0	232.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
17	I	F	418.0	1150.0	27.1	31.0	238.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
18	A	F	381.0	950.0	28.9	28.9	202.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
19	I	F	378.0	910.0	25.7	28.2	227.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
20	I	F	396.0	1010.0	24.0	26.9	227.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
21	I	F	399.0	1070.0	26.4	31.2	222.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
22	I	F	416.0	1190.0	28.8	31.6	237.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
23	I	F	415.0	101.0	24.9	30.7	238.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
24	A	F	392.0	1330.0	27.0	30.3	213.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
25	I	F	380.0	990.0	26.0	30.0	211.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
26	I	M	173.0	100.0	21.801...	26.41...	130.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	SS
27	I	F	399.0	1100.0	26.2	32.5	190.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
28	A	F	401.0	1190.0	28.6	31.6	245.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	RT
29	I	F	205.0	180.0	11.7	15.0	164.0	199.182136...	71.950...	0.7922077...	2.18430...	0.234...	SS

Add Instance Undo OK Cancel

Hình 6. Tập dữ liệu preprocessed_hawks.arff

- **Bước 5:** Do trong thí nghiệm A không sử dụng tập dữ liệu kiểu rời rạc, trong khi thuật toán ID3 thì lại không hoạt động được với dữ liệu liên tục, ta dùng bộ lọc weka.filters.unsupervised.attribute.NumericToNominal để tạm biến đổi các thuộc tính số thành thuộc tính định danh để có thể chạy ID3 cho phần này. Lưu tập dữ liệu này thành file “preprocessed_numeric2nominal_hawks.arff”. Vì tập này chỉ được dùng tạm để chạy ID3 trong thí nghiệm A, kết quả của nó sẽ được ghi cùng dòng với tập “preprocessed_hawks.arff”.

Viewer

Relation: hawks-weka.filters.unsupervised.attribute.Remove-R1-6-weka.filters.unsupervised.attribute.Reorder-R2,3,4,5,6,7,8,9,10,11,12,13,1-weka.filters.unsupervised.attribute.R...

No.	1: Age	2: Sex	3: Wing	4: Weight	5: Culmen	6: Hallux	7: Tail	8: StandardTail	9: Tarsus	10: WingPitFat	11: KeelFat	12: Crop	13: Species
Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	I	F	385	920	25.7	30.1	219	199.182137	71.950...	0.792208	2.184303	0.234...	RT
2	I	F	376	930	21.801...	26.41...	221	199.182137	71.950...	0.792208	2.184303	0.234...	RT
3	I	F	381	990	26.7	31.3	235	199.182137	71.950...	0.792208	2.184303	0.234...	RT
4	I	F	265	470	18.7	23.5	220	199.182137	71.950...	0.792208	2.184303	0.234...	CH
5	I	F	205	170	12.5	14.3	157	199.182137	71.950...	0.792208	2.184303	0.234...	SS
6	I	F	412	1090	28.5	32.2	230	199.182137	71.950...	0.792208	2.184303	0.234...	RT
7	I	F	370	960	25.3	30.1	212	199.182137	71.950...	0.792208	2.184303	0.234...	RT
8	A	F	375	855	27.2	30	243	199.182137	71.950...	0.792208	2.184303	0.234...	RT
9	A	F	412	1210	29.3	31.3	210	199.182137	71.950...	0.792208	2.184303	0.234...	RT
10	I	F	405	1120	26	30.2	238	199.182137	71.950...	0.792208	2.184303	0.234...	RT
11	I	F	393	1010	26.3	30.8	222	199.182137	71.950...	0.792208	2.184303	0.234...	RT
12	I	F	371	1010	25.4	29.7	217	199.182137	71.950...	0.792208	2.184303	0.234...	RT
13	A	F	390	1120	28.9	30.9	213	199.182137	71.950...	0.792208	2.184303	0.234...	RT
14	A	F	393	772.0...	28.2	30.6	238	199.182137	71.950...	0.792208	2.184303	0.234...	RT
15	I	F	416	1170	26.5	34	243	199.182137	71.950...	0.792208	2.184303	0.234...	RT
16	A	F	436	1390	30.5	34	232	199.182137	71.950...	0.792208	2.184303	0.234...	RT
17	I	F	418	1150	27.1	31	238	199.182137	71.950...	0.792208	2.184303	0.234...	RT
18	A	F	381	950	28.9	28.9	202	199.182137	71.950...	0.792208	2.184303	0.234...	RT
19	I	F	378	910	25.7	28.2	227	199.182137	71.950...	0.792208	2.184303	0.234...	RT
20	I	F	396	1010	24	26.9	227	199.182137	71.950...	0.792208	2.184303	0.234...	RT
21	I	F	399	1070	26.4	31.2	222	199.182137	71.950...	0.792208	2.184303	0.234...	RT
22	I	F	416	1190	28.8	31.6	237	199.182137	71.950...	0.792208	2.184303	0.234...	RT
23	I	F	415	101	24.9	30.7	238	199.182137	71.950...	0.792208	2.184303	0.234...	RT
24	A	F	392	1330	27	30.3	213	199.182137	71.950...	0.792208	2.184303	0.234...	RT
25	I	F	380	990	26	30	211	199.182137	71.950...	0.792208	2.184303	0.234...	RT
26	I	M	173	100	21.801...	26.41...	130	199.182137	71.950...	0.792208	2.184303	0.234...	SS
27	I	F	399	1100	26.2	32.5	190	199.182137	71.950...	0.792208	2.184303	0.234...	RT
28	A	F	401	1190	28.6	31.6	245	199.182137	71.950...	0.792208	2.184303	0.234...	RT
29	I	F	205	180	11.7	15	164	199.182137	71.950...	0.792208	2.184303	0.234...	SS

Add Instance Undo OK Cancel

Hình 7. Thuộc tính numeric biến đổi thành nominal để chạy ID3 trong thí nghiệm A

Observation

Which classification method typically has the best result?

Phương pháp phân lớp thường cho kết quả tốt nhất là J48. Theo sau là Naïve Bayes Simple và cuối cùng là ID3.

Which method does not work well and why?

Phương pháp phân lớp **ID3** cho kết quả không tốt, bởi vì tập dữ liệu ban đầu có rất nhiều thuộc tính kiểu số liên tục, trong khi ID3 chỉ hoạt động với kiểu rời rạc. Nếu ta dùng bộ lọc của Weka để chuyển kiểu số thành kiểu định danh (ở thí nghiệm A) hoặc làm rời rạc hóa dữ liệu với độ rộng của bin không đủ lớn (thí nghiệm B), ID3 sẽ bị overfitting rất nặng (kết quả luôn là 100% khi phương pháp test là dùng tập huấn luyện) hoặc fit không tốt như các thuật toán khác. Thuật toán ID3 phụ thuộc rất nhiều vào việc dữ liệu có rời rạc hay không, hoặc phân phối của dữ liệu khi chia bin bằng cách rời rạc hóa. Ngoài ra, do tập dữ liệu ban đầu bị thiếu rất nhiều dữ liệu nên khi rời rạc hóa, các điểm dữ liệu đó sẽ bị trùng nhau, dẫn đến việc chia rổ theo độ rộng trong rời rạc hóa sẽ khiến cho các điểm đó nằm cùng 1 rổ, gây ảnh hưởng rất lớn đến hiệu năng của thuật toán.

Why should we use the discretized version of the data set instead of the original one?

Việc rời rạc hóa các thuộc tính số sẽ giúp các thuộc tính được mô tả đúng với ý nghĩa của nó hơn. Ngoài ra, hiệu quả của các thuật toán như ID3 hay Naïve Bayes Simple phụ thuộc vào sự rời rạc của dữ liệu, nên ta cần phải thực hiện rời rạc hóa dữ liệu.

Do the discretization process and method affect the classification results? If yes then how?

Qua thí nghiệm B và C, ta nhận thấy việc rời rạc và cách rời rạc **có ảnh hưởng** đến kết quả phân lớp. Tuy nhiên, mức độ ảnh hưởng tùy theo từng phương pháp phân lớp.

Ở thí nghiệm A, do không rời rạc hóa dữ liệu nên ID3 không chạy tốt như các phương pháp khác. Ở thí nghiệm B và C thì ID3 đã làm tốt hơn, tuy nhiên thí nghiệm B vẫn bị overfitting khi thử nghiệm bằng tập huấn luyện, do tập dữ liệu gốc thiếu quá nhiều dữ

liệu nên các dữ liệu thiếu được điền bằng các giá trị giống nhau, dẫn đến việc rời rạc hóa theo độ rộng bin mất cân đối.

Thuật toán Naïve Bayes Simple và J48 cũng bị ảnh hưởng bởi việc rời rạc hóa, nhưng không nặng như ID3. Bởi vì Naïve Bayes Simple chỉ quan tâm đến xác suất của các lớp dữ liệu, còn J48 của Weka có thể tự rời rạc hóa dữ liệu trong quá trình chạy (nhưng không đảm bảo tốt hơn việc rời rạc hóa dữ liệu trước khi chạy).

Which evaluation strategy tends to overestimate the accuracy and why?

Phương pháp đánh giá thường hay đưa ra độ chính xác cao hơn là “Use training set”. Bởi vì chúng ta đã dùng tập dữ liệu huấn luyện để cho bộ phân loại học, nên hiển nhiên là bộ phân loại sẽ biết được nhãn của các điểm dữ liệu trong tập huấn luyện đó. Chính vì thế nên khi đánh giá mô hình, ta cần phải dùng một tập dữ liệu khác hoàn toàn với tập dữ liệu dùng để huấn luyện.

Which evaluation strategy tends to underestimate the accuracy and why?

Phương pháp đánh giá thường hay cho ra độ chính xác thấp nhất là “Percentage split 66%”. Điều này là do tập dữ liệu gốc bị thiếu quá nhiều dữ liệu, nên sau khi điền các giá trị bị thiếu, sẽ có rất nhiều điểm dữ liệu bị trùng. Sự thiếu đa dạng dữ liệu khiến cho việc tách bộ dữ liệu gốc gây ảnh hưởng nặng đến quá trình huấn luyện và thử nghiệm.