Pruning Classification Rules with the EkP System

Karol Grudziński GRUDZINSKI.K@GMAIL.COM
Institute of Physics
Kazimierz Wielki University
Plac Weyssenhoffa 11, 85-072 Bydgoszcz, Poland

Editor: Leslie Pack Kaelbling

Abstract

This paper concerns pruning classification rules with our EkP prototype selection system. Experiments on eighteen datasets and with two classifiers indicate that adopting our approach leads to attaining statistically indistinguishable results of generalization of the C4.5 and NNGE systems in comparsion to the ones obtained by training these methods on data sets which size was reduced earlier with the EkP model. The number of rules obtained in such a way is statistically lower than the one obtained by running the C4.5 and NNGE on the original data - i.e. before the training data prunning with the EkP method. We believe that our approach may lead to better data understanding when it is expolored with systems similar to NNGE that tend to induce a large number of classification rules.

Keywords: Induction and Pruning of Classification Rules, Selection of Prototypes, Training Data Reduction

1. Introduction

Induction of classification rules is one of the main data mining methodologies that allows for summarizing data and understanding its structure. There are numerous systems which have been designed for that purpose []. Usually, it is however hard to extract low number of very informative rules without sacrificing generalization ability of the systems with which these rules have been obtained. However there are many methods that perform well on most data and give relatively low number of rules, there is also a large group of classifiers that tend to produce enormous number of rules. Pruning classification rules is therefore a very important issue in data mining and it is not surprising that this subject has been explored by reserarchers for a long time.

In this paper the effect of training data pruning on the number of the generated rules is investigated. In order to reduce the number of training samples the EkP [] system has been used. Two classifiers capable of rule induction have been taken for our experiments. The first one is the NNGE system which is known from generation of a large number of rules. The second method that has been taken for our study is the well known C4.5 decision tree. These classifiers have been used both as underlying methods for the EkP system in order to reduce the size of the training data as well as post-classifiers for classification of the unseen samples with the rules that have been extracted with these systems by training them on the reduced training partitions.

Table 1: NNGE Training Times

Data Set	(NNGE)	(EkP 1 p. + NNGE)	(EkP 2 p. + NNGE)	(EkP 3 p. + NNGE)	(EkP 5 p. + NNGE)
appendicitis	0.018280±0.016462	0.064830±0.053180 ○	0.030900±0.016159	0.032670±0.016560	0.036980±0.017264 ○
audiology	0.325820±0.068313	1.286440±0.092300 ○	2.202330±0.116631 ○	3.124640±0.141720 ○	4.855390±0.171500 ○
breast-cancer	0.056190±0.020598	0.061980±0.019471	0.058990±0.020311	0.065550±0.020111	0.071790±0.014703
horse-colic	0.178230±0.069398	0.091970±0.022301 ●	0.103170±0.023976 ●	0.111930±0.022410 ●	0.135390±0.029760
credit-rating	0.175740±0.047294	0.101390±0.020515 ●	0.112820±0.030241 ●	0.112810±0.027037 ●	0.125510±0.028250 ●
german-credit	1.753240±0.294875	0.151140±0.032946 ●	0.167800±0.032206 ●	0.175560±0.029499 ●	0.211780±0.039254 ●
pima-diabetes	0.228690±0.052866	0.089800±0.032955 ●	0.090200±0.022242 ●	0.096650±0.028137 ●	0.098640±0.024496 ●
glass	0.016470±0.001845	0.062630±0.015828 ○	0.080530±0.022478 ○	0.095360±0.020430 ○	0.125740±0.027138 ○
cleveland-14-heart-disease	0.049280±0.020876	0.075890±0.022004 ○	0.088460±0.020639 ○	0.101020±0.022864 ○	0.129080±0.027115 ○
hungarian-14-heart-disease	0.038950±0.017073	0.074620±0.024348 ○	0.087500±0.029107 ○	0.099660±0.026198 ○	0.124550±0.032985 ○
heart-statlog	0.046090±0.018793	0.063160±0.024578	0.065430±0.017016 ○	0.064950±0.015349 ○	0.073160±0.022355 ○
hepatitis	0.017060±0.002403	0.047520±0.019033 ○	0.054360±0.020285 ○	0.061060±0.022290 ○	0.072270±0.023820 ○
labor	0.004020±0.000876	0.030760±0.014736 ○	0.038550±0.008275 ○	0.043130±0.014416 ○	0.055310±0.017777 ○
lymphography	0.026570±0.014349	0.062400±0.015937 ○	0.080780±0.024785 ○	0.098820±0.024547 ○	0.133520±0.026049 ○
primary-tumor	0.204060±0.034871	0.320770±0.044075 ○	0.569370±0.056287 ○	0.843180±0.070714 ○	1.466150±0.089625 ○
sonar	0.171080±0.044558	0.125610±0.025867 ●	0.175280±0.026681	0.211870±0.033228 ○	0.248320±0.034915 ○
vote	0.033660±0.016193	0.083280±0.020795 ○	0.093680±0.027241 ○	0.098640±0.021795 ○	0.103370±0.023012 ○
zoo	0.011310±0.000895	0.078600±0.023293 ○	0.107160±0.024807 ○	0.134980±0.030210 ○	0.192530±0.064054 ○

○, ● statistically significant improvement or degradation

Table 2: NNGE Testing Times

Data Set	(NNGE)	(EkP 1 p. + NNGE)	(EkP 2 p. + NNGE)	(EkP 3 p. + NNGE)	(EkP 5 p. + NNGE)
appendicitis	0.001110±0.003318	0.000070±0.000256	0.000020±0.000141	0.000030±0.000171	0.000030±0.000171
audiology	0.004000±0.001172	0.001010±0.000301 ●	0.001510±0.000502 ●	0.002030±0.000460 ●	0.002730±0.000529 ●
breast-cancer	0.001770±0.002895	0.000080±0.000307	0.000070±0.000256	0.000120±0.000327	0.000100±0.000302
horse-colic	0.004400±0.000921	0.000150±0.000359 ●	0.000240±0.000429 ●	0.000270±0.000468 ●	0.000200±0.000402 ●
credit-rating	0.008620±0.002295	0.000410±0.000621 ●	0.000220±0.000462 ●	0.000210±0.000478 ●	0.000360±0.000482 ●
german-credit	0.043930±0.016861	0.000380±0.000599 ●	0.000420±0.000516 ●	0.000470±0.000627 ●	0.000530±0.000559 ●
pima-diabetes	0.007550±0.000783	0.000110±0.000314 ●	0.000180±0.000520 ●	0.000140±0.000493 ●	0.000170±0.000587 ●
glass	0.000530±0.000502	0.000110±0.000314 ●	0.000100±0.000302 ●	0.000100±0.000302 ●	0.000120±0.000356
cleveland-14-heart-disease	0.001620±0.000508	0.000090±0.000288 ●	0.000120±0.000356 ●	0.000130±0.000338 ●	0.000190±0.000419 ●
hungarian-14-heart-disease	0.001330±0.000514	0.000100±0.000333 ●	0.000120±0.000327 ●	0.000140±0.000349 ●	0.000210±0.000433 ●
heart-statlog	0.001160±0.001529	0.000090±0.000321	0.000130±0.000367	0.000130±0.000338	0.000110±0.000314
hepatitis	0.000420±0.000496	0.000070±0.000256	0.000080±0.000273	0.000080±0.000273	0.000130±0.000338
labor	0.000020±0.000141	0.000030±0.000171	0.000050±0.000219	0.000070±0.000256	0.000030±0.000171
lymphography	0.000530±0.000502	0.000050±0.000219 ●	0.000150±0.000359	0.000110±0.000314 ●	0.000110±0.000314 ●
primary-tumor	0.004840±0.000507	0.000610±0.000510 ●	0.000840±0.000395 ●	0.001190±0.000394 ●	0.001910±0.000351 ●
sonar	0.002150±0.000592	0.000190±0.000394 ●	0.000150±0.000359 ●	0.000160±0.000368 ●	0.000200±0.000402 ●
vote	0.001150±0.000386	0.000130±0.000338 ●	0.000280±0.000451 ●	0.000190±0.000419 ●	0.000140±0.000349 ●
zoo	0.000110±0.000314	0.000100±0.000302	0.000090±0.000288	0.000080±0.000273	0.000050±0.000219

○, ● statistically significant improvement or degradation

2. The EkP Prototype Selection System

Probabilistic inference has become a core technology in AI, largely due to developments in graph-theoretic methods for the representation and manipulation of complex probability distributions (**?**). Whether in their guise as directed graphs (Bayesian networks) or as undirected graphs (Markov random fields), *probabilistic graphical models* have a number of virtues as representations of uncertainty and as inference engines. Graphical models allow a separation between qualitative, structural aspects of uncertain knowledge and the quan-

Table 3: C4.5 Traning Times

Data Set	(C4.5)	(EkP 1 p. + C4.5)	(EkP 2 p. + C4.5)	(EkP 3 p. + C4.5)	(EkP 5 p. + C4.5)
appendicitis	0.015640±0.031635	0.028230±0.013300	0.025830±0.016882	0.026140±0.010934	0.027720±0.017253
audiology	0.035650±0.021885	0.173020±0.035033○	0.302700±0.066129○	0.420760±0.049737○	0.675520±0.068266○
breast-cancer	0.007300±0.001778	0.035180±0.009258○	0.038910±0.017077○	0.039220±0.017919○	0.042770±0.017567○
horse-colic	0.025280±0.016475	0.050490±0.043090	0.049270±0.017802○	0.053860±0.021459○	0.057810±0.017877○
credit-rating	0.055570±0.024303	0.067090±0.015641	0.072970±0.021744	0.075400±0.024911	0.076330±0.022385
german-credit	0.156590±0.037080	0.090940±0.026585●	0.092570±0.023994●	0.094450±0.022156●	0.100900±0.024901●
pima-diabetes	0.022990±0.005014	0.073810±0.016248○	0.072910±0.023648○	0.073320±0.023395○	0.076200±0.019599○
Glass	0.022110±0.015199	0.051830±0.027061○	0.061560±0.017838○	0.073030±0.017754○	0.112370±0.026428○
cleveland-14-heart-diseas	0.019400±0.002978	0.048530±0.018656○	0.054700±0.020558○	0.060320±0.020345○	0.070750±0.020853○
hungarian-14-heart-diseas	0.015320±0.003698	0.047520±0.017090○	0.052090±0.017421○	0.056590±0.016558○	0.067070±0.019387○
heart-statlog	0.023710±0.009989	0.035930±0.016928	0.040850±0.017157○	0.041850±0.020043○	0.049210±0.022773○
hepatitis	0.008950±0.015827	0.028470±0.015590○	0.031370±0.018127○	0.033590±0.018969○	0.038750±0.016131○
labor	0.001300±0.000522	0.016980±0.006404○	0.019520±0.006001○	0.022860±0.015463○	0.026630±0.006823○
lymphography	0.004890±0.001109	0.033690±0.016263○	0.037950±0.017512○	0.046930±0.017909○	0.059380±0.023500○
primary-tumor	0.025110±0.002399	0.102940±0.024367○	0.149970±0.031797○	0.191960±0.042422○	0.275460±0.039655○
sonar	0.065740±0.024296	0.062660±0.035203	0.061190±0.017364	0.070130±0.023041	0.083240±0.021596
vote	0.005090±0.000933	0.049620±0.018788○	0.052270±0.016305○	0.051960±0.018185○	0.056550±0.023395○
zoo	0.002410±0.002948	0.032180±0.014811○	0.049470±0.023835○	0.062970±0.020332○	0.077360±0.020729○

○, ● statistically significant improvement or degradation

Table 4: C4.5 Testing Times

Data Set	(C4.5)	(EkP 1 p. + C4.5)	(EkP 2 p. + C4.5)	(EkP 3 p. + C4.5)	(EkP 5 p. + C4.5)
appendicitis	0.000450±0.002721	0.000020±0.000141	0.000020±0.000141	0.000020±0.000141	0.000030±0.000171
audiology	0.000750±0.003973	0.000130±0.000506	0.000150±0.000520	0.000200±0.000402	0.000290±0.000518
breast-cancer	0.000100±0.000333	0.000120±0.000327	0.000010±0.000100	0.000070±0.000256	0.000050±0.000219
horse-colic	0.000090±0.000288	0.000080±0.000273	0.000030±0.000171	0.000030±0.000171	0.000040±0.000197
credit-rating	0.000230±0.000423	0.000130±0.000367	0.000130±0.000338	0.000100±0.000302	0.000090±0.000288
german-credit	0.000350±0.000479	0.000110±0.000314	0.000140±0.000377	0.000200±0.000426	0.000200±0.000471
pima-diabetes	0.000150±0.000359	0.000120±0.000356	0.000110±0.000469	0.000070±0.000256	0.000110±0.000490
Glass	0.000050±0.000219	0.000040±0.000197	0.000080±0.000273	0.000050±0.000219	0.000010±0.000100
cleveland-14-heart-diseas	0.000160±0.000368	0.000070±0.000293	0.000040±0.000197	0.000080±0.000273	0.000030±0.000171
hungarian-14-heart-diseas	0.000090±0.000321	0.000040±0.000197	0.000100±0.000389	0.000050±0.000219	0.000090±0.000288
heart-statlog	0.000000±0.000000	0.000030±0.000171	0.000020±0.000141	0.000050±0.000219	0.000030±0.000171
hepatitis	0.000050±0.000219	0.000020±0.000141	0.000030±0.000171	0.000010±0.000100	0.000000±0.000000
labor	0.000000±0.000000	0.000000±0.000000	0.000060±0.000239	0.000000±0.000000	0.000000±0.000000
lymphography	0.000060±0.000239	0.000040±0.000197	0.000030±0.000171	0.000030±0.000171	0.000030±0.000171
primary-tumor	0.000290±0.000478	0.000210±0.000478	0.000250±0.000575	0.000170±0.000451	0.000170±0.000378
sonar	0.000060±0.000239	0.000030±0.000171	0.000010±0.000100	0.000020±0.000141	0.000050±0.000219
vote	0.000020±0.000141	0.000060±0.000239	0.000040±0.000197	0.000070±0.000256	0.000070±0.000256
zoo	0.000030±0.000171	0.000030±0.000171	0.000040±0.000197	0.000030±0.000171	0.000020±0.000141

○, ● statistically significant improvement or degradation

Table 5: NNGE Number of Rules

Data Set	(C4.5)	(EkP 1 p. + C4.5)		(EkP 2 p. + C4.5)		(EkP 3 p. + C4.5)		(EkP 5 p. + C4.5)	
appendicitis	15.99±2.17	1.95±0.22	●	2.02±0.14	●	2.27±0.55	●	2.86±0.96	●
audiology	46.83±3.70	9.97±1.69	●	17.23±2.25	●	23.38±2.76	●	33.94±2.97	●
breast-cancer	86.45±5.06	1.72±0.45	●	2.00±0.00	●	2.08±0.34	●	2.78±0.87	●
horse-colic	97.60±18.80	1.87±0.34	●	2.00±0.00	●	2.00±0.00	●	2.17±0.43	●
credit-rating	142.36±14.63	2.00±0.00	●	2.00±0.00	●	2.01±0.10	●	2.04±0.20	●
german-credit	347.38±26.03	1.06±0.24	●	1.78±0.42	●	1.96±0.37	●	2.81±0.79	●
pima-diabetes	263.80±23.18	1.82±0.39	●	2.00±0.00	●	2.08±0.27	●	2.72±0.82	●
Glass	48.10±4.80	3.91±0.74	●	6.04±1.01	●	8.14±1.26	●	11.92±1.54	●
cleveland-14-heart-diseas	71.61±9.29	2.00±0.00	●	2.88±0.70	●	4.02±0.95	●	6.06±1.30	●
hungarian-14-heart-diseas	61.93±7.20	2.02±0.14	●	2.62±0.75	●	3.58±1.06	●	5.90±1.51	●
heart-statlog	68.27±8.67	2.00±0.00	●	2.00±0.00	●	2.01±0.10	●	2.81±0.76	●
hepatitis	27.56±3.79	1.48±0.50	●	2.01±0.10	●	2.07±0.26	●	2.42±0.61	●
labor	7.91±1.43	1.99±0.10	●	2.00±0.00	●	2.07±0.29	●	2.62±0.72	●
lymphography	32.00±5.20	2.12±0.33	●	2.80±0.67	●	3.84±1.06	●	5.89±1.31	●
primary-tumor	147.78±4.72	13.26±1.73	●	24.37±2.47	●	35.81±2.98	●	58.19±3.91	●
sonar	45.70±7.17	2.00±0.00	●	2.00±0.00	●	2.02±0.14	●	2.40±0.57	●
vote	28.97±3.58	2.00±0.00	●	2.00±0.00	●	2.01±0.10	●	2.06±0.34	●
zoo	7.00±0.00	5.03±0.52	●	6.45±0.54	●	6.91±0.29		7.00±0.00	

○, ● statistically significant improvement or degradation

titative, parametric aspects of uncertainty...

Remainder omitted in this sample. See http://www.jmlr.org/papers/ for full paper.

Acknowledgments

Table 6: C4.5: Number of Rules

Data Set	(C4.5)	(EkP 1 p. + C4.5)		(EkP 2 p. + C4.5)		(EkP 3 p. + C4.5)		(EkP 5 p. + C4.5)	
appendicitis	3.12±0.62	1.00±0.00	●	1.75±0.44	●	1.97±0.17	●	2.02±0.14	●
audiology	19.84±1.73	6.23±0.83	●	9.65±1.22	●	12.08±1.33	●	16.39±1.51	●
breast-cancer	18.43±4.21	1.00±0.00	●	1.46±0.50	●	1.89±0.53	●	2.31±0.51	●
horse-colic	9.08±2.66	1.00±0.00	●	1.90±0.30	●	2.00±0.00	●	2.28±0.47	●
credit-rating	31.36±7.69	1.00±0.00	●	1.97±0.17	●	2.00±0.00	●	2.00±0.00	●
german-credit	69.69±5.81	1.00±0.00	●	1.07±0.26	●	1.15±0.36	●	1.32±0.63	●
pima-diabetes	7.51±1.53	1.00±0.00	●	1.81±0.39	●	1.95±0.22	●	2.02±0.14	●
Glass	15.25±1.56	2.64±0.48	●	3.53±0.59	●	4.54±0.95	●	6.56±1.29	●
cleveland-14-heart-diseas	19.64±2.70	2.00±0.00	●	2.10±0.30	●	2.58±0.70	●	3.86±0.78	●
hungarian-14-heart-diseas	8.21±2.42	2.00±0.00	●	2.11±0.31	●	2.35±0.61	●	3.33±1.21	●
heart-statlog	17.61±2.41	1.00±0.00	●	2.00±0.00	●	2.00±0.00	●	2.16±0.37	●
hepatitis	8.61±1.72	1.00±0.00	●	1.08±0.27	●	1.61±0.49	●	2.10±0.33	●
labor	3.37±0.76	1.00±0.00	●	1.98±0.14	●	2.00±0.00	●	2.08±0.27	●
lymphography	11.32±2.28	1.96±0.20	●	2.17±0.38	●	2.58±0.59	●	3.38±0.60	●
primary-tumor	41.09±3.53	6.13±0.95	●	10.40±1.33	●	13.76±1.91	●	20.38±2.30	●
sonar	7.54±0.96	1.00±0.00	●	2.00±0.00	●	2.00±0.00	●	2.00±0.00	●
vote	6.11±1.06	1.00±0.00	●	2.00±0.00	●	2.00±0.00	●	2.00±0.00	●
zoo	7.65±0.48	3.00±0.00	●	4.89±0.47	●	5.97±0.54	●	6.86±0.38	●

○, ● statistically significant improvement or degradation

We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

Appendix A.

In this appendix we prove the following theorem from Section 6.2:

Theorem *Let u, v, w be discrete variables such that v, w do not co-occur with u (i.e., $u \neq 0 \;\Rightarrow\; v = w = 0$ in a given dataset $\mathcal{D}$). Let N_{v0}, N_{w0} be the number of data points for which $v = 0, w = 0$ respectively, and let I_{uv}, I_{uw} be the respective empirical mutual information values based on the sample $\mathcal{D}$. Then*

$$N_{v0} \;>\; N_{w0} \;\;\Rightarrow\;\; I_{uv} \;\leq\; I_{uw}$$

with equality only if u is identically 0. ■

Proof. We use the notation:

$$P_v(i) \;=\; \frac{N_v^i}{N}, \quad i \neq 0; \quad P_{v0} \;\equiv\; P_v(0) \;=\; 1 - \sum_{i \neq 0} P_v(i).$$

These values represent the (empirical) probabilities of v taking value $i \neq 0$ and 0 respectively. Entropies will be denoted by H. We aim to show that $\frac{\partial I_{uv}}{\partial P_{v0}} < 0$....

Remainder omitted in this sample. See http://www.jmlr.org/papers/ for full paper.