DDML - JeongEun Choi - Feedback for the last meeting

!!!! Have to check some values of the table again and add some more detail. BUt hope this is enough to show how I understood the exercises and result :)

Exercise 1

- Condition:
  - Datasets: contact-lenses, iris, zoo
  - Classifiers: ConjunctiveRule, JRip, Prism
  - Test Option: use training set
  - all other values or parameters are in default values

- Details:

Classifiers - we applied FilteredClassifier for all classifiers.
Ex.for Prism:

FilteredClassifier using **weka.classifiers.rules.Prism** on data filtered through weka.filters.supervised.attribute.Discretize -R first-last -precision 6

Datasets -

	contact-lenses	Iris	zoo
#Instances	24	150	101
# attributes	5	5	18

- Results:

	ConjunctiveRule			JRip			Prism		
Datasets	Contact -lenses	Iris	Zoo	Contact -lenses	Iris	Zoo	Contact -lenses	Iris	Zoo
#Rules	1	1	1	3	3	6	9	16	41
#Bedingungen	0	1	1	3	2	6	26	51	41
#predicted classes	1	1	1	3	3	6	3	3	7
% of correctly classified instances	62.5	66.6	60.396	87.5	96	92	100	94	100

Table 1 - Answer For Exercise. 1a)

- Explanation to Table1: Answer for Exercise 1b)

Note!!
- Bedingungen refers to, in case of (A and B), Bedingungen = {A,B}
- predicted classes refers to the classes defined through the rules produced by the classifiers

Default-Rule:
- **ConjunctiveRule**:
  1. it takes the first instance of the dataset
  2. Create a rule for the class that the instance is being classified to
  3. Check all other instances that are classified to the same class and correct the rule from Step2.

using

i. It prunes the rule uding Reduced Error Prunning (REP)[1]

4. The final rule should be TRUE for all instances of that specific class

So, it is beginning with one very generic rule for one class that is TRUE for all instances of that class. There for the number of rules is always one and sometimes there is no condition (and it is TRUE for all instances in the dataset)

because of prunning?

Ex. (detailed results from weka)

- JRip:[2]
  1. Initialize an "empty" rule (so it is TRUE for every instances) for each class from less prevalent one ont to the more frequent one
  2. Grow one rule by adding conditions to the rule until the rule is perfect (100%). It tires every value possible vale of each attribute and selects the condition with highest information gain.
  3. Prune each rule according to the pruning metric
  4. Repeat Step2 and Step3 until the description length(DL) of the ruleset and examples is 64bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%

Simply stated, it creates a rule that is TRUE for each class and tries to keep the rule as short as possible. However, because the rules of each classes are created subsequently for later classes the rules are not as good as the rule for the very first class.

Ex. (detailed result from weka)

- Prism:[3][4]
  1. calculate the probability of occurence of each attribute-class pair
  2. Select the attribute-value pair with the maximum probability and create a subset of the dataset comprising all instance which contain that attribute-value pair
  3. Repeat Step1 and Step2 for this subset until it contains only instances of that class. The induced rule is a conjunction of all the attribute-value pairs used in creating the homogeneous subset
  4. Remove all instances covered by this rule from the training set and repeat step1-4 until all instances are covered. Can only deal with nominal attributes. Can't deal with missing values. Doesn't do any pruning.==>weka/rules/prism

So this algorithm creates the most precise rules and prunes from that rules. Therefore the number of rules and number of conditions are greater than that of other classifiers.

Ex. (detailed result from weka)

Answer for Exercise 1c)

According to the %of correctly classified instances we cans see that for that dataset 'iris' none of the classifiers were able to reach 100%. Therefor**e, 'iris' is the most difficult to learn.**

The best to learn is 'contact-lenses' because first we know that we can reach 100% correctness without regarding the effectiveness of the rules. However, since the number of rules needed to reach 100% is much smaller in case of contact-lenses and the dataset 'contact-lenses' are much smaller than 'zoo', the dataset **'contact-lenses' is the best to learn.**

Answer for Exercise 1d) allgemeinheit=common point?

So, from all the results and analysis or the results we can predict that the generality of JRip is higher than that of the Prism. Prism is using precision, meaning it does not begin with the most general rule

---

[1] http://weka.sourceforge.net/doc.stable/weka/classifiers/rules/ConjunctiveRule.html

[2] http://weka.sourceforge.net/doc.stable/weka/classifiers/rules/JRip.html

[3] http://weka.sourceforge.net/doc.packages/simpleEducationalLearningSchemes/weka/classifiers/rules/Prism.html

[4] https://ac.els-cdn.com/S0020737387800032/1-s2.0-S0020737387800032-main.pdf?_tid=3201e766-d42f-11e7-8317-00000aab0f27&acdnat=1511868662_8ff55ebf0014bacb60edbac36a205a22

but most precise rule. JRip is using both gaining and prune in order keep the balance between the correctness and the efficiency of rules (i.e. decreasing the size of the rules)

Exercise 2

Answer for 2a)
Conditions:
- change the the test options accordingly: 1x5,1x10,1x20 Cross-Validation, Leave-One-Oue (#Intances = # Folds-Cross-Validation), Training set
- divide the dataset using StratifiedRemoveFolds. The first set is the training set and the second is the validatioset
- in case of 10x10 CV: change the seed to 10
- dataset used: abalone, anneal, anneal.Origin,audiology,autos
- classifier used: JRip
- all other conditions are kept the same

Details:
Dataset -

	abalone	Anneal	Anneal.origin	audiology	autos
# instances trainingset	2089	449	449	113	103
# instances validationset	2088	449	449	113	102
Total	4177	898	898	226	205

Results:

% of correctly classified instances	abalone	Anneal	Anneal.origin	audiology	autos
1x5 CV	17.24	96.43	91.98	63.71	57.2
1x10 CV	18.28	97.32	91.75	69.02	58.2
1x20 CV	17.52	96.88	92.20	70.79	48.54
Leave-One-Out	18.19	97.55	91.98	66.37	48.54
Training Set	19.24	98.66	93.98	76.1	81.55
10x10 CV	17.80	98.66	92.6	69.0	54.3
Valideirungsset	17.76	98.21	93.09	68.14	65.68

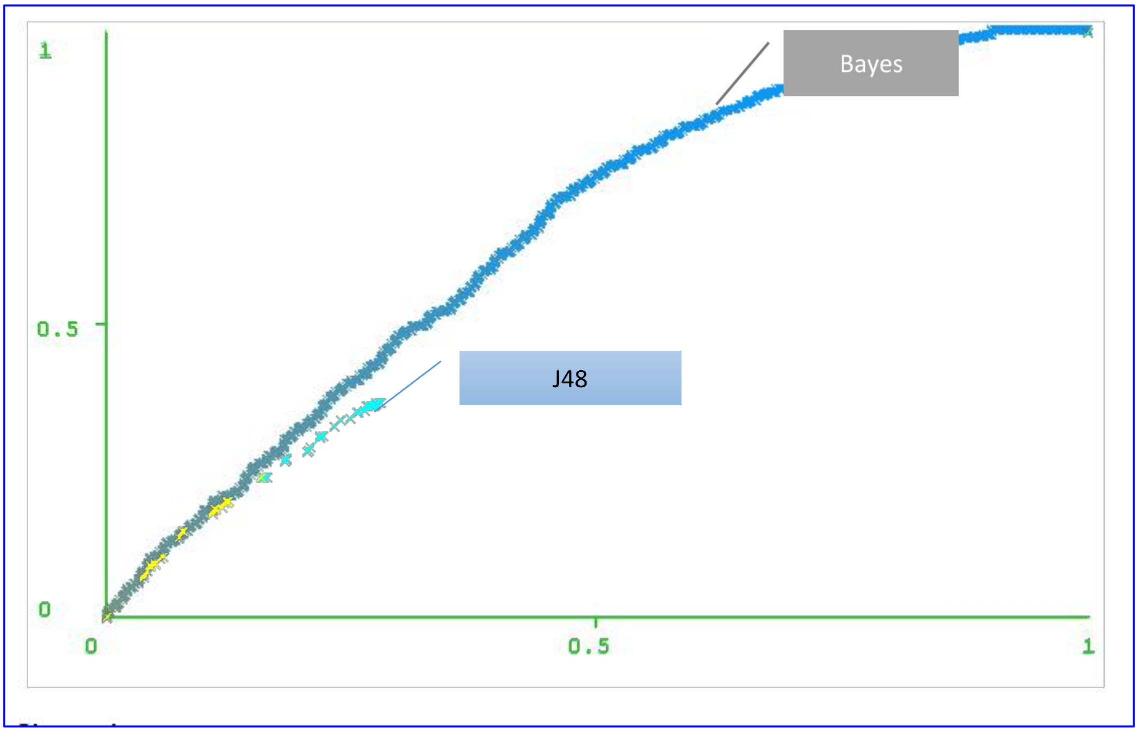
different

Table 2 - Results for 2a and 2b and 2c, % of correctly classified instances

Explanation:
2a)The increase in folds in simple cross validation does not influence the correctness of the rules that much. This is due to how this classifier works. It repeats gaining and pruning several times, so that after reaching "maximum" percentage of correctness we cannot experience much improvement.
Leave-One-Out does as many folds as the number of instances in the dataset, so... (have to do some more research here,,,) best result of CV conditions=>have enough training models
Trainingset obviously received the highest correctness. This is the highest correctness that the classifier could reach using the dataset as trainingset. WE can see that none of the cross validation reached more than the percentage correctness of traning set.
2b)Random seeds could improve the correctness when chosen properly. In our case this applies to every dataset. (more research needed)
2c) The result of the validationset corresponds the results of other crossvalidation results.

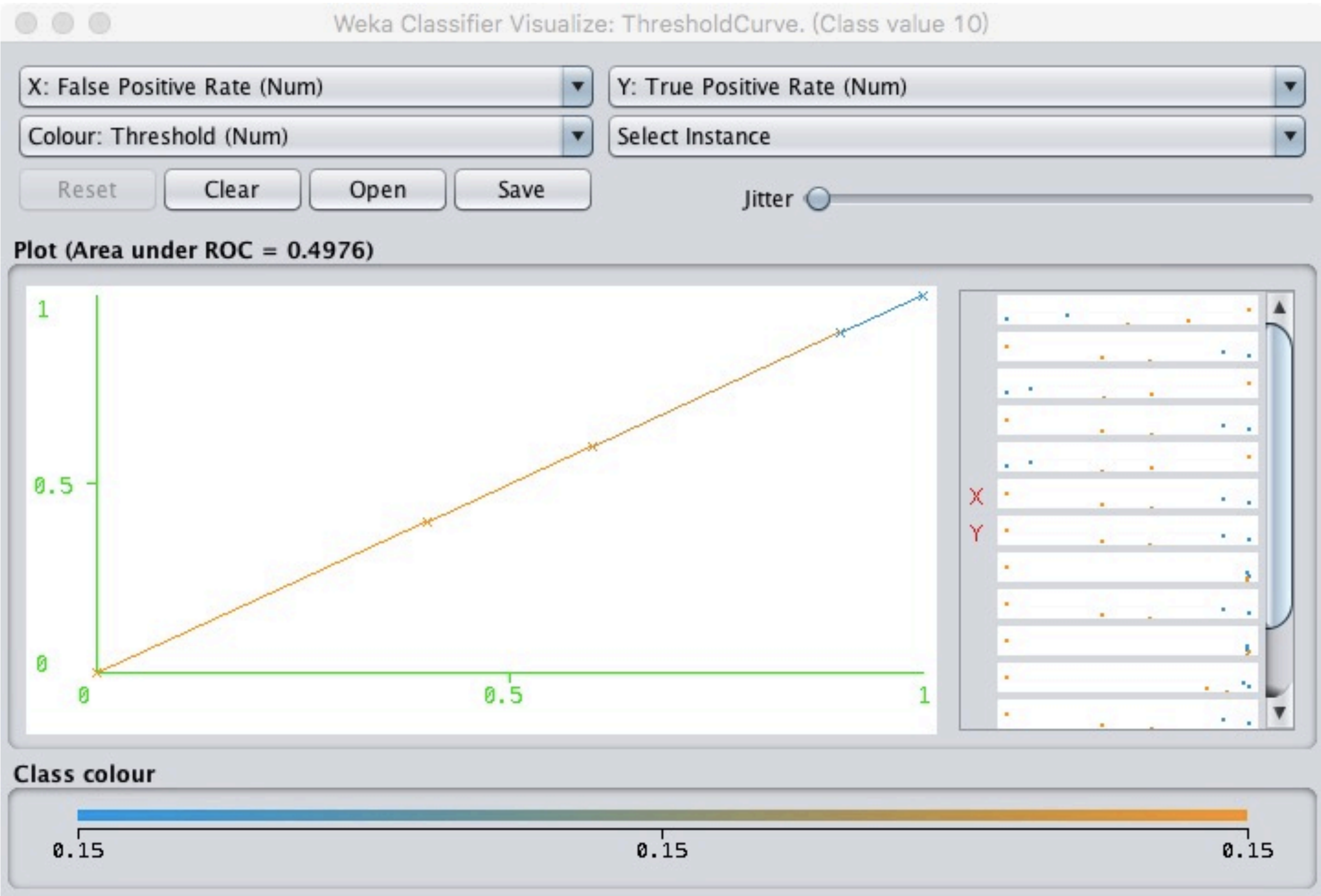
2b) Wenden Sie hierzu zehnmal eine 1x10 Cross-Validation mit 10 unterschiedlichen Random-Seeds an und mitteln die erzielten Genauigkeiten.=>
it not means we should do 10 times 1*10CV use different RS?

Exercise 3



NEED TO DO SOME RESEARCH ON THIS AS WELL

1、have you find a graph-software can display arff-file?
2、ein ARFF-File exportieren, und dieses (nach Löschen des Headers) ?
3、Vergleichen Sie für einen ausgewählten Klassifikationsdatensatz die ROC-Kurven und die Fläche unter diesen Kurven für die Klassifizierer J48 und NaiveBayes.====>insgesamt 3 kurven ?



Weka Classifier Visualize: ThresholdCurve. (Class value 10)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.4976)
1
0.5
0
0
0.5
1
X
Y
Class colour
0.15
0.15
0.15