Rules Induced by Symbolic Machine Learning Algorithms Using Rough Sets Reducts for Selecting Features: An Empirical Comparison with Other Filters

Adriano Donizete Pila and Maria Carolina Monard

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

Abstract Feature Subset Selection — FSS — is an important problem within the Machine Learning — ML — area where the learning algorithm is faced with the problem of selecting relevant features while ignoring the rest. 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*. A reduct is a minimal feature set that preserves the ability to discern each object from the others. This work presents several experiments run one nine natural datasets, results and comparisons using Rough Sets Reducts and other Filters for FSS. After FSS the reduced datasets are used as input to two symbolic ML algorithms that induce *if_then* rules. The purpose of this work is to investigate not only the accuracy of the induced classifiers but also its complexity in terms of the number of rules in each classifier.

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

1 Introduction

In supervised ML an induction algorithm is typically presented with a set of training instances or cases, 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 ML the knowledge induced should be 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 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 [1].

Work supported by FAPESP — Brazil

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 for FSS [2]: Embedded, Filter and Wrapper. In this work we use the Filter approach.

In a previous work [3] 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 [4] 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.

Theses initial experiments only consider the accuracy of the induced classifiers, *i.e.* the classifiers are viewed as black boxes. The objective of this work is to analise further those results taking also into account the number of induced rules in each classifier.

The experiments were conducted using the same nine natural datasets and the four filter methods used in [3]. Afterwards, for each original dataset as well as for the correspondent reduced datasets obtained through FSS, we induced rules using the symbolic ML algorithms $\mathcal{C}4.5$-rules and $\mathcal{CN}2$. 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.

This work is organized as follow: Section 2 presents some important Rough Sets concepts. Section 3 describes the characteristics of the datasets used in the experiments. Section 4 shows the experimental setup used to run the experiments and Section 5 describes the results obtained from these experiments. Section 6 reports analysis and comparison of results. Finally, Section 7 gives some conclusions.

2 Rough Sets

This section deals with fundamental issues of the Rough Sets theory, which is a theory strongly connected with the field of Machine Learning. The theory was introduced by Zdzislaw Pawlak in the early 1980's [4], and based on this theory one can propose a formal framework for the automated transformation of data into knowledge. Pawlak has shown that the principles for learning by examples can be formulated in the basis of his theory [5]. An important result from this theory is that it simplifies the search for dominating attributes leading to specific properties.

Almost inevitably the datasets used for learning will contain imperfection, such as noise, unknown values or errors due to inaccurate measuring equipment. The Rough Set theory comes handy for dealing with these types of problems, as it is a tool for handling vagueness and uncertainty inherent to decision situations.

2.1 Decision System

A *decision system* consists of a set of *objects* where each object has a number of *attributes* with *attribute values* related to it and a special attribute called

decision attribute. The attributes are the same for all objects, but the attribute values may differ. A decision system is thus more or less the same as a relational database (dataset) with a decision attribute.

Definition 1 (Information System, Decision System).
An Information System *— IS — is an ordered pair $\mathcal{A}$ = (U,A) where U is a nonempty finite set of objects — the* Universe*, and A is a nonempty, finite set of elements called* Attributes*. The elements of the Universe will in the following be referred to as Objects. Every attribute $a \in A$ is a total function $a : U \rightarrow V_a$, where V_a is the set of allowed values for the attribute (its range). A* Decision System *— DS — is an IS $\mathcal{A}$ = (U,A) for which the attributes in A are further classified into disjoint sets of* condition *attributes C and* decision *attributes D.* $(A = C \cup D, C \cap D = \emptyset)$.

2.2 Discerning Objects

The next definition introduces the concept of an *indiscernibility relation*. If such a relation exists between two objects, it means that all their attribute values are identical with respect to the attributes under consideration, and thus cannot be discerned (distinguished) between when considering those attributes.

Definition 2 (Indiscernibility Relation).
With every subset of attributes $B \subseteq A$ in the IS $\mathcal{A}$ = (U, A), an equivalence relation IND(B) is associated, called an Indiscernibility Relation*, which is defined as follows:*

$$IND(B) = \{(x, y) \in U^2 \mid \forall a \in B, a(x) = a(y)\} \tag{1}$$

By $U/IND(B)$ is meant the set of all equivalence classes in the relation $IND(B)$.

2.3 Reducing Representation

The data in the information system can be used to discern classes only to a certain degree. However, all attributes may be required in order to be able to do so. This is why the next definition is helpful.

Definition 3 (Reduct).
A Reduct *of B is a set of attributes $B' \subseteq B$ such that all attributes $a \in B - B'$ are dispensable, and IND(B')=IND(B). The term RED(B) is used to denote de family of reducts of B.*

What this implies is that a reduct contains enough information to discern objects from each other in the decision system. The reduct is minimal and thus none of the attributes may be removed without removing the reduct property.

As a reduct is a minimal subset of features that preserves the ability to discern objects from each other, in this work we propose its use as a filter method for Feature Subset Selection.

3 Datasets

Experiments were conducted on several real world domains. Most datasets are from the UCI Irvine Repository [6], 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.

Table 3.1 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.

Table 3.1. Datasets Summary Descriptions

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

4 Experimental Setup

A series of experiments were performed, using these nine datasets and four filter methods: $\mathcal{C}$4.5-rules and ID3 present in $\mathcal{MLC}$++ [7]; the Column Importance facility provided by Mineset [8]; and, the Rough Set tool Rosetta [9].

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

It is important to note that the original datasets were not pre-processed in any way, such as 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 4.1:

1. $\mathcal{C}4.5$, ID3, CI and Rosetta are used as filters for FSS.
2. The reduced datasets, where only features selected by each filter in step 1 are considered, are used by $\mathcal{C}4.5$-rules and $\mathcal{CN}2$ to induce rules.

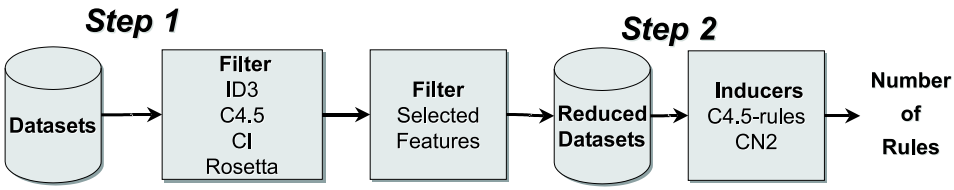

Figure 4.1. Experiment Steps

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 3.

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. 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 RS reduct, the subset of features of that reduct and the subset of features found by filters ID3, $\mathcal{C}4.5$ and CI, respectively, were used to construct the reduced datasets. Afterwards, rules were induced using the ML algorithms $\mathcal{C}4.5$-rules and $\mathcal{CN}2$.

5 Experimental Results

The next sections present the results obtained through these experiments.

5.1 Summary Tables Description

For each dataset two tables are presented:

1. The first table describes each filter selected features. To specify the experiment, the notation FSS(*method,inducer*) is used, where:
   - *method* ∈ {f} indicating that the filter (f) method has been used[2];
   - *inducer* ∈ {$\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).
2. The second 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 reduced dataset.

Table 5.1. TA – Number of Selected Features

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 5.2. TA – Number of Rules

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
Total	323	84
Mean	64.60	16.80
std-dev	3.36	1.79

Table 5.3. Bupa – Number of Selected Features

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 5.4. Bupa – Number of Rules

bupa rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	34	11
FSS(f,CI)	40	2
FSS(f,$\mathcal{C}$4.5)	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

[2] Although in this case we have only used one method for FSS, we decided to maintain the same notation than in [3] and [10] were *method* could be in {wf, wb, f} indicating wrapper-forward, wrapper-backward and filter respectively.

Table 5.5. Pima – Number of Selected Features

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 5.6. Pima – Number of Rules

pima rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	56	6
FSS(f,CI)	58	7
FSS(f,$\mathcal{C}$4.5)	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 5.7. Breast Cancer2 – Number of Selected Features

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 5.8. Breast Cancer2 – Number of Rules

breast cancer2 rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	40	12
FSS(f,CI)	47	17
FSS(f,$\mathcal{C}$4.5)	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 5.9. Cmc – Number of Selected Features

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 5.10. Cmc – Number of Rules

cmc rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	174	36
FSS(f,CI)	180	36
FSS(f,$\mathcal{C}$4.5)	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 5.11. Breast Cancer – Number of Selected Features

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 5.12. Breast Cancer – Number of Rules

breast cancer rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	18	8
FSS(f,CI)	19	8
FSS(f,$\mathcal{C}$4.5)	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 5.13. Smoke – Number of Selected Features

Inducer	Selected Features	#F	%F
FSS(f,CI)	1 2 3 4 5 6 7 8 9 10 12	11	84.62%
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 5.14. Smoke – Number of Rules

smoke rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	426	22
FSS(f,CI)	410	26
FSS(f,$\mathcal{C}$4.5)	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 5.15. Hungaria – Number of Selected Features

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 5.16. Hungaria – Number of Rules

hungaria rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	25	11
FSS(f,CI)	30	8
FSS(f,$\mathcal{C}$4.5)	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 5.17. Hepatitis – Number of Selected Features

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 5.18. Hepatitis – Number of Rules

hepatitis rules	$\mathcal{CN}2$	$\mathcal{C}$4.5-rules
all features	19	10
FSS(f,CI)	25	7
FSS(f,$\mathcal{C}$4.5)	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

6 Results Comparison

It should be observed that for dataset cmc all features seens to be relevant since none of the filters was able to discard any feature.

One important result obtained using the Rough Sets approach as filter is that the number of features selected by RS is always smaller or equal than the number of features selected by $\mathcal{C}$4.5 and ID3, *i.e.*

$$\#\text{FSS(f,RS)} \leq \#\text{FSS(f,}\mathcal{C}\text{4.5)} \text{ and } \#\text{FSS(f,RS)} \leq \#\text{FSS(f,ID3)}$$

Furthermore, the number of features selected by RS is smaller or equal to the number of features selected by CI, except for bupa dataset.

Thus, considering the number of selected features after filtering, Rough Sets can be considered the overall winner

Tables 6.1 and 6.2 summarize for inducers $\mathcal{C}$4.5-rules and $\mathcal{CN}2$ respectively, and for each dataset, the number of induced rules using the features selected by each filter as well as the mean and standard deviation.

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 it can be observed that:

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

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

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

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 6.1 and 6.2 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.*

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

Table 6.1. Number of Rules Induced by $\mathcal{C}4.5$-rules

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			

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 $\text{m}(A_2 - A_1)$ is the mean and $\text{sd}(A_2 - A_1)$ is the standard deviation calculated, respectively, using Equations 2 and 3.

Table 6.2. Number of Rules Induced by $\mathcal{CN}2$

Dataset	Rules								Using Filter		
	All	(f,CI)		(f,$\mathcal{C}$4.5)		(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				

$$\mathrm{m}(A_2 - A_1) = \mathrm{m}(A_2) - \mathrm{m}(A_1) \tag{2}$$

$$\mathrm{sd}(A_2 - A_1) = \sqrt{\frac{\mathrm{sd}(A_2)^2 + \mathrm{sd}(A_1)^2}{2}} \tag{3}$$

Afterwards, the difference in standard deviation, given by Equation 4, 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.

$$\mathrm{ad}(A_2 - A_1) = \frac{\mathrm{m}(A_2 - A_1)}{\mathrm{sd}(A_2 - A_1)} \tag{4}$$

Table 6.3 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 4 is zero are not filled.

Table 6.3. Improved Accuracies

	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) -$\mathcal{CN}2$	# $\triangle$	# $\triangledown$	# +	# $-$
ta	+			+	$-$			+	0	0	3	1
bupa	$\triangledown$			$\triangledown$	$\triangledown$			$-$	0	3	0	1
pima	$-$			$-$	$-$	+		$\triangledown$	0	1	1	3
breast cancer2	$-$	$-$		+	+	+		$-$	0	0	3	3
cmc								+	0	0	1	0
breast cancer				$-$		+	+	$\triangledown$	0	1	2	1
smoke	+			$-$	$\triangledown$			+	0	1	2	1
hungarian	+	+	$-$	$-$	+	$-$	+	$-$	0	0	4	4
hepatitis	+	+	$-$	+	$-$	+	+	$-$	0	0	5	3
# $\triangle$	0	0	0	0	0	0	0	0	0			
# $\triangledown$	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

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);

7 Conclusions

This work describes empirical results using four filter approaches for Feature Subset Selection and two symbolic inducers. The aim is to compare the number of rules induced over each datasets using all features and only the features selected by each filter method. As filters we use $\mathcal{C}4.5$, ID3, the CI MineSet™ facility and Rosetta for the Rough Sets approach. All theses four filters as well as $\mathcal{C}4.5$-rules and $\mathcal{CN}2$ used to induce rules were run using its default options setting, on nine real world datasets.

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

1. R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, pp. 273–324, 1997.
2. A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, pp. 245–271, 1997.
3. A. D. Pila and M. C. Monard, "An empirical comparison of rough sets reducts and other filters approaches for feature subset selection," in *Proceedings of the VI Ibero-American Symposium on Pattern Recognition*, (Florianópolis, SC, Brazil), pp. 41–49, November 2001.

4. Z. Pawlak, "Rough sets," *International Jornal of Computer and Information Sciences*, pp. 341–356, 1982.
5. Z. Pawlak, J. Grzymala-Busse, R. Slowinski, and W. Ziarko, "Rough sets," *Comunications of the ACM*, pp. 89–95, 1995.
6. C. Blake, E. Keogh, and C. Merz, "Uci irvine repository of machine learning databases," 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
7. R. Kohavi, D. Sommerfield, and J. Dougherty, *$\mathcal{MLC}$++: A Machine Learning Library in C++*. IEEE Computer Society Press, 1994.
8. D. Rathjens, "Mineset$^{\text{TM}}$ user's guide," 1996. Silicon Graphics, Inc.
9. A. Øhrn, "Rosetta: Technical reference manual," tech. rep., Knowledge System Group, November 1999.
10. H. D. Lee, M. C. Monard, and J. A. Baranauskas, "Empirical comparison of wrapper and filter approaches for feature subset selection," Tech. Rep. 94, ICMC-USP, oct 1999. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_94.ps.zip.