Marcin Rabiza

Decision Rules Classification Case Study

Table of contents:

Introduction

1. Data preprocessing

2. Chi-squared test of independence

3. Decision rules with JRip classifier

Conclusions

Bibliography

Introduction

The main goal of this task was to discover which attributes affect the prediction of wealth of USA citizens. The interest was particularly in classification whether the income is higher or lower than $50 000. In general, income is defined as the money received on a regular basis before payment of taxes, social security, medical care etc. This information is useful to provide economic information of the welfare of the population.

For this task a data set with the description of wealth of USA citizens was given. The data set consists of 25000 learning examples described with 14 conditional attributes (*age, workclass, demogweight, education, education-num, maritalstatus, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, nativecountry*) and 1 decision attribute (*income*). Among conditional attributes there is 6 numerical and 8 categorical (nominal) ones.

In order to discover which attributes affect the prediction of wealth and to build a simplified model of a “wealthy” USA citizen various different applications and environments were used, including: Statistica, MS Excell ,WEKA, R language and RStudio.

1. Data preprocessing

As a first step of the analysis, a basic data preprocessing was performed in order to check the quality of the data and clean the data if necessary. Using Statistica and Weka applications some missing values were found:

Attribute	Missing values	% of total number
workclass	1399	6%
occupation	1404	6%
Native-country	445	2%

A common strategy for handling missing values is to ignore or delete incomplete instances. When there are few missing values (very roughly, less than 5% of the total number of cases) and those values can be considered to be missing at random - that is, whether a value is missing does not depend upon other values – then the typical method of listwise deletion is relatively "safe" [1].

Regarding the fact, that on this stage of the analysis the one does not now anything specific about randomness of missing values, possible correlations of attributes and that the percentage of incompleteness varies considerably between attributes (from 2 to 6%), missing values were decided to be filled manually.

For **workclass** and **native country** attributes missing values were replaced with the most common values – classes *Private* and *United States*. For **occupation** attribute missing values were replaced with entirely new class – a global constant *Unknow*.

Data preprocessing revealed also that attributes *education* and *education-num* indicate the same variable in differet way: *education-num* is numerical representation of categorical classes of *education*.

2. Chi-squared test of independence

Attributes can be correlated – linearly related to each other. Particular associations may lead to data redundancy, which may be detected by correlation analysis. Handling redundant data is beneficial due to fact that larger number of redundant data may slow-down or confuse knowledge discovery process. In this case, correlation analysis may give an answer whether all conditional attributes are necessary to describe the decision attribute.

However, correlation is about the linear relationship between two variables, which, usually, are both continuous. Because many of attributes in the given dataset are categorical, we are not able to perform general correlation analysis using e.g. Statistica. Due to that, the chi-square dependency test was performed instead, as it resolves the matter of the independence of two variables. Performing chi-square dependency test also enables the analyst to determine a measure of association between two nominal variables (if they are dependent), giving a value between 0 and +1 (inclusive) using Cramér's V formula.

Pearson's chi-squared test (χ^2) is a statistical non-parametric non-parametric test that is used to evaluate association between two qualitative variables. It is applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance, with regards to the chi-squared distribution. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. Because in the given case chi-squared test assesses whether observations consisting of measures on two variables, expressed in a contingency table, are independent of each other, it is called a **test of independency**.

Below a null (H_0) and an alternative (H_1) hypothesises are formulated.

H_0: *ConditionalAttribute* and *income* are independent variables.

H_1: *ConditionalAttribute* and income are dependent

ConditionalAttribute states for any of conditional attributes of the given dataset (*age, workclass, demogweight, education, education-num, maritalstatus, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, nativecountry*).

The one reject null hypothesis if a calculated χ^2 value is bigger than the upper-tail Critical Value of chi-square distribution for certain number of Degrees of Freedom (df) and accepted Significance Level α. In other case, H_0 is accepted. An example of chi-squared distribution was given on fig. 1 below.

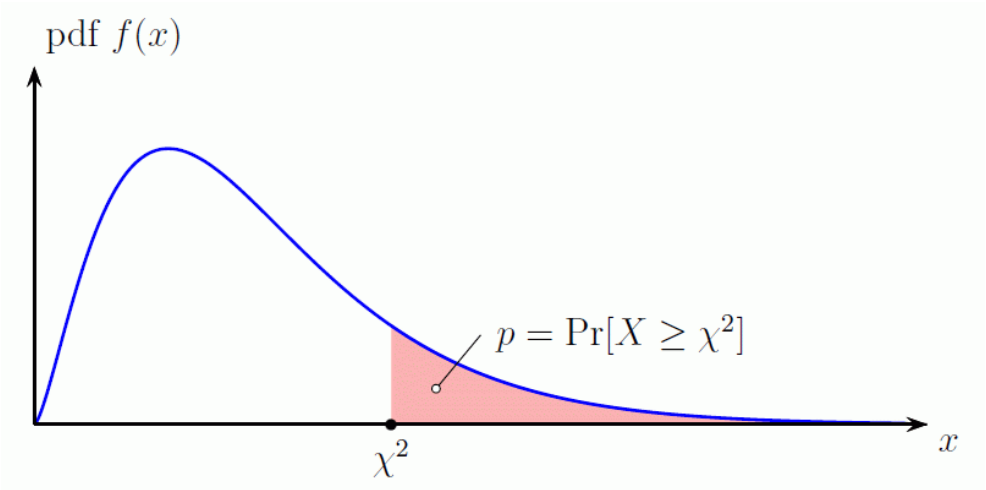

Fig. 1. Exemplary χ^2 distribution [2].

The value of the χ2 independence test-statistic is calculated with formula:

$$\chi^2 = \sum_{i,j=1}^{r,c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{1}$$

where

χ^2 – Pearson's cumulative test statistic,

O_i – the number of observations of type i,

E_i – the expected (theoretical) count of type i,

r – total number of contingency table rows,

c – total number of contingency table columns.

Cramér's V is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1, according to formula:

$$V = \sqrt{\frac{\chi^2}{n * \min(c-1, r-1)}} \tag{2}$$

where

V – Cramér's V coefficient,

χ^2 – derived Pearson's cumulative test statistic,

n – the grand total of observations,

r – total number of contingency table rows,

c – total number of contingency table columns.

For the needs of this task, chi-square independency tests for pairs of different conditional attributes and one decision attribute and Cramér's V calculations were performed using R programming language and software environment for statistical computing. Simplified R code is given below.

Simplified R code:

```
data <- foreign::read.arff("wealthOfUSACitizens.arff")          # Load data
chisq.test(data$'ConditionalAttribute', data$income)            # Calculate χ2 test statistics
DescTools::CramerV(data$'ConditionalAttribute', data$income)    # Calculate Cramér's V coefficient
```

Computed statistics are given in table 1 below. The Significance Level for this task was arbitrary set as $\alpha = 0{,}05$. Critical Value for numbers of Degrees of Freedom of different attributes from the given dataset was read from chi-square distribution tables.

It should be noted that *fnlwgt* (*demogweight*) attribute should be omitted in independency test conclusions due to its definition and key-id character of its values.

In remaining cases the test showed that decision attribute *income* is dependent on all the conditional attributes. Association was strongest in case of citizens *maritial status* (V = 0,45), *relationship* (V = 0,45) and *capital gain* (V = 0,42), and the weakest in case of *workclass* (V=0,17), *race* (V=0,10) and *native country* (V=10).

Table 1. Independence tests statistics computed for all conditional and decision attributes in the dataset.

Conditional / Decision	χ^2	df	p-Value	Critical V.	Reject H_0?	Cramér's V
age / income	2680,5	71	< 2,2e-16	91,67	Yes	0,33
workclass / income	715,82	7	< 2,2e-16	14,07	Yes	0,17
fnlwgt / income	18372	17823	0,002	-	-	0,86
education / income	3389,7	15	< 2,2e-16	25,00	Yes	0,37
edu,-num / income	3389,7	15	< 2,2e-16	25,00	Yes	0,37
maritial status / income	5011,1	6	< 2,2e-16	12,59	Yes	**0,45**
occupation / income	3063,6	14	< 2,2e-16	23,69	Yes	0,35
relationship / income	5149,3	5	< 2,2e-16	11,07	Yes	**0,45**
race / income	253,66	4	< 2,2e-16	9,49	Yes	0,10
sex / income	1173,4	1	< 2,2e-16	3,81	Yes	0,22

capital gain / income	4367	116	< 2,2e-16	142,14	Yes	**0,42**
capital loss / income	1908,2	88	< 2,2e-16	110,90	Yes	0,28
hours p,week / income	2028,7	93	< 2,2e-16	116,51	Yes	0,28
native country / income	248,08	40	< 2,2e-16	55,76	Yes	0,10

3. Decision rules with JRip classifier

Given a population whose members each belong to one of a number of different classes, a **classifier** is a procedure by which the elements of the population set are each predicted to belong to one of the decision attribute classes. A perfect classification is one for which every element in the population is assigned to the class it really belongs to. Usually, a classification is imperfect, what means some records are assigned wrongly.

There are many different approaches to classification including Decision Trees, Bayesian Classifiers, Support Vector Machines, Artificial Neural Networks and so on. For the given dataset it was decided to use Classification Rules.

Rules are the most popular symbolic representation of knowledge derived from data. In standard form a rule corresponding to class K_j is represented as:

$$\text{If } P \text{ Then } Q \tag{3}$$

where

$P = [w_1$ and w_2 and ... $w_n]$ – condition part (a conjunction of conditions)

Q – decision part (object x satisfying P is assigned to class K_j).

Simple syntactic form and an intuitive semantics determines that rules are possible to inspect and interpret by human, and as such they are the most natural, understandable and easy form of knowledge representation.

Amongst the decision rule-style algorithms there is direct rule-based classifier called *RIPPER* (acronym for *Repeated Incremental Pruning to Produce Error Reduction)*. It is one of typical algorithms based on the scheme of a sequential covering and heuristically generate a minimal set of rule covering examples. In general, covering algorithm is a strategy for generating a rule set directly aimed at finding a rule set that covers all examples in each class. Aa each stage a rule is identified that covers some of the examples (then these examples are skipped from consideration for the next rules). Sequential approach means that for a given class algorithm conducts in a stepwise way a general to specific search for the best rules (general to specific – greedy, learn-one-rule) guided by the evaluation measures. The main procedure is iteratively repeated for each class during process [3]. Specifically, RIPPER classifier may be illustrated with the following algorithm [4]:

1 Initialize a set E to be the training set
2 Choose a class C that contains least instances
3 Initialize a rule R to have an empty left-hand side that predicts C
4 Split E into growing and pruning sets
5 While there are positive samples (instances of C) in the growing set, or the description length (DL) is 64 bits greater than the smallest DL found so far, or the error rate is greater than 50%
6 Until R is perfect (or no more attributes to add)
7 For each attribute a not included in R, and for each value v,
8 Consider $a = v$ to add to the left-hand side of R.
9 Choose the a and v that have the highest Foil's information gain
10 Add $a = v$ to R
11 Prune R using Reduced Error Pruning
12 Remove the instances covered by R from the growing set
13 Global optimization strategy is applied to further prune the rule.

To sum up, RIPPER is based on the basic steps:

1) divide training set into growing and pruning sets ,
2) grow a rule adding conditions greedily ,
3) prune rule,
4) go to 2°, stopping criteria: description length, error rate,
5) optimization of rules.

RIPPER learning process is also shown in steps on figure 2 below.

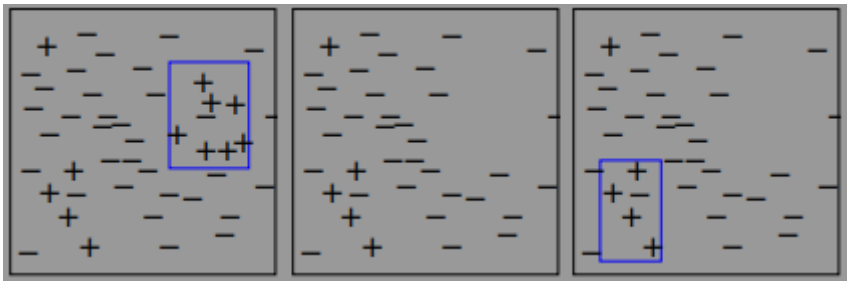

Fig. 2 RIPPER operation. Left: find rule R_n, middle: delete rule R_n instances, Right: find rule R_{n+1} [5].

Propositional rule learner RIPPER has been implemented as a Java class *JRip* in Weka environment. Documentation [6] describes the algorithm in similar manner to [4], but with phases:

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one,
DO:

1. Building stage:
Repeat 1.1 and 1.2 until the descrpition length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

1.1. Grow phase:
Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:
Incrementally prune each rule and allow the pruning of any final sequences of the antecedents; The pruning metric is (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

2. Optimization stage:
After generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. and add resultant ruleset to RS.
ENDDO

A JRip classifier for the wealth of USA citizens problem has been induced in WEKA software (Parameters: 10 cross-validation fold, batchSize = 100, folds = 3, pruning). There were 18 rules induced:

1. (marital-status = Married-civ-spouse) and (education-num >= 12) and (capital-gain >= 5178) => income=>50K (538.0/2.0)
2. (marital-status = Married-civ-spouse) and (education-num >= 14) and (capital-loss >= 1848) => income=>50K (148.0/2.0)
3. (marital-status = Married-civ-spouse) and (education-num >= 12) and (occupation = Exec-managerial) => income=>50K (799.0/200.0)
4. (marital-status = Married-civ-spouse) and (education-num >= 13) and (occupation = Prof-specialty) and (age >= 32) => income=>50K (855.0/232.0)
5. (marital-status = Married-civ-spouse) and (education-num >= 13) and (hours-per-week >= 42) and (workclass = Private) => income=>50K (293.0/100.0)
6. (marital-status = Married-civ-spouse) and (education-num >= 10) and (age >= 41) and (capital-gain >= 5178) => income=>50K (119.0/4.0)
7. (marital-status = Married-civ-spouse) and (education-num >= 10) and (age >= 36) and (capital-loss >= 1848) => income=>50K (117.0/9.0)
8. (marital-status = Married-civ-spouse) and (education-num >= 10) and (age >= 36) and (hours-per-week >= 34) and (fnlwgt >= 168211) and (occupation = Exec-managerial) => income=>50K (135.0/47.0)
9. (marital-status = Married-civ-spouse) and (education-num >= 10) and (age >= 36) and (hours-per-week >= 34) and (fnlwgt >= 119099) and (occupation = Sales) => income=>50K (254.0/97.0)
10. (marital-status = Married-civ-spouse) and (capital-gain >= 5178) => income=>50K (247.0/6.0)
11. (marital-status = Married-civ-spouse) and (education-num >= 10) and (fnlwgt >= 119099) and (age <= 59) and (age >= 44) and (hours-per-week >= 44) => income=>50K (132.0/48.0)
12. (marital-status = Married-civ-spouse) and (education-num >= 13) and (hours-per-week >= 32) and (age >= 29) and (age <= 31) and (fnlwgt <= 170983) => income=>50K (45.0/13.0)
13. (marital-status = Married-civ-spouse) and (education-num >= 10) and (hours-per-week >= 35) and (fnlwgt >= 155659) and (fnlwgt <= 263925) and (age >= 41) => income=>50K (314.0/153.0)
14. (marital-status = Married-civ-spouse) and (age >= 30) and (education-num >= 13) and (age <= 54) and (fnlwgt <= 128798) => income=>50K (133.0/64.0)
15. (marital-status = Married-civ-spouse) and (education-num >= 9) and (age >= 34) and (hours-per-week >= 35) and (capital-loss >= 1848) and (capital-loss <= 1977) => income=>50K (76.0/2.0)
16. (marital-status = Married-civ-spouse) and (education-num >= 9) and (age >= 36) and (hours-per-week >= 35) and (fnlwgt >= 136262) and (occupation = Exec-managerial) => income=>50K (133.0/62.0)
17. (marital-status = Married-civ-spouse) and (education-num >= 9) and (age >= 36) and (hours-per-week >= 38) and (occupation = Adm-clerical) and (workclass = Federal-gov) => income=>50K (61.0/24.0)
18. => income=<=50K (20601.0/2650.0)

In order to evaluate the classifier predictive accuracy, the most important classification statistics were provided based on the generated confusion matrix. They are shown in Tables 2.1 and 2.2 below:

Table 2.1. Correctly and incorrectly classified instances in numbers and %

Correctly Classified Instances	21101	84,404 %
Incorrectly Classified Instances	3899	15,596 %

Table 2.2. TP/FP rates and Precision/Recall indicators per class.

True Positive Rate	False Positive Rate	Precision	Recall	Class
0,937	0,45	0,869	0,937	<=50K
0,55	0,063	0,732	0,55	>50K

Using JRip classifier with pruning disabled ended up inducing only 11 rules, however achieving 79,38% of Correctly Classified Instances.

Conclusions

Due to fact, that Cramer's V is a measure of association between variables, analysis coefficient values for pairs of conditional and decision attributes may indicate attributes having greatest impact on the resultant decision class. The strongest association is noted for attributes: *maritial status, relationship, age, capital gain/loss* then *education*[1], *occupation* and *age*.

Rules induced using learned JRip classifier indicate attributes of high information gain and conjunction of theirs conditions needed to cover decision class outcomes. The most commonly occurring attribute was *maritial status,* which occurred in every conditional rule. Then *education, age, hours per week* and *occupation*. Surprisingly, attributes like *worklass, capital gain* and *capital loss* were rarely seen amongst induced rules. Attributes that took no part in rule induction at all are *race, sex, relationship* or *native country.*

According to results of tests performed on the given data, the one may indicate a few observations about US citizen's wealth listed below:

a) In general, citizens with a complete family are more likely to have income more than $50 000 than those with other types of relationships[2].

[1] *Education* attribute is actually a categorical equivalent of *education-num* number, so it is justified to consider *education* in general rather than numbers that symbolize its classes.

[2] Based on *maritial status* and *reletianship* attributes analysis. "Never-married", "Separated", "Widowed", "Divorced" and "Married-spouse-absent"classes are identified as **incomplete family**, and "Married-civ-spouse" and "Married-AF-spous" are defined as **complete family**.

b) Male USA citizens tend to have income greater than 50% more likely than female USA citizens no matter if they have a complete or an incomplete family[3]. As an example, male and female of complete and incomplete family citizens income % is showed on figure 3 below.

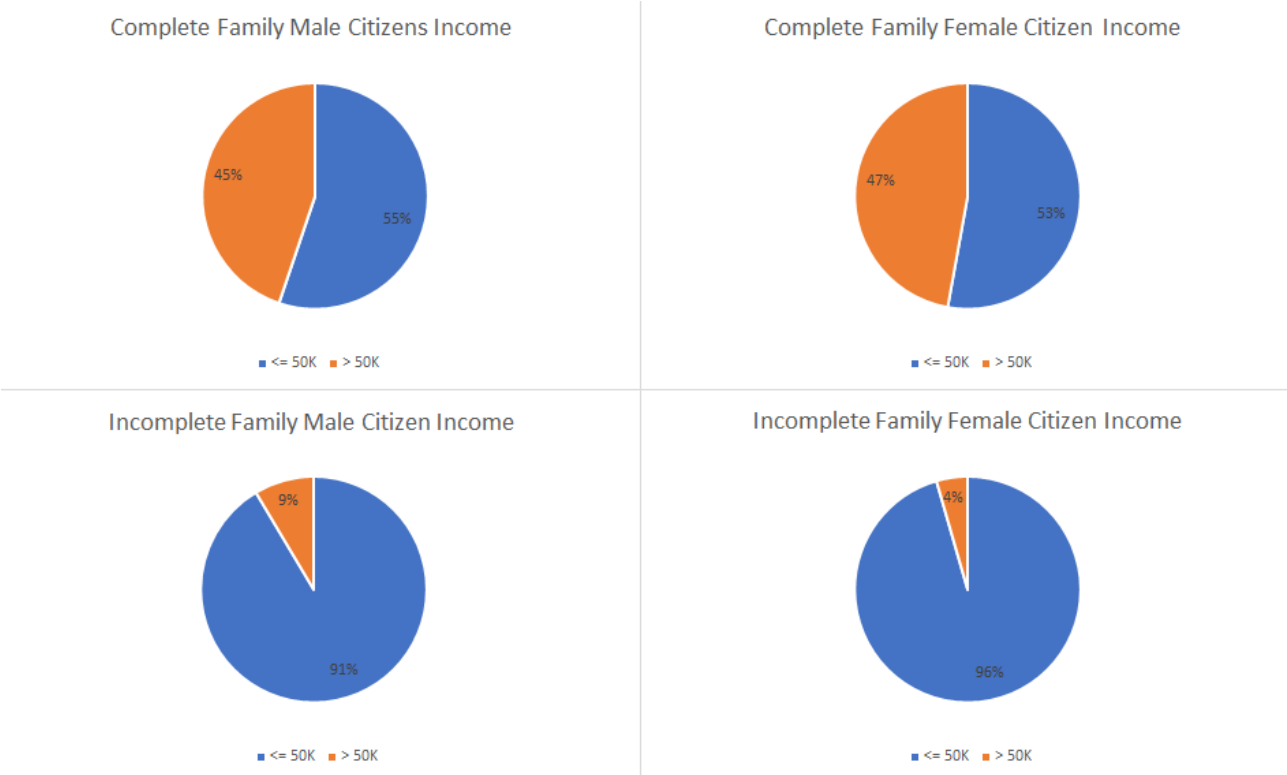

Fig. 3 Male and female of complete and incomplete family citizens income % on pie charts[4].

c) If a citizens of incomplete family lives with his or her own child, the overall probability of having an income more than $50 000 is strongly decreased.

d) The longer education a citizen received, the bigger are his/hers chances of getting an income higher than 50 000 $. The biggest impact on income is associated with receiving education in Professional School of some sort and with having a Doctorate[5].

e) Citizens of white-collar or intellectual-type occupation including: Exec-managerial, Pro-specialty, Sales, Administration, are more likely to have an income greater than $50 000. As if it comes to corresponding workclass, it should be noted that bigger income is usually associated with work at a Private Sector or in Federal Government[6].

[3] Even though *sex* attribute is treated as a less informative parameter in rule induction process.
[4] Pie charts were prepared from preprocessed data using MS Excell.

[5] This observation was made using Logistic Regression model in WEKA software, which theoretical description was not included in this case study.

[6] This observation was made based on the analysis of induced rules.

f) Citizens who have incurred not only capital gain but also capital loss tend to be more likely to earn above $50 000 than those who have not incurred capital loss[7]. Possible explanation of this surprising result is that capital losses indicate citizens who tend to do financial investments – they have losses because they are entrepreneurically active, what pays off.

g) Wealthy citizens people are usually middle-aged (between 36 and 59 years old)[8].

Observations made are useful to provide economic information of the welfare of the population, for example in order to be able to make decisions of national relevance, such as allocation of federal funding or generally to make an expert-level recommendations on policy issues.

However, it should be noted that from epistemological point of view statistical relationships do not necessarily indicate cause-and-effect relationships.

Bibliography

[1] *Introduction to Missing Values* [online], IBM Knowledge Center https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/mva/intro_missing_values_option.html

[2] Chi-square calculator [online], DI Management https://www.di-mgt.com.au/chisquare-calculator.html

[3] Stefanowski J. *Induction of Rules* [online], Data Mining and Analysis, SE Master Course, Poznan University of Technology, Poznan http://www.cs.put.poznan.pl/jstefanowski/sed/DM-6rulesnew.pdf

[4] Pan X., Hu X., Zhang Y.H., Feng K., Wang S.P., Chen L., Huang T., Cai Y.D., *Identifying Patients with Atrioventricular Septal Defect in Down Syndrome Populations by Using Self-Normalizing Neural Networks and Feature Selection* [online] https://www.researchgate.net/figure/The-procedures-of-RIPPER-algorithm_fig2_324486619

[5] Britsch M., Gagunashvili N., Schmelling M., *Application of the rule-growing algorithm RIPPER to particle physics analysis* [online], Max-Planck-Institute for Nuclear Physics, University of Akureyri, Iceland, Erice 2008 https://indico.cern.ch/event/34666/contributions/813578/attachments/683858/939357/talkBritschAcat2008.pdf.pdf

[6] *JRip* Java Class Documentation, WEKA Sourceforge http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/JRip.html

[7] *Capital loss* attribute also occurred more often than *Capital gain* amongst induced rules.
[8] Based on the analysis of induced rules.