I J C T A, 9(28) 2016, pp. 191-195

A LITERATURE REVIEW ON SOFTWARE FAULT PREDICTION USING MACHINE LEARNING AND SOFT COMPUTING TECHNIQUES

R. Sathyaraj* and S. Prabu*

Abstract: This paper presents the comprehensive view on the techniques used for predicting faults in various software developments. Using several classification methodologies we can produce the reliable software by reducing faults and failures. Prediction helps to identify the faults in upcoming modules using past results and training data and it reduce time for debugging.

Method: In this study, we delivers the view about various techniques and methodologies used in fault prediction using soft computing techniques like artificial neural networks, fuzzy logic, genetic algorithm and machine learning algorithm like naïve bayes, random forest, and decision tree. Additionally we summarize the strength and weakness of those methods.

Results: In this paper we acknowledged studies in soft computing and in machine learning algorithms. From this survey we further planned to produce a suitable hybrid algorithm for better fault prediction and to improve the quality in advanced object oriented software systems.

Keywords: Software fault prediction, machine learning, soft computing,

1. INTRODUCTION

In this world lot of innovative things are happening surround us. New technologies and innovations help us to do our work easier and smarter. Nowadays the competence to innovate or to renovate the existing things and the knowledge of research about those processes got improved. By focusing the changes happened in recent years, most of the research focusing on prediction. Prediction is the most important thing, to solve future problems. Prediction is the word which used in multidisciplinary in multi-ways to improve the production and quality of the product or process. In software development and testing, retain the quality will increase the value of the software. To achieve the high productivity, prediction is the significant way to reach.

1.1 Soft Computing Techniques

Soft computing is the best solution finder for real time problems which is more complication and not able to define in crisp values. It provides the randomized search and provides approximate results for the problem which can be critical to define and for intelligent machines.

Overview of techniques in soft computing:

Neural Networks, Fuzzy Logic, Genetic Algorithm, and Hybrid Systems are the techniques in soft computing.

* School of Computer Science and Engineering VIT University Vellore Tamilnadu - 632014
Email: sathyaraj@vit.ac.in, [2]sprabu@vit.ac.in

1.1.1 Neural Networks

Neural networks have been used to crack a wide range of tasks that are hard to solve using ordinary rule-based programming. Simply, we can describe it as, “units “chained as “connections” inter connected to produce the “intelligent results“. Neurons process the information associated with their weights and using activation function and produce the intelligent results for the real time problems. Neural networks mainly focus on learning environment, information processing, fault tolerance and etc.

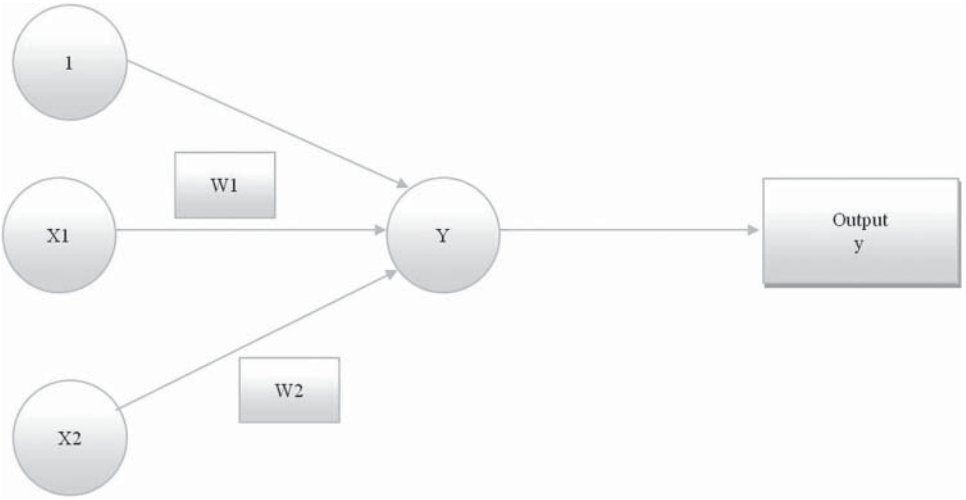

Figure 1. Basic structure of Neural Networks

The figure 1 shows a simple artificial neural net with two input X1, X2 and one output Y. The weights are W1 and W2 associated input neurons.

1.1.2 Fuzzy Logic

Fuzzy logic is a methodology to figure out the approximate results for the linguistic terms. It gives us a language with syntax and local semantics in which we can translate our qualitative domain knowledge. We can take an example; role is labelled by expressions like “hot temperature” or “regular customer”. We apply fuzzification, executed with relevant rules and finally defuzzified to get crisp data.

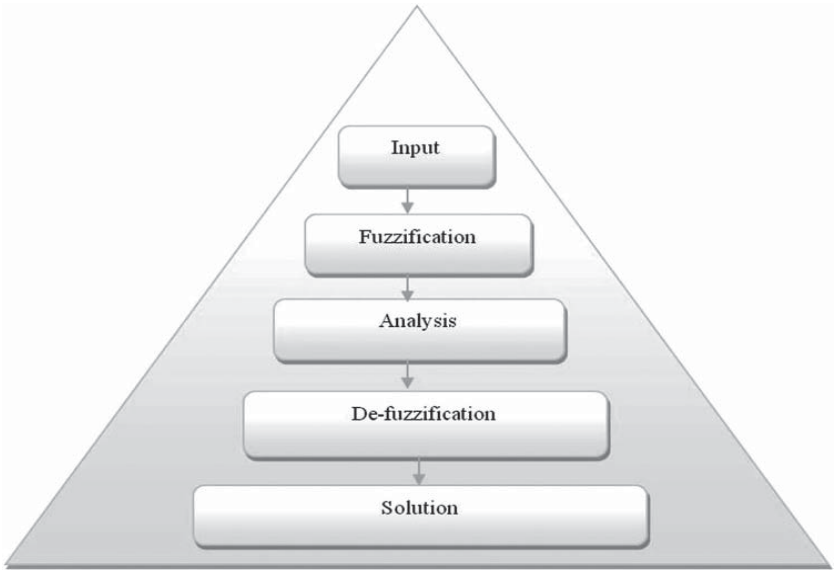

Figure 2. Structure of Fuzzy logic process

1.1.3 Genetic Algorithm

The genetic algorithms follow the evolution process in the nature to find the better solutions of some stimulating problems. In genetic algorithm can produce results based on the natural assortment process and optimize the results. It follow the cycle to generate the solutions, initialization, selection and reproduce with many iteration until the proper solution found.

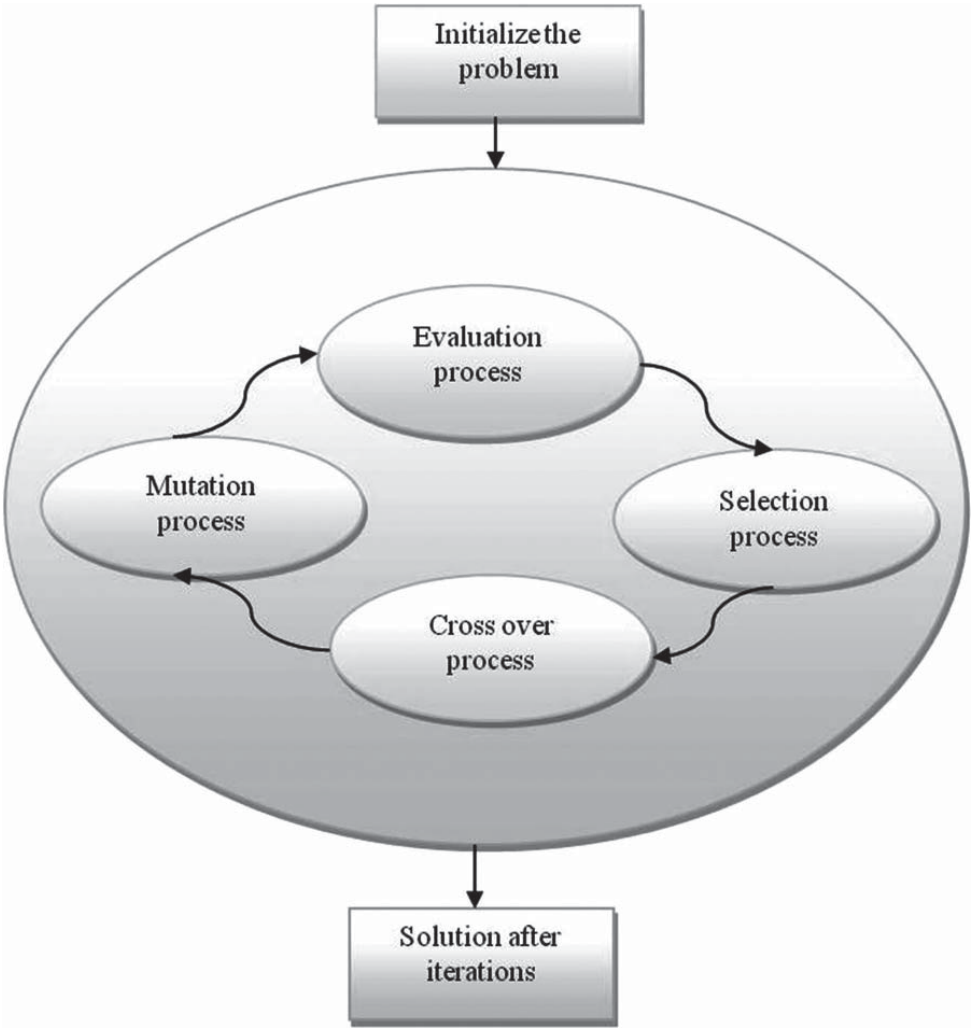

Figure 3. Genetic Algorithm evolution process

1.2 Machine learning algorithms

Machine learning is closely related to statistical learning, data analysis and predicting data from the past results. It is progresses from the computational knowledge and optimization in artificial intelligence. Includes the process of classification, clustering and optimization, also investigate data and extract the features of it using various machine learning algorithms.

1.2.1 Naïve Bayes

Naïve Bayes is the probability classifier, which is suitable for small datasets[1] and it used Bayes theorem which treated all the data as independent from others. Results of the Naïve Bayes mostly depend upon the prior results and the similarities. Using this classifier we can predict the fault to make improvement in software quality.

1.2.2 Random Forest

Random forest is a classifier builds multiple decision trees at training time and majority voting progression applied to finalize the results[2]. Random forest classifier was more suitable for large datasets[1].

2. METHODOLOGIES IN REVIEW:

Artificial Bee Colony (ABC) algorithm finds the subset of attributes independently based on decision attributes and then finds the final reduct. Initially the instances are grouped based on decision attributes. Then the Quick Reduct algorithm is applied to find the reduced feature set for each class. To this set of reducts, the ABC algorithm is applied to select a random number of attributes from each set, based on the Rough Set-Attribute Reduction (RSAR) model, to find the final subset of attributes[3].

In rough set theory, the data is organized in a table called decision table. Rows of the decision table correspond to objects, and columns correspond to attributes. In the data set, a class label indicates the class to which each row belongs. The class label is called as decision attribute, the rest of the attributes are the condition attributes. Here, C is used to denote the condition attributes, D for decision attributes, where C (Intersection) D = K, and t_j denotes the j^{th} tuple of the data table. Rough set theory defines three regions based on the equivalent classes induced by the attribute values: lower approximation, upper approximation and boundary. Lower approximation contains all the objects, which are classified surely based on the data collected and upper approximation contains all the objects, which can be classified probably, while the boundary is the difference between the upper approximation and the lower approximation[4].

Main objective of this work was to identify the nature of faults and avoid failure in web pages. To identify faults in java scripting languages, they used Fault localization approach and they categorized the faults using Transductive support vector machine classification algorithm. It classifies the faults and made subsets with respect to its bug types. Six type's bugs were considered to bring effective categorization. By using TSVM based categorization of faults, we can detect more errors and improve the accuracy of the programming language[5].

Alonso, Belanche and Avresky examined a various machine learning algorithms to predict the crashes in system triggered by software anomalies. Classifiers Rpart decision trees, Support Vector Machines Classifiers (SVM-C), Naive Bayes, K-nearest neighbors, Random Forest were implemented in the R Statistical Language tool and created crashes in e-commerce environment. Five-fold cross validation approach used to calculate the error to compare the models[5]. Based on the results, Random forest having lowest validation error rate and Lasso regularization has analyzed for reducing parameters without changes in accuracy prediction[7].

Software reliability is the main concern while we go for development. By testing linear ensembles and non-linear ensembles, they projected software reliability. To ensure reliability, they established several statistical and intelligent methodologies. Multiple linear regression and multivariate adaptive regression; dynamic evolving neuro fuzzy inference system, backpropagation neural networks and TreeNet are the statistical and intelligent techniques ensemble. Linear ensemble based on weighted mean, average, and weighted median. In non-linear ensemble, backpropagation neural networks used to train and assign weight accordingly. From the results observed, non-linear ensemble performed well compared to all others[8].

Neural networks are used in system's quality estimation with the help of object oriented metrics. In two kinds of exploration, initially they predict class defects and next changes in lines based on class. They used two neural networks, Ward neural network and GRNN (General Regression neural network). Applied

measures like inheritance, complexity, cohesion, coupling and memory allocation as object oriented metrics. Here, the metrics taken as independent variables for prediction. Prediction results summarize GRNN network was better than the Ward network model[9].

Hassan Najadat and IzzatAlsmadi proposed a model RIDOR algorithm with modification and enhanced RIDOR provided better performance with the attributes as extracted rules and accuracy when compare to other classification algorithms. They suggested rule based classification algorithm for classification and using various rule-based classification methodologies they predicted the faulty modules and non-faulty modules in the datasets of the NASA software repository[10].

According to Ludmila the concept of fuzzy rough sets is a deterministic and probabilistic attractive tool to assess the highest classification capacity and to select minimal set of significant features in pattern recognition problems. This approach aims to solve the hypoxic resistance of fuzzy pattern recognition oriented problems. The consideration of positive, negative and boundary regions coincide with original expert conclusions[11].

3. CONCLUSION

This literature study gives the view about various techniques and methodologies used in soft computing and machine learning. We observed the process of various algorithms used for fault prediction and to improve the performance and quality of the software gives better solution in prediction. Further, we interested to improve a hybrid algorithm to produce better performance in software fault prediction.

References

1. Catal C, “Software fault prediction: A literature review and current trends”, Experts systems with applications, 2011; 38(4):PP.4626-36.
2. Moeyersoms J, Fortuny E J, Dejaeger K, Baesens B, “Comprehensible software fault and effort prediction: A data mining approach”, The Journal of Systems and Software, Feb 2015; Volume 100:PP.80-90.
3. Suguna N and Keppana Gowder Thanushkodi, “An Independent Rough Set Approach Hybrid with Artificial Bee Colony Algorithm for Dimensionality Reduction”, American Journal of Applied Sciences, 2011; volume 8 (3): PP.261-266.
4. Thangavel K, Pethalakshmi A, “Dimensionality reduction based on rough set theory: A review”, Applied Soft Computing, 2009: Volume 9, PP.1–12.
5. Fawzia Khan F and Mallika R, “Indian Journal of Science and Technology”,September 2015, Vol 8(21).
6. Monisha M, Samyukta Sherugar, Shobhit Bansal, Rajat Mann and Biju R Mohan, “A Survey on the Application of Machine Learning Algorithms to Predict Software Aging”, International Journal on Advanced Computer Theory and Engineering, 2013, Volume-2, Issue-5, 2319 – 2526.
7. Alonso J, Belanche L, Avresky D , “Predicting Software Anomalies Using Machine Learning Techniques” , Proceedings - 2011 IEEE International Symposium on Network Computing and Applications, NCA 2011 , pp.163 -170.
8. Raj Kiran N , Ravi V, “Software reliability prediction by soft computing techniques”, Journal of Systems and Software, Volume 81, Issue 4, April 2008, Pages 576–583.
9. Mie Mie Thet Thwin, Tong-Seng Quah, “Application of neural networks for software quality prediction using object-oriented metrics”, The Journal of Systems and Software, 76 (2005): PP 147–156.
10. Hassan Najadat, Izzat Alsmadi, “Enhance Rule Based Detection for Software Fault Prone Modules”, International Journal of Software Engineering and Its Applications,January, 2012,Vol. 6, No. 1:PP 75-86.
11. Kuncheva LI. “Fuzzy rough sets: application to feature selection”, Fuzzy sets and Systems. 1992 Oct 26;51(2):147-53.