INFORMATICA, 2014, Vol. 25, No. 3, 385–399

DOI: http://dx.doi.org/10.15388/Informatica.2014.19

Best-Fit Learning Curve Model for the C4.5 Algorithm

Boštjan BRUMEN, Ivan ROZMAN, Marjan HERIČKO,
Aleš ČERNEZEL*, Marko HÖLBL

University of Maribor, Faculty of Electrical Engineering, Computer Science and Informatics
Smetanova 17, SI-2000 Maribor, Slovenia
e-mail: ales.cernezel@uni-mb.si

Received: December 2012; accepted: February 2013

Abstract. *Background*: In the area of artificial learners, not much research on the question of an appropriate description of artificial learner's (empirical) performance has been conducted. The optimal solution of describing a learning problem would be a functional dependency between the data, the learning algorithm's internal specifics and its performance. Unfortunately, a general, restrictions-free theory on performance of arbitrary artificial learners has not been developed yet.

Objective: The objective of this paper is to investigate which function is most appropriately describing the learning curve produced by C4.5 algorithm.

Methods: The J48 implementation of the C4.5 algorithm was applied to datasets ($n = 121$) from publicly available repositories (e.g. UCI) in step wise k-fold cross-validation. First, four different functions (power, linear, logarithmic, exponential) were fit to the measured error rates. Where the fit was statistically significant ($n = 86$), we measured the average mean squared error rate for each function and its rank. The dependent samples T-test was performed to test whether the differences between mean squared error are significantly different, and Wilcoxon's signed rank test was used to test whether the differences between ranks are significant.

Results: The decision trees error rate can be successfully modeled by an exponential function. In a total of 86 datasets, exponential function was a better descriptor of error rate function in 64 of 86 cases, power was best in 13, logarithmic in 3, and linear in 6 out of 86 cases. Average mean squared error across all datasets was 0.052954 for exponential function, and was significantly different at $P = 0.001$ from power and at $P = 0.000$ from linear function. The results also show that exponential function's rank is significantly different at any reasonable threshold ($P = 0.000$) from the rank of any other model.

Conclusion: Our findings are consistent with tests performed in the area of human cognitive performance, e.g. with works by Heathcote *et al*. (2000), who were observing that the exponential function is best describing an individual learner. In our case we did observe an individual learner (C4.5 algorithm) at different tasks. The work can be used to forecast and model the future performance of C4.5 when not all data have been used or there is a need to obtain more data for better accuracy.

Key words: learning curve, learning process, classification, accuracy, assessment, data mining, C4.5, power law.

*Corresponding author.

1. Introduction

In the past, much research has been conducted on the question of a mathematical description of human cognitive performance. The power function is generally accepted as an appropriate description in psychophysics, in skill acquisition, and in retention. Power curves have been observed so frequently, and in such varied contexts, that the term "power law" is now commonplace (Anderson and Schooler, 1991; Anderson, 2001). However, several authors recently argued against the power law (Heathcote *et al.*, 2000), explaining it holds only on an aggregate level; on specific learner's level the exponential law is advantageous.

In the area of artificial learners, not much research on the question of an appropriate description of artificial learners has been conducted (Kotsiantis, 2007). The optimal solution of describing a learning problem would be a functional dependency between the data, the learning algorithm's internal specifics and its performance (e.g. accuracy). This way we could analytically determine the output (error rate) based on input (data, selected learner). Unfortunately, for real-life situations, such an approach does not yet exist. Standard numerical (and other statistical) methods become unstable when using large data sets (Dzemyda and Sakalauskas, 2011). Different theoretical approaches provide estimates for the size of the confidence interval on the training error under various settings of the problem of learning from examples. Vapnik–Chervonenkis theory (Vapnik, 1982) is the most comprehensive description of learning from examples. However, it has some limits (e.g. oracle is never wrong) that make it difficult for real-life implementations, as described in detail (Brumen *et al.*, 2007). The results for a specific learner and for specific type of data can be found in the literature (e.g. Dučinskas and Stabingiene, 2011), but no general analytical solution is available.

The other approach is to empirically measure the learner's performance on as many data sets as possible, and to draw conclusions based on statistical analysis of the results, as in this paper. However, not much research and publications are available for the description of C4.5's performance. Frey and Fisher have measured the decision tree performance and concluded that the results suggest that the error rate of a pruned decision tree generated by C4.5 can be predicted by a power law (Frey and Fisher, 1999). Several authors have either confirmed this or have been building on their results (Last, 2007; Provost *et al.*, 1999). But, a more recent research conducted by Singh has produced some evidence against the power law (Singh, 2005).

Paper contribution. The research question of this paper is the following: which law (power, exponential, linear, and logarithmic) is a better mathematical description of an artificial learner, specifically C4.5, over a larger number of available dataset? Our null hypotheses are as follows:

- The mean difference between function's $f_i()$ and function's $f_j()$ average mean squared error equals 0.
- The median of differences between function's $f_i()$ and function's $f_j()$ average rank equals 0.

Alternative hypotheses are that the mean/median of differences are different.

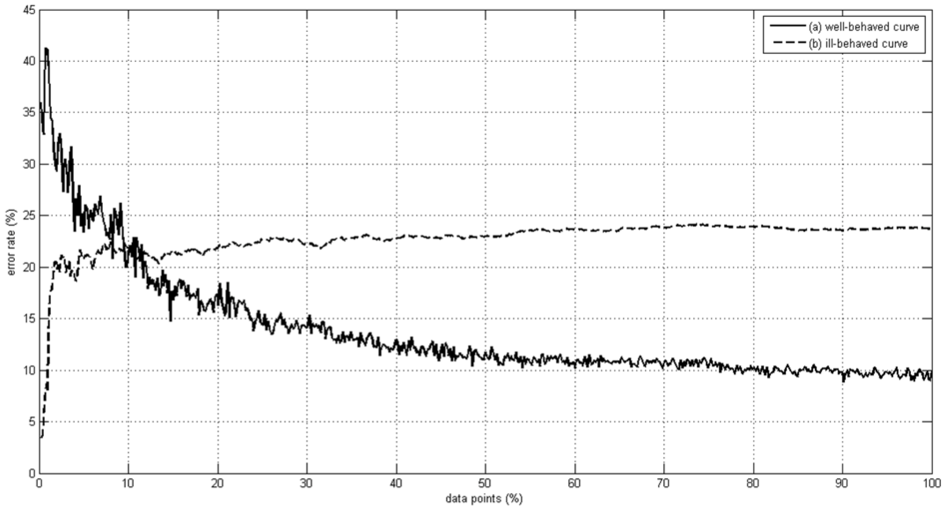

Fig. 1. Typical (a) well-behaved and (b) ill-behaved learning curves.

The main contribution of this paper is the answer to the question: "Which mathematical description is best for C4.5 artificial learner?"

Paper organization. We present the underlying and related work in Section 2. Here, we give an overview of the related work with the description of a learning curve and how to model it. In Section 3, we describe the scientific methods used in our experiment. In Section 4, we present and analyze the results. We conclude the paper with final remarks and comments in Section 5.

2. Related Work

When measuring the accuracy of a classifier (being it a human or a machine), we build a so-called learning curve. The learning curve depicts the relationship between sample size and classifier's performance (see Fig. 1). The horizontal axis represents the number of instances in a given training set. The vertical axis represents the performance (i.e. error rate) of the model produced by a classifying algorithm when given a training set of size l_i.

Learning curves typically have a steeply sloping portion early in the curve, a more gently sloping middle portion, and a plateau late in the curve. This resembles the way humans learn (Anderson and Schooler, 1991). For this reason the curve is called a learning curve. However, some datasets do not produce a typical well-behaved learning curve, where the error rate drops with increased sample size, as can be seen from Fig. 1. The ill-behaved learning curve was produced on IPUMS Census Database (ipums.la.98), and the well-behaved on Optical Recognition of Handwritten Digits dataset (optdigits).

John and Langley have used eleven datasets and a Naïve Bayes classifier, and observed how well the power law in the form of $y = c - a \cdot x^b$ fits the resulting error rate. They did not statistically asses the results (John and Langley, 1996).

Accuracy of a pruned decision tree, generated by C4.5 (Quinlan, 1993) was successfully modeled by the power law by Frey and Fisher (1999). They used the power function $y = a \cdot x^b$ and tried to fit parameters a and b. However, they used only fourteen datasets from the UCI repository.

Gu, Feifang and Liu have used only eight datasets and have observed that the power law in the form of $y = c - a \cdot x^b$ is the best fitting function for C4.5 (Gu *et al.*, 2001), although their work does not include any statistical assessment of the claims and the results.

On the other hand, Singh has observed that the power law is inferior to the logarithm function in three out of four datasets studied (Singh, 2005), and the result was depending more on a dataset than on the classification method (ID3, k-nearest neighbor, support vector machines and neural networks were used for classification). However, the author himself admits the findings are not conclusive and cannot be generalized. Again, no statistical methods were used to assess the results.

3. Method

3.1. *The Toolbox*

For the analytical tool of the experiment, we have chosen the J48 decision tree builder with standard built-in settings and initial values, which is freely available from the Waikato Environment for Knowledge Analysis (WEKA) project toolkit (Witten and Frank, 2005) version 3.6.8. J48 is java-based decision tree builder based on Quinlan's C4.5 tree induction (Quinlan, 1993).

The computer used was equipped with Windows-7 ($\times$64) operating system, an Intel i5-650 processor and 8 GB of DDR3 RAM.

For statistical analyses we used IBM SPSS version 21.

3.2. *Data Collection*

We used publicly available datasets from University of California at Irwine (UCI) Machine Learning Repository (Asuncion and Newman, 2010). We selected the datasets where the problem task is classification; the number of records in a dataset was larger than 200 and the number of instances exceeded the number of attributes (i.e. the task was classification, not feature selection).

The UCI repository contains datasets in ".data" and ".names" format while Weka's native format is ARFF. Therefore we used files available from various sources, such as TunedIT (2012), Håkan Kjellerstrand' weka page (Kjellerstrand, 2012a, 2012b) and Kevin Chai's page (Chai, 2012). We gathered 121 datasets, listed in Table 1.

We used only the original or larger datasets where several ones were available and ignored any separate training or test set, or any associated cost model.

Table 1
Datasets and respective P-values and R^2 values for goodness of fit of a model to data.

Dataset	*df*	POW P	POW R^2	LIN P	LIN R^2	LOG P	LOG R^2	EXP P	EXP R^2
ada_agnostic	449	**0.0000**	0.65	**0.0000**	0.32	**0.0000**	0.47	**0.0000**	0.63
ada_prior	449	**0.0000**	0.55	**0.0000**	0.20	**0.0000**	0.47	**0.0000**	0.67
analcatdata_authorship	77	**0.0000**	0.67	**0.0000**	0.49	**0.0000**	0.61	**0.0000**	0.58
analcatdata_braziltourism	34	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
analcatdata_broadwaymult	21	**0.0001**	0.52	**0.0001**	0.51	**0.0000**	0.52	**0.0000**	0.54
analcatdata_dmft	72	0.8480	0.00	1.0000	0.00	0.8496	0.00	1.0000	0.00
analcatdata_halloffame	126	**0.0000**	0.25	**0.0000**	0.23	**0.0000**	0.25	**0.0000**	0.24
analcatdata_marketing	29	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
analcatdata_reviewer	30	**0.0181**	0.16	0.3450	0.03	0.1309	0.07	**0.0086**	0.20
anneal	82	**0.0000**	0.81	**0.0000**	0.50	**0.0000**	0.72	**0.0000**	0.81
anneal.ORIG	82	**0.0000**	0.94	**0.0000**	0.79	**0.0000**	0.94	**0.0000**	0.93
audiology	15	**0.0000**	0.85	**0.0000**	0.65	**0.0000**	0.77	**0.0000**	0.86
australian	61	**0.0000**	0.47	**0.0000**	0.54	**0.0000**	0.49	**0.0000**	0.55
autos	13	**0.0000**	0.97	**0.0000**	0.87	**0.0000**	0.58	**0.0000**	0.97
badges_plain	22	**0.0000**	0.82	**0.0000**	0.79	**0.0000**	0.82	**0.0000**	0.83
balance-scale	55	0.0001	0.25	**0.0000**	0.25	0.0001	0.25	**0.0000**	0.29
baseball-hitter	25	**0.0000**	0.97	**0.0000**	0.63	**0.0000**	0.82	**0.0000**	0.96
baseball-pitcher	13	**0.0000**	0.98	**0.0000**	0.79	**0.0000**	0.91	**0.0000**	0.98
BC	21	0.1638	0.08	0.9646	0.00	0.7444	0.00	0.9629	0.00
Billionaires92	16	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
biomed	13	1.0000	0.00	1.0000	0.00	N/A	0.00	0.2872	0.08
breast-cancer	21	1.0000	0.00	N/A	0.00	1.0000	0.00	1.0000	0.00
breast-w	62	**0.0000**	0.24	**0.0018**	0.14	**0.0001**	0.21	**0.0000**	0.28
car	165	**0.0000**	0.82	**0.0000**	0.79	**0.0000**	0.88	**0.0000**	0.90
cars_with_names	33	**0.0002**	0.33	**0.0000**	0.49	**0.0001**	0.35	**0.0000**	0.48
CH	312	**0.0000**	0.68	**0.0000**	0.67	**0.0000**	0.77	**0.0000**	0.78
cmc	140	**0.0000**	0.23	**0.0007**	0.08	**0.0000**	0.15	**0.0000**	0.19
colic	29	**0.0059**	0.22	**0.0004**	0.34	**0.0044**	0.23	**0.0004**	0.33
colic.ORIG	29	**0.0001**	0.39	0.0619	0.11	**0.0031**	0.25	**0.0000**	0.46
cps_85_wages	46	**0.0007**	0.21	**0.0032**	0.17	**0.0008**	0.21	**0.0003**	0.24
credit-a	61	**0.0013**	0.15	**0.0004**	0.18	**0.0011**	0.16	**0.0004**	0.18
credit-g	92	**0.0000**	0.65	**0.0000**	0.35	**0.0000**	0.55	**0.0000**	0.65
credit	41	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
csb_ch12	153	**0.0000**	0.72	**0.0000**	0.63	**0.0000**	0.74	**0.0000**	0.75
csb_ch9	316	**0.0000**	0.78	**0.0000**	0.62	**0.0000**	0.78	**0.0000**	0.79
cylinder-bands	46	**0.0273**	0.10	0.0516	0.08	**0.0274**	0.10	**0.0187**	0.11
db3-bf	39	**0.0000**	0.67	**0.0000**	0.77	**0.0000**	0.69	**0.0000**	0.77
dermatology	29	**0.0000**	0.91	**0.0000**	0.71	**0.0000**	0.87	**0.0000**	0.92
diabetes	69	**0.0000**	0.80	**0.0000**	0.68	**0.0000**	0.80	**0.0000**	0.83
ecoli	26	**0.0010**	0.32	**0.0014**	0.31	**0.0010**	0.33	**0.0008**	0.34
eucalyptus	66	**0.0000**	0.54	**0.0000**	0.36	**0.0000**	0.49	**0.0000**	0.57
eye_movements	1086	**0.0000**	0.88	**0.0000**	0.89	**0.0000**	0.91	**0.0000**	0.92
genresTrain	1242	**0.0000**	0.94	**0.0000**	0.66	**0.0000**	0.92	**0.0000**	0.85
gina_agnostic	339	**0.0000**	0.89	**0.0000**	0.73	**0.0000**	0.90	**0.0000**	0.90
gina_prior	339	**0.0000**	0.91	**0.0000**	0.61	**0.0000**	0.86	**0.0000**	0.84
gina_prior2	339	**0.0000**	0.95	**0.0000**	0.76	**0.0000**	0.96	**0.0000**	0.94
GL	14	**0.0150**	0.32	**0.0426**	0.23	**0.0265**	0.27	**0.0131**	0.33
glass	14	**0.0000**	0.83	**0.0000**	0.73	**0.0000**	0.81	**0.0000**	0.85
haberman	23	0.8837	0.00	0.6991	0.01	0.8828	0.00	0.7016	0.01
HD	23	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
heart-c	23	**0.0004**	0.40	0.1747	0.07	**0.0282**	0.18	**0.0001**	0.48

(*continued to next page*)

Table 1
(*Continued.*)

Dataset	*df*	POW *P*	POW R^2	LIN *P*	LIN R^2	LOG *P*	LOG R^2	EXP P	EXP R^2
heart-h	22	0.0000	0.76	0.0008	0.38	0.0000	0.57	0.0000	0.81
heart-statlog	19	1.0000	0.00	1.0000	0.00	N/A	0.00	1.0000	0.00
HO	29	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
HY	309	**0.0000**	0.80	**0.0000**	0.40	**0.0000**	0.72	**0.0000**	0.85
hypothyroid	370	**0.0000**	0.77	**0.0000**	0.40	**0.0000**	0.70	**0.0000**	0.81
ionosphere	28	0.6776	0.01	0.1650	0.06	0.6539	0.01	0.1864	0.06
irish	42	1.0000	0.00	1.0000	0.00	0.9419	0.00	**0.0481**	-0.01
jEdit_4.0_4.2	20	**0.0000**	0.64	**0.0000**	0.63	**0.0000**	0.64	**0.0000**	0.65
jEdit_4.2_4.3	29	1.0000	0.00	1.0000	0.00	1.0000	0.00	N/A	0.00
jm1	1081	1.0000	-0.01	1.0000	0.00	**0.0000**	0.02	**0.0000**	0.13
kc1	203	**0.0000**	0.28	**0.0000**	0.11	**0.0000**	0.23	**0.0000**	0.31
kc2	45	1.0000	0.00	0.7396	0.00	1.0000	0.00	0.7462	0.00
kc3	38	**0.0000**	0.79	**0.0000**	0.35	**0.0000**	0.58	**0.0000**	0.81
kdd_ipums_la_97-small	694	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
kdd_ipums_la_98-small	741	1.0000	0.00	0.9995	0.00	1.0000	0.00	0.9990	0.00
kdd_ipums_la_99-small	877	**0.0000**	0.81	**0.0000**	0.38	**0.0000**	0.65	**0.0000**	0.68
kdd_synthetic_control	52	**0.0000**	0.92	**0.0000**	0.78	**0.0000**	0.92	**0.0000**	0.92
kr-vs-kp	312	**0.0000**	0.94	**0.0000**	0.49	**0.0000**	0.82	**0.0000**	0.89
kropt	2798	**0.0000**	0.91	**0.0000**	0.91	**0.0000**	0.85	**0.0000**	0.96
landsat	636	**0.0000**	0.78	**0.0000**	0.65	**0.0000**	0.79	**0.0000**	0.77
letter	1992	**0.0000**	0.98	**0.0000**	0.64	**0.0000**	0.96	**0.0000**	0.90
liver-disorders	27	**0.0000**	0.51	**0.0000**	0.48	**0.0000**	0.51	**0.0000**	0.50
mc1	939	1.0000	0.00	**0.0000**	0.06	1.0000	0.00	1.0000	0.00
mfeat-factors	192	**0.0000**	0.89	**0.0000**	0.62	**0.0000**	0.84	**0.0000**	0.82
mfeat-fourier	192	**0.0000**	0.91	**0.0000**	0.66	**0.0000**	0.89	**0.0000**	0.89
mfeat-karhunen	192	**0.0000**	0.95	**0.0000**	0.76	**0.0000**	0.95	**0.0000**	0.94
mfeat-morphological	192	**0.0000**	0.41	**0.0000**	0.26	**0.0000**	0.37	**0.0000**	0.43
mfeat-pixel	192	**0.0000**	0.84	**0.0000**	0.71	**0.0000**	0.86	**0.0000**	0.88
mfeat-zernike	192	**0.0000**	0.87	**0.0000**	0.59	**0.0000**	0.82	**0.0000**	0.83
monks-problems-1_test	36	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
monks-problems-2_test	36	**0.0034**	0.20	**0.0045**	0.19	**0.0035**	0.20	**0.0046**	0.19
monks-problems-3_test	36	1.0000	0.00	1.0000	0.00	1.0000	0.00	N/A	0.00
mozilla4	1547	**0.0000**	0.79	**0.0000**	0.53	**0.0000**	0.79	**0.0000**	0.83
MU	805	**0.0000**	0.53	**0.0000**	0.25	**0.0000**	0.46	**0.0000**	0.47
mushroom	805	**0.0000**	0.80	**0.0000**	0.22	**0.0000**	0.58	**0.0000**	0.90
mw1	33	1.0000	0.00	1.0000	0.00	0.7752	0.00	1.0000	0.00
nursery	1288	**0.0000**	0.84	**0.0000**	0.73	**0.0000**	0.88	**0.0000**	0.84
optdigits	554	**0.0000**	0.95	**0.0000**	0.64	**0.0000**	0.94	**0.0000**	0.96
page-blocks	540	**0.0000**	0.76	**0.0000**	0.32	**0.0000**	0.63	**0.0000**	0.84
pc1	103	**0.0000**	0.20	**0.0000**	0.33	**0.0000**	0.21	**0.0000**	0.32
pc3	149	1.0000	0.00	1.0000	0.00	1.0000	0.00	**0.0000**	0.17
pc4	138	**0.0000**	0.31	**0.0000**	0.45	**0.0000**	0.33	**0.0000**	0.45
pendigits	1092	**0.0000**	0.96	**0.0000**	0.52	**0.0000**	0.89	**0.0000**	0.93
primary-tumor	26	**0.0001**	0.45	0.2000	0.06	**0.0240**	0.17	**0.0000**	0.57
prnn_fglass	14	**0.0001**	0.63	**0.0373**	0.24	**0.0059**	0.39	**0.0000**	0.86
prnn_synth	17	**0.0019**	0.40	**0.0016**	0.41	**0.0018**	0.41	**0.0017**	0.41
rmftsa_propores	21	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
schizo	26	**0.0000**	0.53	**0.0000**	0.62	**0.0000**	0.57	**0.0000**	0.61
scopes-bf	55	**0.0000**	0.90	**0.0000**	0.79	**0.0000**	0.91	**0.0000**	0.92
SE	309	**0.0000**	0.57	**0.0000**	0.10	**0.0000**	0.31	**0.0000**	0.60
segment	223	**0.0000**	0.82	**0.0000**	0.72	**0.0000**	0.88	**0.0000**	0.90

(*continued to next page*)

Table 1
(*Continued.*)

Dataset	*df*	POW *P*	POW R^2	LIN *P*	LIN R^2	LOG *P*	LOG R^2	EXP P	EXP R^2
sick	370	**0.0000**	0.20	**0.0000**	0.13	**0.0000**	0.14	**0.0000**	0.24
sonar	13	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
soybean	61	**0.0000**	0.83	**0.0000**	0.55	**0.0000**	0.77	**0.0000**	0.88
spambase	453	**0.0000**	0.83	**0.0000**	0.26	**0.0000**	0.48	**0.0000**	0.81
splice	311	**0.0000**	0.90	**0.0000**	0.59	**0.0000**	0.88	**0.0000**	0.94
sylva_agnostic	1432	**0.0000**	0.85	**0.0000**	0.30	**0.0000**	0.65	**0.0000**	0.76
sylva_prior	1432	**0.0000**	0.73	**0.0000**	0.36	**0.0000**	0.60	**0.0000**	0.59
tic-tac-toe	88	**0.0000**	0.81	**0.0000**	0.69	**0.0000**	0.82	**0.0000**	0.83
ticdata_categ	575	1.0000	0.00	**0.0072**	0.01	1.0000	0.00	1.0000	0.00
titanic	213	**0.0000**	0.44	**0.0000**	0.53	**0.0000**	0.45	**0.0000**	0.32
train	492	**0.0000**	0.59	**0.0000**	0.58	**0.0000**	0.60	**0.0000**	0.63
usp05	13	1.0000	0.00	1.0000	0.00	1.0000	0.00	1.0000	0.00
V1	36	**0.0013**	0.24	**0.0005**	0.27	**0.0011**	0.25	**0.0006**	0.27
vehicle	77	**0.0000**	0.84	**0.0000**	0.51	**0.0000**	0.72	**0.0000**	0.83
visualizing_fly	75	**0.0000**	0.69	**0.0000**	0.51	**0.0000**	0.67	**0.0000**	0.70
VO	36	1.0000	0.00	1.0000	0.00	1.0000	0.00	N/A	0.00
vote	36	0.0580	0.09	**0.0054**	0.19	**0.0459**	0.10	**0.0071**	0.18
vowel	91	**0.0000**	0.87	**0.0000**	0.89	**0.0000**	0.82	**0.0000**	0.95
waveform-5000	492	**0.0000**	0.72	**0.0000**	0.32	**0.0000**	0.60	**0.0000**	0.69

3.3. *Data Pre-Processing*

We followed the following steps for obtaining the error rate curve (i.e. learning curve):

1. Data items in a data set are randomly shuffled;
2. First, $n_{i=1} = 50$ items are chosen;
3. Build decision trees using k-fold cross-validation on sample size of n_i (Cohen, 1995; Weiss and Kulikowski, 1991); k was set to 10 (Cohen, 1995; McLachlan *et al.*, 2004; Provost *et al.*, 1999; Weiss and Kulikowski, 1991);
4. Measure the error rate for each tree in 10-fold run and average the result over 10 runs;
5. Store the pair (n_i = sample size, e_i = error);
6. The number of items in a data set is increased by 10; $n_{i+1} := n_i + 10$;
7. Repeat steps 3–6 until all data items in a dataset are used.

3.4. *Fitting Curve Model to Measured Data*

The next step in our research was to fit a model to the error rate curves. We used four different functions, as in Eqs. (1)–(4):

$$\text{power (POW):} \quad f(x) = p_1 + p_2 x^{p_3}, \tag{1}$$

$$\text{linear (LIN):} \quad f(x) = p_1 + p_2 x, \tag{2}$$

$$\text{logarithm (LOG):} \quad f(x) = p_1 + p_2 \log x, \tag{3}$$

$$\text{exponential (EXP):} \quad f(x) = p_1 + p_2 e^{p_3 x}. \tag{4}$$

The functions do not have the same number of parameters (p_i). They all include the constant p_1 and coefficient p_2, in addition to potent p_3 for power and exponential function. Based on the specifics of the problem and the speed of convergence we limited the parameters to the following intervals:

- p_1 to interval [0, 1] (error rate cannot be less than 0 and more than 1);
- p_2 to interval [0, 100] for power function and to $[-100, 0]$ for the others, and
- p_3 to interval $[-100, 0]$ (error rate is decreasing hence p_3 needs to be negative).

We used the open-source GNU Octave software (Eaton, 2012) and the built-in Levenberg–Marquardt's algorithm (Levenberg, 1944; Marquardt, 1963), also known as the damped least-squares (DLS) method, for fitting the function parameters.

The inputs to the algorithm were vector x (sample sizes n), vector y (error rates e), initial values of parameters p_i ($[0.01; 1; -0.1]$ for POW, $[0.1; -0.001]$ for LIN, $[0.1; -0.01]$ for LOG and $[0.01; 0.1; -0.01]$ for EXP), function to be fit to vectors x, y (power, linear, logarithm, or exponential), partial derivatives of functions with respect to parameters p_i, and limits of parameters p_i (as described above).

The algorithm's output were vector of functional values of fitted function for input x, vector of parameters p_i, where minimum mean squared error was obtained, and a flag whether the convergence was reached or not.

4. Results

For each dataset we tested the claim that the samples can be modeled by the probability density functions POW, LIN, LOG and EXP, respectively. We used the Pearson's chi-squared test (χ^2), also known as the chi-squared goodness-of-fit test or chi-squared test for independence, where the null hypothesis was H_0: $r_\mu = 0$ or there is no correlation between the population and model (Argyrous, 2011), at $\alpha = 0.05$. Table 1 lists the results: the values in bold are P values indicating that the null hypothesis is rejected, the number of degrees of freedom (df), and the coefficient of determination R^2 (indication how well a regression line fits a set of data). N/A indicates that the model was not calculated because the Levenberg–Marquardt's algorithm suggested a constant model and hence χ^2 cannot be computed.

It can be observed that out of 121 datasets, 86 are such that all the models can be used to describe the data. The remaining datasets are such that C4.5 does not capture their internal relations and cannot be used for classification, so we eliminated those from our further study. From the vector of fitted function's values (f) and from the vector y we calculated the mean squared error (MSE) of jth dataset (DS), using Eq. (5):

$$\mathrm{MSE}_{DS_j} = \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n} \tag{5}$$

where n is the number of input points, i.e. the size of a vector, for each individual data set DS_j. MSE describes how well the observed points fit to the modeled function. The average

Table 2
Datasets and the average MSE across function models, and the model's rank.

Dataset	POW avg. (MSE)	POW rank	LIN avg. (MSE)	LIN rank	LOG avg. (MSE)	LOG rank	EXP avg. (MSE)	EXP rank
ada_agnostic	0.000285	2	0.000551	4	0.000421	3	0.000213	1
ada_prior	0.000340	2	0.000402	4	0.000347	3	0.000296	1
analcatdata_authorship	0.000337	2	0.000609	4	0.000489	3	0.000249	1
analcatdata_broadwaymult	0.001161	4	0.001098	2	0.001152	3	0.001088	1
analcatdata_halloffame	0.000609	4	0.000477	3	0.000460	2	0.000289	1
anneal	0.000338	3	0.000421	4	0.000325	2	0.000282	1
anneal.ORIG	0.000361	1	0.000772	4	0.000437	3	0.000369	2
audiology	0.000636	2	0.000952	4	0.000831	3	0.000526	1
australian	0.000656	4	0.000621	1	0.000650	3	0.000621	2
autos	0.000631	2	0.002281	3	0.009721	4	0.000535	1
badges_plain	0.000106	2	0.000116	4	0.000115	3	0.000079	1
balance-scale	0.000986	4	0.000937	2	0.000975	3	0.000936	1
baseball-hitter	0.000125	1	0.000850	4	0.000554	3	0.000126	2
baseball-pitcher	0.000087	2	0.000734	4	0.000681	3	0.000086	1
breast-w	0.001039	4	0.001014	3	0.001006	2	0.000966	1
car	0.000955	4	0.000866	3	0.000836	2	0.000731	1
cars_with_names	0.000998	4	0.000921	1	0.000991	3	0.000927	2
CH	0.000102	4	0.000100	3	0.000095	2	0.000093	1
cmc	0.001263	1	0.001434	4	0.001410	3	0.001266	2
colic	0.000431	4	0.000412	1	0.000429	3	0.000413	2
cps_85_wages	0.002090	4	0.002054	2	0.002083	3	0.002029	1
credit-a	0.000760	4	0.000748	1	0.000759	2	0.000759	3
credit-g	0.000527	2	0.000784	4	0.000602	3	0.000517	1
csb_ch12	0.000181	4	0.000102	3	0.000097	2	0.000081	1
csb_ch9	0.000991	2	0.001109	4	0.001022	3	0.000953	1
db3-bf	0.001726	3	0.001543	2	0.001748	4	0.001531	1
dermatology	0.000489	3	0.000630	4	0.000436	2	0.000377	1
diabetes	0.001178	3	0.001215	4	0.001141	2	0.001074	1
ecoli	0.000371	4	0.000355	2	0.000365	3	0.000353	1
eucalyptus	0.001709	2	0.001949	4	0.001772	3	0.001501	1
eye_movements	0.000402	2	0.000696	4	0.000389	1	0.000420	3
genresTrain	0.000360	1	0.001105	4	0.000476	2	0.000516	3
gina_agnostic	0.000378	3	0.000505	4	0.000366	2	0.000346	1
gina_prior	0.000421	1	0.001017	4	0.000523	3	0.000435	2
gina_prior2	0.000423	3	0.000975	4	0.000407	1	0.000416	2
GL	0.001062	2	0.001180	3	0.001196	4	0.000869	1
glass	0.001148	2	0.001171	3	0.001259	4	0.001024	1
heart-h	0.000935	2	0.001407	4	0.001120	3	0.000814	1
HY	0.000081	3	0.000106	4	0.000076	2	0.000068	1
hypothyroid	0.000052	3	0.000068	4	0.000050	2	0.000044	1
jEdit_4.0_4.2	0.000503	4	0.000473	2	0.000494	3	0.000470	1
kc1	0.000214	2	0.000227	4	0.000216	3	0.000210	1
kc3	0.000222	2	0.000506	4	0.000327	3	0.000218	1
kdd_ipums_la_99-small	0.000075	1	0.000190	4	0.000124	3	0.000078	2
kdd_synthetic_control	0.000560	3	0.000868	4	0.000556	2	0.000503	1
kropt	0.000275	2	0.000611	4	0.000448	3	0.000254	1
kr-vs-kp	0.000072	1	0.000248	4	0.000112	3	0.000078	2
landsat	0.000424	4	0.000384	3	0.000346	2	0.000272	1
letter	0.000263	1	0.003118	4	0.000942	3	0.000860	2
liver-disorders	0.003184	2	0.003269	4	0.003200	3	0.003122	1

(*continued to next page*)

Table 2
(*Continued.*)

Dataset	POW avg. (MSE)	POW rank	LIN avg. (MSE)	LIN rank	LOG avg. (MSE)	LOG rank	EXP avg. (MSE)	EXP rank
mfeat-factors	0.000526	1	0.001425	4	0.000773	3	0.000555	2
mfeat-fourier	0.000896	2	0.001974	4	0.001306	3	0.000844	1
mfeat-karhunen	0.000493	1	0.001331	4	0.000586	3	0.000568	2
mfeat-morphological	0.000499	2	0.000579	4	0.000526	3	0.000462	1
mfeat-pixel	0.000973	3	0.001009	4	0.000941	2	0.000808	1
mfeat-zernike	0.000702	1	0.001384	4	0.000871	3	0.000770	2
monks-problems-2_test	0.002180	2	0.002419	3	0.002421	4	0.002085	1
mozilla4	0.000054	2	0.000062	3	0.000052	1	0.000096	4
MU	0.000080	4	0.000037	3	0.000032	2	0.000019	1
mushroom	0.000040	3	0.000071	4	0.000039	2	0.000023	1
nursery	0.000270	2	0.000405	4	0.000280	3	0.000256	1
optdigits	0.000313	3	0.000864	4	0.000264	2	0.000245	1
page-blocks	0.000111	2	0.000174	4	0.000117	3	0.000083	1
pc1	0.000141	4	0.000130	2	0.000134	3	0.000127	1
pc4	0.000329	3	0.000308	1	0.000327	2	0.000466	4
pendigits	0.000187	1	0.001048	4	0.000255	3	0.000214	2
prnn_fglass	0.002491	2	0.005687	3	0.007976	4	0.001045	1
prnn_synth	0.001522	4	0.001113	3	0.001085	2	0.000803	1
schizo	0.005476	4	0.005141	1	0.005368	3	0.005170	2
scopes-bf	0.000388	3	0.000437	4	0.000364	2	0.000354	1
SE	0.000161	4	0.000086	3	0.000070	2	0.000056	1
segment	0.000214	4	0.000194	3	0.000173	2	0.000136	1
sick	0.000064	4	0.000051	3	0.000050	2	0.000041	1
soybean	0.001903	3	0.002359	4	0.001796	2	0.001364	1
spambase	0.000392	2	0.000797	4	0.000510	3	0.000185	1
splice	0.000544	3	0.000806	4	0.000459	2	0.000396	1
sylva_agnostic	0.000033	3	0.000051	4	0.000029	2	0.000022	1
sylva_prior	0.000049	3	0.000065	4	0.000048	2	0.000035	1
tic-tac-toe	0.001221	3	0.001293	4	0.001183	2	0.001162	1
titanic	0.000087	2	0.000136	4	0.000122	3	0.000078	1
train	0.000335	2	0.000335	3	0.000338	4	0.000302	1
V1	0.000438	4	0.000384	2	0.000385	3	0.000381	1
vehicle	0.001023	2	0.001948	4	0.001409	3	0.000918	1
visualizing_fly	0.000660	2	0.000793	4	0.000734	3	0.000636	1
vowel	0.002071	3	0.001684	2	0.002442	4	0.001411	1
waveform-5000	0.000543	1	0.000801	4	0.000599	3	0.000554	2
Average (MSE)	*0.000697*		*0.000948*		*0.000914*		*0.000616*	
Rank-sum		225		290		230		115

MSEs for each dataset are listed in Table 2, together with the rank of function's model. The model with lowest average MSE gets assigned rank 1. It can be seen that EXP is best fit in 64 of 86 cases, POW is best in 13 out of 86 times, LOG in 3 out of 86 cases, and LIN in 6 out of 86 cases. Average MSEs across all datasets were 0.052954 for EXP, 0.059927 for POW, 0.078639 for LOG, and 0.081561 for LIN.

As can be observed, the EXP had rank-sum of 115 and an average MSE of 0.000616. Please note that the rank is an ordinal value and hence calculating its mean value is inappropriate (Argyrous, 2011, p. 472).

Table 3
Paired samples T-test for average MSE.

Paired samples test	Paired differences					t	df	Sig. (2-tailed)
	Mean	Std. deviation	Std. error mean	95% confidence interval of the difference				
				Lower	Upper			
Pair 1 POW–LIN	−2.515E−004	5.455E−004	5.882E−005	−3.685E−004	−1.346E−004	−4.27	85	**0.000**
Pair 2 POW–LOG	−2.175E−004	1.141E−003	1.231E−004	−4.624E−004	2.725E−005	−1.76	85	0.081
Pair 3 POW–EXP	8.107E−005	2.131E−004	2.298E−005	3.538E−005	1.267E−004	3.528	85	**0.001**
Pair 4 LIN–LOG	3.397E−005	9.154E−004	9.871E−005	−1.622E−004	2.302E−004	0.344	85	0.732
Pair 5 LIN–EXP	3.326E−004	6.076E−004	6.552E−005	2.023E−004	4.629E−004	5.077	85	**0.000**
Pair 6 LOG–EXP	2.986E−004	1.228E−003	1.324E−004	3.532E−005	5.619E−004	2.255	85	0.027

Finally, the main research question was tested: which model was best? To rephrase, was EXP with the rank-sum of 115 and average MSE of 0.000616 significantly better than second-best POW with rank-sum of 225 and average MSE of 0.000697?

To test the significance of difference in MSE we used paired samples T-test for all combinations of models. The null hypotheses, the mean of differences between f_i (MSE) and f_j (MSE) equals 0, were as follows: $H1_0 : \mu_{MSE/power} = \mu_{MSE/linear}$; $H2_0 : \mu_{MSE/power} = \mu_{MSE/logarithmic}$; $H3_0 : \mu_{MSE/power} = \mu_{MSE/exponential}$; $H4_0 : \mu_{MSE/linear} = \mu_{MSE/logarithm}$; $H5_0 : \mu_{MSE/linear} = \mu_{MSE/exponential}$; and $H6_0 : \mu_{MSE/logaritmic} = \mu_{MSE/exponential}$. Because we conducted 6 comparisons, we used the Bonferroni correction to counteract the problem of multiple comparisons (Abdi, 2007). The correction is based on the idea that if an experimenter is testing n dependent or independent hypotheses on a set of data, then one way of maintaining the family-wise error rate is to test each individual hypothesis at a statistical significance level of $1/n$ times what it would be if only one hypothesis were tested. We would normally reject the null hypothesis if $P < 0.05$. However, Bonferroni correction requires a modified rejection threshold for P, $\alpha = (0.05/6) = 0.008$. Table 3 lists the results of statistical analysis for all six comparisons, with values in bold indicating significance.

The results show that exponential function's average mean squared error is significantly different at any reasonable threshold from average MSE power ($P = 0.001$) and linear function ($P = 0.000$), regardless if using the Bonferroni correction or not. Average MSE of exponential function is significantly different from the one of logarithmic function at $\alpha = 0.05$ level, but not at $\alpha = 0.008$ level.

Additionally, we tested whether the ranks of functions are statistically significantly different from each other. We used related samples Wilcoxon signed rank test. The null

Table 4
Wilcoxon signed rank test for different function models.

Pair #	Pair	Sig. (2-tailed)
Pair 1	POW (rank)–LIN (rank)	0.016
Pair 2	POW (rank)–LOG (rank)	0.478
Pair 3	POW (rank)–EXP (rank)	0.000
Pair 4	LIN (rank)–LOG (rank)	0.010
Pair 5	LIN (rank)–EXP (rank)	0.000
Pair 6	LOG (rank)–EXP (rank)	0.000

hypotheses, the median of differences between f_i $(rank)$ and f_j $(rank)$ equals 0, were as follows:

$$H1_0 : \mu_{1/2\text{RANK/power}} = \mu_{1/2\text{RANK/linear}},$$

$$H2_0 : \mu_{1/2\text{RANK/power}} = \mu_{1/2\text{RANK/logarithmic}},$$

$$H3_0 : \mu_{1/2\text{RANK/power}} = \mu_{1/2\text{RANK/exponential}},$$

$$H4_0 : \mu_{1/2\text{RANK/linear}} = \mu_{1/2\text{RANK/logarithm}},$$

$$H5_0 : \mu_{1/2\text{RANK/linear}} = \mu_{1/2\text{RANK/exponential}} \quad \text{and}$$

$$H6_0 : \mu_{1/2\text{RANK/logarithmic}} = \mu_{1/2\text{RANK/exponential}}.$$

Table 4 lists the results of Wilcoxon signed rank test analysis for all six comparisons, with values in bold indicating significance at $\alpha = 0.008$ level.

The results show that exponential function's average rank is significantly different at any reasonable threshold from average rank of any other model ($P = 0.000$).

5. Conclusion

In this paper we presented that the exponential function is most appropriately describing the learning curve produced by C4.5 algorithm, in both average mean squared error and in rank assignment in comparison to logarithmic, power and linear function.

The results are somehow inconsistent with findings of John and Langley (1996), Frey and Fisher (1999) and Gu *et al*. (2001). However, neither of researchers used a substantial number of datasets (in most cases, 14), nor did they perform any statistical tests.

Our findings are consistent with tests performed in the area of human cognitive performance, e.g. with works by Heathcote et al., who were observing that the exponential function is best describing an individual learner (Heathcote *et al*., 2000), and the power law can be observed only at the generalization level. In our case we observed an individual learner (C4.5) at different tasks.

The findings of the present work can be used to forecast and model the future performance of C4.5 when not all data have been used or there is a need to obtain more data for better accuracy. Furthermore, the results show that some datasets exist where modeling of the artificial learner's performance, specifically J48 implementation of C4.5, is not successful due to inability of learner to properly capture the data interrelations. This, however, could be detected early in the learning process.

Our future work will focus on the question whether the averaged performance over several different tasks can be averaged and modeled with power law. Additionally, an early detection of ill-behaved learning curves would also contribute to better performance of artificial learners.

Acknowledgements. This work was partially supported by the Slovenian Research Agency under grant number 1000-11-310138.

References

Abdi, H. (2007). The Bonferonni and Šidák corrections for multiple comparisons. In: Salkind, N.J. (Ed.), *Encyclopedia of Measurement and Statistics*. SAGE Publications, Inc., Thousand Oaks.

Anderson, R.B. (2001). The power law as an emergent property. *Memory & Cognition*, 29(7), 1061–1068.

Anderson, J.R., Schooler, L.J. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6), 396–408.

Argyrous, G. (2011). *Statistics for Research: With a Guide to SPSS*. 3rd ed., SAGE Publications Ltd., Thousand Oaks.

Asuncion, A., Newman, D. (2010). UCI machine learning repository, from `http://archive.ics.uci.edu/ml/datasets.html`. Accessed: 2012-11-09. (Archived by WebCite® at `http://www.webcitation.org/6C2hgsRrX`.)

Brumen, B., Jurič, M.B., Welzer, T., Rozman, I., Jaakkola, H., Papadopoulos, A. (2007). Assessment of classification models with small amounts of data. *Informatica*, 18(3), 343–362.

Chai, K. (2012). Kevin Chai datasets, from `http://kevinchai.net/datasets`. Accessed: 2012-12-12. (Archived by WebCite® at `http://www.webcitation.org/6CqqWlQEp`.)

Cohen, P.R. (1995). *Empirical Methods for Artificial Intelligence*, Vol. 55. MIT Press, Cambridge.

Dučinskas, K., Stabingiene, L. (2011). Expected Bayes error rate in supervised classification of spatial Gaussian data. *Informatica*, 22(3), 371–381.

Dzemyda, G., Sakalauskas, L. (2011). Large-scale data analysis using heuristic methods. *Informatica*, 22(1), 1–10.

Eaton, J.W. (2012). GNU octave, from `http://www.gnu.org/software/octave/`. Accessed: 2012-12-12. (Archived by WebCite® at `http://www.webcitation.org/6CqyEvDKU`.)

Frey, L.J., Fisher, D.H. (1999). Modeling decision tree performance with the power law. In: *The Seventh International Workshop on Artificial Intelligence and Statistics*, San Francisco.

Gu, B., Hu, F., Liu, H. (2001). Modelling classification performance for large data sets. In: Wang, X.S., Yu, G., Lu, H. (Eds.), *Advances in Web-Age Information Management*, Vol. 2118. Springer, Berlin/Heidelberg, pp. 317–328

Heathcote, A., Brown, S., Mewhort, D.J.K. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7(2), 185–207. doi:10.3758/bf03212979.

John, G.H., Langley, P. (1996, August 2–4). Static versus dynamic sampling for data mining. In: *The Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA.

Kjellerstrand, H. (2012a). My Weka page, from `http://www.hakank.org/weka/`. Accessed: 2012-12-12. (Archived by WebCite® at `http://www.webcitation.org/6Cqq5pQtZ`.)

Kjellerstrand, H. (2012b). My Weka page/DASL, from `http://www.hakank.org/weka/DASL/`. Accessed: 2012-12-12. (Archived by WebCite® at `http://www.webcitation.org/6CqqCwPmy`.)

Kotsiantis, S.B. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, 31(3), 249–268.

Last, M. (2007, 28–31 Oct.). Predicting and optimizing classifier utility with the power law. In: *The 7th IEEE International Conference on Data Mining*, ICDM Workshops 2007, Omaha, Nebraska, USA.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2, 164–168.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 431–441. doi:10.2307/2098941.

McLachlan, G.J., Do, K.-A., Ambroise, C. (2004). Analyzing Microarray Gene Expression Data. Wiley, Hoboken.

Provost, F., Jensen, D., Oates, T. (1999). Efficient progressive sampling. In: *The Fifth International Conference on Knowledge Discovery and Data Mining*, San Diego.

Quinlan, R.J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, USA. ISBN:1-55860-238-0.

Singh, S. (2005). *Modeling Performance of Different Classification Methods: Deviation from the Power Law*. Project report, Vanderbilt University, Department of Computer Science, Nashville, Tennessee, USA.

TunedIT (2012). TunedIT research repository, from `http://tunedit.org/search?q=arff&qt=Repository`. Accessed: 2012-12-12. (Archived by WebCite® at `http://www.webcitation.org/6CqplN6Xr`.)

Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer, Berlin.

Weiss, S. M., Kulikowski, C.A. (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo.

Witten, I.H., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed., Morgan Kaufmann, San Mateo. ISBN:0120884070.

B. Brumen received doctor's degree in informatics in 2004. He is an assistant professor at University of Maribor, Faculty of Electrical engineering and computer science. He was Secretary General (Provost) of University of Maribor for two consecutive terms between 2004 and 2011. His research interests include intelligent data analysis, automated learning and learning models, data security and data quality.

I. Rozman is a professor at Faculty of Electrical Engineering and Computer Science, University of Maribor. He graduated in 1977 (electrical engineering), received master's degree (computer science) in 1980 and doctor's degree (computer science) in 1983. His research interests include integration of information systems, software metrics and systems quality. He served two consecutive terms as a rector of University of Maribor.

M. Hölbl is an assistant professor and researcher at the Data Technology Laboratory, Faculty of Electrical Engineering and Computer Science, University of Maribor. His research interests include computer and information security, cryptography, eLearning and databases.

M. Heričko is a professor at the Faculty of electrical Engineering and Computer Science, University of Maribor. He received his PhD in 1998 from the University of Maribor. His research interests include all aspects of information system development with emphasis on design patterns, service-oriented architectures and software metrics.

A. Černezel is a researcher and a PhD student at the Faculty of Electrical Engineering and Computer Science, University of Maribor. His research interests include mobile technologies, user centric design and machine learning.

Geriausiai atitinkančios C4.5 algoritmo mokymosi kreivės modelis

Boštjan BRUMEN, Ivan ROZMAN, Marjan HERIČKO, Aleš ČERNEZEL,
Marko HÖLBL

Dirbtinio besimokančiojo (empirinio) našumo aprašymo klausimai kol kas tyrinėti gana nedaug. Mokymosi problemą geriausiai aprašytų duomenų, mokymosi algoritmo specifinių savybių ir to algoritmo našumo funkcinės priklausomybės. Deja, kol kas bendra, be ribojimų ir tinkanti bet kokio tipo dirbtiniam besimokančiajam našumo teorija dar nesukurta. Šis straipsnis siekia nustatyti, kokia funkcija geriausiai tinka C4.5 algoritmo mokymosi kreivei aprašyti. Atliekant tyrimą, C4.5 algoritmo J48 realizacija žingsnis po žingsnio buvo bandoma viešosios prieigos duomenų saugyklose (pvz. UCI) patalpintiems 121 duomenų rinkiniui keliais būdais kryžmiškai vertinti. Pirmiausiai buvo tikrinta kaip keturios skirtingos funkcijos (laipsninė, tiesinė, logaritminė, ekspnentinė) atitinka santikinių paklaidų matavimams. Tiems atvejams, kuriems atitikimas buvo statistiškai reikšmingas (86 atvejai), kiekvienai funkcijai ir jos rangui buvo matuojama vidutinė kvadratinė santikinė paklaida. Priklausomų imčių T-testas buvo naudotas patikrinti, ar vidutinės kvadratinės paklaidos yra esminiai skirtingos, ir Wilxon rango su ženklu testas buvo naudotas patikrinti, ar rangų skirtumai yra esminiai. Pasirodė, jog sprendimų medžio santikines paklaidas geriausiai modeliuoja eksponentinė funkcija. Ši funkcija geriausiai aprašė santikines paklaidas 64, laipsninė funkcija 13, logaritminė 3 ir tiesinė 6 iš 86 vertinamų duomenų rinkinių. Vidutinė kvadratinė paklaida eksponentinei funkcijai visiems duomenų rinkiniams buvo $P = 0.052954$ ir esminei skyrėsi nuo laipsninės ($P = 0.001$) ir tiesinės ($P = 0.000$) funkcijų. Gauti rezultatai taip pat parodo, kad eksponentinės funkcijos rangas bet kuriam prasmingam slenksčiui esminiai skiriasi ($P = 0.000$) nuo bet kurio kito modelio rango. Gauti rezultatai gerai dera su rezultatais, gautais vertinant žmonių kognityvinį našumą, pavyzdžiui su Heathcote ir jo bendraautorių rezultatais parodančiais, jog eksponentinė funkcija geriausiai aprašo individualų besimokantįjį. Mes stebėjome individualų besimokantįjį (t.y. C4.5 algoritmą) skirtingoms užduotims. Gautieji rezultatai gali būti panaudoti prognozuoti ir modeliuoti C4.5 algoritmo būsimą našumą tais atvejais, kuomet yra panaudojami ne visi duomenys arba kuomet didesniam tikslumui pasiekti reikia daugiau duomenų.