Accuracy of Artificial Neural Network Models of Software Reliability Growth – A Survey

M. K. Saley[1] and Sasikumaran Sreedharan[2]

[1]Research Scholar, ManonmaniamSundaranar University,
Thirunelveli, India
saley_mk@yahoo.com
[2]Department of Computer Engineering,
King Khalid University, KSA.
drsasikumaran@gmail.com

Abstract: Software Engineering data is being analyzed by classical statistical methods and non parametric methods. Performance models are constructed using classical approach as a high maturity practice. Such practices are constrained by data quality and inadequacy of data analysis methods to treat data from real life projects. Data mining techniques can broaden the data analysis capability and improve prediction accuracy even with commonly presented data. Artificial neural networks are found as an improved prediction error estimation method against traditional parametric software reliability growth models. In this paper, we study prediction errors of Artificial Neural Networks (ANN) based Software Reliability Growth Models (ANN SRGM) with the objective of arriving at a criteria for selecting the methods having least prediction errors. All major works in ANN SRGM's are considered and reported errors are analyzed. Accuracy of ANN SRGM's are compared against that of parametric models. Then, inter-comparison of error performances of ANN SRGM's of different applications is made.

Keywords: Software Reliability Growth Model, Artificial Neural Networks, Prediction Accuracy, Root Mean Square Error of ANN, Learning Speed of ANN, Flexibility of ANN.

1 Introduction

Various attempts are being made to apply Artificial Neural Networks (ANN) to substitute traditional parametric Software Reliability Growth Models (SRGM). A new breed of ANN based SRGM's have been introduced by researchers in this field. While parametric models give clues to process composition and behavior, the nonparametric ANN SRGM's offer greater accuracy and flexibility [1].

Typical software reliability data consists of cumulative defects discovered during testing. The pattern may vary depending on how early defects are discovered, how promptly defects are reported, and how well testing process is planned. Because of these reasons data patterns vary from organization to organization, from project to

project. Reliability growth is concerned with the pattern by which defect discovery progresses towards the ideal plateau region of defect free state. This is a time series problem and to study reliability growth models, one does not consider software structure or software development process composition. It is a black box view on defect discovery pattern. Parametric models use mathematical equations that may use test engineering parameters like test efficiency, test effort, software size, review speed etc; such models, when constructed, provide insight into what is happening during testing. ANN SRGM's do not aim at obtaining such an insight; in fact they are incapable of seeing the hidden factors of structural aspects. The purpose of ANN SRGM is to learn from data patterns and predict pending defects.

The inability of ANN SRGM to provide insight into problem structure is richly compensated by the greater prediction accuracy, which emerges as the most important criteria by which we can judge the performance of ANN SRGM's. In this paper, we consider the pioneering work done by several researchers and study the results reported by them, providing an inter-comparison of errors.

2 Comparison of Performance of Parametric Models with ANN SRGM

The key question that concerns users of Artificial Neural Network Software Reliability Growth Models (ANN SRGM) is whether the proposed ANN SRGM will provide better accuracy than existing parametric models such as Goel-Okumoto Model, Delayed S Model, Logarithmic Model, Exponential Model, Power Model, and Inverse Polynomial Model. Every researcher who has proposed ANN SRGM has answered this question in favor of ANN SRGM. [2] Comments and conclusions such as "ANN provides better results" are common. Such conclusions are not objective enough.

Hypothesis Test 1. The first objective error data comparing ANN with parametric models has come from Karunanidhi [3] and is summarized in Table 1. The error metric data Average Error (AE) % has been selected and used to construct Table 1.

Five parametric models and six ANN SRGM's are considered in Table 1. Error data is provided for each model against three data sets. On this data, a hypothesis test has

Table 1. End Point Error from Karunanidhi's Data [3].

		AE %		
CATEGORY	MODEL	DATA 1	DATA 2	DATA 3
PARAMETRIC	LOGARITHMIC	16.84	5.25	12.48
PARAMETRIC	INV POLY	19.4	4.6	13.29
PARAMETRIC	EXP	28.35	11.93	15.87
PARAMETRIC	POWER	18.35	12.16	12.95
PARAMETRIC	DELAYED S	35.78	17.71	27.1
ANN	FFN GEN ENC	26.75	14.51	30.2
ANN	ELMAN	24.73	13.24	19.6
ANN	FFN GEN	6.63	3.52	10.24
ANN	FFN	3.76	2.32	12.32
ANN	JORDAN GEN	3.05	3.11	6.96
ANN	JORDAN	2.68	3.21	9.52

Table 2. Two Sample *t* –Test Results on difference in AE % between ANN SRGM and Parametric Models.

Hypothesis Testing T-Test of difference = 0 (vs ≠)				
MODEL	N	AE % Mean	AE % StDev	SE Mean
ANN SRGM	18	10.91	8.99	2.1
PARAMETRIC MODEL	15	16.80	8.38	2.2
Difference	μ (ANN) - μ (PARAMETRIC)			
Estimate for difference	-5.90			
95% CI for difference	(-12.08, 0.29)			
T-Value	-1.95			
P-Value	0.061			
Degrees of Freedom	30			

been done by the authors to see if ANN SRGMs fare better than parametric models. A group is formed with error data for six ANN SRGMs and three data sets with a total of eighteen data points. Another group is formed with error data for the five parametric models and three data sets, with a total of fifteen data points. A t-Test has been done on the two groups of data and the results are shown in Table 2.

The mean values of average errors are 10.91 and 16.8 for ANN and parametric models, clearly showing an improvement of 5.90 in the mean. This is good enough in normal situations to go in for ANN; however, in the light of skepticism and criticism that prevail upon ANN a more stringent judgment is called for. The t - Test p value provides this information and is only 0.061. Had the p value been less than 0.5 we could have acknowledged, by a popular rule, significant improvement. With a borderline value of 0.061, ANN experiments have not yet delivered convincing results with statistical significance.

In addition to the *t* Test, a visual comparison of the two error data groups has been done using Turkey's box plot as shown in Fig. 1. The most common form of the box plot is constructed with median value at the center line and, quartiles in the box edges and whiskers that stretch from the box to connect with the extreme values. The box plot offers a more robust comparison, compared to a t-Test, and is not dependent on data normality.

Hypothesis Test 2. Karunanidhi's data is typical of ANN performance. Many other researchers support the view that ANN is better than parametric models. There have been many results reported in clear support of ANN. Sultan's data [4] is an example of results that are not so convincing. Sultan uses Summed Square Error (SSE), as the error metric and considers three types of projects, shown in Table 3.

A t-Test done on Sultan's data is shown in Table 4. The mean SSE values are 57.2 and 60.3 for ANN and Parametric Models. ANN scores better, with less mean error. The difference is only marginal. But the p value is 0.970 which suggests that ANN has

Table 3. Test Errors from Sultan's Data [4].

SUM SQUARE ERROR		
PROJECT	PARAMETRIC	ANN
Military	168.54	160.3887
Real Time Control	1.368	1.2542
Operating System	10.885	9.9623

Table 4. *t* -Test Results on Squared Sum of Error difference between Parametric and ANN SRGM.

Hypothesis Testing T-Test of difference = 0 (vs ≠)				
MODEL	N	AE % Mean	AE % StDev	SE Mean
ANN SRGM	3	57.2	89.5	52
PARAMETRIC MODEL	3	60.3	93.9	54
Difference	μ (ANN) - μ (PARAMETRIC)			
Estimate for difference	-3.1			
95% CI for difference	(-241.4, 235.2)			
T-Value	-0.04			
P-Value	0.970			
Degrees of Freedom	3			

not produced significantly different results from Parametric Models. With such a high p value, one would tend to consider them equal.

This is a clear example of a situation where ANN performance is equal to parametric model performance, but offers an advantage, though small.

3 Sensitivity of End Point Errors to Data Patterns – A Hypothesis Test

A limitation of ANN SRGM arises out of its difficulty to cope up with changes in data patterns. Its ability to forecast depends on smoothness of training data. Typically the data is the cumulative defects found during testing and its pattern can vary from project to project, depending on the quality of testing and management of testing process. There could be sudden jumps and micro plateaus; the inflexion point could change if proportion of early discovery changes. The finishing line flatness could also change depending on test case effectiveness. ANN SRGM trained with such turbulent patterns may show larger errors during testing. Also, it has been shown that the number of hidden nodes required for optimum errors vary if the data set changes [5]. As a result software reliability data sets and the actual patterns of data seem to have an influence on ANN SRGM prediction errors (end point errors).

Sultan's [4] results, shown in Table 3, exhibits dramatic variation of SSE when data set changes. SSE could vary from a value as low as 1.25 in one project dataset to a value as high as 160.39 in another project dataset. The change is so dramatic and obvious that without using a sophisticated hypothesis testing one can arrive at the judgment. The problem exists even with parametric models. Basically ANN SRGM has not solved a fundamental problem: reliability models fail to show consistent performance across datasets. If data patterns could dictate errors so much, it is not good news for those who wish switch over to ANN SRGM's, because they do not see any advantage in doing so, from the point of view of data-pattern sensitive performance.

A different situation emerges if we examine Karunanidhi's data shown in Table 1. To test this hypothesis that data sets control errors, Analysis of Variance (ANOVA) has been done on ANN SRGM AE % for three different data sets. ANOVA results are shown in Table 5.

Table 5. ANOVA Results of Dataset Sensitivity of Errors.

Hypothesis Testing: ANOVA					
Source	DF	Adj SS	Adj MS	F-Value	P-Value
DATA	2	200.7	100.33	1.28	0.306
Error	15	1174.7	78.31	-	-
Total	17	1375.4	-	-	-

The p value is 0.306, indicating that data sets do not have significant influence on errors. This means that ANN SRGMs have overcome the problem of data-pattern sensitivity.

4 Inter-comparison of Prediction Errors of ANN SRGM'S

Normalized Data. Prediction errors in six experiments with ANN Software Reliability Growth Models performed from 1992 till 2013 by various researchers have been compiled and shown in Table 6. The error metric selected for this inter-comparison is AE %

Table 6. Normalized Data of ANN SRGM Test Errors.

NORMALIZED DATA OF PREDICTION ACCURACY OF ANN SRGM				
Ref.No	Name of the Author	Year published	Name of NN	AE %
[3]	NachimuthuKarunanidhi, Darrel Whitley and Yaswant K. Malaiya.	1992	FFN GEN ENC DS1	26.7500
			FFN GEN ENC DS2	14.5100
			ELMAN DS1	24.7300
			ELMAN DS2	13.2400
			FFN GEN DS1	6.6300
			FFN GEN DS2	3.5200
			FFN DS1	3.7600
			FFN DS2	2.3200
			FFN DS3	12.3200
			JORDAN GEN DS1	3.0500
			JORDAN GEN DS2	3.1100
			JORDAN DS1	2.6800
			JORDAN DS2	3.2100
[7]	NachimuthuKarunanidhi, Darrell Whitley and Yaswant	1992	FFN generalization	0.0667
			FFN prediction	0.0612
			JN generalization	0.0475
			JN prediction	0.0494
[5]	Jun Zheng	2009	PSNN 1	1.890
			PSNN 2	4.650
			PNNE 1	1.290
			PNNE 2	1.630
[10]	Jung-Hua Lo	2009	MLFFN 1	20.790
			MLFFN 2	7.350
			MLFFN 3	8.900
			MLFFN 4	2.260
[11]	Yogesh Singh, Pradeep Kumar	2010	FFNW 1	0.100
			FFNW 2	0.050
			FFNW 3	0.120
			FFNW 4	0.340
			FFNW 5	0.830
[12]	Rita G. Al gargoor, Nada N Saleem	2013	PSO	7.180
			PSO	6.090
			PSO	4.480

defined in Equation 2. This metric has been used in the pioneering work of Karunanidhi [3], [7] and followed by several researchers. Prediction errors (AE %) for Multi-Layer Perceptron (MLP), Elman, Jordan, PNNE (Ensemble), PSNE (Single Input Single Output Three Layer), PSO networks have been compiled. AE % data shown in Table 6 is normalized data and permits inter-comparison.

Non-Normalized Data. Non-normalized data (scale dependent) from nine experiments is compiled in Table 7. This data does not allow inter-comparison across different experiments performed by different authors. This data is provided just for reference. The metrics used in this data are MSE, RMSE and SSE. Metrics such as NRMSE would permit inter-comparison; however, unfortunately, it is available for only one experiment from a single paper.

Table 7. Non-Normalized Data of ANN SRGM Test Errors.

NON-NORMALIZED DATA OF PREDICTION ACCURACY OF ANN							
Ref. No	**Name of the Author**	**Year published**	**MSE**	**RMSE**	**NRMSE**	**NSE**	**SSE**
[4]	Sultan H, Alaa Sheta and David Rine	2001	9.800	-	-	-	-
			15.826	-	-	-	-
			12.476	-	-	-	-
[8]	N. Raj Kiran and V Ravi	2007	-	-	0.1455	-	-
			-	-	0.1504	-	-
			-	-	0.1579	-	-
			-	-	0.1527	-	-
			-	-	0.1669	-	-
			-	-	0.1479	-	-
			-	-	0.1611	-	-
			-	-	0.1476	-	-
			-	-	0.1162	-	-
			-	-	0.1115	-	-
[9]	Sultan H. Aljahdali and Khalid A. Buragga	2008	-	0.6677	-	-	-
			-	0.1591	-	-	-
			-	0.1394	-	-	-
			-	0.9079	-	-	-
[12]	Lu Zhao, Jian-pei Zhang	2010	0.0486	-	-	-	-
			0.0590	-	-	-	-
[13]	Kewen LI, Lina GONG, Jisong KOU	2010	20.5426	-	-	-	-
			0.3642	-	-	-	-
[14]	V.Ramakrishna,MR Narasinga Rao and TM	2012	-	-	-	0	-
			-	-	-	0.07	-
[15]	Rita G. Al gargoor, Nada N Saleem	2013	-	0.0184	-	-	-
			-	0.0186	-	-	-
			-	0.0180	-	-	-
[16]	Sultan H, Alaa Sheta and David Rine	2001	-	-	-	-	160
			-	-	-	-	1.25
			-	-	-	-	9.96

Inter-comparison of ANN SRGM with normalized data

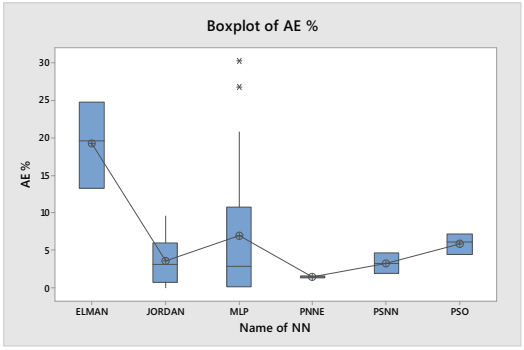

Figure 1. Inter-comparison of Average Errors of Various Types of ANN SRGMs.

It is rather difficult to perform a hypothesis test to find which ANN is better, because of multiplicity of data origin and inequality of sample size. Moreover, data sets used in the six groups of data are different, introducing an element of uncertainty. Instead of the classical hypothesis test a box plot comparison, the next robust alternative, has been performed. Figure 1 shows the inter-comparison results in the form of box plots of data given in Table 6. As usual the box plot comprises of median, quartiles and whiskers and, in our cases, we have included the mean values for reference and connected the mean values by a thin line.

If we compare the medians of the box plots, the ensemble PNNE has the lowest error. If we compare the first quartiles (lower edges of boxes) then MLP shows lowest error. Clearly Elman network shows the highest error. PSO shows larger errors than PNNE and PSNN.

MLP performance requires a special analysis. The dispersion of MLP data is high. One must remember that MLP sample size is also high and data has arrived from heterogeneous sources. The mean value of MLP errors is high because of skew in the data and influence of extreme values. The median shows a fair picture. When data is skewed, mean is not a reliable indicator of central tendency. It is the quartile of the box that makes us think that MLP does show possibility of better performance. To validate this theory we can look at the raw data in Table 6 where we find Yogesh's results [8] with MLP are impressive and are responsible for the lower quartile of the box.

While one has to take the box plot results with a grain of salt because of inherent uncertainties, it is evident that without using recursion and special optimization, geed results can be obtained with simple plain MLP. This evidence motivates one to pursue MLP as an attractive candidate for ANN SRGM. There is no compelling evidence in favor of pursuit of sophistication in ANN topologies.

5 Conclusion

It is seen that several researchers have attempted various realizations of ANN SRGMs in the past 22 years. Each experiment is a discovery an ANN SRGM topology that works better than selected statistical models. Variability of ANN performance with data

sets is a concern. Inadequate attention to characteristics of trained networks leaves ANN less trustworthy in the critical eyes of ANN users. It is noted that input data preparation and structuring extend ANN performance to greater levels. Further refinement is obtained by the choice of judicious activation functions that resemble the empirical data patterns

This survey finds ANN SRGM is better than statistical models for reliability prediction. However, the statistical significance of results is still low, with a p value of 0.61, leaving a wide scope for further research and improvement in ANN SRGM. The errors are comparable and in some cases better than errors achieved in other domains.

References

1. Saley M.K and SasikumaranSreedharan, "A Survey of Software Reliability Growth Models Using Non-Parametric Methods", IEEE International Conference on Computational Intelligence and Computing research", Dec. 2014.
2. Doteswamy and Chanabasayya M Vastrad, "Performance Analysis of Oxazolens and Oxazoles Derivatives Description Dataset", Vol 03, No 6, International Journal of Information Sciences and Techniques, Nov 2013.
3. NachimuthuKarunanidhi, Darrel Whitley and Yaswant K. Malaiya., "Prediction of reliability model using Connectionist Models," IEEE transactions on software engineering, Vol. 18, No. 7, July 1992. pp. 563–574.
4. Sultan H, AlaaSheta and David Rine, "Prediction of Software Reliability: A comparison between Regression and Neural Network Non-Parametric Models," Computer Systems and Applications, ACS/IEEE International Conference, pp. 470–473, June 2001.
5. Jun Zheng, "Predicting software reliability with neural network ensembles", ELSEVIER, Expert Systems with Applications 36, 2116–2122, 2009.
6. ZhigangGan, Nanfeng Xiao, "An Improved Adaptive Boosting Algorithm for Neural Network Ensemble Based on Multi-dimensional Cloud Model", Second International Symposium on Knowledge Acquisition and Modeling, IEEE Computer Society, 2009.
7. NachimuthuKarunanidhi, Darrell Whitley and Yaswant, "Using Neural Networks in Reliability Prediction," IEEE software, 0740-7459/92/0700/0053, July 1992.
8. N. Raj Kiran and V Ravi, "Software Reliability Prediction Using Wavelet Neural Networks", IEEE computer society, pp.195–199, 2007.
9. Sultan Aljahdali and Alaa F sheta, "Predicting the Reliability of software System using Fuzzy Logic", IEEE 2011.
10. Jung-Hua Lo "The Implementation of Artificial Neural Networks Applying to Software Reliability Modeling", IEEE Chinese Control Decision Conference, 2009.
11. Yogesh Singh and Pradeep Kumar, "Prediction of software reliability Using Feed Forward Neural Networks" Computational Intelligence and Software Engineering (CiSE), International Conference, 2010.
12. Lu Zhao, Jian-pei Zhang et al., "Software reliability Growth Model based on Fuzzy Wavelet Neural network", IEEE 2nd International Conference on Future Computer and Communication, Vol. 1, pp. VI 664–VI 668, 2010
13. Kewen LI, Lina GONG, Jisong KOU, "Predicting Software Quality by Fuzzy Neural Network Based in Rough Set", Journal of Computational Information systems, pp. 1439–148, May 2010.

14. V.Ramakrishna,MRNarasingaRao and TM Padmaja "Software reliability prediction using Neural Networks", International journal of Computer Applications(0975 – 887) Vol. 60-No.&, December 2012.
15. Rita G. Al gargoor, Nada N Saleem, "Software Reliability Prediction Using Artificial Techniques", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No. 2, ISSN (Print): 1694-0814, pp. 274–281, July 2013.
16. Sultan H, AlaaShetta, David Rine, "Prediction of Software Reliability: A comparison between Regression and Neural Network Non-Parametric Models," Computer Systems and Applications, ACS/IEEE International Conference, pp. 470–473, June 2001.
17. G.Q. Shang, C.H Sun, "Application of BP Neural Network for Predicting Anode Accuracy in ECM", International Symposium on Information Science and Engineering, IEEE Computer Society 2008. pp. 428–432.
18. Qi Feng, Shengjun Wu, Yun Du, HuaipingnXue, Fei Xiao Ban, and Xiaodong Li, "Improving Neural network Prediction Accuracy for PM Individual Air Quality Index Pollution Levels", Environmental Engineering Science, Volume 30, 2013.
19. Lihua Chen, Yuchen Wang, "Application of Radial Basis Function Neural Network on the Prediction of Urban Built-up Area", Applied Mechanics and Materials, Trans tech Publications, Switzerland, Vols. 556–562, pp. 5308–5311. 2014.
20. Di Lu and Jing Liu, "Application of Improved Neural Network in the Automotive Engine Fault Diagnosis", International journal of Control and Application, Vol. 7, No. 4, pp. 262–355, 2014.
21. Fuqiang Zhou, Zhen Su, Xinghua Chai and Lipeng Chen, "Detection of Foregin Matter in Transfusion Solution Based on Gaussian Background Modeling and an Optimized BP Neural Network", Sensors ISSN 1424-8220, 2014.
22. A. Noriega, D Blanco, B.J. Alvarez, A.Garcia, "Dimensional Accuracy improvement of FDM square cross-section parts using artificial neural networks and an optimization algorithm", Int J Adv Manuf Technol, Springer, 2013.
23. Kunwar Singh Vaisla and Ashuthosh Kumar Bhatt, "An Analysis of the Performance of Artificial Neural Network Technique for Stock Market Forecasting", Vol 02, No. 06, International Journal on Computer Science and Engineering, 2010.