Software Fault Estimation using Fuzzy C-Means and Neuro-Fuzzy Classification

Jyoti Nagpal
M. Tech. Scholar, Computer Science Dept.
Poornima college of Engineering, Jaipur (Rajasthan.), India
jyotibhugra2010@gmail.com

Dr. Ajay Khuteta
Associate professor, Computer Science Dept.
Poornima college of Engineering, Jaipur (Rajasthan.), India
khutetaajay@poornima.org

Abstract — The fault-proneness of a software module is the probability that the module contains faults and a software fault is a defect that causes software failures in an executable project. Early detection of fault prone software components enables verification experts to concentrate their time and resources on the problem areas of the software systems under development. In this paper, the performance comparison of a Software Fault Prediction System is done using two techniques; the first technique is Fuzzy C-Means clustering and another one is hybrid method which is combination of Fuzzy c-means clustering and Neural Networks approach (Neuro-Fuzzy). Both of the methods have been performed with the real time data set named PC1, taken from NASA MDP software projects. The performance is recorded on the basis of accuracy, net reliability, and RMSE and MAE values.

Keywords – Fault-Proneness, Fuzzy C-Means, Neural Networks, Neuro-Fuzzy, NASA MDP.

I. INTRODUCTION

Faults are major problem in software systems that need to be resolved. Fault is a flaw that results in failure. We should have to know the clear difference between bug, fault and failure. Failure is deviation of software actions from the expected outcomes. A fault in software is a flaw that results in failure. Bug occurs when specified requirements of the software do not conform. There are many number of software having number of faults are delivered to the market [1].

A fault is a defect, an error in source code that causes failures when executed. A fault prone software module is the one containing more number of expected faults. Accurate prediction of fault prone modules enables the verification and validation activities focused on the critical software components.

A software fault is a defect that causes software failure in an executable product. For each execution of the software program where the output is incorrect, we observe a failure. Software engineers distinguish software faults from software failures.

Faults in software systems continue to be a major problem. Various systems are delivered to users with excessive faults. This is despite a huge amount of development effort going into fault reduction in terms of quality control and testing. It has long been recognized that seeking out fault-prone parts of the system and targeting those parts for increased quality control and testing is an effective approach to fault reduction. An inadequate amount of valuable work in this area has been carried out previously. Regardless of this it is difficult to identify a reliable approach to identifying fault-prone software components. Using software complexity measures, the techniques build models, which categorize components as likely to contain faults or not.

Till now there are proposed numerous methods for data clustering methods. The algorithms provide a satisfying measure for the classification and mining of data. The software fault prediction is also now using the data clustering techniques because of the features and the functions they are expected to deliver. The clustering techniques till now have solved many purposes yet the satisfying result could not be guaranteed. In this research work, we have tried to modify the previous algorithms for the better results. We do not say that it is the end of research in this segment but it will definitely provide the new researchers with the scope to bring new considerations that could serve the future demands.

The main objective of this paper is to design a Software Fault Prediction System using Fuzzy c-means clustering approach and a hybrid technique (combination of Fuzzy c-means clustering and Neural Networks approach). The results after classification of software fault data come in terms of certain efficiency parameters like Accuracy, Net Reliability, Mean Absolute Error, and Root Mean Squared Error in order to compare both approaches.
II. SOFTWARE FAULT PREDICTION

A software system is the composition of number of modules dependent on each other. Any module with faults in its functionality adverse the output and lowers its reliability. In this scenario, the detection of faulty modules in early stage (development stage) is mandatory to minimize faults in operation phase. Hence, the systems are classified in two categories i.e. with faulty/non-faulty modules in their testing phase. This classification diverts the focus to neutralize faulty sections to achieve high reliability and accuracy. A software fault or error is reason of failure in execution stage. The error message at each stage of executing the program indicates the fault in programming. Generally speaking the errors are the logical errors in programming of software. Out of the heavy researches explored in this section a limited amount of valuable work is illustrated in literature section. The prediction models of software fault proneness technique estimate the amount of faulty modules in a program. The software metrics are the attributes for process, execution and product of the software system. Various other attributes like defect density, normalized work, fault proneness, maintain ability, reusability etc. determines the quality of software. Different fault prediction techniques are stated below:

Decision Tree: Decision trees are great and standard tools for classification and prediction. It produces classifiers in a form of structure of tree where each leaf node represents decision node. In this technique, classification starts from root of the tree and continues to move down until leaf node is reached. Classification helps in classifying faulty and non-faulty modules. Prediction helps in predicting faulty and non-faulty modules. Decision tress helps in developing fault prediction models that predicts faults.

Neural Network: Neural Network helps in recognizing patterns from the data set. An artificial neural network is composed of many artificial neurons that are interconnected together according to specific network architecture. The goal of the neural network is to transform the inputs into meaningful outputs. Adaptive Resonance Neural Network is generally used for defect prediction in software systems. It helps in identifying faulty modules very excellently. The benefit of using this technique is that it assists in decreasing effort and cost of developing software.

Clustering approach: Density Based Clustering is a clustering algorithm. It can be used to estimate the number of faulty and non-faulty modules in software system. Clusters are defined as areas of higher density.

Bagging method: It creates base learners on many data subsets that are uniformly sampled from the original data, and then uses a linear combination to aggregate them. It is also referred as Bootstrap Aggregating. Combination technique can be majority voting. It also helps in identifying faulty and non-faulty modules with data sets that suffers from imbalance problem. This method can increase the performance of the defect data predictions.

Naïve Bayes: It is a classifier based on Bayes theorem used in software fault prediction. It resolves the several difficulties like spam classification (to predict whether email is spam or not), medical diagnosis (given list of symptoms, predict whether patient has cancer or not) and so on. This method can be used to predict faulty and non-faulty modules.

III. METHODOLOGY

1. Find the structural code and requirement attributes

The first step is to find the structural code and requirement attributes of software systems i.e. software metrics. The real time defect data sets are taken from the NASA’s MDP (Metric Data Program) data repository, [online] Available: http://mdp.ivv.nasa.gov.in named as PC1 dataset which is collected from a flight software from an earth orbiting satellite coded in C programming language, containing 1107 modules and only 109 have their requirements specified. PC1 has 320 requirements available and all of them are associated with program modules. All these data sets varied in the percentage of defect modules, with the PC1 dataset containing the least number of defect modules.

2. Select the suitable metric values as representation of statement

The Suitable metric values used are fault and without fault attributes, we set these values in database create in MATLAB R2010 A as 0 and 1. Means 0 for data with fault and 1 for data without fault. The metrics in these datasets (NASA MDP dataset) describe projects which vary in size and complexity, programming languages, development processes, etc. When reporting a fault prediction modelling experiment, it is important to describe
the characteristics of the datasets. Each data set contains twenty-one software metrics, which describe product’s size, complexity and some structural properties. We use only fault and without attributes to classify the selected NASA MDP PC1 dataset. Also the product metrics and product module metrics available in dataset which can also be use are the product requirement metrics are as follows:

- Module
- Action
- Conditional
- Continuance
- Imperative
- Option
- Risk_Level
- Source
- Weak_Phrase

The product module metrics are as follows:

1. Module
2. Loc_Blank
3. Branch_Count
4. Call_Pairs
5. LOC_Code_and_Comment
6. LOC_Comments
7. Condition_Count
8. Cyclomatic_complexity
9. Cyclomatic_Density
10. Decision_Count
11. Edge_Count
12. Essential_Complexity
13. Essential_Density
14. LOC_Executable
15. Parameter_Count
16. Global_Data_Complexity
17. Global_Data_Density
18. Halstead_Content
19. Halstead_Difficulty
20. Halstead_Effort
21. Halstead_Error_EST
22. Halstead_Length
23. Halstead_Prog_Time
24. Halstead_Volume
25. Normalized_Cyclomatic_Complexity
26. Num_Operands
27. Num_Operators
28. Num_Unique_Operands
29. Num_Unique_Operators
30. Number_Of_Lines
31. Pathological_Complexity
32. LOC_Total

Figure 1 and 2 show flow diagrams for Fuzzy c-means clustering approach and hybrid (Neuro-Fuzzy) approach respectively.

Initially data is passed to Fuzzy C-Means clustering block in order to classify data according to attributes (true / false attributes)

The classified data from Fuzzy C-Means clustering block is given to Neural Network block to train neural network for these attributes in order to classify any fault in the system.

In this paper we have developed a software fault prediction module using two methods:

- Fuzzy c-means clustering (FCM) approach.
Hybrid technique or Neuro-Fuzzy (Fuzzy c-means clustering (FCM) + Neural Network) technique.

PC1 software fault database is used available at NASA’s research website. In the first method Fuzzy c-means clustering approach is used to detect any fault present in the data. In the hybrid method, Fuzzy c-means clustering approach is used to classify data according to attributes. The classified data from Fuzzy C-Means clustering is trained by neural network in order to predict any fault in the system.

Fuzzy C-Means Clustering

Fuzzy C-Means iteratively moves the cluster centers to the “right” location within a data set. Objective function based fuzzy clustering algorithms such as the fuzzy c-means (FCM) algorithm has been used extensively for different tasks such as pattern recognition, data mining, and image processing and fuzzy modeling.

In general, cluster analysis refers to a broad spectrum of methods which try to subdivide a data set X into c subsets (clusters) which are pairwise disjoint, all nonempty, and reproduce X through union. The clusters then are termed a hard (i.e., non-fuzzy) c-partition of X.

Parameters of the FCM Algorithm

Number of Clusters: The number of clusters c is the most important parameter, in the sense that the remaining parameters have less influence on the resulting partition. When clustering real data without any a priori information about the structures in the data, one usually has to make assumptions about the number of underlying clusters. The chosen clustering algorithm then searches for c clusters, regardless of whether they are really present in the data or not. Two main approaches to determining the appropriate number of clusters in data can be distinguished:

A. Validity measures: Validity measures are scalar indices that assess the goodness of the obtained partition. Clustering algorithms generally aim at locating well-separated and compact clusters. When the number of clusters is chosen equal to the number of groups that actually exist in the data, it can be expected that the clustering algorithm will identify them correctly. When this is not the case, misclassifications appear, and the clusters are not likely to be well separated and compact. Hence, most cluster validity measures are designed to quantify the separation and the compactness of the clusters.

B. Iterative merging or insertion of clusters: The basic idea of cluster merging is to start with a sufficiently large number of clusters, and successively reduce this number by merging clusters.

Fuzziness Parameter: The weighting exponent m is a rather important parameter as well, because it significantly influences the fuzziness of the resulting partition.

Termination Criterion: The FCM algorithm stops iterating when the norm of the difference between U in two successive iterations is smaller than the termination parameter ϵ. For the maximum norm \( \max_{i,k}(\mu_{i,k}^{(t)} - \mu_{i,k}^{(t-1)}) \). The usual choice is \( \epsilon = 0.001 \), even though \( \epsilon = 0.01 \) works well in most cases, while drastically reducing the computing times.

Norm-Inducing Matrix: The shape of the clusters is determined by the choice of the matrix A in the distance measure. A common choice is \( A = I \), which gives the standard Euclidean norm:

\[
D_{ik}^2 = (z_k - v_i)^T (z_k - v_i) \tag{2}
\]

Where \( v_i \) are ordinary means of the clusters.

Let \( \{x_1, x_2, ..., x_N\} \) be a set of N data objects represented by n-dimensional feature vectors.

\[
x_k = [x_{1k}, ..., x_{nk}]^T \in \mathbb{R}^N \tag{3}
\]

A set of N feature vectors is then denoted as a data matrix of \( n \times N \).

\[
X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix} \tag{4}
\]

A fuzzy clustering algorithm partitions the data X into M fuzzy clusters, forming a fuzzy partitioning. A fuzzy partition can be conveniently represented as a matrix, \( U \), whose elements \( u_{ik} \in [0; 1] \) represents the membership degree of \( x_k \) in cluster ‘i’. Hence, the \( i^{th} \) row of \( U \) contains values of the \( i^{th} \) membership function in the fuzzy partition. Objective function based fuzzy clustering algorithms minimize an objective function of the type

\[
J(X; U, V) = \sum_{i=1}^{M} \sum_{k=1}^{N} (u_{ik})^m d^2(x_k, v_i) \tag{5}
\]

Where,

\[
V = [v_1, ..., v_M]^T \in \mathbb{R}^N \tag{6}
\]
is an M-tuple of cluster prototypes which have to be determined, and \( m \in (1; \infty) \) is a weighting exponent which determines the fuzziness of the clusters in order to avoid the trivial solution, constraints must be forced on \( U \).

\[
\sum_{i=1}^{M} u_{ik} = 1, \forall k \quad (7)
\]
\[
0 < \sum_{i=1}^{N} u_{ik} < N, \forall i \quad (8)
\]

These constraints imply that the sum of each column of \( U \) is 1. Further, there may be no empty clusters, but the distribution of membership among the \( M \) fuzzy subsets is not constrained. The prototypes are typically selected to be idealized geometric forms such as linear varieties (e.g. FCV algorithm) or points (e.g. GK or FCM algorithms). When point prototypes are used, the general form of the distance measure is given by

\[
d^2(x_k, v_i) = (x_k - v_i)^T A_i (x_k - v_i) \quad (9)
\]

Where the norm matrix \( A_i \) is a positive definite symmetric matrix. The FCM algorithm uses the Euclidian distance measure, i.e. \( A_i = I \), while the GK algorithm uses the Mahalonibisdistance, i.e. \( A_i = P_i^{-1} \), with \( P_i \) the covariance matrix of cluster \( i \), and the additional volume constraint \( |A_i| = \rho_i \).

The FCM algorithms are best described by recasting conditions in matrix-theoretic terms [3]. Towards this end, let \( U \) be a real \( c \times N \) matrix, \( U = [u_{ik}] \). \( U \) is the matrix representation of the partition \{ \( Y_i \) \} in the situation

\[
u_i(y_k) = u_{ik} = \begin{cases} 1; & y_k \in Y_i \\ 0; & \text{otherwise} \end{cases} \quad (10)
\]

\[
\sum_{i=1}^{M} u_{ik} > 0 \quad \text{for all } i \quad (11)
\]
\[
\sum_{i=1}^{M} u_{ik} = 1 \quad \text{for all } k \quad (12)
\]

In equation (10), \( u_i \) is a function such that: \( u_i: Y \rightarrow \{0, 1\} \). In conventional models, \( u_i \) is the characteristic function of, \( Y_i \); in fact, \( u_i \) and \( Y_i \) determine one another, so there is no harm in labelling \( u_i \); the \( i^{th} \) hard subset of the partition (It is unusual, of course, but is important in terms of understanding the term “fuzzy set”). Conditions of equations (11) and (12) are equivalent, so \( U \) is termed a hard \( c \)-partition of \( Y \). Generalizing this idea, we refer to \( U \) as a fuzzy \( c \)-partition of \( Y \) when the elements of \( U \) are numbers in the unit interval \([0, 1]\) that continue to satisfy both equations (11) and (12). The basis for this definition are \( c \) functions \( u_i: Y \rightarrow [0, 1] \) whose values \( u_i(y_k) \in [0, 1] \) are interpreted as the grades of membership of the \( y_k \)'s in the “fuzzy subsets” \( u_i \) of \( Y \).
Transfer Function

The whole behavior of Neural Network totally depends on both the weights and the input-output function i.e. transfers function which is specified in the all units. There are basically three categories of Transfer Functions:

- Linear (or ramp)
- Threshold
- Sigmoid

For linear units or for the linear transfer function, the output activity is directly proportional to the total weighted output units. For threshold units or for threshold transfer function, the output unit outputs are set at one of two levels, which totally depending on whether the total input of output unit is greater than or less than some predefined threshold value. For sigmoid units or for sigmoid transfer function, the output varies or changes continuously but not linearly as the inputs of input unit changes.

IV. Simulation and Results

Simulation is carried out using MATLAB 2010a:

- Figure 4: Graphical User Interface (GUI) for proposed work
- Figure 5: Input PC1 dataset with attributes (fault and without fault)
- Figure 6: Input PC1 dataset with fault attributes when separating fault attributes from input data
- Figure 7: Input PC1 dataset with without fault attributes when separating without fault attributes from input data
- Figure 8: Data Input (PC1 database without attributes)
Table 1: Performance comparison for Fuzzy c-means and proposed (Adaptive-Neuro-Fuzzy) technique

<table>
<thead>
<tr>
<th>Evaluation Parameter</th>
<th>Fuzzy C-Means Clustering</th>
<th>Hybrid (Adaptive Neuro-Fuzzy) technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>75</td>
<td>87</td>
</tr>
<tr>
<td>Net Reliability</td>
<td>60.07</td>
<td>47.20</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>0.25</td>
<td>0.13</td>
</tr>
<tr>
<td>Root Mean Squared Error (RMSE)</td>
<td>0.0833</td>
<td>0.0194</td>
</tr>
</tbody>
</table>

The disadvantage of the methods based on genetic algorithms is that the number of fuzzy sets must be informed. Genetically optimized fuzzy clustering method [9] have high computational complexity and large time spending.

V. CONCLUSION

In this paper, a software Fault Prediction System is implemented using Fuzzy C-Means clustering and hybrid (Neuro-Fuzzy) techniques. Performance of both the techniques is discussed in the comparative analysis in order to predict level of impact of faults in NASA’s public domain defect dataset. Predicting faults in the software life cycle can be used to improve software process control and achieve high software reliability. On the basis of evaluation parameters, it was found that the hybrid technique shows more accuracy and less errors as compared to Fuzzy C-Means clustering method.

REFERENCES


