Lecture Notes in Computer Science Vol. 6113, pp. 347-354, 2010.

Pruning Classification Rules with Reference Vector Selection Methods

Karol Grudziński[1], Marek Grochowski[2] and Włodzisław Duch[2]

[1] Institute of Physics, Kazimierz Wielki University, Bydgoszcz, Poland
[2] Dept. of Informatics, Nicolaus Copernicus University, Grudziądzka 5, Poland

Contact: grudzinski.k@gmail.com, grochu@is.umk.pl, Google: W. Duch

Abstract. Attempts to extract logical rules from data often lead to large sets of classification rules that need to be pruned. Training two classifiers, the C4.5 decision tree and the Non-Nested Generalized Exemplars (NNGE) covering algorithm, on datasets that have been reduced earlier with the EkP instance compressor leads to statistically significantly lower number of derived rules with non-significant degradation of results. Similar results have been observed with other popular instance filters used for data pruning. Numerical experiments presented here illustrate that it is possible to extract more interesting and simpler sets of rules from filtered datasets. This enables a better understanding of knowledge structures when data is explored using algorithms that tend to induce a large number of classification rules.

1 Introduction

Induction of classification rules is one of the main data mining tasks, allowing for summarization of data and understanding their structure. Numerous systems have been designed for that purpose [8]. However, it is usually hard to extract low number of very informative rules without sacrificing their generalization ability. A few methods perform well on most data generating relatively low number of rules, but most rule-based systems tend to induce quite large number of rules, making the solution obtained difficult to understand. Reducing the number of classification rules is therefore an important issue in data mining.

In this paper the effect of instance selection (pruning training data) on the number of generated rules and their generalization ability is investigated. The EkP method [11, 12] has been used to select reduced reference vectors. Two classifiers capable of rule induction have been taken for our experiments, but results should generalize to other types of covering algorithms and decision trees. The first one is the NNGE system, available in the Weka package [20], which usually generates a large number of covering hyperrectangles, or logical rules. The second method, called PART [9] is based on C4.5 decision tree [17], used recursively to generate rules (taken from largest node, and removing the data covered so far). These classifiers have been trained on original data and on training partitions reduced by EkP. Additional computational experiments with other popular instance compressors have also been performed, but only the C4.5 decision tree results are reported here to save space.

2 Pruning Classification Rules with Reference Vector Selection Methods

Three types of classification rules may be used for data understanding [8]. Propositional logical rules use hyperboxes to define decision borders between classes, they are generated using either univariate decision trees or covering methods. Second, threshold logic rules, equivalent to hyperplanes that may be generated either by multivariate trees or by linear discriminants, for example linear support vector machines. Third and most general, rules based on similarity to prototypes may provide complex decision regions, including hyperboxes and fuzzy decision regions. Prototype-based rules (P-rules) are comprehensible if similarity functions are sufficiently simple. The study of prototype-based rules has been much less popular than of the other forms of rules [7, 3, 6].

Below a short description of the instance pruning algorithms which have been employed in our numerical experiments is provided.

The EkP Prototype Selection System has been used in all experiments conducted in this paper [11, 12]. Simplex method [16] for minimization of cost function (number of errors), as implemented by M. Lampton and modified by one of us (K.G.), has been used, despite its rather high computational cost. The advantage of simplex method is that it essentially does not require any user intervention to control the minimization process. The pseudocode for the simplex initialization algorithm used in our experiments is given in Algorithm 1 and the cost function procedure of the EkP system is given in Algorithm 2.

Algorithm 1 Simplex initialization algorithm for EkP

Require: A vector of training set instances **trainInstances**[]
Require: A vector p[] of optimization parameters (**numProtoPerClass** * **numClasses** * **numAttributes** dimensional)
Require: A matrix **simplex** to construct a simplex
  numPoints, the number of points to build the simplex on
  for i = 0 to **numPoints** - 1 **do**
    randomize(**trainInstances**[])
    for j = 0 to **numClasses** * **numProtoPerClass** - 1 **do**
      for k = 0 to **numAttributes** - 1 **do**
        simplex[i][k] := p[k + **numAttributes** * j] := **trainInstances**[i][k]
      end for
    end for
    simplex[i][**numAttributes**] := **costFunction**(p[])
  end for

Algorithm 2 The EkP cost function algorithm

Require: A training set **trainInstances**[], a vector p[] of optimization parameters.
  tmpTrain, empty training set.
  for i = 0 to **numClasses** * **numProtoPerClass** - 1 **do**
    for j = 0 to **numAttributes** - 1 **do**
      Extract the prototype which is stored in p[] and add it to **tmpTrain**
    end for
  end for
  Build (train) the classifier on **tmpTrain** and test it on **trainInstances**
  Remember the optimal p[] value and the lowest value of **numClassificationErrors** associated with it.
  return numClassificationErrors

Other Reference Vector Selection Methods Used. Only a very concise description of the instance pruning algorithms that have been used in our experiments is given below. For in-depth review of these algorithms see [15, 10, 19].

- **Condensed Nearest Neighbor Rule (CNN)** [13] method starts with a reference set containing one vector per class and adds incrementally to this set each instance from the training data that is wrongly classified when that reference set is used for learning and the training instances are used for testing.
- **DROP3** [19] removes instance $\boldsymbol{x}$ from the training set if it does not change classification of instances associated with $\boldsymbol{x}$. Vectors associated with $\boldsymbol{x}$ are defined as a set of instances for which instance $\boldsymbol{x}$ is one of the k nearest neighbors.
- **Edited Nearest Neighbor (ENN)** [18] removes a given instance from the training set if it's class does not agree with the majority class of its neighbors.
- **Edited NRBF** [15] uses Normalized RBF [14] to estimate probability $P(C_k|\boldsymbol{x}, T)$ of C_k class for a given vector $\boldsymbol{x}$ and the training set T. Each vector inconsistent with its class density estimation is treated as noise and is removed from the dataset. Probability that a vector from correct class will be removed is low.
- **Iterative Case Filtering (ICF)** [4] starts from DROP3 and creates hyperspheres that contain only single-class instances, removing instances which are located inside clusters of vectors from the same class.
- **Gabriel Editing (GE)** [2] method is based on graph theory. It uses the Gabriel graph to define neighbors and removes from the training dataset all instances that belong to the same class as all their neighbors.

3 Numerical Experiments

The EkP prototype selection method is compared here with other instance compressors, and examine rules obtained from classifiers which have been trained on pruned data. In the first part experiments with the NNGE and C4.5 systems trained on a large number of datasets filtered using the EkP method have been described. In the second part EkP algorithm has been matched against several instance compressors. Finally explicit example of the benefit of this approach is demonstrated by presenting greatly simplified and highly accurate rules for non-trivial dataset.

Pruning C4.5 and NNGE Rules with the EkP Instance Selector. Numerical experiments have been performed on 17 real-world problems taken mainly from the UCI repository of machine-learning databases [1], described in Table 1. The EkP system has been used for instance selection and the rules have been extracted with the C4.5 and NNGE classifiers. All experiments have been performed using SBLWeka, an extension of Weka system [20], done by one of us (K.G.)

In all experiments 10 simplex points for the EkP system are used and j=1, 2, 3 or 5 prototypes per class selected, denoted EkPs10pj. In the first experiment the influence of data pruning on classification generalization has been examined (Tables 2 and 3). One prototype per class is not sufficient, but increasing the number of prototypes to 5 per class leads to results that are statistically equivalent to training on the whole dataset, with the NNGE method on all 17 problems, and with the C4.5 system on 16 problems.

Table 1. Datasets used in numerical experiments

#	Dataset	# Instances	# Attributes	# Numeric	# Nominal	# Classes	Base Rate [%]	Rnd. Choice [%]
1	Appendicitis	106	8	7	1	2	80.2	50.0
2	Breast C.W.	286	10	0	10	2	70.3	50.0
3	Horse Colic	368	23	7	16	2	63.0	50.0
4	Credit rating	690	16	6	10	2	55.5	50.0
5	German credit	1000	21	8	13	2	70.0	50.0
6	Pima I.D.	768	9	8	1	2	65.1	50.0
7	Glass	214	10	9	1	6	35.5	16.7
8	Cleveland heart	303	14	6	8	2	54.4	50.0
9	Hungarian heart	294	14	6	8	2	63.9	50.0
10	Heart Statlog	270	14	13	1	2	55.6	50.0
11	Hepatitis	155	20	2	18	2	79.4	50.0
12	Labor	57	17	8	9	2	64.7	50.0
13	Lymphography	148	19	0	19	4	54.8	25.0
14	Primary Tumor	339	18	0	18	21	24.8	4.8
15	Sonar	208	61	60	1	2	53.4	50.0
16	Voting	435	17	0	17	2	61.4	50.0
17	Zoo	101	18	0	18	7	40.6	14.3
Average		337.8	18.2	8.2	9.9	3.8	58.4	41.8

Table 2. NNGE – Generalization Ability

Data Set	NNGE	EkPs10p1	EkPs10p2	EkPs10p3	EkPs10p5
Appendicitis	83.7 ± 11.3	85.6 ± 9.9	86.0 ± 9.3	85.7 ± 9.2	86.5 ± 10.0
Breast C.W.	67.8 ± 7.1	70.8 ± 5.5	72.4 ± 5.9	72.5 ± 6.2	71.1 ± 6.7
Horse Colic	79.0 ± 6.5	66.4 ± 7.5 •	76.7 ± 7.2	79.9 ± 6.5	81.5 ± 6.3
Credit rating	82.8 ± 4.7	72.3 ± 5.9 •	80.4 ± 5.0	84.5 ± 4.1	85.2 ± 4.1
German credit	69.2 ± 4.5	69.9 ± 0.7	70.1 ± 2.0	70.0 ± 2.6	70.2 ± 1.9
Pima I.D.	72.8 ± 4.6	67.5 ± 5.0 •	70.3 ± 4.8	72.2 ± 4.5	72.6 ± 5.0
Glass	68.0 ± 9.3	53.6 ± 9.3 •	57.9 ± 9.5 •	61.1 ± 7.7	62.5 ± 9.0
Cleveland heart	77.8 ± 7.7	79.0 ± 7.1	77.6 ± 7.6	78.6 ± 7.7	77.6 ± 7.0
Hungarian heart	79.6 ± 6.8	81.6 ± 6.0	82.5 ± 6.5	81.7 ± 7.3	81.0 ± 7.2
Heart Statlog	77.3 ± 8.1	74.3 ± 9.2	79.0 ± 7.4	77.9 ± 7.5	77.2 ± 8.0
Hepatitis	81.9 ± 8.1	78.5 ± 6.1	82.7 ± 8.4	83.1 ± 8.0	82.9 ± 7.5
Labor	86.2 ± 15.2	85.9 ± 15.8	83.1 ± 17.4	82.8 ± 14.8	84.2 ± 13.8
Lymphography	77.1 ± 10.1	75.3 ± 10.7	73.3 ± 10.3	74.6 ± 9.5	75.7 ± 9.2
Primary Tumor	39.1 ± 7.2	34.7 ± 6.7	36.6 ± 6.9	38.4 ± 7.1	39.1 ± 6.6
Sonar	71.1 ± 9.2	63.2 ± 11.9	67.6 ± 10.3	68.6 ± 9.3	69.3 ± 9.8
Voting	95.1 ± 3.1	88.8 ± 4.5 •	93.0 ± 4.1	94.8 ± 3.6	94.9 ± 3.1
Zoo	94.1 ± 6.4	83.6 ± 8.5 •	90.0 ± 8.0	93.1 ± 6.2	95.4 ± 6.2
Average	76.6 ± 12.6	72.4 ± 13.4	75.2 ± 13.2	76.4 ± 13.0	76.9 ± 13.1
Win/Tie/Lose		0/11/6	0/16/1	0/17/0	0/17/0

○, • statistically significant improvement or degradation

With these 5 prototypes per class statistically lower number of rules was obtained in 16 cases for NNGE and 17 times in case of C4.5 (see Table 4 and 5).

Comparison of pruning rules with various data compressors. The efficiency of EkP algorithm has also been compared with several vector selection methods listed in section 2. Table 6 presents average accuracy of classification estimated using 10-fold stratified cross-validation tests repeated 10 times. The average number of rules generated by the C4.5 algorithm is reported in Tab. 7. First column contains results for C4.5 trained on entire training dataset, and each successive column represent results for C4.5 trained on data reduced by one of the vector selection methods. Columns are sorted according to the average compression achieved by these methods, as shown in the last row of Tab. 6. The number of resulting prototypes in EkP method was set for each training to the value that corresponds to about 10% of the size of the original training set. On

Table 3. C4.5 – Generalization Ability

Data Set	C4.5	EkPs10p1	EkPs10p2	EkPs10p3	EkPs10p5
Appendicitis	84.6 ± 10.3	80.2 ± 2.6	83.7 ± 8.7	84.4 ± 11.4	85.5 ± 9.7
Breast C.W.	69.4 ± 7.6	70.3 ± 1.4	71.0 ± 5.1	69.1 ± 6.7	69.5 ± 7.1
Horse Colic	84.4 ± 5.9	63.0 ± 1.1 •	78.7 ± 8.6	81.5 ± 5.8	82.2 ± 5.9
Credit rating	84.4 ± 4.3	55.4 ± 1.2 •	73.5 ± 11.7•	85.5 ± 4.0	85.5 ± 3.9
German credit	70.5 ± 4.2	70.0 ± 0.0	69.8 ± 1.1	69.6 ± 1.9	69.6 ± 1.9
Pima I.D.	73.4 ± 4.5	65.1 ± 0.3 •	69.5 ± 5.7	71.5 ± 5.9	73.4 ± 5.0
Glass	68.7 ± 10.6	48.9 ± 9.1 •	56.1 ± 7.1 •	58.1 ± 8.8 •	60.1 ± 9.6 •
Cleveland heart	78.0 ± 7.1	74.1 ± 7.6	73.3 ± 7.1	73.5 ± 7.3	77.1 ± 8.1
Hungarian heart	81.1 ± 6.7	80.6 ± 7.9	80.8 ± 7.1	80.0 ± 6.7	79.3 ± 7.6
Heart Statlog	77.3 ± 7.8	55.6 ± 0.0 •	71.0 ± 9.5	72.8 ± 8.3	72.6 ± 7.8
Hepatitis	79.8 ± 8.5	79.4 ± 2.3	79.5 ± 3.8	81.4 ± 6.4	82.1 ± 9.0
Labor	77.7 ± 15.5	64.7 ± 3.1 •	79.1 ± 14.8	77.7 ± 15.8	79.8 ± 15.1
Lymphography	76.4 ± 9.3	68.7 ± 12.	72.7 ± 11.0	72.8 ± 10.9	76.5 ± 11.4
Primary Tumor	40.9 ± 6.4	31.1 ± 6.4 •	36.9 ± 6.7	36.9 ± 7.7	38.0 ± 7.6
Sonar	77.4 ± 9.4	53.4 ± 1.6 •	59.6 ± 12.2•	68.6 ± 10.3•	69.6 ± 10.1
Voting	96.0 ± 3.2	61.4 ± 0.8 •	94.4 ± 4.2	95.6 ± 2.8	95.6 ± 2.8
Zoo	93.4 ± 7.3	71.6 ± 5.7 •	81.7 ± 6.2 •	87.3 ± 7.4 •	91.5 ± 7.7
Average	77.3 ± 12.0	64.3 ± 12.8	72.4 ± 12.8	74.5 ± 13.1	75.8 ± 13.1
Win/Tie/Lose		0/7/10	0/13/4	0/14/3	0/16/1

○, • statistically significant improvement or degradation

Table 4. NNGE – Number of Rules

Data Set	NNGE	EkPs10p1	EkPs10p2	EkPs10p3	EkPs10p5
Appendicitis	16.0 ± 2.2	1.9 ± 0.2•	2.0 ± 0.1•	2.3 ± 0.5•	2.9 ± 1.0 •
Breast C.W.	86.4 ± 5.1	1.7 ± 0.4•	2.0 ± 0.0•	2.1 ± 0.3•	2.8 ± 0.9 •
Horse Colic	97.6 ± 18.8	1.9 ± 0.3•	2.0 ± 0.0•	2.0 ± 0.0•	2.2 ± 0.4 •
Credit rating	142.4 ± 14.6	2.0 ± 0.0•	2.0 ± 0.0•	2.0 ± 0.1•	2.0 ± 0.2 •
German credit	347.4 ± 26.0	1.1 ± 0.2•	1.8 ± 0.4•	2.0 ± 0.4•	2.8 ± 0.8 •
Pima I.D.	263.8 ± 23.2	1.8 ± 0.4•	2.0 ± 0.0•	2.1 ± 0.3•	2.7 ± 0.8 •
Glass	48.1 ± 4.8	3.9 ± 0.7•	6.0 ± 1.0•	8.1 ± 1.3•	11.9 ± 1.5 •
Cleveland heart	71.6 ± 9.3	2.0 ± 0.0•	2.9 ± 0.7•	4.0 ± 0.9•	6.1 ± 1.3 •
Hungarian heart	61.9 ± 7.2	2.0 ± 0.1•	2.6 ± 0.7•	3.6 ± 1.1•	5.9 ± 1.5 •
Heart Statlog	68.3 ± 8.7	2.0 ± 0.0•	2.0 ± 0.0•	2.0 ± 0.1•	2.8 ± 0.8 •
Hepatitis	27.6 ± 3.8	1.5 ± 0.5•	2.0 ± 0.1•	2.1 ± 0.3•	2.4 ± 0.6 •
Labor	7.9 ± 1.4	2.0 ± 0.1•	2.0 ± 0.0•	2.1 ± 0.3•	2.6 ± 0.7 •
Lymphography	32.0 ± 5.2	2.1 ± 0.3•	2.8 ± 0.7•	3.8 ± 1.1•	5.9 ± 1.3 •
Primary Tumor	147.8 ± 4.7	13.3 ± 1.7•	24.4 ± 2.5•	35.8 ± 3.0•	58.2 ± 3.9 •
Sonar	45.7 ± 7.2	2.0 ± 0.0•	2.0 ± 0.0•	2.0 ± 0.1•	2.4 ± 0.6 •
Voting	29.0 ± 3.6	2.0 ± 0.0•	2.0 ± 0.0•	2.0 ± 0.1•	2.1 ± 0.3 •
Zoo	7.0 ± 0.0	5.0 ± 0.5•	6.4 ± 0.5•	6.9 ± 0.3	7.0 ± 0.0
Average	88.3 ± 92.9	2.8 ± 2.8	3.9 ± 5.4	5.0 ± 8.1	7.2 ± 13.4
Win/Tie/Lose		0/0/17	0/0/17	0/1/16	0/1/16

○, • statistically significant improvement or degradation

average EkP produced smallest size data among all methods compared here. Table 6 shows that the EkP algorithm, despite such high reduction of the training data size, was able to achieve good accuracy in comparison to other pruning methods tested here. For two datasets (Horse Colic and Breast Cancer Wisconsin) paired corrected t-test shows significant improvement in favor of EkP, each time producing about 6 times less rules than C4.5 with all training data. Only GE and ENN methods can compete in generalization with EkP, giving no significant difference in comparison with original C4.5. However these pruning techniques produce training data with average size reduced only to 85% in case of ENN, and 95% for GE, respectively, while the EkP method creates much smaller datasets.

Table 5. C4.5 – Number of Rules

Data Set	C4.5	EkPs10p1	EkPs10p2	EkPs10p3	EkPs10p5
Appendicitis	3.1 ± 0.6	1.0 ± 0.0●	1.7 ± 0.4●	2.0 ± 0.2●	2.0 ± 0.1●
Breast C.W.	18.4 ± 4.2	1.0 ± 0.0●	1.5 ± 0.5●	1.9 ± 0.5●	2.3 ± 0.5●
Horse Colic	9.1 ± 2.7	1.0 ± 0.0●	1.9 ± 0.3●	2.0 ± 0.0●	2.3 ± 0.5●
Credit rating	31.5 ± 7.7	1.0 ± 0.0●	2.0 ± 0.2●	2.0 ± 0.0●	2.0 ± 0.0●
German credit	69.7 ± 5.8	1.0 ± 0.0●	1.1 ± 0.3●	1.1 ± 0.4●	1.3 ± 0.6●
Pima I.D.	7.5 ± 1.5	1.0 ± 0.0●	1.8 ± 0.4●	1.9 ± 0.2●	2.0 ± 0.1●
Glass	15.2 ± 1.6	2.6 ± 0.5●	3.5 ± 0.6●	4.5 ± 0.9●	6.6 ± 1.3●
Cleveland heart	19.6 ± 2.7	2.0 ± 0.0●	2.1 ± 0.3●	2.6 ± 0.7●	3.9 ± 0.8●
Hungarian heart	8.2 ± 2.4	2.0 ± 0.0●	2.1 ± 0.3●	2.3 ± 0.6●	3.3 ± 1.2●
Heart Statlog	17.6 ± 2.4	1.0 ± 0.0●	2.0 ± 0.0●	2.0 ± 0.0●	2.2 ± 0.4●
Hepatitis	8.6 ± 1.7	1.0 ± 0.0●	1.1 ± 0.3●	1.6 ± 0.5●	2.1 ± 0.3●
Labor	3.4 ± 0.8	1.0 ± 0.0●	2.0 ± 0.1●	2.0 ± 0.0●	2.1 ± 0.3●
Lymphography	11.3 ± 2.3	2.0 ± 0.2●	2.1 ± 0.4●	2.6 ± 0.6●	3.4 ± 0.6●
Primary Tumor	41.1 ± 3.5	6.1 ± 0.9●	10.4 ± 1.3●	13.8 ± 1.9●	20.4 ± 2.3●
Sonar	7.5 ± 1.0	1.0 ± 0.0●	2.0 ± 0.0●	2.0 ± 0.0●	2.0 ± 0.0●
Voting	6.1 ± 1.1	1.0 ± 0.0●	2.0 ± 0.0●	2.0 ± 0.0●	2.0 ± 0.0●
Zoo	7.6 ± 0.5	3.0 ± 0.0●	4.9 ± 0.5●	6.0 ± 0.5●	6.9 ± 0.4●
Average	16.8 ± 16.9	1.7 ± 1.3	2.6 ± 2.2	3.1 ± 3.0	3.9 ± 4.5
Win/Tie/Lose		0/0/17	0/0/17	0/0/17	0/0/17

○, ● statistically significant improvement or degradation

Table 6. Average classification accuracy

Data Set	C4.5	EkP	ENRBF	DROP3	ICF	CNN	ENN	GE
Appendicitis	85.3 ± 10.4	84.1 ± 10.3	80.3 ± 18.5	83.0 ± 11.3	82.0 ± 12.1	79.9 ± 14.1	83.6 ± 10.9	84.1 ± 11.1
Breast C.W.	68.6 ± 8.7	73.6 ± 7.8 ○	58.8 ± 12.0 ●	65.3 ± 12.1	66.0 ± 9.3	61.7 ± 10.0	70.3 ± 8.1	68.0 ± 8.2
Horse Colic	79.6 ± 6.0	84.6 ± 5.9 ○	77.0 ± 7.8	80.3 ± 6.5	76.0 ± 7.4	74.6 ± 7.2	81.2 ± 6.3	79.6 ± 6.0
Credit rating	84.2 ± 4.0	85.5 ± 3.7	70.4 ± 11.8 ●	81.0 ± 6.0	83.2 ± 4.6	73.8 ± 5.2 ●	85.4 ± 4.0	84.3 ± 4.1
German credit	71.1 ± 4.1	70.7 ± 4.7	60.1 ± 6.1 ●	66.6 ± 6.9	66.8 ± 4.8 ●	64.6 ± 5.1 ●	73.5 ± 4.1	71.1 ± 4.1
Pima I.D.	73.2 ± 4.1	73.7 ± 4.0	70.7 ± 7.1	71.1 ± 5.7	69.6 ± 7.1	71.1 ± 6.1	74.9 ± 4.4	73.5 ± 4.1
Glass	68.6 ± 10.5	58.4 ± 10.6 ●	59.1 ± 11.5 ●	51.4 ± 14.1 ●	61.7 ± 10.5	60.6 ± 12.0	68.2 ± 9.6	69.0 ± 9.9
Cleveland heart	78.1 ± 7.3	77.5 ± 7.0	72.3 ± 9.4	76.2 ± 9.4	74.2 ± 9.6	70.6 ± 8.8 ●	80.7 ± 7.4	78.0 ± 7.2
Hungarian heart	80.6 ± 7.3	78.6 ± 7.6	73.4 ± 10.9	74.0 ± 11.3	75.0 ± 9.9	73.6 ± 9.5 ●	78.5 ± 7.9	80.5 ± 7.4
Heart Statlog	77.4 ± 7.7	78.6 ± 8.3	72.1 ± 9.3	73.4 ± 9.1	73.2 ± 9.2	71.8 ± 8.6	79.3 ± 6.6	77.9 ± 8.0
Hepatitis	81.3 ± 10.6	80.2 ± 9.9	64.6 ± 17.0 ●	79.9 ± 11.1	78.1 ± 10.8	67.4 ± 16.0 ●	82.1 ± 9.4	81.5 ± 10.7
Labor	82.8 ± 12.6	82.1 ± 13.6	56.1 ± 25.4 ●	77.8 ± 18.8	76.9 ± 19.3	81.0 ± 17.5	81.5 ± 13.7	83.5 ± 12.6
Lymphography	75.8 ± 11.5	76.9 ± 11.0	69.3 ± 13.7	72.0 ± 12.5	73.0 ± 12.4	72.5 ± 11.1	76.6 ± 11.3	75.8 ± 11.5
Primary Tumor	40.3 ± 7.8	32.7 ± 8.2 ●	34.6 ± 8.4	27.3 ± 7.9 ●	37.4 ± 8.1	36.4 ± 8.8	39.6 ± 8.5	40.3 ± 7.8
Sonar	74.8 ± 10.7	71.3 ± 9.7	59.3 ± 12.0 ●	66.9 ± 10.6	71.7 ± 11.3	66.5 ± 11.1	76.5 ± 9.5	74.7 ± 10.9
Voting	95.0 ± 3.1	95.2 ± 3.0	91.0 ± 11.2	95.2 ± 3.3	94.7 ± 3.2	90.9 ± 5.8 ●	95.7 ± 2.7	95.0 ± 3.1
Zoo	92.8 ± 8.7	71.1 ± 14.7 ●	70.0 ± 15.4 ●	67.3 ± 15.7 ●	84.0 ± 14.1	92.2 ± 8.9	84.9 ± 13.6	92.8 ± 8.7
Average	77.0 ± 12.0	75.0 ± 13.5	67.0 ± 12.2	71.1 ± 14.7	73.1 ± 12.1	71.1 ± 12.5	77.2 ± 11.6	77.0 ± 12.0
Win/Tie/Lose		2/12/3	0/9/8	0/14/3	0/16/1	0/11/6	0/17/0	0/17/0
Wilcoxon p-value		0.181	0.000	0.000	0.000	0.000	0.136	0.274
Average compression [%]		9.0 ± 1.0	11.0 ± 4.9	12.4 ± 6.2	27.7 ± 9.5	46.0 ± 14.8	85.4 ± 9.8	94.6 ± 10.7

○, ● statistically significant improvement or degradation

Rules generated for the Mushroom dataset may serve as an interesting example to see the influence of the EkP selection of reference vectors on C4.5 decision tree rules. Direct application of the C4.5 algorithm creates quite complex set of rules with odor, gill-size, ring-number, spore-print-color, stalk-shape, stalk-surface-below-ring, population and bruises used in their conditions. So far the best published set of rules that distinguish poisonous and edible mushrooms was [5]:

1. odor=NOT(almond.OR.anise.OR.none): Poisonous
2. spore-print-color=green: Poisonous
3. Else: Edible

Table 7. Average number of rules created by C4.5.

Data Set	C4.5	EkP	ENRBF	DROP3	ICF	CNN	ENN	GE
Appendicitis	3.2 ± 0.7	2.0 ± 0.2 ●	1.8 ± 0.4 ●	2.0 ± 0.0 ●	2.3 ± 0.5 ●	1.9 ± 0.8 ●	3.6 ± 1.1	3.1 ± 0.6
Breast C.W.	18.7 ± 4.1	3.1 ± 0.4 ●	4.2 ± 1.9 ●	2.4 ± 0.8 ●	7.2 ± 2.2 ●	15.2 ± 3.7	11.2 ± 3.8 ●	17.5 ± 3.9
Horse Colic	20.0 ± 3.3	3.5 ± 0.7 ●	3.0 ± 1.0 ●	2.4 ± 0.7 ●	5.9 ± 2.0 ●	16.3 ± 3.1 ●	11.3 ± 1.9 ●	20.0 ± 3.3
Credit rating	30.0 ± 3.4	5.1 ± 0.9 ●	4.3 ± 1.1 ●	3.9 ± 1.5 ●	11.4 ± 1.7 ●	23.9 ± 3.1 ●	14.7 ± 2.0 ●	30.3 ± 3.2
German credit	69.1 ± 6.2	8.5 ± 2.1 ●	11.8 ± 2.3 ●	10.4 ± 2.2 ●	23.9 ± 3.4 ●	54.2 ± 5.5 ●	38.9 ± 3.4 ●	69.1 ± 6.2
Pima I.D.	7.7 ± 1.7	4.8 ± 1.4 ●	5.1 ± 1.5 ●	5.7 ± 1.9 ●	3.7 ± 1.3 ●	3.6 ± 1.6 ●	9.5 ± 2.0	7.4 ± 1.6
Glass	15.5 ± 1.8	3.8 ± 0.7 ●	5.2 ± 0.6 ●	6.8 ± 1.4 ●	9.5 ± 0.9 ●	13.8 ± 2.2	10.5 ± 0.9 ●	15.6 ± 2.2
Cleveland heart	20.3 ± 2.8	3.9 ± 0.6 ●	4.1 ± 1.0 ●	4.4 ± 0.9 ●	8.4 ± 1.6 ●	16.1 ± 2.3 ●	12.1 ± 1.8 ●	20.4 ± 3.0
Hungarian heart	16.1 ± 2.5	3.8 ± 1.2 ●	3.6 ± 1.1 ●	3.1 ± 1.1 ●	6.1 ± 1.7 ●	13.2 ± 2.3 ●	9.6 ± 2.2 ●	15.9 ± 2.6
Heart Statlog	18.0 ± 2.5	4.0 ± 0.6 ●	3.4 ± 0.9 ●	3.6 ± 1.0 ●	8.3 ± 1.5 ●	14.5 ± 2.2 ●	12.3 ± 1.9 ●	18.0 ± 2.5
Hepatitis	9.4 ± 1.6	2.2 ± 0.5 ●	2.4 ± 0.6 ●	2.3 ± 0.6 ●	3.1 ± 1.1 ●	8.0 ± 1.7	4.7 ± 1.2 ●	9.5 ± 1.6
Labor	3.4 ± 0.7	2.0 ± 0.0 ●	1.7 ± 0.6 ●	2.0 ± 0.1 ●	2.2 ± 0.5 ●	2.7 ± 0.7 ●	3.0 ± 0.7	3.4 ± 0.7
Lymphography	11.3 ± 2.1	3.0 ± 0.2 ●	3.5 ± 0.9 ●	2.7 ± 0.7 ●	6.0 ± 1.5 ●	8.4 ± 2.0 ●	8.8 ± 1.9 ●	11.3 ± 2.1
Primary Tumor	46.7 ± 4.0	6.5 ± 1.2 ●	12.2 ± 1.7 ●	5.8 ± 1.7 ●	22.5 ± 2.7 ●	45.0 ± 3.9	24.7 ± 2.8 ●	46.7 ± 4.0
Sonar	7.3 ± 1.1	2.3 ± 0.4 ●	2.7 ± 0.7 ●	3.2 ± 0.7 ●	4.4 ± 0.6 ●	5.3 ± 0.9 ●	6.4 ± 0.9	7.3 ± 1.1
Voting	8.8 ± 2.6	3.8 ± 0.7 ●	3.2 ± 1.2 ●	2.0 ± 0.1 ●	5.0 ± 1.5 ●	7.6 ± 1.7	6.1 ± 1.8 ●	8.8 ± 2.6
Zoo	7.7 ± 0.5	3.0 ± 0.0 ●	3.9 ± 0.4 ●	3.7 ± 0.7 ●	6.3 ± 0.7 ●	7.6 ± 0.5	6.4 ± 0.7 ●	7.7 ± 0.5
Average	18.4 ± 16.9	3.8 ± 1.7	4.5 ± 3.0	3.9 ± 2.2	8.0 ± 6.2	15.1 ± 14.3	11.4 ± 8.7	18.3 ± 16.9
Win/Tie/Lose		0/0/17	0/0/17	0/0/17	0/0/17	0/6/11	0/4/13	0/17/0
Wilcoxon p-value		0.000	0.000	0.000	0.000	0.000	0.001	0.296

○, ● statistically significant improvement or degradation

This set of rules has 99.4% accuracy in overall data reclassification, but it is quite robust and may also be found in crossvalidation tests. Slightly more accurate set of 3 rules was found combining EkP selection with C4.5 decision tree rules:

1. odor = none AND spore-print-color NOT green: Edible
2. odor NOT almond AND odor NOT anise: Poisonous
3. Default: Edible

These rules summarize the entire Mushroom dataset at the 99.7% accuracy level. For several other datasets similar excellent results may be demonstrated but are omitted here due to the lack of space.

4 Conclusions

Selection of training vectors is a powerful method that should be used more often, especially for very large datasets. The experiments presented in this paper compared results of training two inductive methods for generation of logical rules, NNGE and PART, on the full training data and on the data reduced by using several algorithms or vector selection. In particular recently introduced EkP system proved to be quite competitive, reducing the original dataset sometimes even by an order of magnitude, simplifying subsequent training and reducing the number of rules also by an order of magnitude, without significant reduction of accuracy. Rules generated in this way for the Mushroom database are surprisingly compact and easy to comprehend, the best found so far for this dataset.

It is clear that training on appropriately pruned data will be especially useful for very large datasets, giving hope to find solutions that are compact, accurate and easy to understand.

References

1. A. Asuncion and D.J. Newman. UCI machine learning repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2009.
2. B. Bhattacharya, K. Mukherjee, and G. Toussaint. Geometric decision rules for instance-based learning problems. *LNCS*, 3776:60–69, 2005.
3. M. Blachnik and W. Duch. *Prototype rules from SVM*, volume 80 of *Springer Studies in Computational Intelligence*, pages 163–184. Springer, 2008.
4. H. Brighton and C. Mellish. Advances in instance selection for instance-based learning algorithms. *Data Mining and Knowledge Discovery*, 6(2):153–172, 2002.
5. W. Duch, R. Adamczak, and K. Grąbczewski. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, 12:277–306, 2001.
6. W. Duch and M. Blachnik. Fuzzy rule-based systems derived from similarity to prototypes. *Lecture Notes in Computer Science*, 3316:912–917, 2004.
7. W. Duch and K. Grudziński. Prototype based rules - new way to understand the data. In *IEEE International Joint Conference on Neural Networks*, pages 1858–1863, Washington D.C, 2001. IEEE Press.
8. W. Duch, R. Setiono, and J. Zurada. Computational intelligence methods for understanding of data. *Proceedings of the IEEE*, 92(5):771–805, 2004.
9. Eibe Frank and Ian H. Witten. Generating accurate rule sets without global optimization. In J. Shavlik, editor, *Machine Learning: Proceedings of the Fifteenth International Conference*. Morgan Kaufmann Publishers, San Francisco, CA., 1998.
10. M. Grochowski and N. Jankowski. Comparison of instance selection algorithms. II. Results and comments. *Lecture Notes in Computer Science*, 3070:580–585, 2004.
11. K. Grudzinski. E*k*P: A fast minimization–based prototype selection algorithm. In *Intelligent Information Systems XVI*, pages 45–53. Academic Publishing House EXIT, Warsaw, Poland, 2008.
12. K. Grudzinski. Selection of prototypes with the E*k*P system. *Control and Cybernetics*, submitted.
13. P. E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 114:515–516, 1968.
14. N. Jankowski. Data regularization. In L. Rutkowski and R. Tadeusiewicz, editors, *Neural Networks and Soft Computing*, pages 209–214, 2000.
15. N. Jankowski and M. Grochowski. Comparison of instance selection algorithms. I. Algorithms survey. *Lecture Notes in Computer Science*, 3070:598–603, 2004.
16. J. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.
17. J.R. Quinlan. *C 4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA, 1993.
18. D.L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Systems, Man and Cybernetics*, 2:408–421, 1972.
19. D.R. Wilson and T.R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):257–286, 2000.
20. I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd Ed, 2005.