Marcin Rabiza

Decision Trees Case Study

MANUAL CALCULATIONS

1. Age attribute

Original data:

Age	38	53	21	84	42	55	17	5	88	61	10	7	13
Class	T	T	N	N	T	N	T	N	N	T	N	T	T

$$p_N = -\frac{6}{13} log_2 \frac{6}{13} = 0{,}51$$

$$p_T = -\frac{7}{13} log_2 \frac{7}{13} = 0{,}48$$

$$E(S) = p_N + p_T = 0{,}99$$

Sorted records:

Age	5	7	10	13	17	21	38	42	53	55	61	84	88
Class	N	T	N	T	T	N	T	T	T	N	T	N	N

Entropy only needs to be evaluated between points of different classes (Fayyad & Irani, 1992).

Potential splits: 5|7, 7|10, 10|13, 17|21, 21|38, 53|55, |55|61, 61|84.

Gain(5|7) = Info([6,7]) – Info([1,0],[5,7]) = 0,99 – 0,90 = 0,09

Gain(7|10) = Info([6,7]) – Info([1,1],[5,6]) = 0,99 – 0,98 = 0,01

Gain(10|13) = Info([6,7]) – Info([2,1],[4,6]) = 0,99 – 0,96 = 0,03

Gain(17|21) = Info([6,7]) – Info([2,3],[4,4]) = 0,99 – 0,99 = 0,00

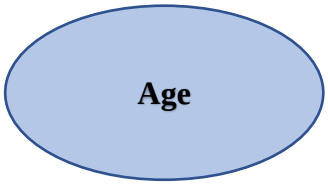
Gain(21|38) = Info([6,7]) – Info([3,3],[3,4]) = 0,99 – 0,99 = 0,00

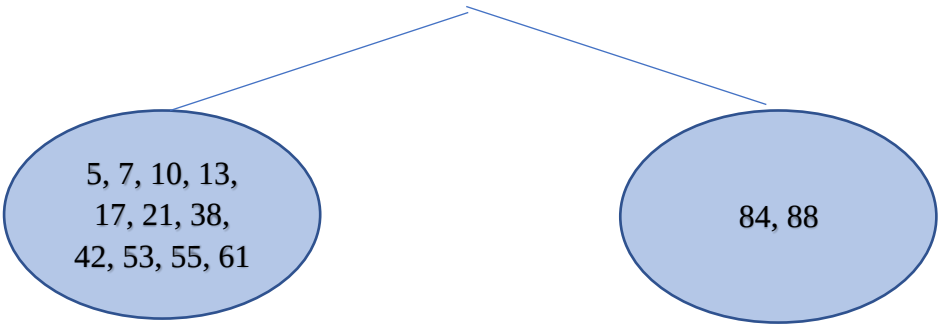
Gain(53|55) = Info([6,7]) – Info([3,6],[3,1]) = 0,99 – 0,89 = 0,10

Gain(55|61) = Info([6,7]) – Info([4,6],[2,1]) = 0,99 – 0,96 = 0,03

Gain(61|84) = Info([6,7]) – Info([4,7],[2,0]) = 0,99 – 0,80 = 0,19 → **The highest Info Gain!**

Because of the highest information gain the best attribute split is between 61 and 84 values.





2. Party attribute

Original data:

Hangover	Exam	Weekend	Party
no	easy	no	yes
no	hard	no	no
no	no	no	yes
no	no	yes	yes
yes	easy	no	no
yes	hard	no	no
yes	no	no	no
yes	no	yes	no

Evaluation of accuracy and coverage:

Hangover = no	→	**Party = yes 3/3**
Exam = easy	→	Party = yes 1/1
Exam = hard	→	Party = yes 0/1
Exam = no	→	**Party = yes 2/2**
Weekend = no	→	Party = yes 2/3
Weekend = yes	→	Party = yes 1/1

Induced rules:

If Hangover = no *and Exam = no* *then Party = yes*	*If Hangover = no* *and Exam = easy* *then Party = yes*

BREAST CANCER

The aim of breast cancer modelling is to enhance the accuracy in identifying breast cancer patients. Classifiers such as decision trees are used as the prediction models for decision-making system in the prognosis of breast cancer survivability.

1. Introduction

WEKA breast cancer data set contains10 attributes and total 286 rows. The attributes are:

- *Age* - patient's age at the time of diagnosis;
- *Menopause* - menopause status of the patient at the time of diagnosis;
- *Tumor size* - tumor size (in mm);
- *Inv-nodes* - range 0 - 39 of axillary lymph nodes showing breast cancer at the time of histological examination;
- *Node caps* - penetration of the tumor in the lymph node capsule or not;
- *Degree of malignancy* - range 1-3 the histological grade of the tumor. That are grade: 1 predominantly that consist of cancer cells, grade: 2 neoplastic that consist of usual characteristics of cancer cells, grade: 3 predominately that consist of cells that are highly affected;
- *Breast* - breast cancer may occur in either breast;
- *Breast quadrant* - if the nipple consider as a central point the breast may be divided into four quadrants;
- *Irradiation* - patient's radiation (x-rays) therapy history.
- *Class* - no-recurrence or recurrence depending reappearing symptoms of breast cancer in the patients after treatment.

Attributes and their possible values are shown in Table 1. Visual form of breast cancer survivals using all attributes is shown on Figure 1.

Attributes	**Values**
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50- 54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33- 35, 36-39
node-caps	yes, no

deg-malig	1, 2, 3
breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiation	yes, no
class	no-recurrence-events, recurrence-events

Table 1. Attributes and their possible values.

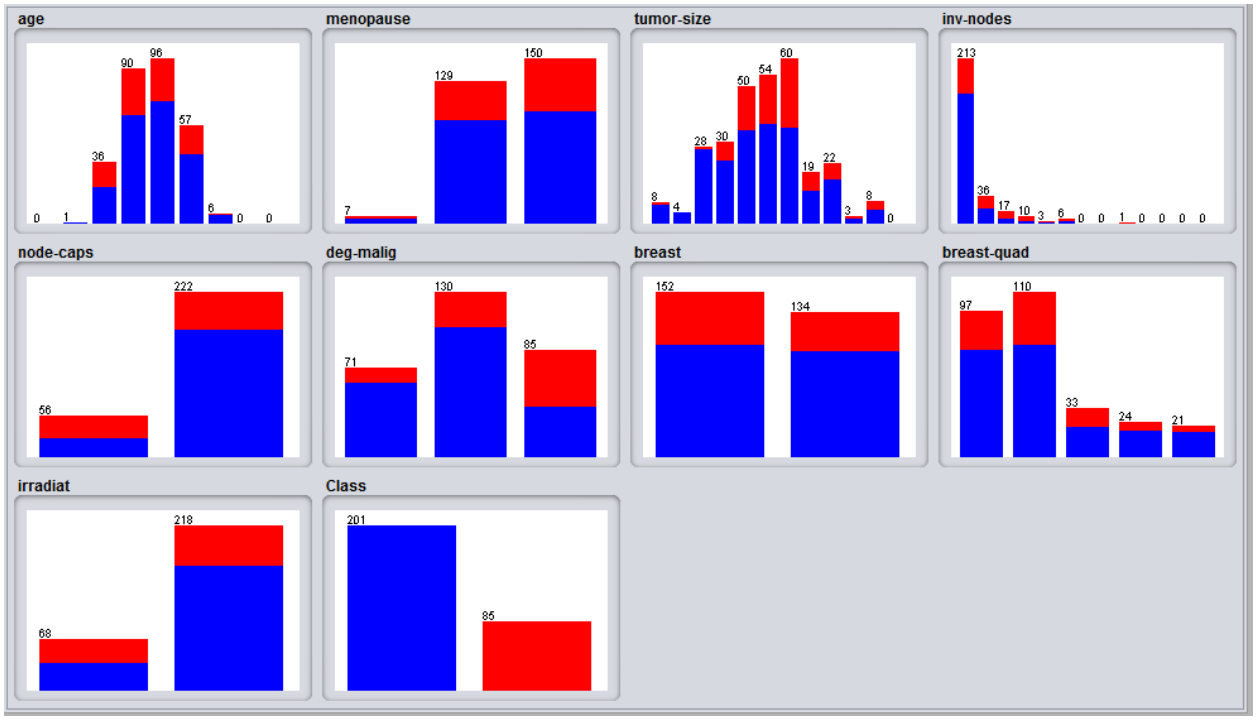


Figure 1. Visual form of breast cancer survivals using all attributes.

2. Decision trees

A decision tree classifier for the breast cancer problem has been induced in WEKA software using J48 classifier (10 cross-validation fold, Confidence Factor for pruning of 0.25). The decision tree has 4 leaves and the size of 6. It is shown on Figure 2.1.

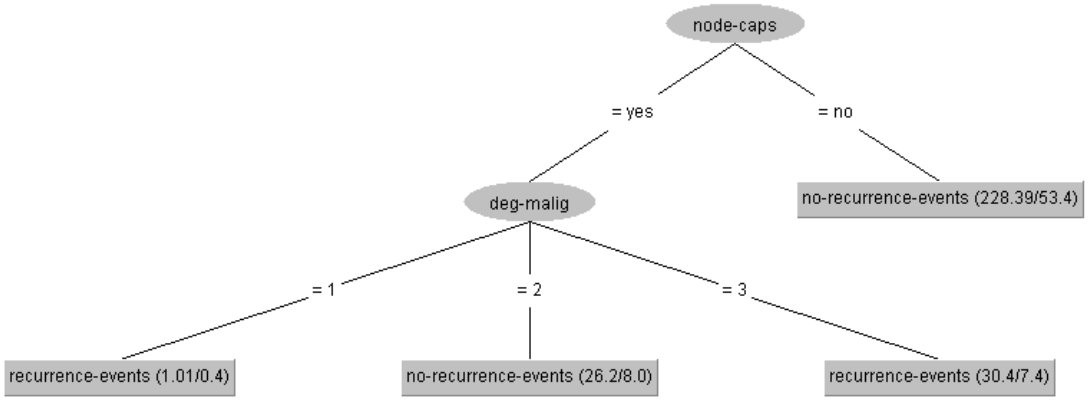


Figure 2.1. Pruned decision tree induced for breast cancer problem.

The most important statistics of the classification are shown in Tables 2.1 and 2.2 below.

Correctly Classified Instances	216	75.5245 %
Incorrectly Classified Instances	70	24.4755 %

Table 2.1. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,960	0,729	no-recurrence events
0,271	0,040	recurence events

Table 2.2. TP and FP rates per class.

J48 classifier algorithm was modified in order to induce an **unpruned** decision tree. The tree has a much more complex and unreadable structure: has 152 leaves and the size of 179. It is shown on Figure 2.2.

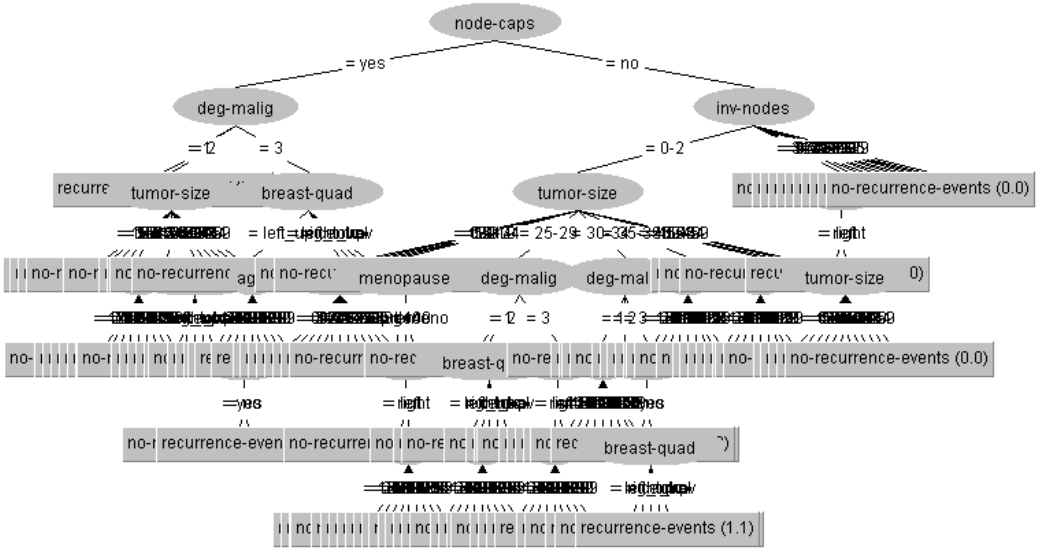


Figure 2.2. Unpruned decision tree induced for breast cancer problem.

Unpruned tree algorithm also results in lower correctly classified instances rate and lower true positive rate of no-reccurence events classification. However, true positive rate of recurrence events and is significantly higher (0,271 vs 0,341) and false positive rate of no-recurrence evets is significantly lower (0,659 vs 0,729) than in case of pruned tree classification. Classification statistics are show in Tables 2.3 and 2.4.

Annotation: Higher recurrence events TP rate is a desirable result from the medical point of view – it is better for a patient to classify cancer recurrence correctly than no-recurrence, because the former case is seen as a life-threatening scenario and requires medical intervention. Also lower FP rate of no-recurrence events is advantageous, because high false positive outcome of no-recurrence means that patient may be diagnosed as healthy, while she actually has cancer recurrence.

Correctly Classified Instances	199	69,5804 %
Incorrectly Classified Instances	87	30,4196 %

Table 2.3. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,846	0,659	no-recurrence events
0,341	0,154	recurrence events

Table 2.4. TP and FP rates per class.

Confidence Factor parameter of tree induction is a parameter of pruning – smaller values incur more pruning. Classification results analysis has been performed in function of a Confidence Factor value. Values of TP Rate of cancer recurrence events classification and FP Rate of cancer no-recurrence events classification were researched according to Confidence Factor value. The results are shown on Figure 2.3.

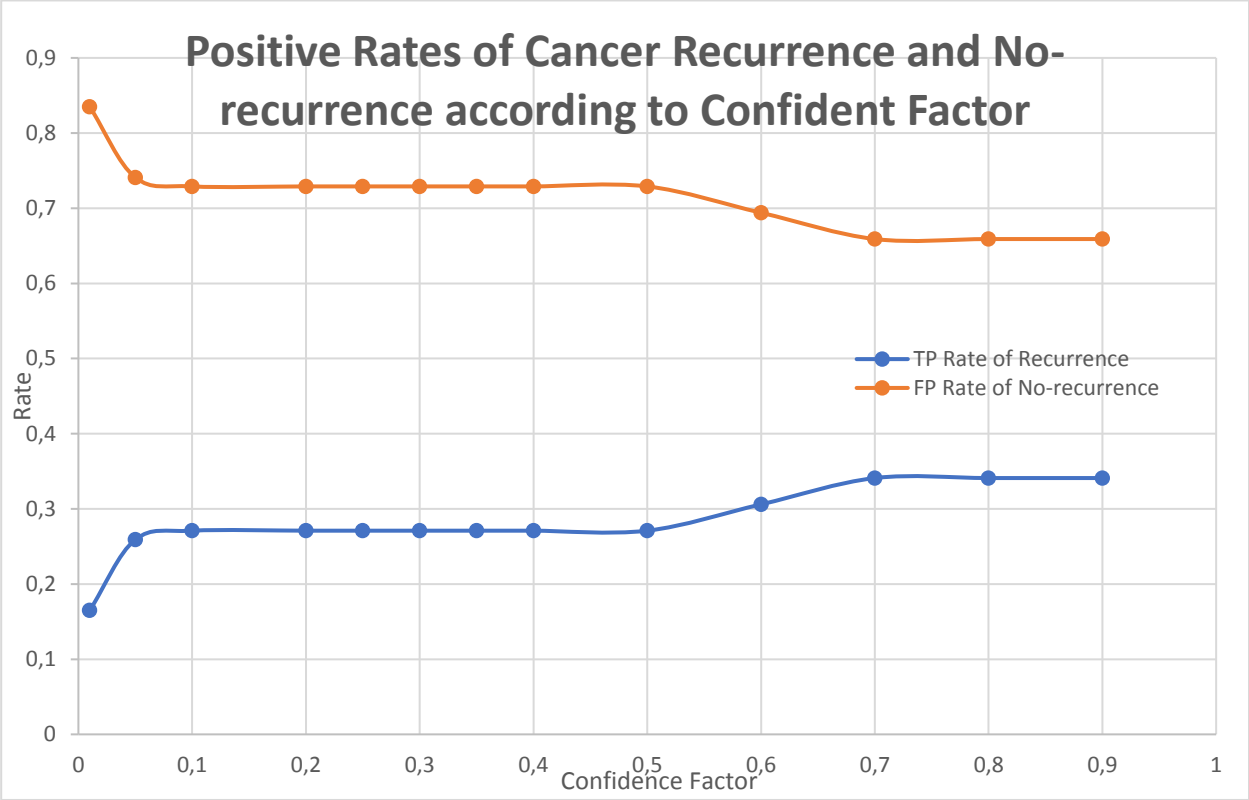


Figure 2.3. Positive Rates of Cancer Recurrence and No-recurrence according to Confident Factor

Chart lines indicate that the higher Confidence Factor (so less pruning) the higher TP Rate of Recurrence classification and the lesser FP Rate of No-recurrence classification are.

To conclude, one may say that using unpruned tree results in lower Correctly Classified score, but at the same time seems to be better choice from medical point of view, since higher TP recurrence and FP no-recurrence rates help identify patient with cancer recurrence. Having that in mind, it must be said that unpruned tree visualisation is much less readable than the pruned ones.

3. Induction of rules

An induction rule classifier for the breast cancer problem has been induced in WEKA software using JRip classifier (10 cross-validation fold, batchSize = 100, folds = 3, pruning). There were 3 rules induced (shown on Figure 3.1):

```
(deg-malig = 3) and (node-caps = yes) => Class=recurrence-events (30.0/7.0)
(inv-nodes = 3-5) and (breast = left) => Class=recurrence-events (11.0/4.0)
 => Class=no-recurrence-events (245.0/55.0)
```

Figure 3.1. Rules induced using JRip

The most important statistics of the classification are shown in Tables 3.1 and 3.2 below.

Correctly Classified Instances	203	70.979 %
Incorrectly Classified Instances	83	29.021 %

Table 3.1. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,856	0,635	no-recurrence events
0,365	0,144	recurrence events

Table 3.2. TP and FP rates per class.

JRip classifier algorithm was modified in order to induce an **unpruned** rules. Classification resulted in 5 rules shown on Figure 3.2.

```
(deg-malig = 3) and (node-caps = yes) and (breast = left) and (irradiat = yes) => Class=recurrence-events (11.0/0.0)
(deg-malig = 3) and (node-caps = yes) and (irradiat = no) and (menopause = premeno) => Class=recurrence-events (8.0/0.0)
(deg-malig = 3) and (inv-nodes = 3-5) and (node-caps = no) and (menopause = ge40) => Class=recurrence-events (4.0/0.0)
(deg-malig = 3) and (age = 60-69) and (breast = right) and (tumor-size = 30-34) => Class=recurrence-events (2.0/0.0)
 => Class=no-recurrence-events (261.0/60.0)
```

Figure 3.2. Rules induced using JRip

Rules with pruning disabled result in higher Correctly Classified Instances Rate, but in this case True Positive Rate of recurrence events classification is significantly lower (0,271 vs 0,465) and False Positive Rate of no-recurrence events classification is significantly higher (0,729 vs 0,659). Prediction measures are shown in Tables 3.3 and 3.4.

Correctly Classified Instances	210	73.4266 %
Incorrectly Classified Instances	76	26.5734 %

Table 3.3. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,930	0,729	no-recurrence events
0,271	0,070	recurrence events

Table 3.4. TP and FP rates per class.

Folds parameter determines the amount of data used for pruning. Classification results analysis has been performed in function of a Folds value. Values of TP Rate of recurrence events and FP Rate of no-recurrence events were research according to pruning Folds value. The results are shown on Figure 3.3.

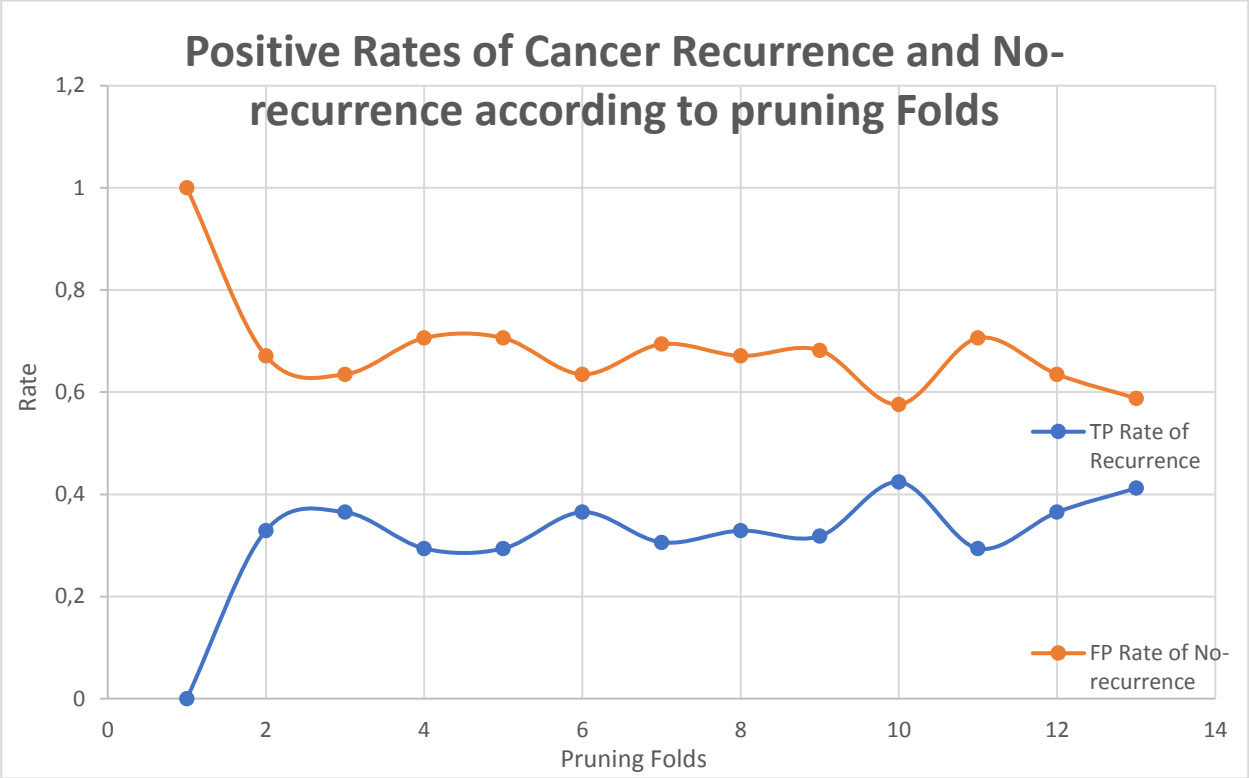


Figure 3.3. Positive Rates of Cancer Recurrence and No-recurrence according to pruning Folds.

Chart lines are not very regular, but it seems that certain Folds values are better in terms of recurrence TP and no-recurrence FP classification than others. The best results in this analysis in terms of cancer diagnostics were generated for Folds value of 10. Additional observation is, that for the last Fold value of 13, only 2 rules were produced.

To conclude, one may say that using pruned rules results in lower Correctly Classified score, but at the same time seems to be better choice from medical point of view, since higher TP recurrence and FP no-recurrence rates help identify patient with cancer recurrence. Also, pruning seems to constraint the number of generated rules.

IONOSPHERE

Ionosphere data set is aimed at classification of radar returns from the ionosphere. The radar data was collected by a system in Goose Bay, Labrador. The targets were free electrons in the ionosphere.

1. Introduction

WEKA ionosphere data set contains 34 continuous attributes, 1 nominal attribute (“Good”, “Bad”) and total 351 rows.

"Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. Received signals

were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers for the Goose Bay system. Instances in the database are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal.

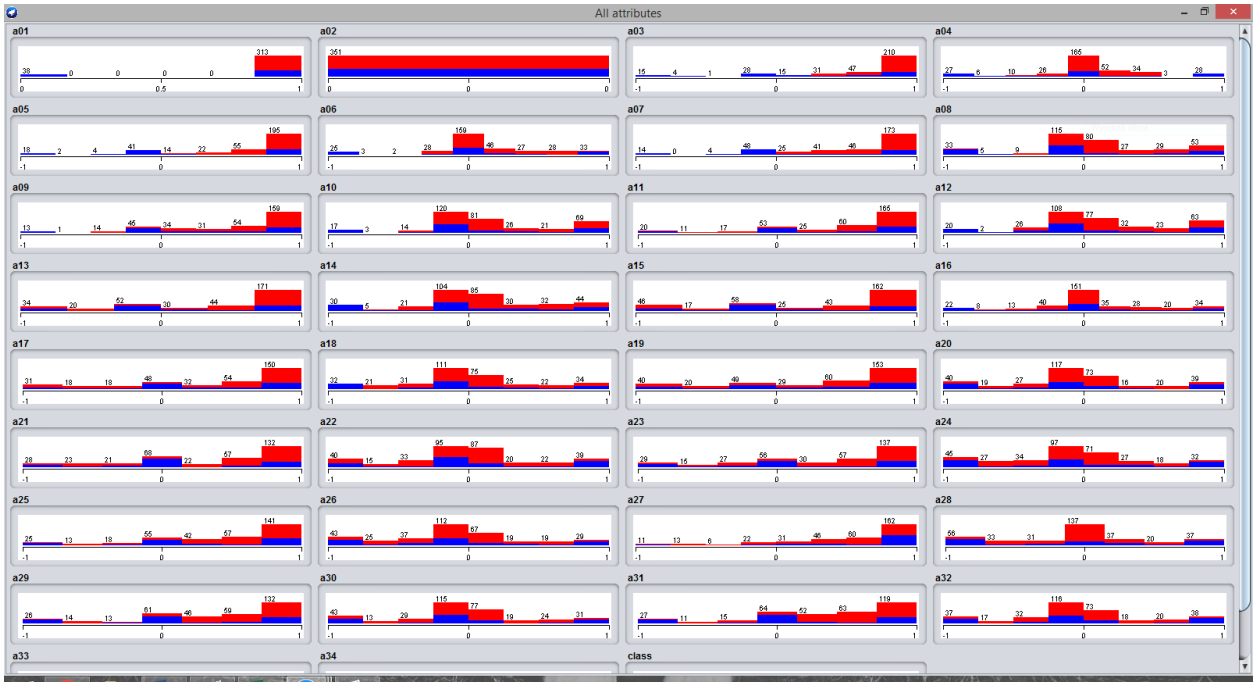


Figure 1. Visual form of ionosphere radar signals properties using almost all attributes.

2. Decision trees

A decision tree classifier for the ionosphere problem has been induced in WEKA software using J48 classifier (10 cross-validation fold, Confidence Factor for pruning of 0.25). The decision tree has 18 leaves and the size of 35. It is shown on Figure 2.1.

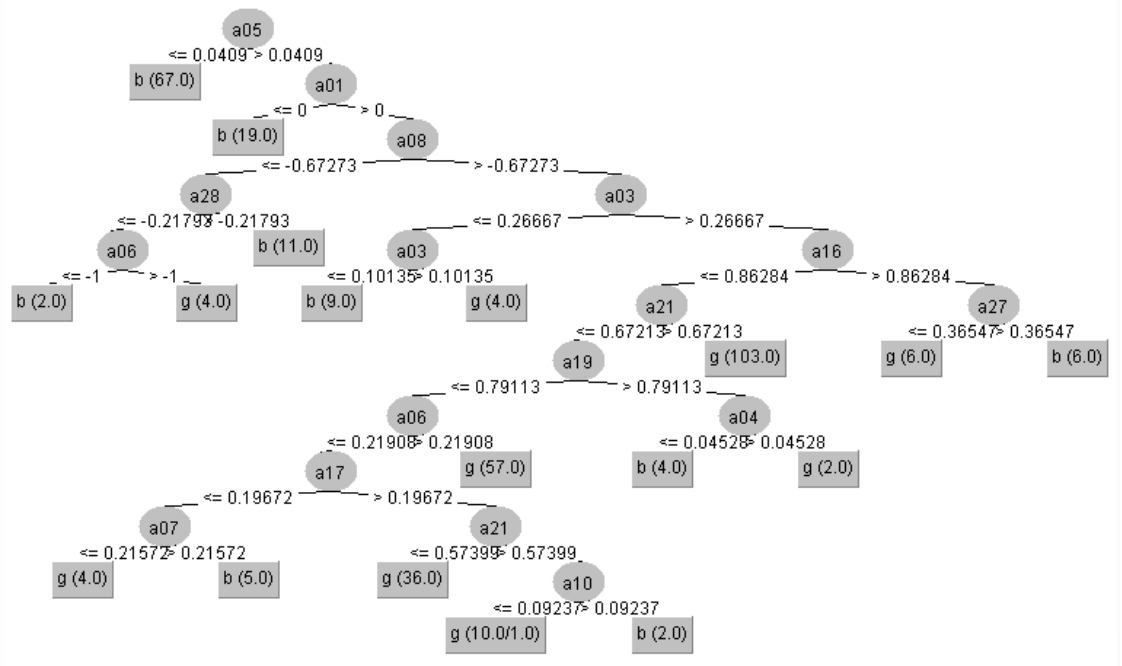


Figure 2.1. Pruned decision tree induced for ionosphere problem.

The most important statistics of the classification are shown in Tables 2.1 and 2.2 below.

Correctly Classified Instances	321	91.453 %
Incorrectly Classified Instances	30	8.547 %

Table 2.1. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,825	0,036	Bad
0,964	0,175	Good

Table 2.2. TP and FP rates per class.

J48 classifier algorithm was modified in order to induce an **unpruned** decision tree. The tree has exactly the same structure as the pruned one: it has the same number of leaves and size. Unpruned tree algorithm also results in exactly the same values of indicators of classification: Correctly Classified Instances, TP Rates, FP Rates etc. There are only some minor differences in classification errors results.

Classification results analysis has been performed in function of a Confidence Factor value. Values of TP Rate of "Good" radar returns (showing structure in the ionosphere) classification and FP Rate of "Bad" " radar returns (passing through the ionosphere) classification were researched according to Confidence Factor value. The results are shown on Figure 2.3.

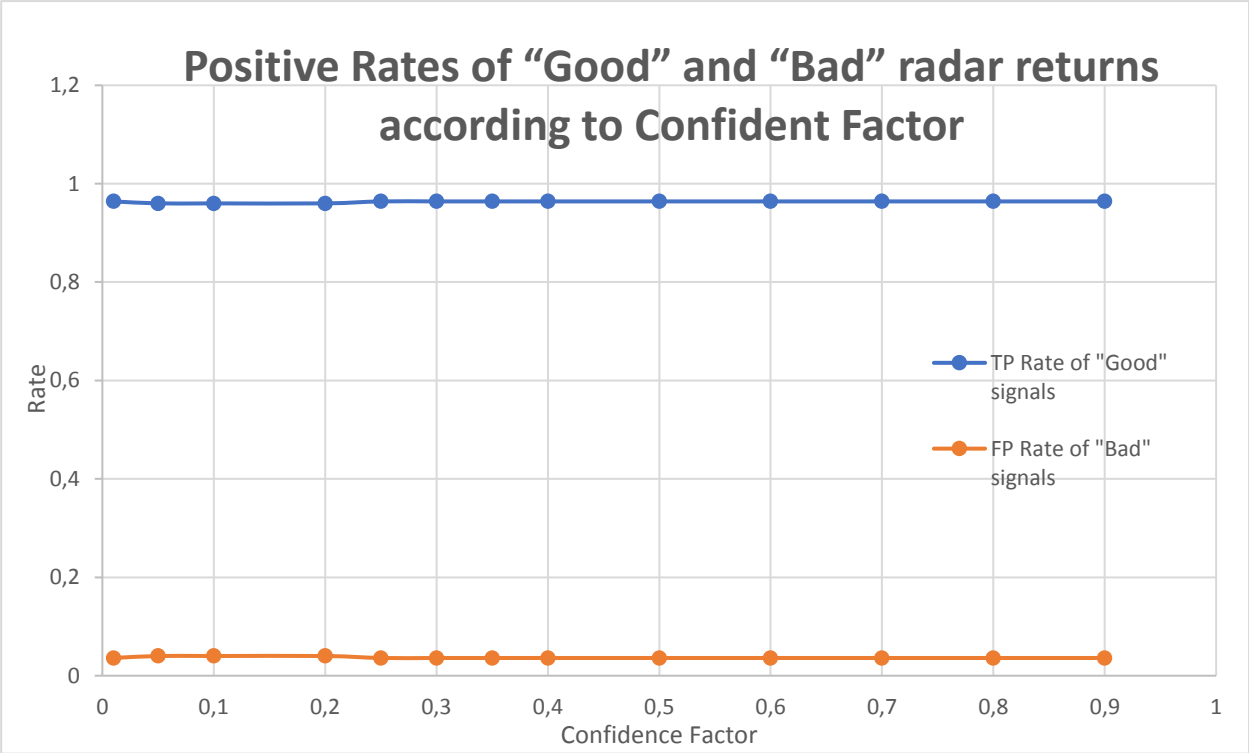


Figure 2.3. Positive Rates of "Good" and "Bad" radar returns according to pruning Confident Factor.

Chart lines are almost exactly flat and parallel.

To conclude, one may say that using unpruned tree gives the same results as the pruned one. In this case pruning has no effect on positive rates of "Good" and "Bad" signals classification. Having that in mind, it must be said that overall decision tree classification produce very good results and only a small number of radar signals was wrongly classified.

3. Induction rules

An induction rule classifier for the ionosphere signals problem has been induced in WEKA software using JRip classifier (10 cross-validation fold, batchSize = 100, folds = 3, pruning). There were 3 rules induced (shown on Figure 3.1):

```
(a27 >= 1) => class=b (88.0/13.0)
(a05 <= 0.23) => class=b (41.0/4.0)
 => class=g (222.0/14.0)
```

Figure 3.1. Rules induced using JRip

The most important statistics of the classification are shown in Tables 3.1 and 3.2 below.

Correctly Classified Instances	315	89.7436 %
Incorrectly Classified Instances	36	10.2564 %

Table 3.1. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,865	0,084	Bad
0,916	0,135	Good

Table 3.2. TP and FP rates per class.

JRip classifier algorithm was modified in order to induce an **unpruned** rules. Classification resulted in 10 rules shown on Figure 3.2.

```
(a05 <= 0.0409) => class=b (67.0/0.0)
(a27 >= 1) and (a01 <= 0) => class=b (19.0/0.0)
(a27 >= 1) and (a03 <= 0.85271) and (a05 >= 0.82809) => class=b (10.0/0.0)
(a08 <= -1) => class=b (7.0/0.0)
(a07 <= 0.2825) and (a20 >= 0.05455) => class=b (7.0/0.0)
(a06 <= -0.23067) and (a21 <= 0.64883) => class=b (6.0/0.0)
(a04 <= -0.05) and (a16 >= 0.38869) => class=b (4.0/0.0)
(a03 <= -0.65625) => class=b (2.0/0.0)
(a10 <= -0.11765) and (a10 >= -0.1209) => class=b (2.0/0.0)
 => class=g (227.0/2.0)
```

Figure 3.2. Rules induced using JRip

Rules with pruning disabled result in higher Correctly Classified Instances Rate and True Positive Rate of Good signals classification is slightly higher (0,947 vs 0,916) and False Positive Rate of

Bad signals classification is slightly lower (0,053 vs 0,084). Classification statistics are show in Tables 3.3 and 3.4.

Correctly Classified Instances	318	90.5983 %
Incorrectly Classified Instances	33	9.4017 %

Table 3.3. Correctly and incorrectly classified instances in %.

True Positive Rate	False Positive Rate	Class
0,833	0,053	Bad
0,947	0,167	Good

Table 3.4. TP and FP rates per class.

Classification results analysis has been performed in function of a Folds value. Values of TP Rate of good signals measures and FP Rate of bad signals measures were research according to pruning Folds value. The results are shown on Figure 3.3.

Another observation was made in terms of number of rules according to different Folds values. The biggest number of rules with pruning enabled is 8 and it corresponds to 5 pruning Folds. The chart is shown on Figure 3.4.

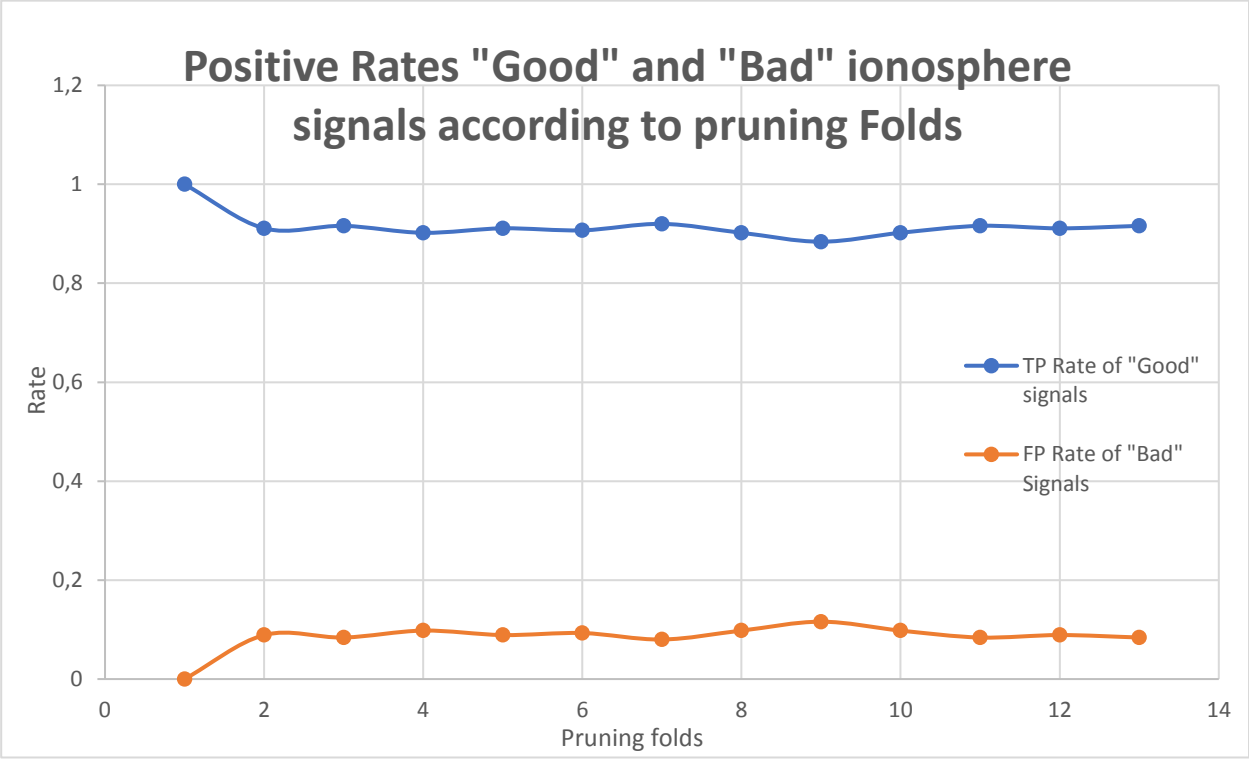


Figure 3.3. Positive Rates "Good" and "Bad" ionosphere signals according to pruning Folds.

Chart lines are not very regular, but they seem to be more or less parallel. It means that positive rates of radar signals classification may be independent with pruning Folds.

Another observation was made in terms of number of rules according to different Folds values. The biggest number of rules with pruning enabled is 8 and it corresponds to 5 pruning Folds. The chart is shown on Figure 3.4.

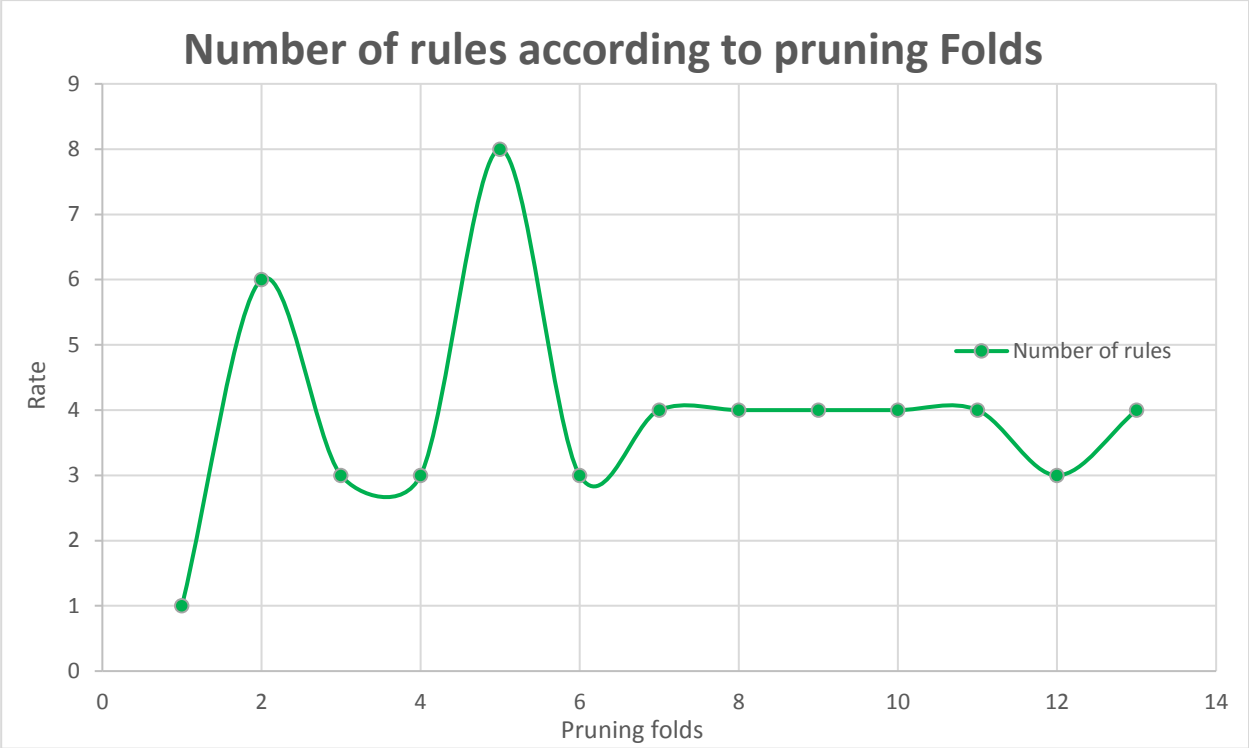


Figure 3.4. Number of rules generated according to pruning Folds.

To conclude, one may say that pruning has no effect on rates of classification with regards to ionosphere data set. Nevertheless, different pruning settings resulted in different numbers of rules induced, so pruning definitely has an overall effect decision algorithms such as J48 or JRip in given data set – it effects in "size" of generated output.

TREES VS. RULES

General task of classification is assigning a decision class label to a set of unclassified objects described by a fixed set of attributes. A classifier enables us to discover the classification knowledge representation with given set of pre-classified examples. There are many different approaches to learn classifiers and the task aimed at discussing **decision trees** and **rules**. Thera are many evaluation criteria and amongst them:

- *Predictive accuracy* - ability of the model to correctly predict the class label of new or previously unseen data. Accuracy is % of testing set examples correctly classified by the classifier;
- *Speed* - computation costs involved in generating and using the model;

- *Robustness* - ability of the model to make correct predictions given noisy data or data with missing values;
- *Scalability* - ability to construct the model efficiently given large amount of data ;
- *Interpretability* - level of understanding and insight that is provided by the model ;
- *Simplicity*:
  - decision tree size,
  - rule compactness;
- Domain-dependent quality indicators.

For the **breast cancer** dataset decision tree predictive accuracy turned out to be higher than in decision rules case. The best classification accuracy of value of 75% was achieved for pruned decision tree. However, true positive rates of cancer recurrence events and false positive rates of cancer no-recurrence events of pruned decision tree and unpruned rules classifiers were exactly the same. Generated pruned tree has 4 leaves and a total size of 6 while there was 5 rules induced.

For the **ionosphere** dataset decision tree predictive accuracy also turned out to be higher than in decision rules case. The best classification accuracy of value of 91% was achieved for decision tree and was independent with regards to pruning abled or disabled. However, the one should notice that generated decision tree has 18 leaves and o total size of 35, while there were only 3 rules induced with pruning abled, what implicated 2% loss in predictive accuracy.

It might be pointed out that, for these particular datasets and problems decision trees turned out to be as good as or better than rules in terms of predictive accuracy. On the other hand, in terms of simplicity decision trees get unreadable because of its size at some point, while rules seem to be more natural and easy form of knowledge representation. In general, rules are more comprehensive than any other knowledge representation. There were no difference in computations speed between decision trees and decision rules classifiers noticed, however it may be assumed that rules are computed with a better performance than trees, because trees are in general a graphical interpretation of similar *if…then* clause rules.