IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. IV (Nov.-Dec. 2016), PP 122-125
www.iosrjournals.org

Feature Selection Based Neural Networks for Software Defect Prediction

M Satya Srinivas[1], Dr. A Yesubabu[2], Dr. G Pradeepini[3]
[1]Research Scholar CSE, K L University, Guntur
[2]Prof. & HOD CSE, Sir C R Reddy Engg. College, Eluru, W.G. Dist
[3] Prof. CSE Dept, K L University, Guntur

Abstract: *Predicting software defects in advance greatly reduces the software development cost. To predict the software defects already various methodologies are proposed like decision trees, Ada boost, SVM and Neural networks. The datasets are obtained from NASA and PROMISE. These datasets contain large number of attributes. Applying mining techniques on large number of attributes reduces the performance of classifier and hence, In this paper we are proposing Principle component analysis for feature selection, then Neural networks are used for best fitting non linear relationships between attributes. In this paper we considered AR1 dataset from PROMISE and presented the results of applying Neural Networks with feature selection & without feature selection.*

Keywords: *Software defect Prediction, Feature selection, Principle component analysis, Neural Networks.*

I. Introduction

The software development process is mainly driven by the factors like Cost, Schedule and Quality. These factors are in turn effected by defects in the software. As the defects in a software increases the cost, schedule and decreases Quality. Hence predicting a software bug in advance causes great reduction in cost , schedule and improves the quality. During 1971 Akiyama indicated the software defects depends on Lines of code(LOC). Later McCube and Halstead found correlation between software metrics and defects by constructing predictive models.

The software metrics used for construction of predictive models are Total LOC, Blank LOC, Comment LOC, Code & Comment LOC, Executable LOC, Unique operands, Unique operators, Total operands, Total operators, Halstead Vocabulary, Halstead Length, Halstead Volume, Halstead level, Halstead Difficulty, Halstead Effort, Halstead Error, Halstead time, Branch count, Condition count, Decision Count, Call pairs, Multiple condition count, Cyclometic complexity, Cyclometic density, Decision density, Design components, Design density, Normalized cyclometic complexity and Formal parameters.

Data Pre processing

1. Conversion of symbolic value attributes into numerical attributes
   Some of the attributes in Software defect dataset are symbolic attributes. But neural networks can be constructed on nominal data. So symbolic valued attributes must be converted to numerical attributes.
2. Feature set selection
   Software defect dataset contains 30 attributes. Some them are irrelevant to our task. Hence feature set selection has to be done in order to import only required attributed bases on information gain of attributes.
3. Normalization
   High valued attributes dominate low value attributes. To eliminate this domination, each attributed must be normalized to [0 1]. The advantage of normalization is neural networks can operate best with normalized data.

II. Related work

A wide range of metrics have been proposed for identification of faults[1]. Requirement metrics achieved success in predicting faulty modules[2]. Design metrics have proved their utility for fault prediction[3].Static code metrics also proven their effectiveness in many studies[4]. Weyuker et.al, [5] found addition of developer information improves the accuracy of fault prediction models. Mohammad Alshayeb proposed software defect prediction using ensemble learning on selected features[11]. A Systematic Review of Software Defect Prediction using data mining techniques was done by Romi Satria Wahono[12]. He also applied Metaheuristic Optimization approach in feature selection for Software Defect Prediction[13].

III. Methodology

3.1 Feature Selection:

There are various method for identifying the features that best fits the data into a model. Feature selection algorithms are broadly categorized into three categories.

1. Filter Methods: Filter Feature selection methods assigns a score value for each feature by applying statistical techniques. The features are selected based on these scoring values. Examples are Chi Square test, Co relation coefficient , Information gain and principle component analysis
2. Wrapper methods: Wrapper methods considers the feature selection problem as a search problem. A different combinations of features are prepared, evaluated & compared. A combination of features with high model accuracy is considered for feature selection. For example recursive feature elimination algorithm.
3. Embedded Methods: Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods.

   In this paper we are using principle component analysis for feature selection. Principle component analysis is used to emphasize variation and bring out strong patterns in a dataset.

3.2 Model Construction:

There are various algorithms for constructing the model for classification task. For example Decision trees, Random Forests, Bayesian classifiers, Neural Networks and Support vector machines. In this paper we are using Neural networks for constructing the model.

Neural Network is a collection of nodes(neurons) which are used to hold the knowledge by training the network with known data. The performance of network is evaluated by testing with unseen data. Training the network is repeated until it reaches the steady state in number of epochs. In each epoch a different training set in used to train the network. The process will be repeated until it reaches steady state or maximum number of epochs are completed.

There are various kinds of neural networks that can be best fitted for various applications. For classification or regression of data multi layer perceptrons or radial basis function networks can be best fitted. For clustering Self organized feature maps are used. In this paper I am using multi layer pereptrons for fitting non linear relationships between the attributes.

The complexity of training neural networks depends on number of features. The training is said to be efficient if the data set containing only relevant attributes. Irrelevant attributes can be eliminated by identifying with principle component analysis or co-relation coefficient.

IV. Results

Feature selection using Principle components:

Principle component analysis is applied on software defect dataset to identify weakly relevant attributes . Any numeric variables with relatively large rotation values (negative or positive) in any of the first few components are generally variables that you may wish to include in the modelling.

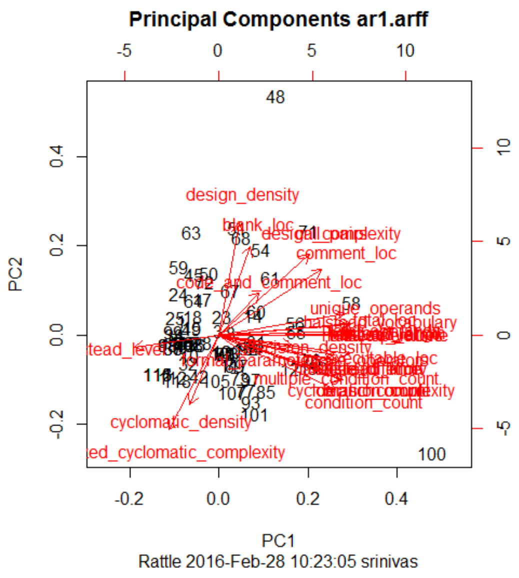
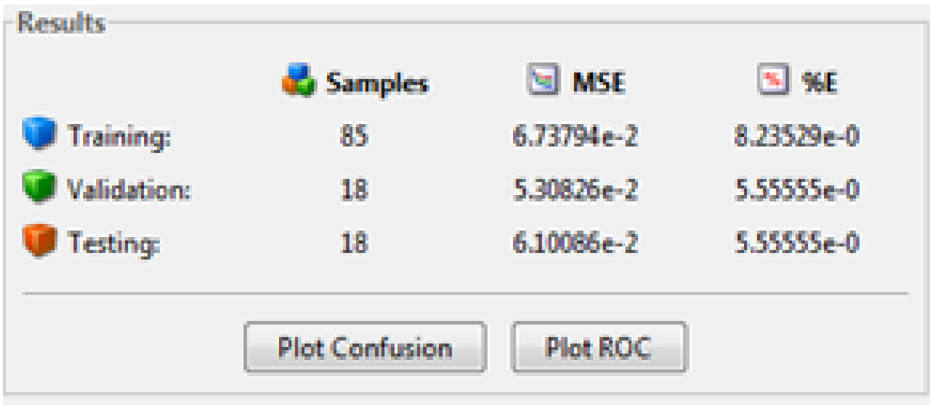


Fig 1: Principle component analysis on SDP dataset

From the above analysis the identified principle components are

- blank_loc
- call_pairs
- design_complexity
- design_density
- normalized_cyclomatic_complexity
- halstead_level
- cyclomatic_density
- code_and_comment_loc
- blank_loc
- multiple_condition_count
- decision_density
- comment_loc

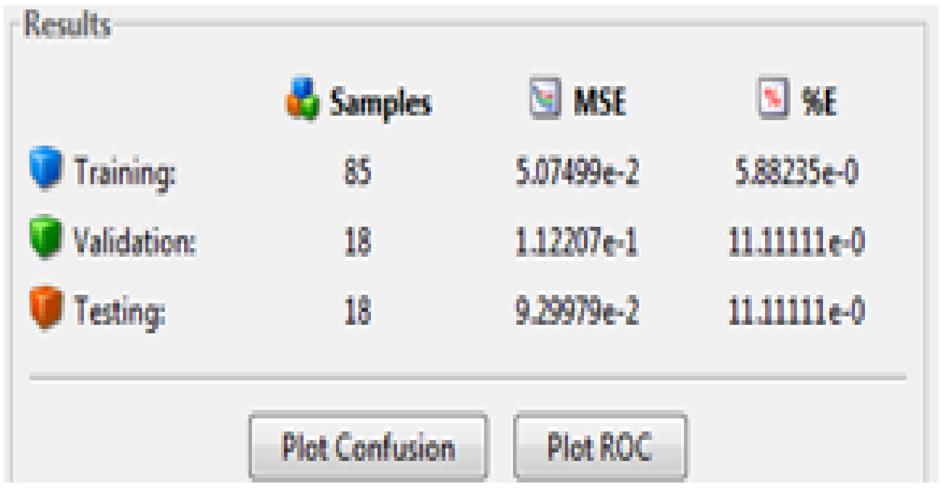
A model was constructed using multi layer perceptron using pattern recognition tool box in MAT Lab. The Average classification error on training, Validation and testing data are presented in the figure 2 & Figure 3.



Results

	Samples	MSE	%E
Training:	85	6.73794e-2	8.23529e-0
Validation:	18	5.30826e-2	5.55555e-0
Testing:	18	6.10086e-2	5.55555e-0

Fig 2: Neural Networks with PCA



Results

	Samples	MSE	%E
Training:	85	5.07499e-2	5.88235e-0
Validation:	18	1.12207e-1	11.11111e-0
Testing:	18	9.29979e-2	11.11111e-0

Fig 3: Neural networks without PCA

V. Conclusion

There are various methods for predicting the software defects. All those methods constructs the model without performing Data Pre processing. But applying Data Pre processing improves the classifier accuracy. Applying Principle component analysis identifies relevant attributes. Then model will be constructed using neural networks. The results are presented for constructing the model with feature selection & without feature selection. The results shows that Feature selection(Principle component analysis) based neural networks reduces the average classification error on software defect prediction.

References

[1]. Jiang Y et.al Fault prediction using early lifecycledata. In:18th IEEE International Symposium on Software Reliablity.2007.p 237-246.

[2]. Emam KE, Melo W, Machado JC. The prediction of faulty classes using object oriented designmtrics. Journal of System & Software 2001;p 63-75

[3]. Sandhu PS et.al A model for early prediction of faults in Software systems. In : 2nd International conference on Computer & Automation engineering. 2010.p.281-285.

[4]. Yanmin Suna,∗ , Mohamed S. Kamela, Andrew K.C.Wongb,YangWangc "Cost-sensitive boosting for classification of imbalanced data" in pattern recognition society, Elsevier publication,2007.

[5]. Yang Yi, Jiansheng Wu., Wei Xu "Incremental SVM based on reserved set for network intrusion detection" in Expert Systems with Applications 38 ,2011.

[6]. Qinglei Zhang,Wenying Feng, "Network Intrusion Detection by Support Vectors and Ant Colony" in Proceedings of the 2009 International Workshop on Information Security and Application"

[7]. S.V.Shirbhate, Dr V.M.Thakare & Dr S.S.Sherekar "Data Mining Approaches For Network Intrusion Detection System" in International Journal of Computer Technology and Electronics Engineering (IJCTEE)

[8]. Rukshan Batuvita, vasile palade, "FSVM-CIL Fuzzy support vector machines for class imbalance learning" in IEEE Transactions in Fuzzy systems (2010)

[9]. Jang, Tsai sun and Mijutani " Supervised learning Neural networks " Neuro-Fuzzy soft computing, pp. 231-241, 1997.

[10]. Jang, Tsai sun and Mijutani "ANFIS: Advanced Neuro Fuzzy Inference system" Neuro-Fuzzy modelling, pp. 335-340, 1997.

[11]. Mohammad Alshayeb, Lahouari Ghouti,"software Defect Prediction using Ensemble Learning on Selected Features", Information and Software Technology 58(2015)pp 388-402.

[12]. Romi Satria Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks" Journal of Software Engineering 1(2015).

[13]. Romi Satria Wahono, Nnna Suryana, Sabrina Ahmad, "Metaheuristic Optimization based Feature Selection for Software Defect Prediction", Journal of Software Engineering, 9(2014).