A Comparison of Efficiency and Robustness of ID3 and C4.5 Algorithms Using Dynamic Test and Training Data Sets

Payam Emami Khoonsari and AhmadReza Motie

Abstract—**In the machine learning world making a decision is very important. Several approaches have been invented for doing so. Among the most efficient ones is the decision tree. ID3 and C4.5 algorithms have been introduced by J.R Quinlan which produce reasonable decision trees. In this paper we evaluate robustness of these algorithms against the training and test data set changes. At first an introduction has been presented, in the second part, we take a look at the algorithms and finally unique experimentations and findings are submitted.**

Index Terms—**ID3 algorithm, C4.5 algorithm, ID3 and C4.5 comparison, robustness of ID3 and C4.5, an empirical comparison of ID3 and C4.5.**

I. Introduction of the Decision Trees

The decision trees which have been known as classification trees are used perfectly in machine learning and data mining. The reasons for using such trees are:

- Easy to implement.
- Easy to comprehend.
- Don't need preparation methods like normalization.
- This structure works on both numerical and categorical data and works well with huge databases.

There are numerous algorithms for creating such trees; two of the popular ones are ID3 [1] and C4.5 [2] by J.R Quinlan.

II. ID3 vs. C4.5

ID3 algorithm selects the best attribute based on the concept of entropy [3] [4] and information gain [5] [6] [7] for developing the tree.

C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors:

- A possibility to use continuous data.
- Using unknown (missing) values which have been marked by "?".
- Possibility to use attributes with different weights.
- Pruning the tree after being created.

III. Experimentations and Comparison of the Two Algorithms

In this section we use nine data sets [8] in ascending order (Table I).

We use two approaches to evaluate the algorithms:

A. Constant Sets

In the first method we hold number of test set members constant and decrease number of training set members in a way training set members decline 1/12 (rounded off) of total number of the data set members in each step and until number of the training set members has not reached less than 1/3 (rounded off) of total number of the data set members and after each step we calculate the error rate (charts 1 through 9).

B. Dynamic Sets

In this approach we repeat the same process but we do not freeze the test sets, we instead increase the test set members by 1/12 (rounded off) of total number of the data set members in each step and until number of the training set members has not reached less than 1/3 (rounded off) of total number of the data set members and After each step we calculate the error rate (Charts 9 through 18).

At the end of all steps we evaluate difference of the most and the least error rates for each set (charts 19 and 20).

The error rate and the instability of the classifications correctness in each of the two methods are thoroughly simulated under various conditions (the training sets and the test sets).All of the selection process was performed randomly using a computer program that we developed for this purpose, which led us to some interesting results as shown in the charts 1 through 20.

IV. Conclusion

The final results as shown in each set (charts 1 through 18) and comparison of difference of the most and the least rate for the two methods (charts 19 and 20) point to the fact that robustness and accuracy of the C4.5 exceeds that of ID3.

TABLE I: Data Sets Information.

Name	Number of Instances	Number of Attributes	Missing Attribute	Type
adult + stretch	40	4	None	Categorical
Hayes Roth	132	4	None	Categorical
Monk1	556	7	None	Categorical
Monk2	601	7	None	Categorical
Balance Scale	625	4	None	Categorical
Car	1798	6	None	Categorical
Chess	3196	36	None	Categorical
nursery	12960	8	None	Categorical
connect-4	67557	42	None	Categorical

Manuscript received July 29, 2012; revised September 15, 2012.

Payam Emami Khoonsari is a master student in bioinformatics, University of Tampere, Finland (e-mail: payam.emamy@ gmail.com).

AhmadReza Motie is a lecturer in Jahad daneshgahi institute of higher education Esfahan, Iran (e-mail: motiee@acecr.ac.ir).

Chart 1.The error rate in adult data set, number of test set members is three and is fixed.

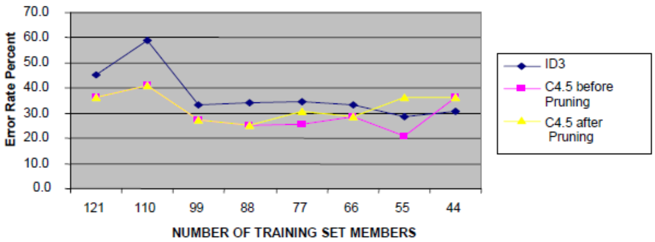

Chart 2.The error rate in Hayes data set, number of test set members is 11 and is fixed.

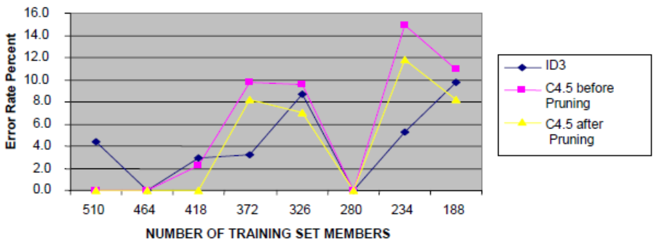

Chart 3.The error rate in Monk1 data set, number of test set members is 46 and is fixed.

Chart 4.The error rate in Monk2 data set, number of test set members is 50 and is fixed.

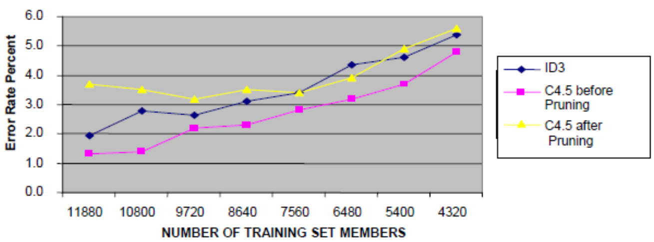

Chart 5.The error rate in Nursery data set, number of test set members is 1080 and is fixed.

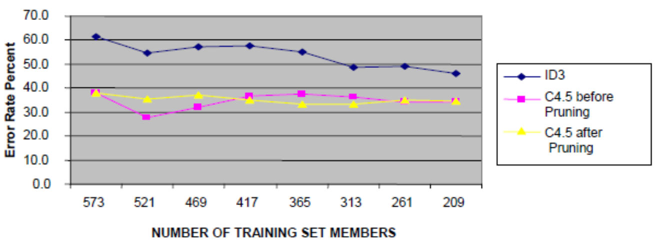

Chart 6.The error rate in Balance data set, number of test set members is 52 and is fixed.

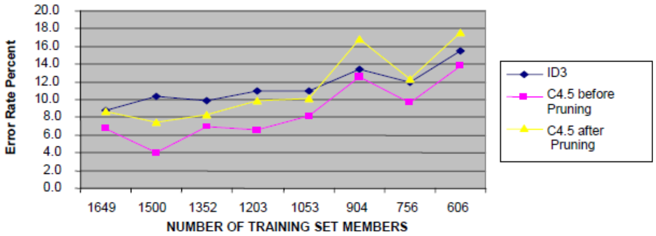

Chart 7.The error rate in Car data set, number of test set members is 149 and is fixed.

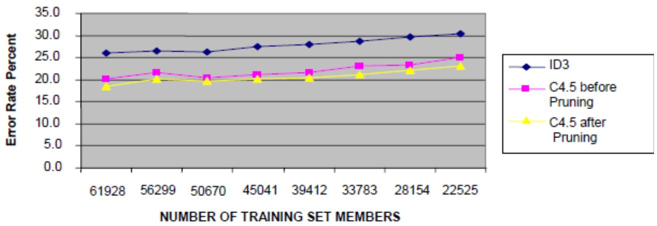

Chart 8.The error rate in Connect4 data set, number of test set members is 5629 and is fixed.

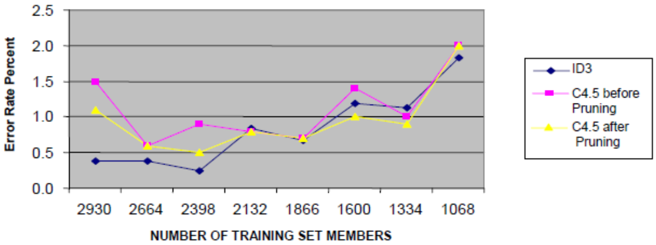

Chart 9.The error rate in Chess data set, number of test members set is 266 and is fixed.

Chart 10.The error rate in Adult data set, number of test set members is according to table (2) and is dynamic.

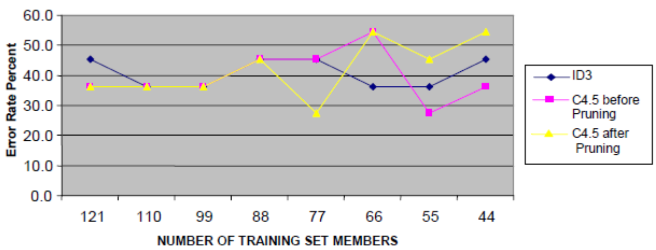

Chart 11.The error rate in Hayes data set, number of test set members is according to table (3) and is dynamic.

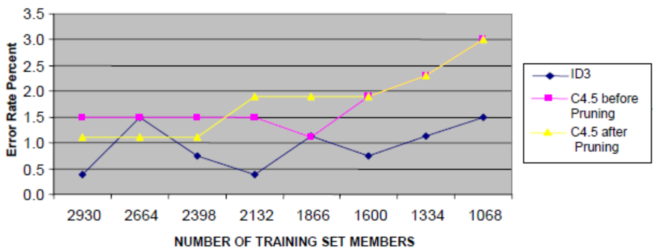

Chart 12.The error rate in Chess data set, number of test set members is according to table (4) and is dynamic.

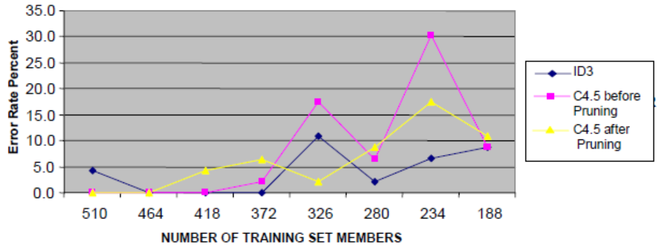

Chart 13.The error rate in Monk1 data set, number of test set members is according to table (5) and is dynamic.

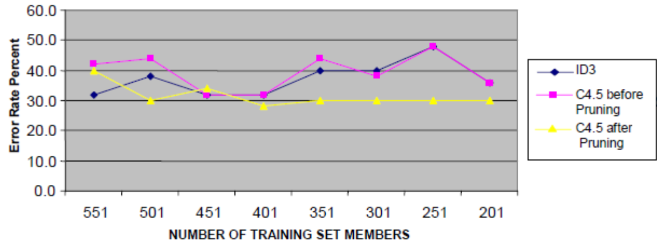

Chart 14.The error rate in Monk2 data set, number of test set members is according to table (6) and is dynamic.

Chart 15.The error rate in Balance data set, number of test set members is according to table (7) and is dynamic.

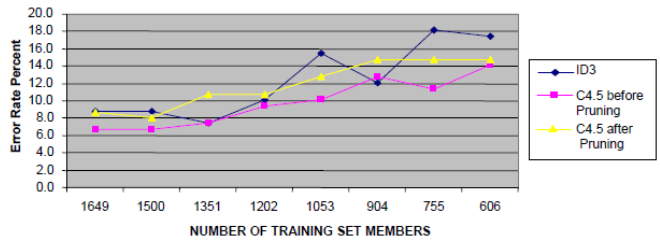

Chart 16.The error rate in Car data set, number of test set members is according to table (8) and is dynamic.

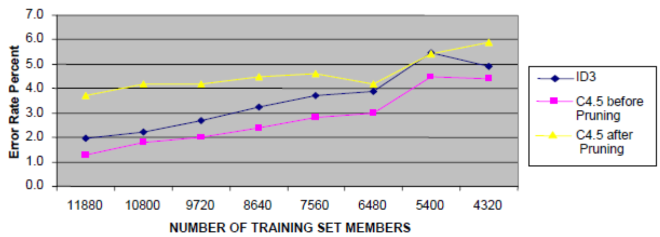

Chart 17.The error rate in Nursery data set, number of test set members is according to table (9) and is dynamic.

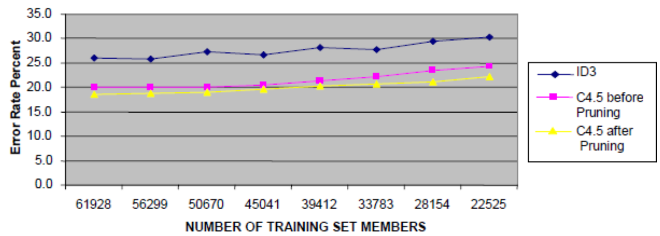

Chart 18.The error rate in Connect 4 data set, number of test set members is according to table (10) and is dynamic.

TABLE II: NUMBER OF TRAINING AND TEST SET MEMBERS IN ADULT DATA SET WHEN THE TEST SET IS DYNAMIC.

Training set	Test set
37	3
34	6
31	9
28	12
25	15
22	18
19	21
16	24
13	27
10	30

TABLE III: NUMBER OF RAINING AND TEST SET MEMBERS IN HAYES DATA SET WHEN THE TEST SET IS DYNAMIC.

Training set	Test set
121	11
110	22
99	33
88	44
77	55
66	66
55	77
44	88

TABLE IV: NUMBER OF TRAINING AND TEST SET MEMBERS IN CHESS DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
2930	266
2664	532
2398	798
2132	1064
1866	1330
1600	1596
1334	1862
1068	2128

TABLE V: NUMBER OF TRAINING AND TEST SET MEMBERS IN MONK1 DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
510	46
464	92
418	138
372	184
326	230
280	276
234	322
188	368

TABLE VI: NUMBER OF TRAINING AND TEST SET MEMBERS IN MONK2 DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
551	50
501	100
451	150
401	200
351	250
301	300
251	350
201	400

TABLE VII: NUMBER OF TRAINING AND TEST SET MEMBERS IN BALANCE DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
573	52
521	104
469	156
417	208
365	260
313	312
261	364
209	416

TABLE VIII: NUMBER OF TRAINING AND TEST SET MEMBERS IN CAR DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
1649	149
1500	298
1352	446
1203	595
1053	745
904	894
756	1042
606	1192

TABLE IX: NUMBER OF TRAINING AND TEST SET MEMBERS IN NURSERY DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
11880	1080
10800	2160
9720	3240
8640	4320
7560	5400
6480	6480
5400	7560
4320	8640

TABLE X: NUMBER OF TRAINING AND TEST SET MEMBERS IN CONNCEC4 DATA SET WHEN THE TEST SET MEMBERS ARE DYNAMIC.

Training set	Test set
61928	5629
56299	11258
50670	16887
45041	22516
39412	28145
33783	33774
28154	39403
22525	45032

REFERENCES

[1] Quinlan, J. R., Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), pp.81-106.

[2] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[3] Robert B. Ash. Information Theory. New York: Interscience, 1965.

[4] Raymond W. Yeung. *Information Theory and Network Coding Springer* 2008, 2002

[5] Kullback, S.; Leibler, R.A. (1951). "On Information and Sufficiency". *Annals of Mathematical Statistics* 22 (1): pp.79–86.

[6] S. Kullback (1959) Information theory and statistics (John Wiley and Sons, NY).

[7] Kullback, S. (1987). "Letter to the Editor: The Kullback–Leibler distance". *The American Statistician* 41 (4): pp.340-341.

[8] Blake C, Merz C. UCI repository of machine learning databases. University of California, Irvine, Department of Information and ComputerSciences,http://www.ics.uci.edu/|mlearn/MLRepository.html,1998.

Payam Emami Khoonsari was born on 09.05.1986 in Esfahan, Iran. He earned a bachelor's degree in computer science in 2010 at Jahad Daneshgahi institute of higher education, Esfahan, Iran. egrees should be listed with type of degree in what field, which institution, city, state or country, and year degree was earned. The author's major field of study should be lower-cased. He is currently studying MSc. in bioinformatics in university of Tampere in Finland. Mr. Emami is a member of International Association of Computer Science and Information Technology (IACSIT) and also a member bioinformatics society in Finland.