CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

1.1. Introduction:

For this assignment, we were tasked to use the Weka library to discover patterns and/or irregularities in three data sets. After reading through the API for selected classes in Weka, I was able to write a program that intakes a dataset, an algorithm name, and the number of folds for cross validation, and prints out the Weka evaluation as well as the rules created by the algorithm.

1.2. Algorithms used:

I used three algorithms, J48, JRip, and PART. J48 is a java implementation of the C4.5 algorithm, which is a recursive algorithm that creates a decision making tree. The algorithm follows 5 steps: first, it checks for base cases. If all of the samples belong to the same class, it creates a leaf node for the decision tree that says to choose that class. If none of the features provide any information gain, it creates a decision node higher up the tree using the expected value of the class. If an instance of a previously unseen class is encountered, it again creates a decision node higher up the tree using the expected value of the class. Then, for each attribute a, it finds the normalized information gain ratio from splitting on a. After that, it sets a_best as the attribute with the highest information gain. The algorithm then creates a decision node that splits on a_best. Finally, the algorithm recurs on the sublists obtained from splitting on a_best (https://en.wikipedia.org/wiki/C4.5_algorithm). JRip is a java implementation of the RIPPER algorithm, which builds rules one at a time. It begins by growing one rule by adding antecedents (conditions) until the rule is 100% accurate. The algorithm then prunes every single rule at the same time. These two steps are repeated until the ruleset is deemed "good enough" by a few conditions. Then, in the optimization stage, the ruleset is optimized by changing the weighting of each rule until error is minimized (https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/JRip). PART is a divide and conquer algorithm that creates a partial C4.5 tree in each iteration and makes the "best" leaf into a rule (http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/PART.html).

1.3. Program Structure:

My program consisted solely of a main method, in which I begin by prompting the user through the console for the .arff file to analyze, the algorithm to use, and the number of folds to use in cross-validation. I used a while loop to ensure that the input for the number of folds was an integer greater than or equal to 2. The program then closes the scanner and instantiates a null BufferedReader. The BufferedReader is then constructed as a new BufferedReader with the parameter being a FileReader, in which the parameter is the .arff file. It then creates an Instances object, with the parameter being the BufferedReader. Then, the class index of the Instances is set to the last attribute in the .arff file. The BufferedReader is closed and then the program splits into three if statements depending on which algorithm the user input. In each one, I create an algorithm object (J48, JRip, or PART), then set the classifier to the Instances object. An Evaluation object is then created with the parameter being the Instances object. The model is then cross-validated using the Weka crossValidateModel() method, with the parameters being the algorithm object, the Instances object, the integer input by the user for number of folds, and new Random(1), which means the validation can start anywhere. Finally, the evaluation summary is printed along with the rules for the decision tree.

1.4. Data set 1 -- Iris:

1.4.1. Dataset Description:

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

Iris contains 150 instances, each a certain flower. The attributes include: sepallength, a REAL, sepalwidth, a REAL, petallength, a REAL, petalwidth, a REAL, and class, a nominal attribute for which the choices are “Iris-setosa”, “Iris-versicolor”, and “Iris-virginica”.

1.4.2. Rules:

For J48, the tree decision was as follows:

```
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)
```

This means that the tree’s first decision is based on petalwidth; if petalwidth is less than or equal to 1.5, continue along the upper part of the tree, and if petalwidth is greater than 1.5, continue along the lower part of the tree. One continues to make decisions in such a manner until the class is determined. As one can see from the tree, this tree classified 147/150 flowers correctly.

The rules from JRip are as follows:

```
(petallength >= 3.3) and (petalwidth <= 1.6) and (petallength <= 4.9) => class=Iris-versicolor (46.0/0.0)
(petallength <= 1.9) => class=Iris-setosa (50.0/0.0)
 => class=Iris-virginica (54.0/4.0)
```

This is a pretty straightforward ruleset. If the three conditions in the first line are met, the flower is identified as “Iris-versicolor”. With the remaining flowers, if they meet the conditions in the second line, they are identified as “Iris-setosa”. The remaining unclassified flowers are then all classified as “Iris-virginica”. This algorithm classified 146/150 flowers correctly, with all misidentifications occurring as “Iris-virginica” by nature of JRip.

The rules from PART are as follows:

```
petalwidth <= 0.6: Iris-setosa (50.0)

petalwidth <= 1.7 AND
petallength <= 4.9: Iris-versicolor (48.0/1.0)

: Iris-virginica (52.0/3.0)
```

This ruleset begins by checking if petalwidth is less than 0.6. If it is, the flower is identified as “Iris-setosa”. Then, if the remaining flowers have petalwidth <= 1.7 and petallength <= 4.9, they are classified as “Iris-versicolor”. The rest of the flowers are then classified as “Iris-virginica”. This algorithm classified 146/150 flowers correctly.

1.5. Data set 2 -- Diabetes:

1.5.1. Dataset Description:

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

Diabetes contains 768 instances, each a certain person. The dataset contains 9 attributes: a real preg, a real plas, a real pres, a real skin, a real insu, a real mass, a real pedi, and the class attribute, with nominal choices “tested_negative” and “tested_positive”.

1.5.2. Rules:

The decision tree for J48 was as follows:

```
plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   age <= 30: tested_positive (4.0)
|   |   |   |   |   |   age > 30
|   |   |   |   |   |   |   age <= 34: tested_negative (7.0/1.0)
|   |   |   |   |   |   |   age > 34
|   |   |   |   |   |   |   |   mass <= 33.1: tested_positive (6.0)
|   |   |   |   |   |   |   |   mass > 33.1: tested_negative (4.0/1.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
```

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

| | | | age > 30: tested_positive (60.0/17.0)
| | plas > 157: tested_positive (92.0/12.0)

This decision tree starts at the rightmost point, making comparisons based on mass. As one continues to traverse the tree, they will eventually reach a classification. This algorithm correctly classified 563/768 people.

For JRip, the ruleset was as follows:

(plas >= 132) and (mass >= 30) => class=tested_positive (182.0/48.0)
(age >= 29) and (insu >= 125) and (preg <= 3) => class=tested_positive (19.0/4.0)
(age >= 31) and (pedi >= 0.529) and (preg >= 8) and (mass >= 25.9) => class=tested_positive (22.0/5.0)
=> class=tested_negative (545.0/102.0)

Like before, this ruleset works kind of as a series of else if statements based on the conditions of each line. This algorithm correctly classified 575/768 people.

For PART, the ruleset was as follows:

plas <= 127 AND
mass <= 26.4 AND
preg <= 7: tested_negative (117.0/1.0)

plas > 154 AND
mass > 29.8: tested_positive (100.0/14.0)

plas <= 99 AND
age <= 25 AND
age <= 22: tested_negative (33.0)

age <= 28 AND
skin > 0 AND
skin <= 34 AND
age > 22 AND
preg <= 3 AND
plas <= 127: tested_negative (61.0/7.0)

plas <= 99 AND
insu <= 88 AND
insu <= 18 AND
skin <= 21: tested_negative (26.0/1.0)

age <= 24 AND
skin > 0 AND
mass <= 33.3: tested_negative (37.0)

pres <= 40 AND
plas > 130: tested_positive (10.0)

plas <= 107 AND
pedi <= 0.229 AND
pres <= 80: tested_negative (23.0)

preg <= 6 AND
plas <= 112 AND
pres <= 88 AND
age <= 35: tested_negative (44.0/8.0)

age > 61 AND
preg > 4: tested_negative (11.0)

age <= 30 AND
pres > 72 AND
mass <= 42.8: tested_negative (41.0/7.0)

plas <= 89 AND
plas > 0: tested_negative (13.0/1.0)

: tested_positive (252.0/105.0)

This ruleset can also be read as a series of else if statements, though in comparison to JRip, PART can have multiple if statements for each class. This algorithm correctly classified 557/768 people.

1.6. Data set 3 -- Soybean:

1.6.1. Dataset Description:

Soybean contains 683 soybean plants, each with 35 nominal attributes. The attributes and their choices are listed below.

1. Date: “april”, “may”, “june”, “july”, “august”, “september”, “october”
2. Plant-stand: “normal”, “lt-normal”
3. Precip: “lt-norm”, “norm”, “gt-norm”
4. Temp: “lt-norm”, “norm”, “gt-norm”
5. Hail: “yes”, “no”
6. Crop-hist: “diff-lst-year”, “same-lst-year”, “same-lst-two-years”, “same-lst-sev-years”
7. Area-damaged: “scattered”, “low-areas”, “upper-areas”, “whole-field”
8. Severity: “minor”, “pot-severe”, “severe”
9. Seed-tmt: “none”, “fungicide”, “other”
10. Germination: “90-100”, “80-89”, “lt-80”
11. Plant-growth: “norm”, “abnorm”
12. Leaves: “norm”, “abnorm”
13. Leafspots-halo: “absent”, “yellow-halos”, “no-yellow-halos”
14. Leafspots-marg: “w-s-marg”, “no-w-s-marg”, “dna”
15. Leadspot-size: “lt-⅛”, “gt-⅛”, “dna”

16. Leaf-shred: “absent”, “present”
17. Lead-malf: “absent”, “present”
18. Leaf-mild: “absent”, “upper-surf”, “lower-surf”
19. Stem: “norm”, “abnorm”
20. Lodging: “yes”, “no”
21. Stem-cankers: “absent”, “below-soil”, “above-soil”, “above-sec-nde”
22. Canker-lesion: “dna”, “brown”, “dk-brown-blk”, “tan”
23. Fuiting-bodies: “absent”, “present”
24. External-decay: “absent”, “firm-and-dry”, “watery”
25. Mycelium: “absent”, “present”
26. Int-discolor: “none”, “brown”, “black”
27. Sclerotia: “absent”, “present”
28. Fruit-pods: “norm”, “diseased”, “few-present”, “dna”
29. Fruit-spots: “absent”, “colored”, “brown-w/blk-specks”, “distort”, “dna”
30. Mold-growth: “absent”, “present”
31. Seed-discolor: “absent”, “present”
32. Seed-size: “norm”, “lt-norm”
33. Shriveling: “absent”, “present”
34. Roots: “norm”, “rotted”, “galls-cysts”
35. Class: “diaporthe-stem-canker”, “charcoal-rot”, “rhizoctonia-root-rot”, “phytophthora-rot”, “brown-stem-rot”, “powdery-mildew”, “downy-mildew”, “brown-spot”, “bacterial-blight”, “bacterial-pustule”, “purple-seed-stain”, “anthracnose”, “phyllosticta-leaf-spot”, “alternarialeaf-spot”, “frog-eye-leaf-spot”, “diaporthe-pod-&-stem-blight”, “cyst-nematode”, “2-4-d-injury”, “herbicide-injury”

1.6.2. Rules:

The J48 decision tree was as follows:

```
leafspot-size = lt-1/8
|   canker-lesion = dna
|   |   leafspots-marg = w-s-marg
|   |   |   seed-size = norm: bacterial-blight (21.0/1.0)
|   |   |   seed-size = lt-norm: bacterial-pustule (3.23/1.23)
|   |   leafspots-marg = no-w-s-marg: bacterial-pustule (17.91/0.91)
|   |   leafspots-marg = dna: bacterial-blight (0.0)
|   canker-lesion = brown: bacterial-blight (0.0)
|   canker-lesion = dk-brown-blk: phytophthora-rot (4.78/0.1)
|   canker-lesion = tan: purple-seed-stain (11.23/0.23)
leafspot-size = gt-1/8
|   roots = norm
|   |   mold-growth = absent
|   |   |   fruit-spots = absent
|   |   |   |   leaf-malf = absent
```

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

```
| | | | | fruiting-bodies = absent
| | | | | | date = april: brown-spot (5.0)
| | | | | | date = may: brown-spot (24.0/1.0)
| | | | | | date = june
| | | | | | | precip = lt-norm: phyllosticta-leaf-spot (4.0)
| | | | | | | precip = norm: brown-spot (5.0/2.0)
| | | | | | | precip = gt-norm: brown-spot (21.0)
| | | | | | date = july
| | | | | | | precip = lt-norm: phyllosticta-leaf-spot (1.0)
| | | | | | | precip = norm: phyllosticta-leaf-spot (2.0)
| | | | | | | precip = gt-norm: frog-eye-leaf-spot (11.0/5.0)
| | | | | | date = august
| | | | | | | leaf-shread = absent
| | | | | | | | seed-tmt = none: alternarialeaf-spot (16.0/4.0)
| | | | | | | | seed-tmt = fungicide
| | | | | | | | | plant-stand = normal: frog-eye-leaf-spot (6.0)
| | | | | | | | | plant-stand = lt-normal: alternarialeaf-spot (5.0/1.0)
| | | | | | | | seed-tmt = other: frog-eye-leaf-spot (3.0)
| | | | | | | leaf-shread = present: alternarialeaf-spot (2.0)
| | | | | | date = september
| | | | | | | stem = norm: alternarialeaf-spot (44.0/4.0)
| | | | | | | stem = abnorm: frog-eye-leaf-spot (2.0)
| | | | | | date = october: alternarialeaf-spot (31.0/1.0)
| | | | | fruiting-bodies = present: brown-spot (34.0)
| | | | leaf-malf = present: phyllosticta-leaf-spot (10.0)
| | | fruit-spots = colored
| | | | fruit-pods = norm: brown-spot (2.0)
| | | | fruit-pods = diseased: frog-eye-leaf-spot (62.0)
| | | | fruit-pods = few-present: frog-eye-leaf-spot (0.0)
| | | | fruit-pods = dna: frog-eye-leaf-spot (0.0)
| | | fruit-spots = brown-w/blk-specks
| | | | crop-hist = diff-lst-year: brown-spot (0.0)
| | | | crop-hist = same-lst-yr: brown-spot (2.0)
| | | | crop-hist = same-lst-two-yrs: brown-spot (0.0)
| | | | crop-hist = same-lst-sev-yrs: frog-eye-leaf-spot (2.0)
| | | fruit-spots = distort: brown-spot (0.0)
| | | fruit-spots = dna: brown-stem-rot (9.0)
| | mold-growth = present
| | | leaves = norm: diaporthe-pod-&-stem-blight (7.25)
| | | leaves = abnorm: downy-mildew (20.0)
| roots = rotted
| | area-damaged = scattered: herbicide-injury (1.1/0.1)
```

```
|  |  area-damaged = low-areas: phytophthora-rot (30.03)
|  |  area-damaged = upper-areas: phytophthora-rot (0.0)
|  |  area-damaged = whole-field: herbicide-injury (3.66/0.66)
|  roots = galls-cysts: cyst-nematode (7.81/0.17)
leafspot-size = dna
|  int-discolor = none
|  |  leaves = norm
|  |  |  stem-cankers = absent
|  |  |  |  canker-lesion = dna: diaporthe-pod-&-stem-blight (5.53)
|  |  |  |  canker-lesion = brown: purple-seed-stain (0.0)
|  |  |  |  canker-lesion = dk-brown-blk: purple-seed-stain (0.0)
|  |  |  |  canker-lesion = tan: purple-seed-stain (9.0)
|  |  |  stem-cankers = below-soil: rhizoctonia-root-rot (19.0)
|  |  |  stem-cankers = above-soil: anthracnose (0.0)
|  |  |  stem-cankers = above-sec-nde: anthracnose (24.0)
|  |  leaves = abnorm
|  |  |  stem = norm
|  |  |  |  plant-growth = norm: powdery-mildew (22.0/2.0)
|  |  |  |  plant-growth = abnorm: cyst-nematode (4.3/0.39)
|  |  |  stem = abnorm
|  |  |  |  plant-stand = normal
|  |  |  |  |  leaf-malf = absent
|  |  |  |  |  |  seed = norm: diaporthe-stem-canker (21.0/1.0)
|  |  |  |  |  |  seed = abnorm: anthracnose (9.0)
|  |  |  |  |  leaf-malf = present: 2-4-d-injury (3.0)
|  |  |  |  plant-stand = lt-normal
|  |  |  |  |  fruiting-bodies = absent: phytophthora-rot (50.16/7.61)
|  |  |  |  |  fruiting-bodies = present
|  |  |  |  |  |  roots = norm: anthracnose (11.0/1.0)
|  |  |  |  |  |  roots = rotted: phytophthora-rot (12.89/2.15)
|  |  |  |  |  |  roots = galls-cysts: phytophthora-rot (0.0)
|  int-discolor = brown
|  |  leaf-malf = absent: brown-stem-rot (35.73/0.73)
|  |  leaf-malf = present: 2-4-d-injury (3.15/0.68)
|  int-discolor = black: charcoal-rot (22.22/2.22)
```

As one would expect, J48 returns a decision tree that can be traversed starting at the rightmost point. In this example, the first decision is based on plant-stand. J48 correctly classified 613/683 soybean plants.

The ruleset for JRip was as follows:

```
(leaf-malf = present) and (stem = abnorm) => class=herbicide-injury (8.0/0.0)
(fruit-pods = few-present) => class=cyst-nematode (14.0/0.0)
(shriveling = present) and (stem-cankers = absent) => class=diaporthe-pod-&-stem-blight (15.0/0.0)
(leaf-malf = present) and (leafspots-halo = absent) => class=2-4-d-injury (16.0/0.0)
```

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

(seed-discolor = present) and (canker-lesion = tan) => class=purple-seed-stain (20.0/0.0)
(leaf-malf = present) and (seed = norm) and (leafspot-size = gt-1/8) => class=phyllosticta-leaf-spot (10.0/0.0)
(precip = lt-norm) and (date = june) => class=phyllosticta-leaf-spot (4.0/0.0)
(precip = norm) and (leafspot-size = gt-1/8) and (plant-stand = lt-normal) and (seed-tmt = none) and (hail = yes) => class=phyllosticta-leaf-spot (4.0/0.0)
(fruiting-bodies = present) and (fruit-spots = dna) => class=diaporthe-stem-canker (20.0/0.0)
(leafspot-size = lt-1/8) and (leafspots-marg = w-s-marg) and (seed-size = norm) => class=bacterial-blight (21.0/1.0)
(leaf-mild = upper-surf) => class=powdery-mildew (20.0/0.0)
(leaf-mild = lower-surf) => class=downy-mildew (20.0/0.0)
(int-discolor = black) => class=charcoal-rot (20.0/0.0)
(stem-cankers = below-soil) and (canker-lesion = brown) => class=rhizoctonia-root-rot (20.0/0.0)
(leafspot-size = lt-1/8) => class=bacterial-pustule (19.0/0.0)
(fruit-spots = brown-w/blk-specks) and (leafspots-halo = absent) => class=anthracnose (38.0/0.0)
(stem-cankers = above-soil) and (fruit-pods = norm) => class=anthracnose (5.0/0.0)
(int-discolor = brown) => class=brown-stem-rot (44.0/0.0)
(plant-growth = abnorm) and (canker-lesion = dk-brown-blk) => class=phytophthora-rot (88.0/0.0)
(fruit-pods = diseased) => class=frog-eye-leaf-spot (66.0/2.0)
(date = august) and (germination = 90-100) and (seed-tmt = fungicide) => class=frog-eye-leaf-spot (5.0/0.0)
(date = august) and (seed-tmt = other) => class=frog-eye-leaf-spot (3.0/0.0)
(temp = gt-norm) and (date = september) => class=alternarialeaf-spot (28.0/0.0)
(severity = minor) and (leaf-shread = present) => class=alternarialeaf-spot (11.0/1.0)
(date = august) and (stem = norm) => class=alternarialeaf-spot (23.0/6.0)
(date = october) => class=alternarialeaf-spot (27.0/1.0)
(crop-hist = diff-lst-year) and (precip = gt-norm) => class=alternarialeaf-spot (6.0/1.0)
(date = september) and (crop-hist = same-lst-two-yrs) and (stem = norm) => class=alternarialeaf-spot (3.0/0.0)
 => class=brown-spot (105.0/15.0)
Again, the JRip results can be read as a series of else if statements. JRip correctly classified 617/683 soybean plants.

The ruleset for PART was as follows:

leafspot-size = lt-1/8 AND
canker-lesion = dna AND
leafspots-marg = w-s-marg AND
seed-size = norm: bacterial-blight (21.0/1.0)

int-discolor = none AND
plant-growth = abnorm AND
leaves = abnorm AND
stem = abnorm AND

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

plant-stand = lt-normal AND
area-damaged = low-areas AND
fruiting-bodies = absent: phytophthora-rot (81.29/0.76)

leafspot-size = lt-1/8 AND
canker-lesion = dna: bacterial-pustule (20.31/1.31)

leafspot-size = gt-1/8 AND
mold-growth = present AND
leaves = abnorm: downy-mildew (20.61/0.61)

leafspot-size = gt-1/8 AND
fruit-pods = diseased AND
leaves = abnorm: frog-eye-leaf-spot (66.69/2.69)

fruit-pods = dna AND
leaf-malf = absent: rhizoctonia-root-rot (21.27/1.27)

fruit-pods = diseased AND
stem-cankers = above-sec-nde: anthracnose (39.71/1.71)

fruit-pods = diseased AND
canker-lesion = dna: diaporthe-pod-&-stem-blight (14.67/0.49)

leafspot-size = lt-1/8: purple-seed-stain (11.5/0.5)

leafspots-marg = w-s-marg AND
roots = norm AND
int-discolor = none AND
leaf-malf = absent AND
fruiting-bodies = present: brown-spot (35.0)

stem = norm AND
leafspots-halo = absent AND
leaves = abnorm AND
leaf-malf = absent: powdery-mildew (21.41/1.41)

stem = norm AND
fruit-pods = few-present: cyst-nematode (12.21/0.5)

leafspots-marg = w-s-marg AND
int-discolor = none AND

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

leaf-malf = absent AND
date = october: alternarialeaf-spot (31.0/1.0)

leafspots-marg = w-s-marg AND
int-discolor = none AND
leaf-malf = absent AND
canker-lesion = dna AND
date = september AND
temp = gt-norm: alternarialeaf-spot (28.0)

leafspots-marg = w-s-marg AND
int-discolor = none AND
leaf-malf = absent AND
canker-lesion = dna AND
date = may: brown-spot (24.0/1.0)

leafspots-marg = w-s-marg AND
int-discolor = none AND
canker-lesion = dna AND
leaf-malf = absent AND
date = june AND
precip = gt-norm: brown-spot (21.0)

leafspots-marg = w-s-marg AND
int-discolor = none AND
precip = lt-norm: phyllosticta-leaf-spot (9.0)

leaf-malf = present AND
leafspots-marg = dna: 2-4-d-injury (17.21/4.31)

int-discolor = brown: brown-stem-rot (44.88/0.88)

leafspots-marg = w-s-marg AND
canker-lesion = dna AND
leaf-malf = present: phyllosticta-leaf-spot (6.0)

leafspots-marg = w-s-marg AND
canker-lesion = dna AND
precip = gt-norm AND
date = september AND
leaf-shread = present: alternarialeaf-spot (5.0)

CS 2110: Data Mining and Machine Learning Extra Credit
Christian Kinzer
cfk5ax
12/5/17

leafspots-marg = w-s-marg AND
canker-lesion = dna AND
precip = gt-norm AND
seed = norm AND
date = august AND
hail = yes AND
temp = norm AND
leaf-shread = absent AND
plant-stand = lt-normal: alternarialeaf-spot (8.0/2.0)

leafspots-marg = w-s-marg AND
canker-lesion = dna AND
precip = gt-norm AND
seed = abnorm: alternarialeaf-spot (5.0)

leafspots-marg = w-s-marg AND
canker-lesion = brown: frog-eye-leaf-spot (4.0/1.0)

stem = norm AND
leaves = norm: purple-seed-stain (6.0)

stem = norm AND
precip = gt-norm AND
date = september AND
crop-hist = same-lst-yr: frog-eye-leaf-spot (3.0/1.0)

stem = norm AND
date = april: brown-spot (5.0)

stem = norm AND
precip = gt-norm AND
hail = yes AND
leaf-shread = absent AND
seed-tmt = none AND
area-damaged = scattered: alternarialeaf-spot (7.0/3.0)

stem = norm AND
date = june: brown-spot (5.0/2.0)

stem = norm AND
hail = yes AND
precip = gt-norm AND

leaf-shread = absent AND
seed-tmt = none AND
date = august: alternarialeaf-spot (6.0/1.0)

stem = norm AND
leaf-shread = absent AND
date = august: frog-eye-leaf-spot (10.0/1.0)

stem = norm AND
precip = gt-norm AND
area-damaged = low-areas: alternarialeaf-spot (5.0/1.0)

stem = norm AND
precip = gt-norm: frog-eye-leaf-spot (9.0/3.0)

plant-stand = normal AND
precip = lt-norm: charcoal-rot (20.0)

roots = rotted AND
area-damaged = low-areas: phytophthora-rot (4.45)

leafspots-halo = no-yellow-halos AND
temp = lt-norm: herbicide-injury (2.78)

fruiting-bodies = present: diaporthe-stem-canker (20.0)

precip = gt-norm AND
leaves = abnorm: anthracnose (6.0)

plant-stand = normal: purple-seed-stain (3.0)

: phyllosticta-leaf-spot (2.0)
Much like JRip, the PART ruleset can be read as a series of else if statements. For this dataset, PART correctly classified 628/683 soybean plants.

1.7. Results:

Table 1: J48

		Folds (k)	Folds (k)	Folds (k)	Folds (k)	Folds (k)
Dataset		3	5	10	20	50
Iris	% accuracy	93.3333	96.0000	96.0000	96.0000	96.0000

Diabetes	% accuracy	73.3073	71.2240	73.8281	75.1302	75.7813
Soybean	% accuracy	89.7511	90.7760	91.5081	92.3865	92.3865

Table 2: JRip

		Folds (k)	Folds (k)	Folds (k)	Folds (k)	Folds (k)
Dataset		3	5	10	20	50
Iris	% accuracy	95.3333	94.6667	95.3333	95.3333	95.3333
Diabetes	% accuracy	74.8698	74.4792	76.0417	74.6094	77.0833
Soybean	% accuracy	90.3367	92.5329	92.2401	90.7760	91.9473

Table 3: PART

		Folds (k)	Folds (k)	Folds (k)	Folds (k)	Folds (k)
Dataset		3	5	10	20	50
Iris	% accuracy	94.6667	94.0000	94.0000	95.3333	95.3333
Diabetes	% accuracy	72.5260	74.2168	75.2604	72.6563	72.1354
Soybean	% accuracy	91.9473	90.9224	91.9473	91.3616	92.2401

1.8. Conclusion:

When observing the tables in section 1.7., one can see that, generally, as k increases (moving to the right across a row), % accuracy also increases. However, there are a number of instances where increasing k actually decreases % accuracy. This can be excused by the fact that every cross-validation is random, and that it is possible to receive a more accurate ruleset if the cross-validation just so happens to be effective, regardless of the magnitude of k. Also, regarding rules generated by each algorithm for each dataset, it was clear to see that as the number of attributes in a dataset increased, so did the size of the ruleset. For example, the decision trees made by J48 for iris were considerably smaller than those generated for soybean. This correlation can be resolved to the notion that some characterizations are more complicated than other. There were many conditions required for each class in soybean, whereas there were only 5 total attributes for iris. Regarding accuracy, one can tell by comparing the results for diabetes and soybean that as the number of classes increases, so too does the % accuracy. As diabetes only had 2 classes, it had to resolve a multitude of different variables into a binary decision, which was predictably less accurate than soybean. For each dataset, it appears as if there are outliers. For example, in the iris dataset, each

algorithm reached an upper limit on % accuracy regardless of how large k got. These kinds of upper limits appeared for every dataset in every algorithm; this notion goes to show that it is sometimes incredibly difficult to classify an object based on a limited number of variables, and that a computer is probably not best suited to make those classifications.