International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

Polynomial Curve Fitting of Execution Time of Binary Search in Worst Case in Personal Computer

Dipankar Das
Assistant Professor, The Heritage Academy, Kolkata, India

Abstract—Curve fitting is a well known method of data mining. This method can be used to identify the hidden patterns of any data set and thus may lead us to knowledge discovery. The present study fit the polynomial curves to the execution time (*in nano-seconds*) of binary search in worst case *versus* data size in a personal computer. Data size varies from one thousand (*1000*) to twenty thousand (*20000*) with an interval of five hundred (*500*). For each data size one thousand (*1000*) observations have been collected and to discard the outliers from the observations, two-step clustering algorithm have been employed. The mean value of the largest cluster for each data size gives us the mean execution time (*in nano-seconds*) for the respective data size. Polynomial curve fitting has been employed on the data points and candidate models are identified based on the values of Adjusted R-Squared, Residual Standard Error and Root Mean Square Error. Two (*2*) separate information criteria – AIC and BIC have been used to find out the best polynomial curves that fit the data points. It has been observed that polynomial of degree 24 comes out to be the best curve using AIC and polynomial of degree 22 becomes best curve using BIC.

Keywords—Curve fitting, AIC, BIC, Two-Step clustering, Binary search

I. INTRODUCTION

Binary search algorithm is a very common searching algorithm in computer science. Literature reviews has shown that analysis of binary search is an interesting research question amongst the researchers. In this present work, the researchers have tried to identify the best fit polynomial curve which can be fitted to the data points (Execution Time *versus* Data Size). Curve fitting is the process of constructing a curve that has the best fit to a series of data points, possibly subject to constraints [18]. Curve fitting is considered as a classical statistical method for data mining [19]. The present work only focuses on polynomial curve fitting leaving aside other popular curves *e.g.* Fourier, Gaussian and Exponential.

In this paper, the researchers have used *Akaike information criterion (AIC) and Bayesian information criterion (BIC) for identifying the best model.* AIC was introduced by Hirotugu Akaike [20]. AIC is one of the most widely used and known model selection tools [21]. The BIC was developed by Gideon Schwarz [22]. The model which has lowest BIC value is considered as the best model [8]. BIC tends to favor smaller models than AIC [23].

II. RELATED WORK

Binary Search algorithm had been analyzed in different ways some of which are as follows:

- Comparative analysis between binary search and linear search [9],[10],[14],[15].
- Comparative analysis between linear search, binary search and interpolation search [12].
- Comparative analysis between the performances of binary search in the worst case on two personal computers [16].
- Visualization and analysis of performance of binary search in the worst case on a personal computer [17].
- Modified binary search algorithm had been proposed [13].
  Randomized searching algorithm had been proposed whose performance lies between binary search and linear search [11].

III. OBJECTIVES OF THE STUDY

- To find out the best polynomial curve that can be fitted to the data points (Execution Time *versus* Data Size) generated by simulating Binary Search algorithm in worst case.
- To find the equation of the best model that can be fitted to the data points (Execution Time *versus* Data Size) generated by simulating Binary Search algorithm in worst case.
- To visualize the best model (polynomial curve) that can be fitted to the data points (Execution Time *versus* Data Size) generated by simulating Binary Search algorithm in worst case.

IV. METHODOLOGY

Step1. Generating the experimental dataset by running the java program of Binary Search algorithm (worst case scenario) on Linux (Ubuntu 12.04.4 LTS) operating system and OpenJDK (Java version 1.6.0_36). The researchers had run the java program of Binary Search algorithm for data size one thousand *(1000)* to twenty thousand (*20000*) with an interval of five hundred (*500*). Here, we had taken sorted array and the binary search had been conducted on the sorted array. In total, execution times of thirty nine (*39*) data sizes had been calculated. For each data size one thousand (*1000*) execution times had been noted. All the execution times had been observed in nano-seconds.

Step2. Using Two-Step clustering algorithm to identify the largest cluster for each data size and finding the mean execution time for each data size. In this research work, Log – Likelihood distance measure had been used and Schwarz's Bayesian Criterion (BIC) had been used as clustering criterion.

Step3. Using curve fitting techniques to identify the best polynomial curve(s) that can be fitted to the data set. In this research work, we had started our testing with polynomial of degree 2 and step by step increased the testing with higher order polynomials (*e.g. polynomial of degree 3, degree 4, degree 5, …*) until we had achieved a predefined high Adjusted R-Squared value (*0.95 in this case*). Initially, the following Goodness of fit statistics of all the models was noted: (i) Adjusted R-Squared, (ii) Residual Standard Error and (iii) Root Mean Square Error. The models which had high Adjusted R-Squared [1][2] along with low Residual Standard Error [3][4] and low Root Mean Square Error [1][2] were identified as the candidate models.

Step4. Calculating the F-statistics of the candidate models. In this research work, we had set the alpha level equal to 0.05. The decision rule is that - if the significance (p value) of the F-test is less than the alpha level then it can be concluded that the model is significant [5][6].

Step5. Calculating the *Akaike information criterion (AIC*) and *Bayesian information criterion (BIC*) values of the candidate models. We had used two different criteria (*AIC and BIC*) to select the best models. The model which has lowest AIC value is considered as the best model [7][8] using AIC criterion. Again, the model which has lowest BIC value is considered as the best model [8] using BIC criterion.

Step6. Plotting the data points (Execution Time *versus* Data Size) along with the (i) best fit model using AIC criterion and (ii) best fit model using BIC criterion.

Step7. Noting the equations of the best fit models using (i) AIC criterion and (ii) BIC criterion.

Hardware used for data generation: Intel(R) Core(TM)2 Duo CPU T5870 @2.00 GHz.
Software used for data analysis and curve fitting: SPSS 17.0 and R version 3.1.0.

V. DATA ANALYSIS & FINDINGS

Table 1. Goodness of fit statistics of all the tested models

Model	Adjusted R-Squared	Residual Standard Error	Root Mean Square Error
Polynomial of degree 2	0.2323	80.84	77.66965
Polynomial of degree 3	0.339	75.01	71.0626
Polynomial of degree 4	0.452	68.3	63.77326
Polynomial of degree 5	0.4645	67.52	62.10741
Polynomial of degree 6	0.4528	68.25	61.82269
Polynomial of degree 7	0.4411	68.98	61.49838
Polynomial of degree 8	0.4235	70.05	61.44088
Polynomial of degree 9	0.4041	71.22	61.41523
Polynomial of degree 10	0.3902	72.05	61.04962
Polynomial of degree 11	0.3852	72.35	60.19632
Polynomial of degree 12	0.3966	71.67	58.52035
Polynomial of degree 13	0.4486	68.51	54.85442
Polynomial of degree 14	0.5228	63.74	50.00013
Polynomial of degree 15	0.6134	57.37	44.05823
Polynomial of degree 16	0.7552	45.65	34.28406
Polynomial of degree 17	0.8472	36.07	26.46891
Polynomial of degree 18	0.8745	32.69	23.40817
Polynomial of degree 19	0.8965	29.69	20.72287
Polynomial of degree 20	0.8964	29.69	20.17264
Polynomial of degree 21	0.9211	25.91	17.10838
# Polynomial of degree 22	0.9502	20.59	13.18653
# Polynomial of degree 23	0.9506	20.5	12.71308
# Polynomial of degree 24	0.9525	20.1	12.0436

Candidate models

Table 2. F – Statistics of the candidate models

Model	F – statistics	p – value
Polynomial of degree 22	33.97	1.353e-09
Polynomial of degree 23	32.82	4.758e-09
Polynomial of degree 24	32.77	1.343e-08

Findings: The p-values in all the three (*3*) cases are much lower than 0.05. Therefore, we may conclude that all the three (*3*) models are significant.

Table 3. AIC value of the candidate models

Model	AIC value
Polynomial of degree 22	359.8545
Polynomial of degree 23	359.0025
Polynomial of degree 24	356.7828

Findings: Among the three (*3*) candidate models the last model *i.e.* the Polynomial of degree 24 model is having less AIC value. Therefore this model is chosen as the best model using AIC criterion.

Table 4. BIC value of the candidate models

Model	BIC value
Polynomial of degree 22	399.7799
Polynomial of degree 23	400.5915
Polynomial of degree 24	400.0354

Findings: Among the three (*3*) candidate models the first model *i.e.* the Polynomial of degree 22 model is having less BIC value. Therefore this model is chosen as the best model using BIC criterion.

The Plot of the best fit model using AIC criterion is given below:

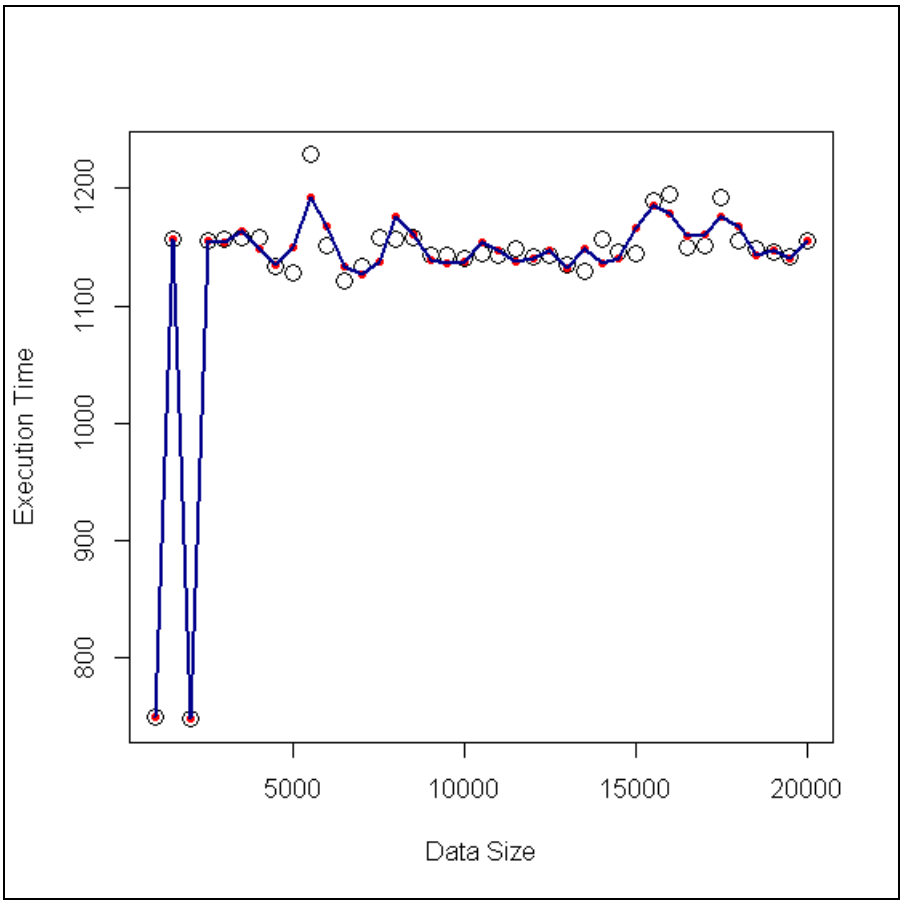

Figure 1. Polynomial of degree 24 model for Binary Search worst case

In the above figure the black circles represent the observed data points, the red dots represent the predicted points obtained by using the model equation and the dark blue line represents the curve of polynomial of degree 24.

Equation of the proposed model:

y = 1131.722 + 212.698*x - 207.315*x^2 + 195.767*x^3 - 195.785*x^4 + 90.434*x^5 - 37.096*x^6 + 39.494*x^7 + 16.604*x^8 + 11.087*x^9 - 41.787*x^10 + 63.521*x^11 - 88.089*x^12 + 127.316*x^13 - 140.893*x^14 + 147.637*x^15 - 172.811*x^16 + 136.077*x^17 - 77.161*x^18 + 67.983*x^19 - 29.624*x^20 + 66.748*x^21 - 68.071*x^22 + 21.868*x^23 + 25.424*x^24

Where, y is Execution time in nano-seconds and x is Data size.

The Plot of the best fit model using BIC criterion is given below:

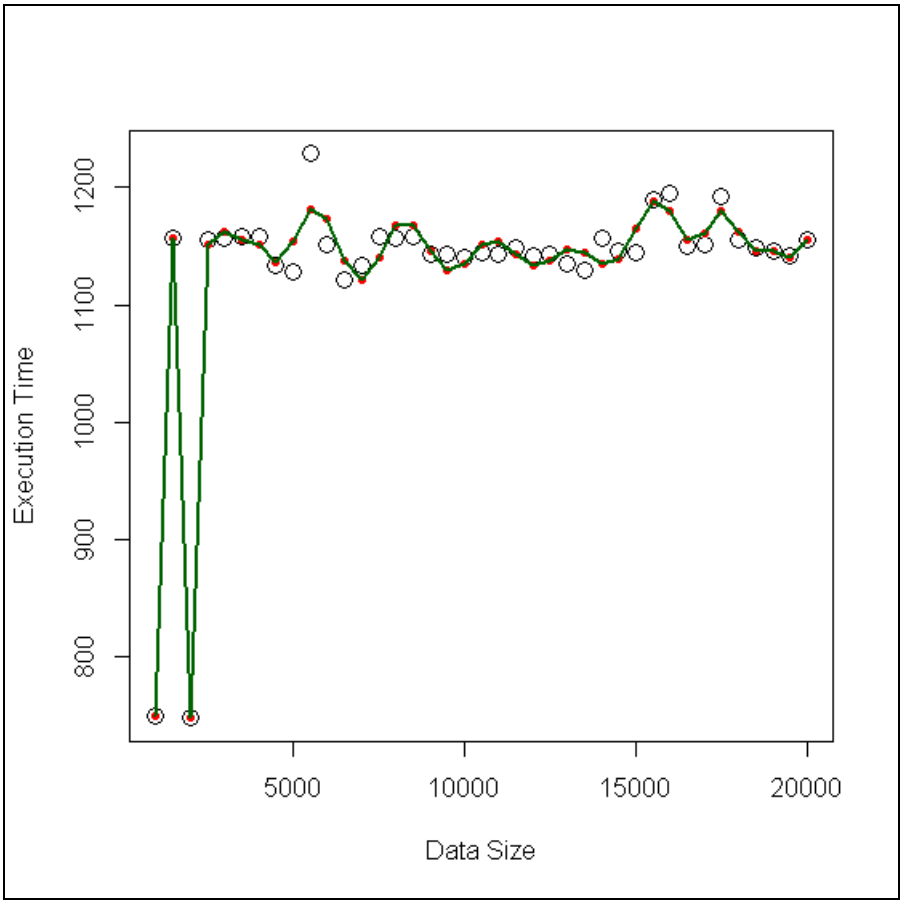

Figure 2. Polynomial of degree 22 model for Binary Search worst case

In the above figure the black circles represent the observed data points, the red dots represent the predicted points obtained by using the model equation and the dark green line represents the curve of polynomial of degree 22.

Equation of the proposed model:

y = 1131.722 + 212.698*x - 207.315*x^2 + 195.767*x^3 - 195.785*x^4 + 90.434*x^5 - 37.096*x^6 + 39.494*x^7 + 16.604*x^8 + 11.087*x^9 - 41.787*x^10 + 63.521*x^11 - 88.089*x^12 + 127.316*x^13 - 140.893*x^14 + 147.637*x^15 - 172.811*x^16 + 136.077*x^17 - 77.161*x^18 + 67.983*x^19 - 29.624*x^20 + 66.748*x^21 - 68.071*x^22

Where, y is Execution time in nano-seconds and x is Data size.

VI. CONCLUSION

In this paper, the researchers have tried to find out the best polynomial curve that can be fitted to the data points (Execution Time *versus* Data Size) generated by simulating binary search in worst case in a personal computer on Linux operating system and OpenJDK. For this purpose twenty three (*23*) different polynomial models have been tested and out of them three (*3*) models have been identified as the candidate models. All these three (*3*) models are having high Adjusted R-Squared values (more than 0.95) along with low Residual Standard Error and Root Mean Square Error values. We have used two different well known techniques (*AIC & BIC*) to identify the best model amongst the candidate models. The polynomial of degree 24 model shows lowest AIC value and thus identified as the best model for the given data points using AIC criterion and polynomial of degree 22 model shows lowest BIC value and thus identified as the best model for the given data points using BIC criterion. The researchers found that as expected the BIC favors smaller model than AIC. It is to be noted that the main objective of this paper is to fit a polynomial curve to the data points

and therefore we have not tried to fit other popular models to the data points which will remain our future scope of study.

REFERENCES

1. Goodness-of-Fit Statistics. (n.d.). Retrieved June 1, 2016, from http://web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html
2. Documentation. (n.d.). Retrieved June 1, 2016, from http://in.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html?s_tid=gn_loc_drop
3. What is residual standard error? (n.d.). Retrieved June 1, 2016, from http://stats.stackexchange.com/questions/57746/what-is-residual-standard-error
4. Buechler, S. (2007). *Statistical Models in R Some Examples* [PDF]. Retrieved June 1, 2016, from https://www3.nd.edu/~steve/Rcourse/Lecture8v1.pdf
5. Documentation. (n.d.). Retrieved June 1, 2016, from http://in.mathworks.com/help/stats/f-statistic-and-t-statistic.html
6. Documentation. (n.d.). Retrieved June 1, 2016, from http://in.mathworks.com/help/stats/understanding-linear-regression-outputs.html
7. MAZEROLLE, M. J. (n.d.). *APPENDIX 1: Making sense out of Akaike's Information Criterion (AIC): It...and interpretation in model selection and inference from ecological data* [PDF]. Retrieved June 1, 2016, from http://avesbiodiv.mncn.csic.es/estadistica/senseaic.pdf
8. Maydeu-Olivares, A., & Garcı´a-Forero, C. (2010). *Goodness-of-Fit Testing* [PDF]. Elsevier Ltd. Retrieved June 1, 2016, from http://www.ub.edu/gdne/amaydeusp_archivos/encyclopedia_of_education10.pdf
9. Kumari, A., Tripathi, R., Pal, M., & Chakraborty, S. (2012). Linear Search Versus Binary Search: A Statistical Comparison For Binomial Inputs. *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, *2*(2). Retrieved October 10, 2015, from http://www.airccse.org/journal/ijcsea/papers/2212ijcsea03
10. Sapinder, Ritu, Singh, A., & Singh, H.L. (2012). Analysis of Linear and Binary Search Algorithms. *International Journal of Computers & Distributed Systems*, *1*(1).
11. Das, P., & Khilar, P. M. (2013). A Randomized Searching Algorithm and its Performance analysis with Binary Search and Linear Search Algorithms. *International Journal of Computer Science & Applications (TIJCSA)*, *1*(11). Retrieved October 10, 2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.7058&rep=rep1&type=pdf
12. Roy, D., & Kundu, A. (2014). A Comparative Analysis of Three Different Types of Searching Algorithms in Data Structure. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(5). Retrieved October 10, 2015, from http://www.ijarcce.com/upload/2014/may/IJARCCE6C%20a%20arnab%20A%20Comparative%20Analysis%20of%20Three.pdf
13. Chadha, A. R., Misal, R., & Mokashi, T. (2014). Modified Binary Search Algorithm. *arXiv preprint arXiv:1406.1677*.
14. Parmar, V. P., & Kumbharana, C. (2015). Comparing Linear Search and Binary Search Algorithms to Search an Element from a Linear List Implemented through Static Array, Dynamic Array and Linked List. *International Journal of Computer Applications*, *121*(3). Retrieved October 10, 2015, from http://search.proquest.com/openview/a9b016911b033e1e8dd2ecd4e7398fdd/1?pq-origsite=gscholar
15. Pathak, A. (2015). Analysis and Comparative Study of Searching Techniques. *International Journal of Engineering Sciences & Research Technology*, *4*(3), 235-237. Retrieved October 10, 2015, from http://www.ijesrt.com/issues pdf file/Archives-2015/March-2015/33_ANALYSIS AND COMPARATIVE STUDY OF SEARCHING TECHNIQUES.pdf
16. Das, D., Kole, A., Mukhopadhyay, S., & Chakrabarti, P. (2015). Empirical Analysis of Binary Search Worst Case on Two Personal Computers Using Curve Estimation Technique. *International Journal of Engineering and Management Research*, 5(5), 304 – 311. Retrieved November 10, 2015, from http://www.ijemr.net/DOC/EmpiricalAnalysisOfBinarySearchWorstCaseOnTwoPersonalComputersUsingCurveEstimationTechnique%28304-311%29.pdf
17. Das, D., Kole, A., & Chakrabarti, P. (2015, November). Sample Based Visualization and Analysis of Binary Search in Worst Case Using Two-Step Clustering and Curve Estimation Techniques on Personal Computer. *International Research Journal of Engineering and Technology*, *02*(08), 1508-1516.
18. Curve fitting. (n.d.). Retrieved June 1, 2016, from https://en.wikipedia.org/wiki/Curve_fitting
19. Becerra-Fernandez., et al. (2004). Knowledge Management 1/e, Retrieved May 11, 2016, from https://home.cse.ust.hk/~dekai/523/notes/KM_Slides_Ch12.pdf
20. Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (pp. 199-213). Springer New York.
21. Cavanaugh, J. E. (2012, August 28). *171:290 Model Selection Lecture II: The Akaike Information Criterion* [PDF]. The University of Iowa. Retrieved May 11, 2016, from http://myweb.uiowa.edu/cavaaugh/ms_lec_2_ho.pdf
22. Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.
23. Cavanaugh, J. E. (2012, September 25). *171:290 Model Selection Lecture V: The Bayesian Information Criterion* [PDF]. The University of Iowa. Retrieved May 11, 2016, from http://myweb.uiowa.edu/cavaaugh/ms_lec_2_ho.pdf

