Software Fault Prediction based on CURE Clustering Algorithm and Artificial Intelligence

Margi Patel, Upama Vachhani
Computer Engineering,
Silver Oak College of Engineering & Technology, Ahmedabad, Gujarat, India

***Abstract*– There has been rapid growth of software development. During Transmission of Data faults are Created. However software Fault Prediction Techniques are used to Detect Fault. Software Fault Prediction improve the quality and reliability of software by predicting faults .Quality of Software measure in term of fault proneness of data .These software defect may lead to degradation of the quality which might be the cause of failure. We show a comparatively analysis of software fault prediction based on clustering technique, neural network method, statistical method. Fault prediction reduce the overall time and less data processing. In this paper, hybrid approach based on CURE clustering and Neural Network based approach has 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, MAE, RMSE values. This paper focus on clustering with large dataset and predicting faults efficiently.**

***Index Terms*- CURE Clustering, Neural Network, Fault Prediction**

I. INTRODUCTION

SOFTWARE quality and reliability are main concerns in modern era.It is widely accepted that software with defects lacks quality.Real time software application and complex software systems demands high quality.A software system contain many modules and any of these can contain faults[5].

Clustering is a division of data into groups of similar objects. Each group called cluster consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Classification and prediction that can be used to extract models describing significant defect data classes or to predict future defect trends. Classification predicts categorical or discrete, and unordered labels, whereas prediction models predict continuous valued functions. Such analysis can help us for providing better understanding of the software defect data at large.

The underlying software engineering assumption is that the faultprone software modules will have similar software measurements, and hence are likely to be grouped together in the same cluster(s). Similarly, the not fault-prone modules will likely be grouped in the same cluster[2]. Clustering is an approach that uses software measurement data consisting of limited or no fault-proneness data for analyzing software quality[3].Clustering algorithms are being successfully applied for solving both classification and regression problems. It is therefore important to investigate the capabilities of this algorithm in predicting software quality[10]. Early prediction of software fault at coding phase can result in decrease cost and effort for software development. So, it is better to categorize the software module in faulty / non -faulty module just after completing the coding phase. module contain error derived as fault prone module.

A software fault is a defect that causes software failure in an executable product. In software engineering, the nonconformance of software to its requirements is commonly called a bug. Software Engineers distinguish between software faults, software failures and software bugs. In case of a failure, the software does not do what the user expects but on the other hand fault is a hidden programming error that may or may not actually manifest as a failure and the non-conformance of software to its requirements is commonly called a bug[3].

II. RELATED WORK

Software fault prediction uses historical and development data to identify fault in software. Various techniques have been applied for software

fault prediction like partional clustering, hierarchical clustering, neural network, naive bayes, support vector machine and many more.

Other papers include research work using various classification technique as Dhankhar, Swati, Himani Rastogi, and Misha Kakkar Proposed Software Fault Prediction Performance using Bayesian network,Naïve bayes,Neural Network.By using this method improve software quality and testing efficiency by early identification of fault.Neural Network classification model are more superior to other network model[1].

Gupta, Deepika, Vivek K. Goyal, and Harish Mittal Proposed Estimating of Software Quality with Clustering Techniques. this paper focus on clustering with very large dataset and very many attribute of different types. Effective result can be produced by using fuzzy c-mean clustering[2] .

Kaur, Arashdeep, Parvinder S. Sandhu, and Amanpreet Singh Bra proposed Early software fault prediction using real time defect data.Predicting fault early in software life cycle can be used to improve software process control and achieve high software reliability.best prediction model is fusion of requirement and code metric model[3].

Kaur, Arashdeep, Amanpreet Singh Brar, and Parvinder S. Sandhu Proposed An empirical approach for software fault prediction. In this paper we investigate that whether the metrics available in the early lifecycle can be used to predict fault prone area or not. fuzzy c mean is better than k-mean in in case of requirement and combination metric model[4].

Shyna Kakkar, Amanpreet Singh Dhanoa Proposed Software Fault Prediction using hybrid k-mean feed forward neural network. this paper used hybrid approach to predict faults in software system .k-mean feed forward neural network has better accuracy than fuzzy cmeans feed forward neural network. It can help in directing testeffort,reducing cost,increase quality of software and its reliability[5]

III. CURE CLUSTERING

CURE (Clustering Using REpresentatives) is an efficient data clustering algorithm for large databases that is more robust to outliers and identifies clusters having non-spherical shapes, size and densities[7].

CURE is a hierarchical clustering algorithm for large datasets proposed by Guha, Rastogi and Shim in 1998. This algorithm is agglomerative hierarchical approach. CURE employ a novel hierarchical clustering algorithm that adopt a middle ground between the centroid based and the all-point extremes. CURE can identify non-spherical shaped clusters and wide variances in size with the help of well scattered representative points and centroid shrinking. CURE can handle large databases by combining random sampling and partitioning method.

Cure combines centroid and single linkage approaches by choosing more than one representative points from each cluster. At the end of each step, the clusters with the closest representative points are clustered together. Cure represents each cluster by a fixed number of points that are generated by selecting well scattered points from the cluster, then shrink them toward the center of the cluster by a specified faction. This enables CURE to correctly identify the clusters and makes it less sensitive to outliers. We cannot apply this algorithm directly to large datasets, instead we have to apply random sampling, partitioning for speedup – the advantage of partitioning the inputs is to reduce the execution time, labeling on disk.

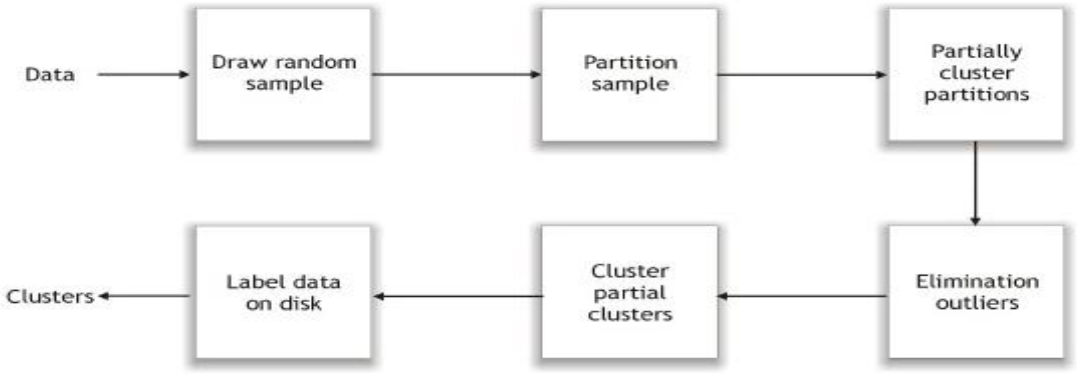

Fig 1: Architecture of CURE [7]

IV. ARTIFICIAL NEURAL NETWORK

Artificial Neural Network is used to obtain Proper Recognition procedure due to its ability to learn from feature data[8].

A feed forward neural network is an artificial neural network where connections between the units do not form a directed cycle. This is different from recurrent neural networks.

The feed forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network.

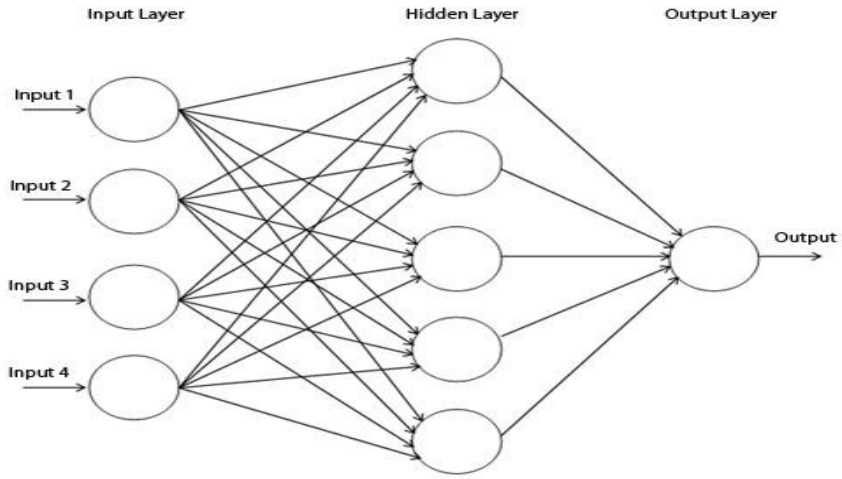

Fig 2: Feed forward Neural Network [8]

V. PROPOSED METHOD

In this paper we are going to develop a software fault prediction model based on CURE data clustering Algorithm and Feed-Forward Neural Network based approach, we are going to use NASA PC1 software fault database available Nasa's research website. First, we classify the large dataset using CURE data clustering Algorithm, these classified results and inputs are further pass through the neural network as an input and output in order to train neural network in order to make an efficient and supervised classification model. Implement the model and test the performance of the model using following criteria:

- Perform the training of the dataset.
- After training, test it on the basis of error values MAE and RMSE, and efficiency parameters like accuracy and net reliability in percentage.

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, 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 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. Each data set contains twenty-one software metrics, which describe product's size, complexity and some structural properties. The product metrics and product module metrics available in dataset.

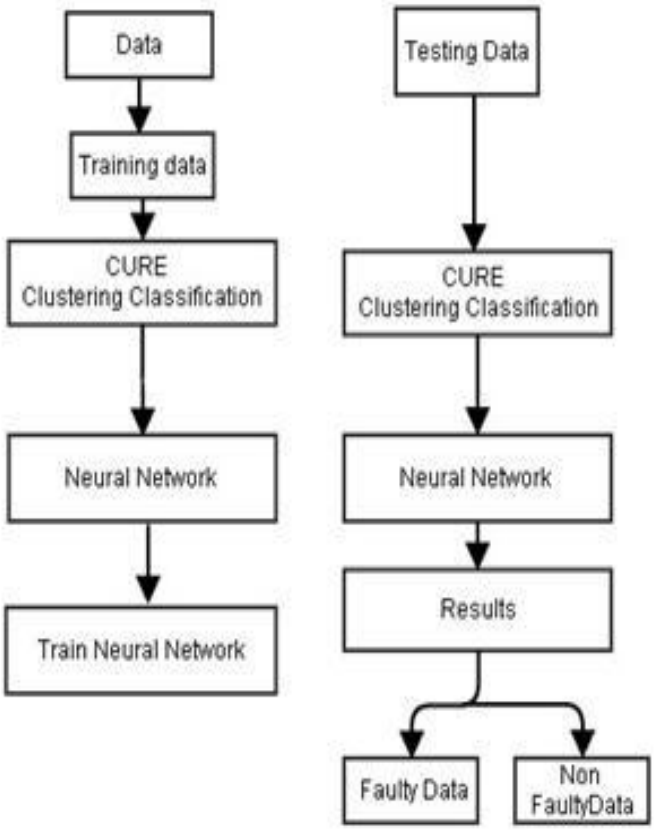

Fig 3: Proposed Flow Diagram

VI. RESULTS AND DISCUSSION

The implementation of "Software Fault Prediction based on CURE Clustering Algorithm and Artificial Intelligence" Using CURE Algorithm and Neural Network is implemented in MATLAB.

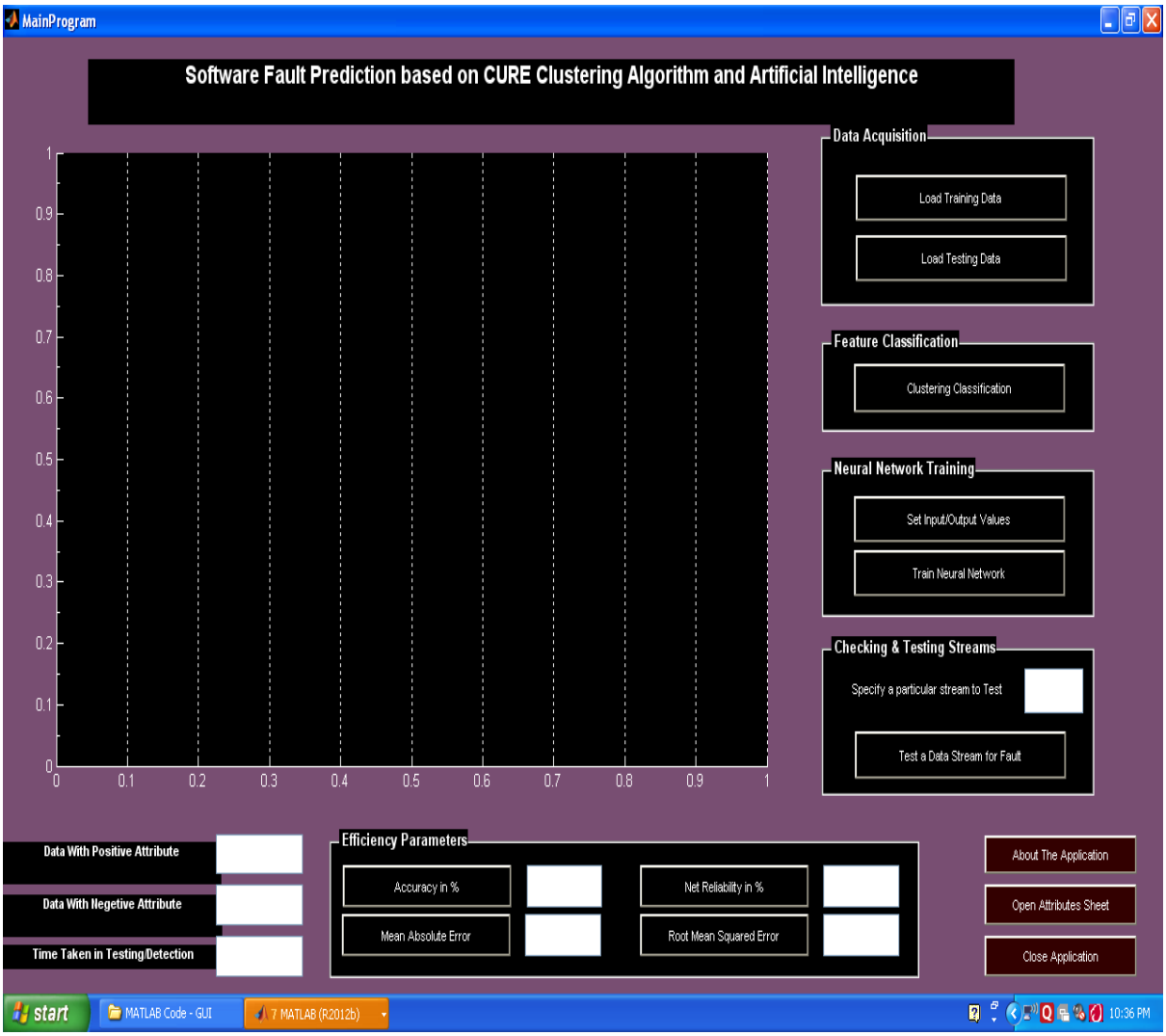

Fig 4: Graphical User Interface for Proposed work

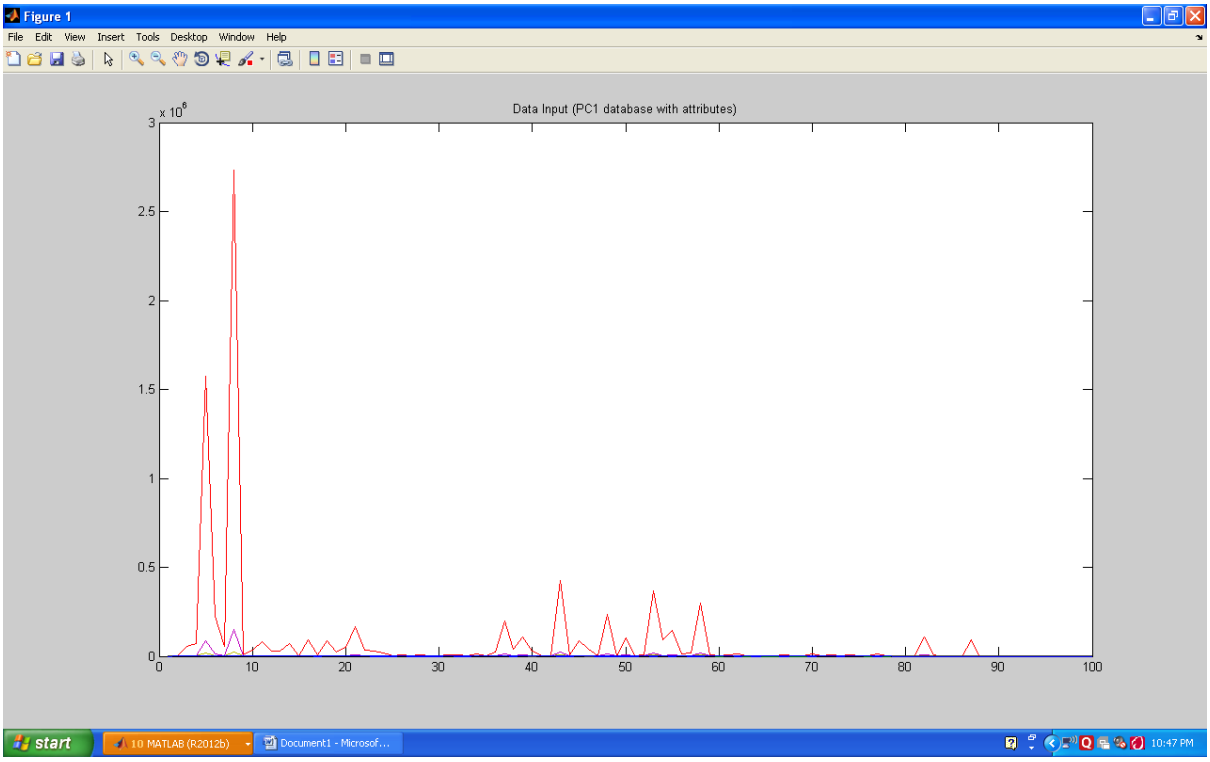

Fig 5: Input PC1 dataset with attributes (faults and without fault)

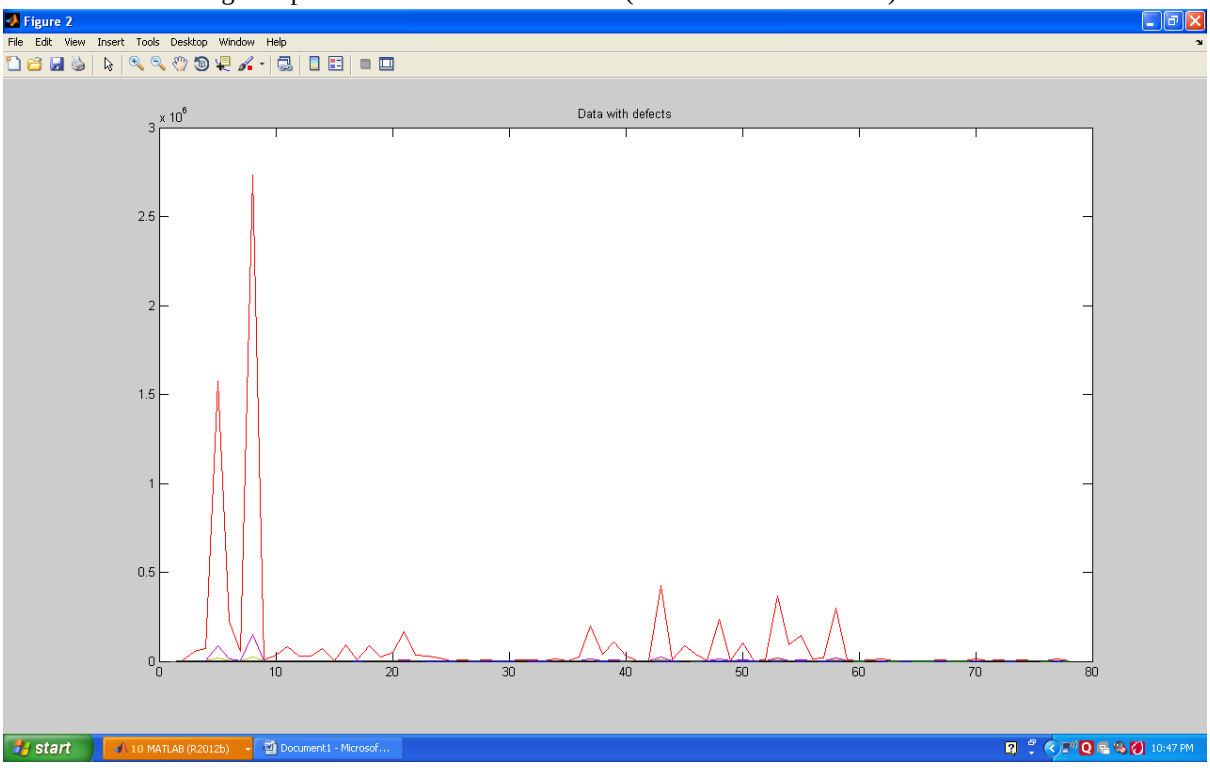

Fig 6: Input PC1 dataset with fault attributes when separating fault attribute from input data

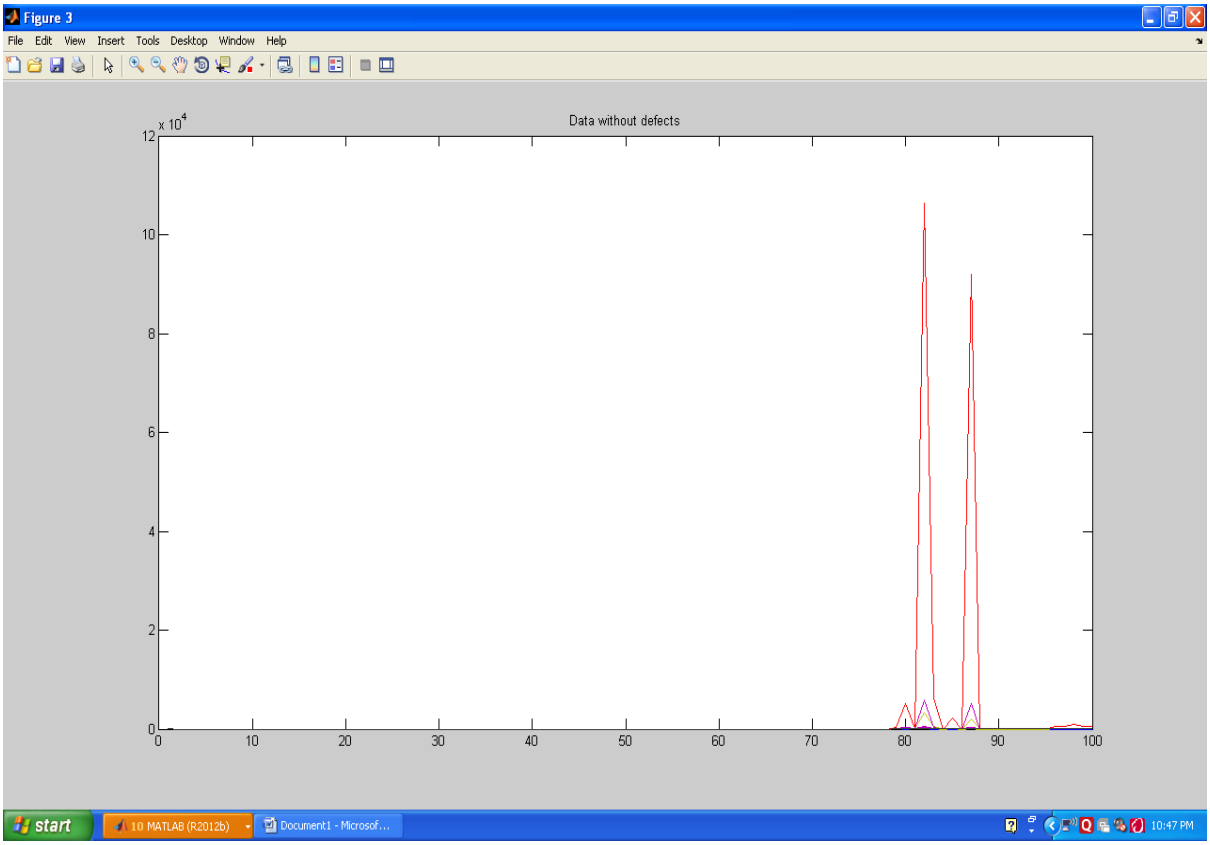

Fig 7: Input PC1 dataset without fault attributes when separating without fault attribute from input data

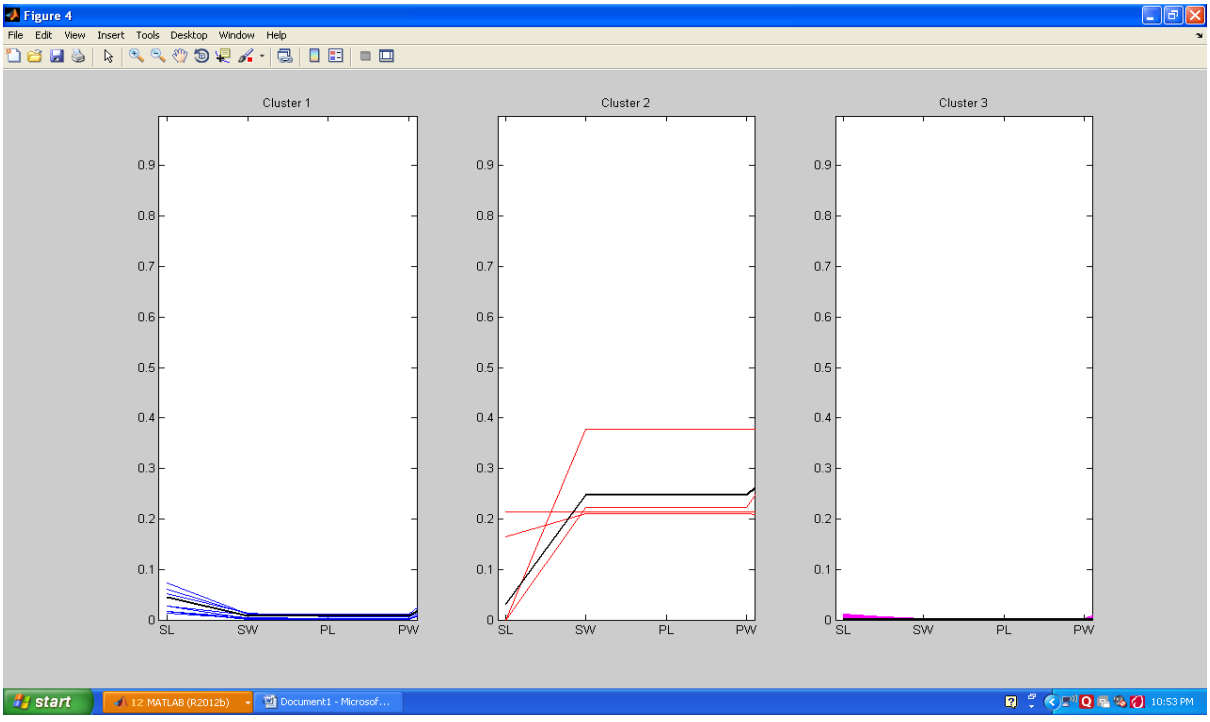

Fig 8: Grouping of different data values in clusters

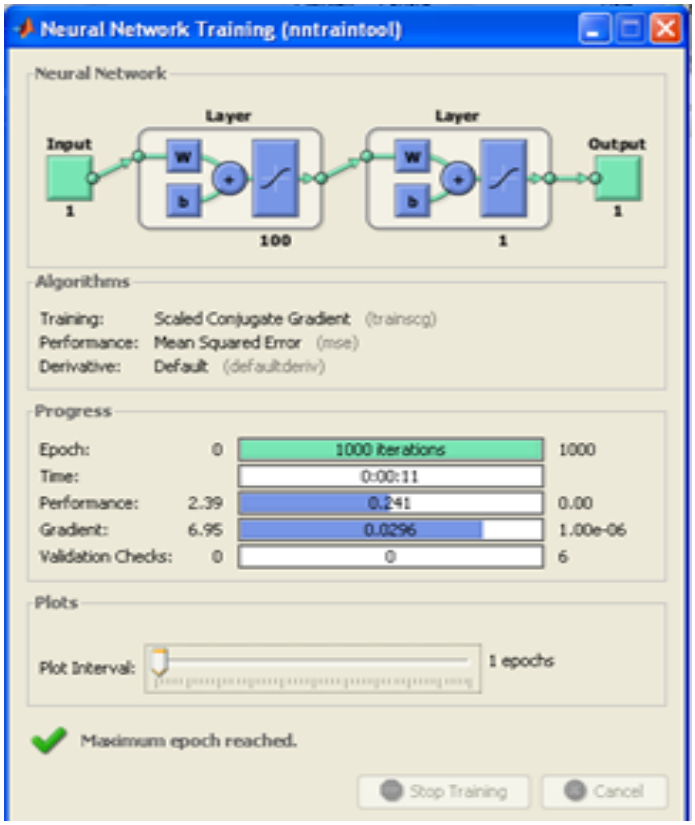

Fig 9 : Training of Neural Network

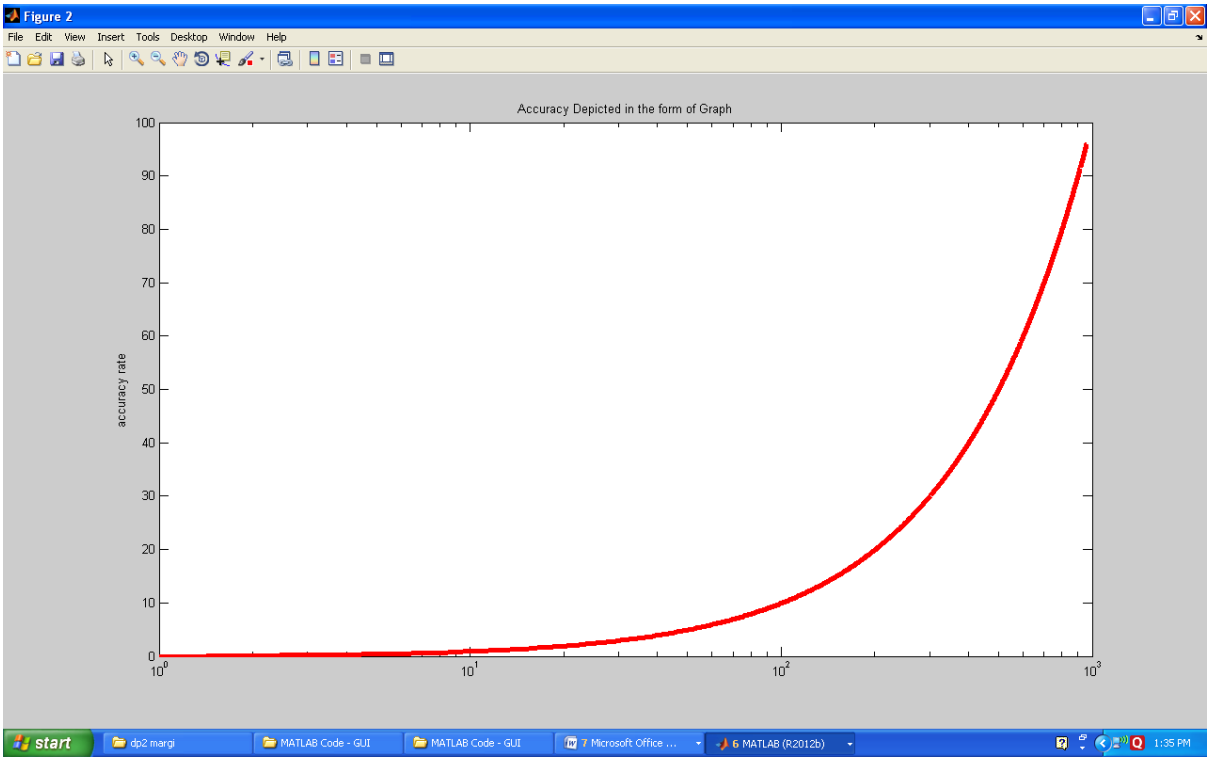

Fig 10: Accuracy graph of Hybrid approach

Table 1: Computed Result for Hybrid Approaches

Technique	Accuracy %	MAE	RMSE	Reliability %
Neural Network	66.2338	0.337662	0.414899	70.8683
Hybrid Approach	96.1039	0.038961	0.0397429	92.0478

VII. CONCLUSION

In this paper, a Software Fault Prediction System is implemented using CURE clustering and Neural Network Techniques. A variety of software fault prediction technique have been proposed,but none has proven to be consistently accurate. We used the training and testing methodology.By analyzing the results it is clear that Hybrid approach based on CURE clustering and neural network gives more accuracy and less error as compared to Neural Network on the basis of evaluation parameters: accuracy, reliability,MSE and RMSE.

REFERENCE

[1] Dhankhar, Swati, Himani Rastogi, and Misha Kakkar. "SOFTWARE FAULT PREDICTION PERFORMANCEIN SOFTWARE ENGINEERING.",IEEE-2015.

[2] Gupta, Deepika, Vivek K. Goyal, and Harish Mittal. "Estimating of Software Quality with Clustering Techniques." Advanced Computing and Communication Technologies (ACCT), 2013 Third International Conference on. IEEE, 2013.

[3] Kaur, Arashdeep, Parvinder S. Sandhu, and Amanpreet Singh Bra. "Early software fault prediction using real time defect data." Machine Vision, 2009. ICMV'09. Second International Conference on. IEEE, 2009.

[4] Kaur, Arashdeep, Amanpreet Singh Brar, and Parvinder S. Sandhu. "An empirical approach for software fault prediction." Industrial and Information Systems (ICIIS), 2010 International Conference on. IEEE, 2010.

[5] Shyna Kakkar, Amanpreet Singh Dhanoa." Software Fault Prediction using hybrid k-mean feed forward neural network" IJCSEE-2015.

[6] PSO Optimized Software Fault Prediction system using Fuzzy C –Means,IJDACR- Volume 3, Issue 6, January 2015.

[7] Mary, Sandra Sagaya. "A Study of K-Means and Cure Clustering Algorithms." International Journal of Engineering Research and Technology. Vol. 3. No. 2 (February-2014). ESRSA Publications, 2014.

[8] Gayathri, M., and A. Sudha. "Software Defect Prediction System using Multilayer Perceptron Neural Network with Data Mining." International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878.

[9] Mahajan, Er Rohit, Dr Sunil Kumar Gupta, and Rajeev Kumar Bedi. "Comparison of Various Approaches of Software Fault Prediction: A Review."International Journal of Advanced Technology & Engineering Research (IJATER) (2014).

[10] Kaur, Simranjit, Manish Mahajan, and Dr Parvinder S. Sandhu. "Identification of Fault Prone Modules in Open Source Software Systems using Hierarchical based Clustering." ISEMS, Bangkok, July (2011).