Instituto de Ciências Matemáticas e de Computação
ISSN - 0103-2569

Rule Induction using Rough Sets Reducts as Filter for Selecting Features: An Empirical Comparison with Other Filters

Adriano Donizete Pila

Maria Carolina Monard/ILTC

Nº 139

RELATÓRIOS TÉCNICOS DO ICMC

USP – São Carlos
Abril de 2001

SYSNO 1212087
DATA / /
ICMC - SBAB

Rule Induction using Rough Sets Reducts as Filter for Selecting Features: An Empirical Comparison with Other Filters*

Adriano Donizete Pila
Maria Carolina Monard/ILTC

Department of Computer Science and Statistics
Institute of Mathematics and Computer Sciences
University of São Paulo – Campus So Carlos
Laboratory of Computational Intelligence
P.O. Box 668, 13560-970 - São Carlos, SP, Brazil
e-mail: {pila, mcmonard}@icmc.sc.usp.br

Abstract The Feature Subset Selection is an important problem within the Machine Learning area where the learning algorithm is faced with the problem of selecting relevant features while ignoring the rest. Another important problem within this area is the complexity of the knowledge acquired (hypotheses) though rules induction. Rough Sets Theory is a mathematical tool to deal with vagueness and uncertainty information. One of the main features of this approach are the *reducts*, which is a minimal feature set that preserves the ability to discern each object from the others. This work presents in detail several experiments, results and comparisons using Rough Sets Reducts and other Filters for feature subset selection and rule induction. The purpose of this work is to investigate the reduction of the complexity of the rules induced in terms of the Feature Subset Selection problem, considering as measure of rules complexity the number of rules induced. All the experiments where run on natural datasets, most of them obtained from the UCI Irvine Repository.

Keywords: Feature Selection; Rough Set; Data Mining; Machine Learning; Filter.

April 2001

*Work supported by FAPESP (98/16172-3) www.fapesp.br

Contents

List of Figures

List of Tables

1 Introduction

In supervised Machine Learning — ML — an induction algorithm is typically presented with a set of training instances, where each instance is described by a vector of feature values and a class label. The task of the induction algorithm (inducer) is to induce a classifier that will be useful in classifying new cases.

In Symbolic Machine Learning the knowledge extracted should be presented in a form that humans can understand such as rules or decision trees.

One of the main problems in ML is the Feature Subset Selection — FSS — problem, *i.e.* the learning algorithm is faced with the problem of selecting some subset of features upon which to focus its attention, while ignoring the rest (Kohavi & John, 1997).

There are several reasons for doing FSS, such as improving the accuracy of the classifiers, improving the comprehensibility of rules induced by symbolic ML algorithms as well as reducing the cost of processing huge quantity of data. Basically, there are three approaches — Embedded, Filter and Wrapper — for FSS (Blum & Langley, 1997).

In (Pila & Monard, 2001) a serie of experiments using the filter approach for FSS are presented including the Rough Sets Reducts. Rough Sets is a theory introduced by Zdzislaw Pawlak (Pawlak, 1982) in the early 1980s where the main feature is the *reduct*. A reduct is a minimal subset of features that preserves the ability to discern the examples from each other.

The experiments in (Pila & Monard, 2001) were run using nine datasets, most of them from UCI Irvine Repository (Blake et al., 1998), only considering the accuracy of the classifiers. The objetive of this work is to analise further those results considering the number of induced rules.

In order to compare previous results with the new results presented in this work, we selected the same datasets, inducers and tools used in (Pila & Monard, 2001), *i.e.* the inducers $\mathcal{C}4.5$ (Quinlan, 1993), $\mathcal{C}4.5$-rules (Quinlan, 1993), $\mathcal{CN}2$ (Clark & Boswell, 1991; Clark & Niblett, 1989), ID3 (Quinlan, 1986) implemented in $\mathcal{MLC}$++ (Kohavi et al., 1994; Felix et al., 1998) as well as the Column Importance facility (Rathjens, 1996) provided by MineSet™ and Rosetta (Øhrn, 1999) — Rough Set Toolkit for Analysis of Data — for the Rough Sets approach.

Afterwards, for each original dataset and using all features we induced rules using $\mathcal{C}4.5$-rules and $\mathcal{CN}2$ as well as using only the features selected by each filter, *i.e.* filters $\mathcal{C}4.5$, ID3, CI and RS. Finally, the number of rules induced by $\mathcal{C}4.5$-rules and $\mathcal{CN}2$ using all features and the filtered features are compared. Still, in order to facilitate reading we include in Section 2 the discription also found in (Pila & Monard, 2001) of the datasets used in the experiments. Section 3 shows the experimental setup used to run the experiments and Section 4 describes the results obtained from these experiments. Section 5 reports analysis and comparison of results. Finally, Section 6 gives some conclusions.

2 Datasets

Experiments were conducted on several real world domains. Most datasets are from the UCI Irvine Repository (Blake et al., 1998), except Smoke and TA datasets. This two datasets can

be obtained respectively from

- http://lib.stat.cmu.edu/datasets/csb/ and
- http://www.stat.wisc.edu/p/stat/ftp/pub/loh/treeprogs/datasets/.

To assist comparisons, the datasets chosen also have different type of attributes. They involve continuous attributes, either alone or in combination with nominal attributes, as well as unknown values. Section 2.2 summarizes datasets characteristics. It follows a basic datasets description.

2.1 General Description

As all datasets used in this work are described in detail in (Lee et al., 1999), a more simple description is presented here.

TA This dataset consists of evaluation of teaching performance over 3 regular semesters and 2 summer semesters of 151 teaching assistant assignments at the Statistics Department of the University of Wisconsin – Madison.

Bupa This dataset consists of predicting whether or not a male patient has liver disorders based on various blood tests and the amount of alcohol consumption.

Pima In this dataset all patients are females at least 21 years old of Pima Indian heritage living near Phoenix, Arizona, USA. The problem is to predict whether a patient would test positive for diabetes.

Breast-cancer2 This dataset is one of the breast cancer datasets at UCI, where the problem is to predict the recurrence or not of breast cancer.

CMC The examples in this dataset are married women who were either not pregnant or do not know if they were at the time of the interview. The problem is to predict the current contraceptive method choice (none, long-term methods or short-term methods) of a woman based on her demographic and socio-economic characteristics.

Breast-cancer In this dataset the problem is to predict whether a tissue sample taken from a patient's breast is malignant or benign.

Smoke This survey dataset is concerned with the problem of predicting attitude toward restrictions on smoking in the workplace (prohibited, restricted or unrestricted) based on by-law-related, smoking-related and sociodemographic covariates.

Hepatitis This dataset is for predicting life expectation of patients with hepatitis.

Hungaria This dataset is for diagnosing heart diseases.

2.2 Datasets Summary

Table 2.1 summarizes the datasets employed in this study. It shows, for each dataset, the number of instances (#Instances), number and percentage of duplicate (appearing more than once) or conflicting (same attribute-value but different class) instances, number of features (#Features) continuous and nominal, class distribution, the majority error and if the dataset have at least one missing value[1].

Datasets are presented in ascending order of the number of features, as will be in the remaining tables and graphs. Figure 2.1 shows datasets dimensionality, *i.e.* number of features and number of instances of each dataset. Observe that due to large variation, the number of instances in Figure 2.1 is represented as $\log_{10}$(#Instances).

Dataset	# Instances	#Duplicate or conflicting (%)	# Features (cont.,nom.)	Class	Class %	Majority Error	Missing Values
ta	151	45 (39.13%)	5 (1,4)	1 2 3	32.45% 33.11% 34.44%	65.56% on value 3	N
bupa	345	4 (1.16%)	6 (6,0)	1 2	42.03% 57.97%	42.03% on value 2	N
pima	769	1 (0.13%)	8 (8,0)	0 1	65.02% 34.98%	34.98% on value 0	N
breast-cancer2	285	2 (0.7%)	9 (4,5)	recurrence no-recurrence	29.47% 70.53%	29.47% on value no-recurrence	Y
cmc	1473	115 (7.81%)	9 (2,7)	1 2 3	42.70% 22.61% 34.69%	57.30% on value 1	N
breast-cancer	699	8 (1.15%)	9 (9,0)	2 4	65.52% 34.48%	34.48% on value 2	Y
smoke	2855	29 (1.02%)	13 (2,11)	0 1 2	5.29% 25.18% 69.53%	30.47% on value 2	N
hungaria	294	1 (0.34%)	13 (13,0)	presence absence	36.05% 63.95%	36.05% on value absence	Y
hepatitis	155	0 (0%)	19 (6,13)	die live	20.65% 79.35%	20.65% on value live	Y

Table 2.1: Datasets Summary Descriptions

3 Experimental Setup

A series of experiments were performed, using the datasets described in Sections 2. It is important to observe that the results about selected features were extracted from our previous work (Pila & Monard, 2001).

It is also important to note that the original data has not been pre-processed in any way trying to remove or replace missing values or transform continuous attributes in categorical attributes. Furthermore, each individual inducer was run with default setting for all parameters, *i.e.* no attempt was made to tune any inducer.

For each filter used, the performed experiments can be divided into two main steps — Figure 3.2:

[1]This information has been obtained using the $\mathcal{MLC}$++ *info* utility.

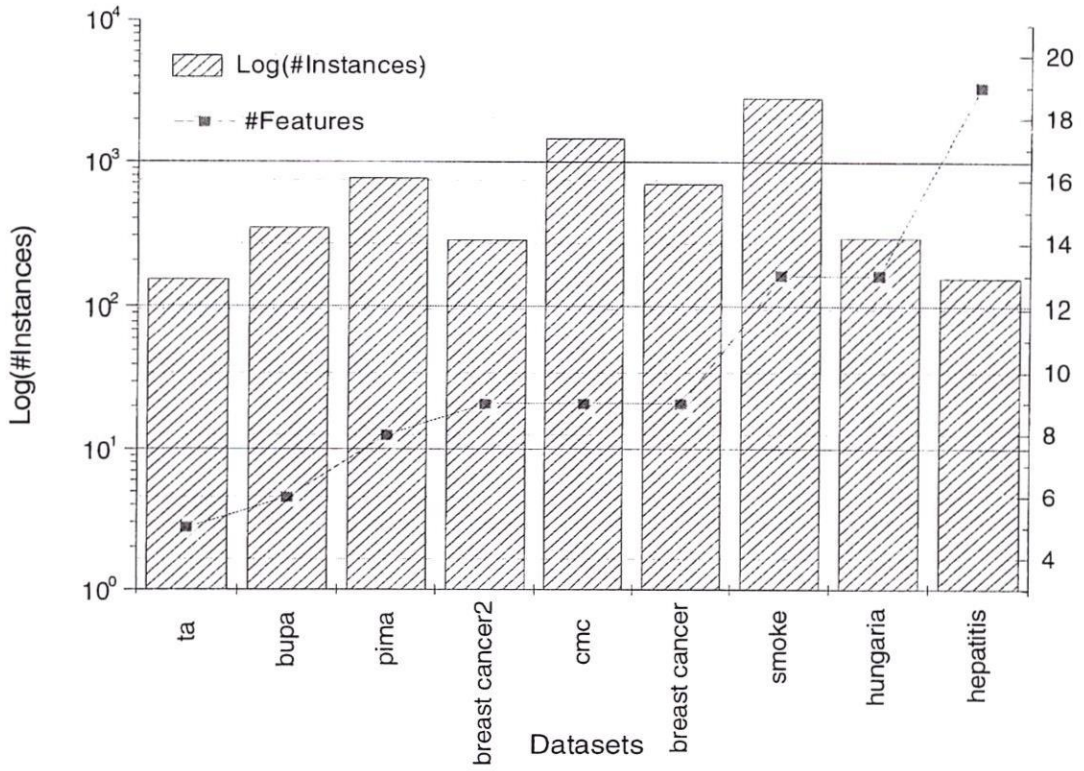

Figure 2.1: Datasets Dimensionality

1. The first step runs $\mathcal{C}4.5$, ID3, CI and Rosetta as filters for FSS
2. The second step uses features selected by the filter in step 1 to compute the number of rules induced by $\mathcal{C}4.5$-rules and $\mathcal{CN}2$ inducers

The filter process was conducted as follows: ID3, $\mathcal{C}4.5$, CI and Rosetta were applied as filters for the datasets described in Section 2.

It is important to note that when using Rosetta as a filter the result is a set of subsets where each subset is a set of selected features (reducts) and there can be several reducts[2]. Rosetta has a default setting to compute a set of reducts where all resulting reducts have the same ability to discern the examples from each other. So each reduct is a subset of selected features where the number of selected features may be different. In this work we decided to choose the reduct with the smallest number of features.

After selecting the smallest reduct, the subset of features of the reduct — similarly to the subset of features found by (Lee et al., 1999) using ID3, $\mathcal{C}4.5$ and CI — were used to compute the number of rules induced by $\mathcal{C}4.5$-rules and $\mathcal{CN}2$ inducers.

[2]More on reducts can be found in (Pila & Monard, 2001).

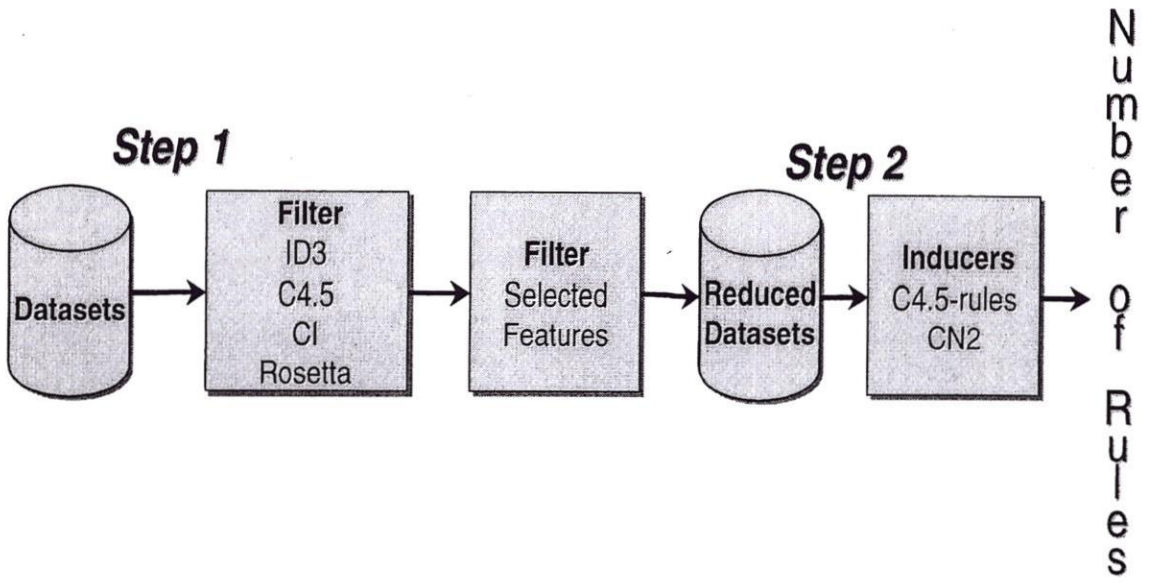

Figure 3.2: Experiments Steps

4 Experimental Results

The next sections present the results obtained through these experiments. We also include the experimental results published in (Pila & Monard, 2001).

4.1 Summary Tables Description

For each dataset four tables are presented:

1. The first table describes each feature in the dataset: feature number (features numbering starts at zero), feature name and type (continuous or nominal). For nominal features, the maximum possible number of values (as described in the *names* file) and the actual number of values (the one really found in the dataset through the $\mathcal{MLC}$++ *info* utility) are shown. It should be observed that a number of actual nominal values greater than the possible number of values indicates that there are missing values for that specific attribute. The reverse is not true.

2. The second table describes filter selected features. To specify the experiment, it is used the notation FSS(*method,inducer*) where:

   - *method* $\in$ {f} indicating that the filter (f) method for selecting features has been used[3];

[3]Although in this case there is only one method we decided to maintain the same notation used in (Pila & Monard, 2001) were *method* could be in {wf, wb, f} indicating wrapper-forward, wrapper-backward and filter respectively.

- *inducer* $\in$ {$\mathcal{C}$4.5, ID3, CI, RS} indicating the algorithm or tool that has been used as filter.

This table shows, for each FSS(*method,inducer*), the features subset selected, the number of features in the selected subset (#F) as well as the proportion of selected features (%F).

3. The third table shows similar information than the second one, but in a different way such that it is easy to visualize common features selected by every FSS(*method,inducer*) tested.

4. The fourth table shows the number of rules induced by $\mathcal{CN}2$ and $\mathcal{C}$4.5-rules, as well as the mean and standard deviation. The first column indicates the feature subset used. The second and third column indicates the number of rules induced by $\mathcal{CN}2$ and $\mathcal{C}$4.5-rules respectively, using the correspondent feature subset.

4.2 TA

Feature Number	Feature Name	#Distinct Values possible	actual	type
#0	Eng-speaker	-	2	Nominal
#1	Course-inst	-	25	Nominal
#2	Course	-	26	Nominal
#3	Sem	-	2	Nominal
#4	Class-size	-	46	Continuous

Table 4.2.1: TA – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	0 1 2 3	4	80.00%
FSS(f,C4.5)	0 1 2 3 4	5	100.00%
FSS(f,ID3)	0 1 2 3 4	5	100.00%
FSS(f,RS)	1 2 4	3	60.00%

Table 4.2.2: TA – Number of Selected Features

Feature Number	FSS (f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0	◇	◇	◇	
#1	◇	◇	◇	◇
#2	◇	◇	◇	◇
#3	◇	◇	◇	
#4		◇	◇	◇
Total 5 100%	4 80.00%	5 100.00%	5 100.00%	3 60.00%

Table 4.2.3: TA – Filter Selected Features

ta rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	61	17
FSS(f,CI)	65	14
FSS(f,$\mathcal{C}$4.5)	70	17
FSS(f,ID3)	63	17
FSS(f,RS)	64	19

continued on next page

continued from previous page

	$CN2$	C4.5-rules
Total	323	84
Mean	64.60	16.80
std-dev	3.36	1.79

Table 4.2.4: TA – Number of Rules

4.3 Bupa

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	mcv	-	26	continuous
#1	alkphos	-	78	continuous
#2	sgpt	-	67	continuous
#3	sgot	-	47	continuous
#4	gammagt	-	94	continuous
#5	drinks	-	16	continuous

Table 4.3.1: Bupa – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	4	1	16.67%
FSS(f,C4.5)	0 1 2 3 4 5	6	100.00%
FSS(f,ID3)	0 1 2 3 4 5	6	100.00%
FSS(f,RS)	0 1 2	3	50.00%

Table 4.3.2: Bupa – Number of Selected Features

Feature Number	FSS			
	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0		◇	◇	◇
#1		◇	◇	◇
#2		◇	◇	◇
#3		◇	◇	
#4	◇	◇	◇	
#5		◇	◇	
Total 6	1	6	6	3
100%	16.67%	100.00%	100.00%	50.00%

Table 4.3.3: Bupa – Filter Selected Features

bupa rules	$CN2$	C4.5-rules
all features	34	11
FSS(f,CI)	40	2
FSS(f,C4.5)	34	11
FSS(f,ID3)	37	11
FSS(f,RS)	46	3
Total	191	38
Mean	38.20	7.60
std-dev	5.02	4.67

Table 4.3.4: Bupa – Number of Rules

4.4 Pima

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	Number	-	17	continuous
#1	Plasma	-	136	continuous
#2	Diastolic	-	47	continuous
#3	Triceps	-	51	continuous
#4	Two	-	186	continuous
#5	Body	-	248	continuous
#6	Diabetes	-	517	continuous
#7	Age	-	52	continuous

Table 4.4.1: Pima – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	0 1 4 5 6 7	6	75.00%
FSS(f,C4.5)	0 1 2 4 5 6 7	7	87.50%
FSS(f,ID3)	0 1 2 3 4 5 6 7	8	100.00%
FSS(f,RS)	1 2 6	3	37.50%

Table 4.4.2: Pima – Number of Selected Features

Feature Number	FSS			
	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0	◇	◇	◇	
#1	◇	◇	◇	◇
#2		◇	◇	◇
#3			◇	
#4	◇	◇	◇	
#5	◇	◇	◇	
#6	◇	◇	◇	◇
#7	◇	◇	◇	
Total 8	6	7	8	3
100%	75.00%	87.50%	100.00%	37.50%

Table 4.4.3: Pima – Filter Selected Features

pima rules	$\mathcal{CN}2$	C4.5-rules
all features	56	6
FSS(f,CI)	58	7
FSS(f,C4.5)	53	8
FSS(f,ID3)	56	6
FSS(f,RS)	88	4
Total	311	31
Mean	62.20	6.20
std-dev	14.53	1.48

Table 4.4.4: Pima – Number of Rules

4.5 Breast Cancer2

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	Age	-	44	continuous
#1	Age-at-meno	-	3	nominal
#2	Tumor-size	-	23	continuous
#3	Involved-nodes	-	18	continuous
#4	Node-capsule	3	3	nominal
#5	Degree-of-malig	-	3	continuous

continued on next page

continued from previous page

Feature Number	Feature Name	#Distinct Values possible	actual	type
#6	Breast	-	2	nominal
#7	Breast-Quadrant	6	6	nominal
#8	Irradiation	-	2	nominal

Table 4.5.1: Breast Cancer2 – Feature Description

Inducer	Selected Features	# F	%F
FSS(f,CI)	1 2 3 4 5 6 7 8	8	88.89%
FSS(f,C4.5)	0 1 3 4 5 6 7 8	8	88.89%
FSS(f,ID3)	0 1 2 3 4 5 6 7 8	9	100.00%
FSS(f,RS)	0 2 3 5 7	5	55.56%

Table 4.5.2: Breast Cancer2 – Number of Selected Features

Feature Number	FSS (f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0		◇	◇	◇
#1	◇	◇	◇	
#2	◇		◇	◇
#3	◇	◇	◇	◇
#4	◇	◇	◇	
#5	◇	◇	◇	◇
#6	◇	◇	◇	
#7	◇	◇	◇	◇
#8	◇	◇	◇	
Total 9	8	8	9	5
100%	88.89%	88.89%	100.00%	55.56%

Table 4.5.3: Breast Cancer2 – Filter Selected Features

breast cancer2 rules	$\mathcal{CN}2$	C4.5-rules
all features	40	12
FSS(f,CI)	47	17
FSS(f,C4.5)	48	6
FSS(f,ID3)	40	12
FSS(f,RS)	44	9
Total	219	56
Mean	43.80	11.20
std-dev	3.77	4.09

Table 4.5.4: Breast Cancer2 – Number of Rules

4.6 Cmc

Feature Number	Feature Name	#Distinct Values possible	actual	type
#0	Wage	-	34	continuous
#1	Wedu	-	4	nominal
#2	Hedu	-	4	nominal
#3	Nchi	-	15	continuous
#4	Wrel	-	2	nominal
#5	Work	-	2	nominal
#6	Hocu	-	4	nominal
#7	Stdliv	-	4	nominal
#8	Medexp	-	2	nominal

continued on next page

continued from previous page				
Feature Number	Feature Name	#Distinct Values		
		possible	actual	type

Table 4.6.1: Cmc – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	0 1 2 3 4 5 6 7 8	9	100.00%
FSS(f,C4.5)	0 1 2 3 4 5 6 7 8	9	100.00%
FSS(f,ID3)	0 1 2 3 4 5 6 7 8	9	100.00%
FSS(f,RS)	0 1 2 3 4 5 6 7 8	9	100.00%

Table 4.6.2: Cmc – Number of Selected Features

Feature Number	FSS			
	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0	◇	◇	◇	◇
#1	◇	◇	◇	◇
#2	◇	◇	◇	◇
#3	◇	◇	◇	◇
#4	◇	◇	◇	◇
#5	◇	◇	◇	◇
#6	◇	◇	◇	◇
#7	◇	◇	◇	◇
#8	◇	◇	◇	◇
Total 9	9	9	9	9
100%	100%	100%	100%	100%

Table 4.6.3: Cmc – Filter Selected Features

cmc rules	$\mathcal{CN}2$	C4.5-rules
all features	174	36
FSS(f,CI)	180	36
FSS(f,C4.5)	176	36
FSS(f,ID3)	174	37
FSS(f,RS)	173	35
Total	877	180
Mean	175.40	36.00
std-dev	2.79	0.31

Table 4.6.4: Cmc – Number of Rules

4.7 Breast Cancer

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	Clump Thickness	-	10	continuous
#1	Uniformity of Cell Size	-	10	continuous
#2	Uniformity of Cell Shape	-	10	continuous
#3	Marginal Adhesion	-	10	continuous
#4	Single Epithelial Cell Size	-	10	continuous
#5	Bare Nuclei	-	10	continuous
#6	Bland Chromatin	-	10	continuous
#7	Normal Nucleoli	-	10	continuous
#8	Mitoses	-	9	continuous

Table 4.7.1: Breast Cancer – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	0 1 2 3 4 5 6 7 8	9	100.00%
FSS(f,C4.5)	0 1 2 3 4 5 6 8	8	88.89%
FSS(f,ID3)	0 1 2 3 4 5 6 7	8	88.89%
FSS(f,RS)	0 3 5 6	4	44.44%

Table 4.7.2: Breast Cancer – Number of Selected Features

Feature Number	FSS			
	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0	◇	◇	◇	◇
#1	◇	◇	◇	
#2	◇	◇	◇	
#3	◇	◇	◇	◇
#4	◇	◇	◇	
#5	◇	◇	◇	◇
#6	◇	◇	◇	◇
#7	◇		◇	
#8	◇	◇		
Total 11	9	8	8	4
100%	100.00%	88.89%	88.89%	44.44%

Table 4.7.3: Breast Cancer – Filter Selected Features

breast cancer rules	$\mathcal{CN}2$	C4.5-rules
all features	18	8
FSS(f,CI)	19	8
FSS(f,C4.5)	14	7
FSS(f,ID3)	18	8
FSS(f,RS)	31	7
Total	100	38
Mean	20.00	7.60
std-dev	6.44	0.55

Table 4.7.4: Breast Cancer – Number of Rules

4.8 Smoke

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	Weight	-	128	continuous
#1	Time	-	2	nominal
#2	Work1	-	2	nominal
#3	Work2	-	2	nominal
#4	Residence	-	2	nominal
#5	Smoking1	-	2	nominal
#6	Smoking2	-	2	nominal
#7	Smoking3	-	2	nominal
#8	Smoking4	-	2	nominal
#9	Knowledge	-	13	nominal
#10	Sex	-	2	nominal
#11	Age	-	73	continuous
#12	Education	-	5	nominal

Table 4.8.1: Smoke – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	1 2 3 4 5 6 7 8 9 10 12	11	84.62%

continued on next page

continued from previous page

Inducer	Selected Features	#F	%F
FSS(f,C4.5)	0 1 2 3 4 5 6 7 8 9 10 11 12	13	100.00%
FSS(f,ID3)	0 1 2 3 4 5 6 7 8 9 10 11 12	13	100.00%
FSS(f,RS)	0 2 3 4 5 6 7 8 9 11 12	11	84.62%

Table 4.8.2: Smoke – Number of Selected Features

Feature Number	FSS			
	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0		◇	◇	◇
#1	◇	◇	◇	
#2	◇	◇	◇	◇
#3	◇	◇	◇	◇
#4	◇	◇	◇	◇
#5	◇	◇	◇	◇
#6	◇	◇	◇	◇
#7	◇	◇	◇	◇
#8	◇	◇	◇	◇
#9	◇	◇	◇	◇
#10	◇	◇	◇	
#11		◇	◇	◇
#12	◇	◇	◇	◇
Total 13	11	13	13	11
100%	84.62%	100.00%	100.00%	84.62%

Table 4.8.3: Smoke – Filter Selected Features

smoke rules	$\mathcal{CN}2$	C4.5-rules
all features	426	22
FSS(f,CI)	410	26
FSS(f,C4.5)	423	22
FSS(f,ID3)	426	22
FSS(f,RS)	474	37
Total	2159	129
Mean	431.80	25.80
std-dev	24.50	6.50

Table 4.8.4: Smoke – Number of Rules

4.9 Hungaria

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	age	-	38	continuous
#1	sex	-	2	continuous
#2	cp	-	4	continuous
#3	trestbps	-	31	continuous
#4	chol	-	153	continuous
#5	fbs	-	2	continuous
#6	restecg	-	3	continuous
#7	thalach	-	71	continuous
#8	exang	-	2	continuous
#9	oldpeak	-	10	continuous
#10	slope	-	3	continuous
#11	ca	-	2	continuous
#12	thal	-	3	continuous

Table 4.9.1: Hungaria – Feature Description

Inducer	Selected Features	#F	%F

continued on next page

continued from previous page

Inducer	Selected Features	#F	%F
FSS(f,CI)	1 2 4 5 6 7 8 9 11 12	10	76.92%
FSS(f,C4.5)	0 1 2 3 4 5 6 7 8 9 10	11	84.62%
FSS(f,ID3)	0 1 2 3 4 5 7 8 9 10 12	11	84.62%
FSS(f,RS-b)	4 7 9	3	23.07%

Table 4.9.2: Hungaria – Number of Selected Features

Feature Number	FSS			
	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0		◇	◇	
#1	◇	◇	◇	
#2	◇	◇	◇	
#3		◇	◇	
#4	◇	◇	◇	◇
#5	◇	◇	◇	
#6	◇	◇		
#7	◇	◇	◇	◇
#8	◇	◇	◇	
#9	◇	◇	◇	◇
#10		◇	◇	
#11	◇			
#12	◇		◇	
Total 13	10	11	11	3
100%	76.92%	84.62%	84.62%	23.07%

Table 4.9.3: Hungaria – Filter Selected Features

hungaria rules	$\mathcal{CN}2$	$\mathcal{C}4.5$-rules
all features	25	11
FSS(f,CI)	30	8
FSS(f,C4.5)	25	12
FSS(f,ID3)	25	11
FSS(f,RS)	43	2
Total	148	44
Mean	29.60	8.80
std-dev	7.80	4.09

Table 4.9.4: Hungaria – Number of Rules

4.10 Hepatitis

Feature Number	Feature Name	#Distinct Values		
		possible	actual	type
#0	age	-	49	continuous
#1	female	2	2	nominal
#2	steroid	2	3	nominal
#3	antivirals	2	2	nominal
#4	fatigue	2	3	nominal
#5	malaise	2	3	nominal
#6	anorexia	2	3	nominal
#7	liver-big	2	3	nominal
#8	liver-firm	2	3	nominal
#9	spleen-palpable	2	3	nominal
#10	spiders	2	3	nominal
#11	ascites	2	3	nominal
#12	varices	2	3	nominal
#13	bilirubin	-	34	continuous
#14	alk-phosphate	-	83	continuous
#15	sgot	-	84	continuous
#16	albumin	-	29	continuous
#17	protime	-	44	continuous

continued on next page

continued from previous page				
Feature	Feature	#Distinct Values		
Number	Name	possible	actual	type
#18	histology	2	2	nominal

Table 4.10.1: Hepatitis – Feature Description

Inducer	Selected Features	#F	%F
FSS(f,CI)	2 3 5 8 10 11 13 16 17 18	10	52.63%
FSS(f,C4.5)	0 1 3 4 5 7 8 10 11 15 16 17	12	63.16%
FSS(f,ID3)	0 3 7 10 11 13 14 16 17	9	47.37%
FSS(f,RS)	0 10 16	3	15.79%

Table 4.10.2: Hepatitis – Number of Selected Features

Feature	FSS			
Number	(f,CI)	(f,C4.5)	(f,ID3)	(f,RS)
#0		◇	◇	◇
#1		◇		
#2	◇			
#3	◇	◇	◇	
#4		◇		
#5	◇	◇		
#6				
#7		◇	◇	
#8	◇	◇		
#9				
#10	◇	◇	◇	◇
#11	◇	◇	◇	
#12				
#13	◇		◇	
#14			◇	
#15		◇		
#16	◇	◇	◇	◇
#17	◇	◇	◇	
#18	◇			
Total 19	10	11	9	3
100%	52.63%	57.89%	47.37%	15.79%

Table 4.10.3: Hepatitis – Filter Selected Features

hepatitis rules	$\mathcal{CN}2$	C4.5-rules
all features	19	10
FSS(f,CI)	25	7
FSS(f,C4.5)	20	10
FSS(f,ID3)	22	6
FSS(f,RS)	28	2
Total	114	35
Mean	22.80	7.00
std-dev	3.70	3.32

Table 4.10.4: Hepatitis – Number of Rules

5 Results Comparison

Tables 5.5 and 5.6 summarize for inducers C4.5-rules and $\mathcal{CN}2$ respectively, and for each dataset, the number of induced rules using the features selected by each filter[4] as well as the average and standard deviation.

[4]The % of selected features is indicated in brackets.

One immediate result is that for inducer $\mathcal{C}4.5$-rules the total number of rules induced using the features selected by RS (118) is smaller than the total number of rules induced using the others filters. Let *#TotalRules(Inducer,Filter)* be the total number of rules induced by *Inducer* using the subset of selected features using filter *Filter*,then:

$$\begin{array}{ll} \#\text{TotalRules}(\mathcal{C}4.5\text{-rules,FSS(f,RS)}) & \leq \\ \#\text{TotalRules}(\mathcal{C}4.5\text{-rules,FSS(f,CI)}) & \leq \\ \#\text{TotalRules}(\mathcal{C}4.5\text{-rules,FSS(f,ID3)}) & \leq \\ \#\text{TotalRules}(\mathcal{C}4.5\text{-rules,FSS(f,}\mathcal{C}4.5\text{))} & \leq \\ \#\text{TotalRules}(\mathcal{C}4.5\text{-rules,All}) & \end{array}$$

On the other hand, it is interesting to observe that the opposite result holds for $\mathcal{CN}2$, *i.e.*

$$\begin{array}{ll} \#\text{TotalRules}(\mathcal{CN}2\text{,All}) & \leq \\ \#\text{TotalRules}(\mathcal{CN}2\text{,FSS(f}\mathcal{C}4.5\text{))} & \leq \\ \#\text{TotalRules}(\mathcal{CN}2\text{,FSS(f,ID3))} & \leq \\ \#\text{TotalRules}(\mathcal{CN}2\text{,FSS(f,CI))} & \leq \\ \#\text{TotalRules}(\mathcal{CN}2\text{,FSS(f,RS))} & \end{array}$$

This later results confirms that $\mathcal{CN}2$ works better if we let the algorithm do its own feature selection. In fact, it seems that the number of rules induced by $\mathcal{CN}2$ increases when the number of selected features decreases. For example, FSS(f,RS) selected, on the average, the smallest number of features, and $\mathcal{CN}2$ induced the greatest number of rules (991) considering all datasets. On the other hand, $\mathcal{C}4.5$-rules induced the smallest number of rules (118) in this case. Also, from Tables 5.5 and 5.6 it can be seen that $\mathcal{CN}2$ has a tendency to induce a much greater number of rules than $\mathcal{C}4.5$-rules does. In fact, for all datasets and filters results show that the number of rules induced by $\mathcal{CN}2$ is greater than the number of rules induced by $\mathcal{C}4.5$-rules, *i.e.*

$$\#\text{TotalRules}(\mathcal{CN}2\text{,All or FSS features}) > \#\text{TotalRules}(\mathcal{C}4.5\text{-rules,All or FSS features})$$

Dataset	Rules									Using Filter		
	All	(f,CI)		(f,$\mathcal{C}4.5$)		(f,ID3)		(f,RS)		Total	Average	Std-dev
ta	17	14	(80.00%)	17	(100.00%)	17	(100.00%)	19	(60.00%)	67	16.75	2.06
bupa	11	2	(16.67%)	11	(100.00%)	11	(100.00%)	3	(50.00%)	27	6.75	4.92
pima	6	7	(75.00%)	8	(87.50%)	6	(100.00%)	4	(37.50%)	25	6.25	1.71
breast cancer2	12	17	(88.89%)	6	(88.89%)	12	(100.00%)	9	(55.56%)	44	11.00	4.69
cmc	36	36	(100.00%)	36	(100.00%)	36	(100.00%)	36	(100.00%)	144	36.00	0.00
breast cancer	8	8	(100.00%)	7	(88.89%)	8	(88.89%)	7	(44.44%)	30	7.50	0.58
smoke	22	26	(84.62%)	22	(100.00%)	22	(100.00%)	37	(84.62%)	107	26.75	7.09
hungaria	11	8	(76.92%)	12	(84.62%)	11	(84.62%)	2	(23.07%)	33	8.25	4.50
hepatitis	10	7	(52.63%)	10	(63.16%)	6	(47.37%)	2	(15.79%)	25	6.25	3.30
Total	133	125		129		129		119				
Average	14.78	13.89		14.33		14.44		13.11				
Std-dev	9.28	10.90		9.58		9.91		14.01				

Table 5.5: Number of Rules Induced by $\mathcal{C}4.5$-rules

Dataset	Rules								Using Filter		
	All	(f,CI)		(f,C4.5)		(f,ID3)		(f,RS)	Total	Average	Std-dev
ta	61	65	(80.00%)	63	(100.00%)	63	(100.00%)	64 (60.00%)	255	63.75	0.96
bupa	34	40	(16.67%)	34	(100.00%)	37	(100.00%)	46 (50.00%)	157	39.25	5.12
pima	56	58	(75.00%)	53	(87.50%)	56	(100.00%)	88 (37.50%)	255	63.75	16.30
breast cancer2	40	47	(88.89%)	48	(88.89%)	40	(100.00%)	44 (55.56%)	179	44.75	3.59
cmc	174	180	(100.00%)	176	(100.00%)	174	(100.00%)	173 (100.00%)	703	175.75	3.10
breast cancer	18	19	(100.00%)	14	(88.89%)	18	(88.89%)	31 (44.44%)	82	20.50	7.33
smoke	426	410	(84.62%)	423	(100.00%)	426	(100.00%)	474 (84.62%)	1743	435.75	25.62
hungaria	25	30	(76.92%)	25	(84.62%)	25	(84.62%)	43 (23.07%)	123	30.75	8.50
hepatitis	19	25	(52.63%)	20	(63.16%)	22	(47.37%)	28 (15.39%)	95	23.75	3.50
Total	853	874		884		861		991			
Average	94.78	95.11		98.22		95.67		110.11			
Std-dev	133.15	132.26		130.03		132.71		143.60			

Table 5.6: Number of Rules Induced by $\mathcal{CN}2$

However it is important to consider not only the number of rules induced using FSS but also the performance of the induction algorithms on new cases.

In order to compare if the difference between two algorithms — say A_1 and A_2 — is significant or not, we applied the following significance test, where $\mathrm{m}(A_2 - A_1)$ is the mean and $\mathrm{sd}(A_2 - A_1)$ is the standard deviation calculated, respectively, using Equations 1 and 2.

$$\mathrm{m}(A_2 - A_1) = \mathrm{m}(A_2) - \mathrm{m}(A_1) \tag{1}$$

$$\mathrm{sd}(A_2 - A_1) = \sqrt{\frac{\mathrm{sd}(A_2)^2 + \mathrm{sd}(A_1)^2}{2}} \tag{2}$$

Afterwards, the difference in standard deviation, given by Equation 3, is calculated. If that difference is positive then A_2 (or A_1 depending on the result being considered) outperforms A_1, the other way around if the difference is negative. However, one result outperforms the other at the 95% level of confidence only if that difference is grater (less) than 2.

$$\mathrm{ad}(A_2 - A_1) = \frac{\mathrm{m}(A_2 - A_1)}{\mathrm{sd}(A_2 - A_1)} \tag{3}$$

Table 5.7 shows improved accuracies of $\mathcal{CN}2$ and $\mathcal{C}4.5$-rules at the significance level (95% confidence) for filter selection compared with the inducers using all features on the datasets. Improvements bellow 2 standard deviations are reported with $\triangle$, *i.e.* the filter approach outperforms the standard inducer at the 95% confidence level. Improvements bellow zero (but not bellow 2 standard deviation) are reported with $+$. The opposite case where the standard inducer outperforms the filter approach at the 95% confidence level are reported with $\triangledown$, the others with $-$. Cases where Equation 3 is zero are not filled.

	FSS											
Dataset	(f,CI) -$\mathcal{C}$4.5-rules	(f,$\mathcal{C}$4.5) -$\mathcal{C}$4.5-rules	(f,ID3) -$\mathcal{C}$4.5-rules	(f,RS) -$\mathcal{C}$4.5-rules	(f,CI) -$\mathcal{CN}$2	(f,$\mathcal{C}$4.5) -$\mathcal{CN}$2	(f,ID3) -$\mathcal{CN}$2	(f,RS) -$\mathcal{CN}$2	# $\triangle$	# ∇	# +	# −
ta	+			+	−			+	0	0	3	1
bupa	∇			∇	∇			−	0	3	0	1
pima	−			−	−	+		∇	0	1	1	3
breast cancer2	−	−		+	+	+		−	0	0	3	3
cmc								+	0	0	1	0
breast cancer				−		+	+	∇	0	1	2	1
smoke	+			−	∇			+	0	1	2	1
hungarian	+	+	−	−	+	−	+	−	0	0	4	4
hepatitis	+	+	−	+	−	+	+	−	0	0	5	3
# $\triangle$	0	0	0	0	0	0	0	0	0			
# ∇	1	0	0	1	2	0	0	2		6		
# +	4	2	0	3	2	4	3	3			21	
# −	2	1	2	4	3	1	0	4				17

Table 5.7: Improved Accuracies

Not considering cases where all the features were selected by the filter, and concentrating in improvements reported with + and where the number of rules induced are at most 20% more than the number of rules induced by the standard inducer, we can see that there is a gain on the following datasets:

- Using $\mathcal{C}$4.5-rules as standard inducer:
  - ta using (f,CI) and (f,RS);
  - breast cancer2 using (f,RS);
  - smoke using (f,CI);
  - hungarian using (f,CI) and (f,$\mathcal{C}$4.5);
  - hepatitis using (f,CI), (f,$\mathcal{C}$4.5) and (f,RS);
- Using $\mathcal{CN}$2 as standard inducer:
  - ta using (f,RS);
  - pima using (f,$\mathcal{C}$4.5);
  - breast cancer2 using (f,CI) and (f,$\mathcal{C}$4.5);
  - breast cancer using (f,$\mathcal{C}$4.5) and (f,ID3);
  - smoke using (f,RS);
  - hungarian using (f,CI) and (f,ID3);
  - hepatitis using (f,$\mathcal{C}$4.5) and (f,ID3);

6 Conclusions

This work describes empirical results using filter approaches for Feature Subset Selection and two inducers to induce the rules. The aim is to compare the number of rules induced by each inducer over each datasets using only those features selected by each filter method. As standard inducers for the filter approach, is was used $\mathcal{C}4.5$, ID3, the CI MineSet™ facility and Rosetta for the Rough Sets approach. All these inducers were run using its default options setting, on nine real world datasets. Previous results related with accuracy were extracted from (Pila & Monard, 2001).

In this work we investigated how the reduction on the number of features — FSS — affects the number of rules induced by $\mathcal{CN}2$ and $\mathcal{C}4.5$-rules. An overall result is that $\mathcal{C}4.5$-rules induces a smaller number of rules when using a small number of features. Opposite result was observed when using $\mathcal{CN}2$, confirming that $\mathcal{CN}2$ works better if it is allowed to do its own feature selection.

References

Blake, C., Keogh, E., & Merz, C. (1998). Uci irvine repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Blum, A. L. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, pages 245–271.

Clark, P. & Boswell, R. (1991). Rule induction with cn2: Some recent improvements. In Kodratoff, Y., editor, *Proceedings of the 5th European Conference EWSL 91*, pages 151–163. Springer-Verlag.

Clark, P. & Niblett, T. (1989). The cn2 induction algorithm. *Machine Learning*, 3(4):261–283.

Øhrn, A. (1999). Rosetta: Technical reference manual. Technical report, Knowledge System Group.

Felix, L. C. M., Rezende, S. O., Doi, C. Y., de Paula, M. F., & Romanato, M. J. (1998). MLC^{++} biblioteca de aprendizado de máquina em C^{++}. Technical Report 72, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_72.ps.zip.

Kohavi, R. & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, pages 273–324.

Kohavi, R., Sommerfield, D., & Dougherty, J. (1994). *MLC++: A Machine Learning Library in C++*. IEEE Computer Society Press.

Lee, H. D., Monard, M. C., & Baranauskas, J. A. (1999). Empirical comparison of wrapper and filter approaches for feature subset selection. Technical Report 94, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_94.ps.zip.

Pawlak, Z. (1982). Rough sets. *International Jornal of Computer and Information Sciences*, pages 341–356.

Pila, A. D. & Monard, M. C. (2001). Rough sets reducts as a filter approach for feature subset selection: An empirical comparison with wrapper and other filter approaches. Technical Report 134, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_134.ps.zip.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA.

Rathjens, D. (1996). Mineset™ user's guide. Silicon Graphics, Inc.

A Scripts used to Run the Experiments

The scripts used to run the experiments described in this work are listed in this Appendix.

A.1 $\mathcal{CN}2$ Rule Induction

```
script-rules <loglevel>

#!/bin/csh
#
# Author: Adriano Donizete Pila (pila@icmc.sc.usp.br)
#         LABIC-ICMC-USP --- Modified from a previous script from
#         Jose Augusto Baranauskas (jaugusto@icmc.sc.usp.br)
#
# Summary: This script runs the MLC++ CN2 inducer in
# several datasets with several filters. Rules are induced
# for each dataset and kept in files for later user evaluation.
#
# arguments:
#     a) MLC++ loglevel (optional)
#
# pre:
#     a) file "datasets" containing in each line one dataset name,
#        without extension (.names, .data and .test assumed)
#     b) file "filters" containing in each line one
#        Filter name to be used in the rule induction.
#
# pos:
#     a) files $dataset.$filter.cn2.out, for each $dataset in the
#        "dataset" file and for each $filter in the "filters" file. Each
#        output file contains the rules induced
#        for each feature set present in the
#        $dataset file
#
# NOTE: There is no value checking for datasets and filters to be used.
#       The user must check them for valid values before running this script.

# Search path for MLC++ libraries unalias rm alias libinfo 'setenv
LD_LIBRARY_PATH /lib:/usr/mlclib/mlc' alias libAccEst 'setenv
LD_LIBRARY_PATH /usr/lib:/lib:/usr/mlclib/mlc' alias libproject
libinfo

# Define default MLC++ loglevel as 1 if it was not user supplied
set loglevel = 1
if ($1 != "") then        # has been supplied by the user?
  set loglevel = $1       # yes, set it up
```

```
endif
setenv LOGLEVEL $loglevel

foreach dataset (`cat datasets`)
  foreach filter (`cat filters`)
    echo "================================================================="
    echo "Working on $dataset.$filter with Inducer CN2 ..."
    echo "================================================================="
    set outfile = $dataset.$filter.cn2.out
    set stime=`date`
    echo "Start time ..: $stime" > $outfile
    echo "Inducer .....: CN2" >> $outfile
    echo "Dataset  ....: $dataset" >> $outfile
    echo "Working dir .: `pwd`" >> $outfile
    echo "Output file .: $outfile" >> $outfile
    setenv INDUCER cn2
    setenv DATAFILE $dataset.$filter.all
    setenv NAMESFILE $dataset.$filter.names
    setenv TESTFILE $dataset.$filter.all
    set et  = `time Inducer >> & $outfile`
    echo "Start time ......: $stime " >> $outfile
    echo "Stop time .......: `date`" >> $outfile
    echo "Execution time ..: $et ">> $outfile
  end
end
```

A.2 $C4.5$-rules Rule Induction

```
script-rules2 <loglevel>

#!/bin/csh
#
# Author: Adriano Donizete Pila (pila@icmc.sc.usp.br)
#         LABIC-ICMC-USP --- Modified from a previous script from
#         Jose Augusto Baranauskas (jaugusto@icmc.sc.usp.br)
#
# Summary: This script runs the MLC++ C4.5 and C4.5-rules inducers in
# several datasets with several filters. Decision trees and rules are induced
# for each dataset and kept in files for later user evaluation.
#
# arguments:
#     a) MLC++ loglevel (optional)
#
# pre:
#     a) file "datasets" containing in each line one dataset name,
#        without extension (.names, .data and .test assumed)
```

```
#     b) file "filters" containing in each line one
#        Filter name to be used in the rule induction.
#
# pos:
#     a) files $dataset.$filter.c4.5-rules.out and $dataset.$filter.c4.5-rules.out,
#        for each $dataset in the "dataset" file and for each $filter in the "filters"
#        file. Each output file contains the rules induced
#        for each feature set present in the
#        $dataset file
#
# NOTE: There is no value checking for datasets and filters to be used.
#       The user must check them for valid values before running this script.
# IMPORTANT: As the rules induced by C4.5-rules depends on the tree induced by
#            C4.5 a file named $dataset.$filter.c4.5.out is also generated.

# Search path for MLC++ libraries unalias rm alias libinfo 'setenv
LD_LIBRARY_PATH /lib:/usr/mlclib/mlc' alias libAccEst 'setenv
LD_LIBRARY_PATH /usr/lib:/lib:/usr/mlclib/mlc' alias libproject
libinfo

# Define default MLC++ loglevel as 1 if it was not user supplied
set loglevel = 1
if ($1 != "") then        # has been supplied by the user?
  set loglevel = $1       # yes, set it up
endif
setenv LOGLEVEL $loglevel

foreach dataset (`cat datasets`)
  foreach filter (`cat filters`)
    echo "================================================================="
    echo "Working on $dataset.$filter with Inducer C4.5 ..."
    echo "================================================================="
    set outfile = $dataset.$filter.c4.5.out
    set stime=`date`
    echo "Start time ..: $stime" > $outfile
    echo "Inducer .....: C4.5" >> $outfile
    echo "Dataset  ....: $dataset" >> $outfile
    echo "Working dir .: `pwd`" >> $outfile
    echo "Output file .: $outfile" >> $outfile
    setenv DATAFILE $dataset.$filter.all
    setenv NAMESFILE $dataset.$filter.names
    setenv TESTFILE $dataset.$filter.all
    set et  = `time c4.5 -f $dataset.$filter >> & $outfile`
    echo "Start time ......: $stime " >> $outfile
    echo "Stop time .......: `date`" >> $outfile
    echo "Execution time ..: $et ">> $outfile
  end
end
```

```
foreach dataset (`cat datasets`)
  foreach filter (`cat filters`)
    echo "============================================================="
    echo "Working on $dataset.$filter with Inducer C4.5-rules ..."
    echo "============================================================="
    set outfile = $dataset.$filter.c4.5-rules.out
    set stime=`date`
    echo "Start time ..: $stime" > $outfile
    echo "Inducer .....: C4.5-rules" >> $outfile
    echo "Dataset  ....: $dataset" >> $outfile
    echo "Working dir .: `pwd`" >> $outfile
    echo "Output file .: $outfile" >> $outfile
    setenv DATAFILE $dataset.$filter.all
    setenv NAMESFILE $dataset.$filter.names
    setenv TESTFILE $dataset.$filter.all
    set et  = `time c4.5rules -f $dataset.$filter >> & $outfile`
    echo "Start time ......: $stime " >> $outfile
    echo "Stop time .......: `date`" >> $outfile
    echo "Execution time ..: $et ">> $outfile
  end
end
```