International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference & workshop on Advanced Computing 2013 (ICWAC 2013) – www.ijais.org

Software Quality Analysis with Clustering Method

P.V.Ingle
Dept of Computer Engineering
A.C.Patil College of Engineering
Kharghar, Navi Mumbai

M.M.Deshpande
Dept.of Computer Engineering
A.C.Patil College of Engineering
Khargher, Navi Mumbai

ABSTRACT

Software development team tries to increase the software quality by decreasing the number of defects as much as possible. A major concern for managers of software project are the triple constraint of cost, schedule and quality due to the difficulties to quantify accurately the trade-off between them .number of defects remaining in a system provides an insight into the quality of the system. Software defects are one of the major factors that can decide the time of software delivery. The proposed system will analyze the software defects. We are trying to categorize the software defects using some clustering approach and then the software defects will be measured in each clustered separately.

Keywords

Software Defect, Quality, Clustered, Kmeans, Cmeans

1. INTRODUCTION

Software defects are not only an inherent component of software product but also significant factors of software Quality. It is impossible to completely eliminate software defects .People can attempt to avoid or minimize defects in software development as much as possible. To deliver a defect free software it is imperative to predict and fix the defects as many as possible before the product delivers to the customer.

In order to take full advantage of a large number of defects data collected in the software development process, improve the efficiency and quality of software testing .This paper applied data mining technology in the analysis of defects data based on its advantages in data processing. The process of grouping a set of objects into classes of similar objects is called Clustering. By clustering one can identify dense and sparse region and overall distribution pattern. Clustering technique categorize the defects correction effort is used as the criterion for classifying the defects thereby providing the project managers with general idea about the nature of complexity of defects. This Clustering technique helps the project managers to understand about the nature of defects and allocate resources [4].

This paper is organized as follows: In section 2 we discuss about Software Defect. It involves Software defect management and Software defect prediction and according to this we describe various types of defects with description in section 3.section 4 describe Data Mining Techniques as Classification, Association, and Clustering.

2. SOFTWARE DEFECT

2.1 Software Defect Definition

A software defect is an error, flaw, mistake, failure or fault in a computer program or system that produces an incorrect or unexpected result or causes it to behave an unintended ways[3] Defects , like quality can be defined as deviations from specification or expectation which might lead to failure in operation. In despite of small program or large scale software system, they all have defects some of these defects can be found easily, some conceal deeply hard finding and some could make enormous loss of property even lives. In order to guarantee software quality we must effectively predict defects which exist in software.

2.2 Software Defect Management

The main goal of defect management is to increase the software quality by finding and fixing the defects as early as possible. Software defect Management is a process which tracks defect found in software develop process and ensure these defects to be closed.Inorder to investigate the defects and to find way to fix them, certain information needs to be recorded for each defect. According to Fenton [2], a number of key attributes are needed in problem reports on defects for measurement purpose, such as location timing and severity.

2.3 Software Defect Prediction

In the domain of software defect prediction, people have developed many software defect prediction Model. These models are mostly described in two classes. One class is in the later period of the software life cycle (testing phase), the other class is before developing the software, predicts how many defects will be in the software develop process by analyzing on defect data in formerly projects.

Software defect prediction is the process of locating defective modules in software to produce high quality software; the final product should have as few defects as possible. Early detection of software defects could lead to reduced development costs and rework effort and more reliable software. So the study of the defect prediction is important to achieve software quality [4].

3. DEFECT TYPE DESCRIPTION

1. Computational Defect

Computational defects are those that cause a computation to erroneously evaluate a variable’s value. These defects could be equation that are incorrect not because of the incorrect use of a data structure within the statement but by miscalculation.

2. Data Value Defect

Data value defects are those that are a result of the incorrect use of a data structure. Example of this type of defects errors are the use of incorrect subscripts for an array ,the use of the wrong variable in an equation, the use of the wrong unit of measurement ,or the inclusion of an incorrect declaration of a variable local to the module.

3. Internal interface defect:
Internal interface defects are those that were associated with internal structure of a module.

4. External interface defect:
External interface defects are those that were associated with structures existing outside the module's local environment but which the module used.

5. Initialization defect:
Initialization defects are those that result from an incorrectly initializes variable, failure to re-initialized a variable, or because a necessary initialization was missing; failure to initialize or re-initialize a data structure properly upon a module's entry/exit is also considered an initialization defect.

6. Logic/control structure defect:
Logic/control structure defects are those that cause an incorrect path in a module to be taken. Such a control defect might be a conditional statement causing control to be passed to an incorrect path.

4. DATA MINING TECHNIQUES

4.1Data Mining Definition

The development of information Technology has generated large amount of database and huge data in various areas. The research in database and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data information and pattern from a mining is a process of extraction of useful information and pattern from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data/pattern analysis.

Data mining is a logical process that is used to search through large amount of data. The goal of this technique is to find pattern that were previously unknown. Once these patterns are found they can further be used to making certain decision for development of their business.
Clustering is a research area in data mining and specifically it locates indirect data mining subgroups. Data mining is grouped into two parts as follows:

1. Direct data mining:
Classification, estimation, prediction researches locates here, for data mining some variable are single out as targets.

2. Indirect data mining:
Clustering, association rules, visualization researches are in this research group. For indirect data mining, no variable is singled out as a target and the goal is to discover relationships among all the variables.

4.2 Data Mining Algorithms

4.2.1 Classification:

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network based classification algorithm. In classification test data are Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network based classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples for a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record by record basis. The classifier training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination the algorithm then encodes these parameters into a model called a classification.

4.2.2. Association rule:

Association and correlation is usually to find frequent item set findings among large data sets. This type of findings among large data sets. This type of finding helps businesses to make certain decisions such as catalogue, design, cross marketing and customer shopping behavior analysis Association rule algorithms need to be able to generate rules with confidence value less than one. However the number of possible Association rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

4.2.3. Clustering:

Clustering can be said as identification of similar objects by using clustering technique we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes classification approach can also be used for effective means of distinguishing groups or classes of objects but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification .for example to form group of customers based on purchasing patterns to categories genes with similar functionality.
Components of a Clustering
Typical pattern clustering activity involves the following steps [1].
1.Pattern representation (optionally including feature extraction or selection),
2. Definition of pattern proximity measure appropriate to the data domain,
3. Clustering or grouping,
4. Data abstraction (if needed), and
5. Assessment of output (if needed).

Clustering is applied to the identified defects along with their effort in order to group the defects into three categories based on their isolation and correction effort. The three categories are [4]:

1. COMPLEX: Defects that require a large amount of isolation and correction time.
2. MODERATE: Defects that require a considerable amount of time for isolation and correction, but not large enough to be complex.

3. SIMPLE: Defects that are easy to isolate and correct and which require a very small amount of time for isolation and correction.

This categorization is not based on any fixed boundaries. The boundaries are dynamic and are formed based on the data presents in that particular system. Hence the clustering is specific to the system to system. In this project, clustering is performed based on the following attributes-
Defect Isolation Effort, Defect Correction Effort, Typological Error, Bad Code and Omission Error.

4.3 Defect Cluster Algorithms

1. The number of clusters is fixed as 3(SIMPLE, MODERATE and COMPLEX)

2. Initial 3 clusters are formed by choosing one defect sets each cluster from database.

3. The defect set with the lowest isolation and correction effort forms the first defect set in the SIMPLE cluster while the defect set with the maximum isolation and correction effort forms the first defect set in the COMPLEX cluster.

4. The average of the isolation effort and correction effort of the defects in the SIMPLE and COMPLEX cluster is computed and the defect set with effort nearest to the average forms the first defect set in the MODERATE cluster. Now each cluster consists of one defect set.

5. Next each defect set is assigned to only one of the clusters. Each defect set is assigned to the nearest cluster by computing its distance using the Euclidean Distance Method.

6. Each defect set is assigned to the most similar cluster based on the distance method and the arithmetic mean is calculated for all the defect sets in a cluster.

7. The arithmetic mean forms the new centre for this cluster. Similarly centers are computed for each cluster.

8. The defect sets are reassigned to the cluster based on their distance from the new centers.

9. This process is repeated until stable clusters are formed.
There are number of existing clustering approaches we will try to perform one of such approach that can be implemented in this method are:

4.3.1 K-Means Clustering

K-means clustering is an algorithm to classify or to group your objects based on attributes/features into K number of group is positive integer number. The grouping is done by minimizing the sum of squares of centroid. Thus the purpose of k-mean clustering is to classify the data.
The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroid or the first K objects in sequence can also serve as the initial centroids [1].
Step 1.Begin with a decision on the value of K=number of clusters.
Step 2.Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:
1. Take the first k training sample as single- element clusters.
2. Assign each of the remaining (N-K) training samples to the clusters with the nearest centroid .After each assignment, recomputed the centroid of the gaining cluster.
Step 3.Take each sample in sequence and compute its distance from the centroid of each of the clusters, if a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing sample.

4.3.2. Hierarchical Clustering:

A hierarchical algorithm yields a dendrogram representing the nested group of pattern and similarity levels at which grouping change dandles at which corresponding can be broken at different levels to yield different clustering of the data.

Most hierarchical clustering algorithm is variants of the single-link Complete-link and minimum-variance algorithm. Of these the single-link and complete-link algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the minimum of the distance between all pairs of pattern drawn from the two clusters. In the complete-link algorithms, the distance between two clusters is the maximum of all pair wise distance between patterns the two clusters [9].

4.3.3 Mountain clustering:

The mountain clustering approach is a simple way to find cluster centers based on a density measure called the mountain function. This method is a simple way to find approximate cluster centers and can be used as a preprocessor for other sophisticated clustering method.

4.3.4 Subtractive clustering:

The problem with the previous clustering method, mountain clustering is that its computation grows exponentially with the dimension of the problem; that is because the mountain function has to be evaluated at each grid point subtractive clustering solves this problem by using data points as the candidates forcluster centers, instead of grid points as in mountain clustering. This means that the computation is now proportional to the problem size instead of the problem dimension.

4.3.5 Self-Organizing Map:

The Self-Organizing Map is an unsupervised neural network. This type of network can be used to map a high-dimensional data space onto a usually one-or-two dimensional data lattice of neurons, all of which have a reference model weight vector.SOM learns to recognize groups of similar input vectors in such a way that neurons near each other respond to similar input vectors. The utilization of SOM in real world application ranges from large document collection. In the latter, the application of SOM has been the ability to arrange documents with similar content in neigh boring regions [9].

4.3.6 C-Means Clustering

Partitioned clustering is an important part of cluster analysis .Based on various theories, numerous clustering algorithms have been developed, and new clustering algorithms continue to appear in the literature. It is known that Occam's razor plays a pivotal role in data-based models, and partitioned clustering is categorized as a data-based model.

It minimize the index of quality defines as sum of squared distances for all points included in the cluster space to the center of the cluster [1].
Algorithm:

1. Fix the number of cluster.
2. Randomly assign all training input vector to a cluster. this creates partition.
3. Calculate the cluster center as the mean of each vector component of all vectors assigned to that cluster. Repeat for all cluster.
4. Compute all Euclidean distances between each cluster center and each input vector.
5. Update partitioned by assigning each input vector to its nearest cluster minimum Euclidean distance.
6. Stop if the center do not move any more otherwise loop to step, where the calculation of a cluster center.

5. CONCLUSION

In order to improve quality of software development, we can make use of Data mining Clustering technique. This paper reviewed the software defect management based on different types of defects by using clustering algorithms. The resulting data is used as the basis for determining the nature of defects.

6. REFERENCES

[1]PuneetDhiman, Manish, RakeshChawla" A Clustered Approach to Analyze the Software Quality using Software Defects"2012

[2]Omar Alshathry, HelgeJanicke,"Optimizing Software Quality Assurance", 2012 34th Annual IEEE computer software and applications conference workshops.

[3] NaheedAzeem,ShaziaUsmani,"Analysis of Data Mining Based software defect prediction Techniques", Global of computer Science and technology Volume 11 Issue 16 Version 1.0 september,2011.

[4]R.Karthik, N.Manikandan,"Defect Association and complexity prediction by mining Association and Clustering Rules" Volume 7, 2010.

[5]Yuan chen,PengDu,Xiang-hengshen,BingGe,"Research on software Defect prediction based on data mining" volume 1,2010.

[6]Boehm, B Basili, V., (2001) software Defect Reduction Top 10 List Published in Journal computer archive Volume 34 Issue 1, January 2001, IEEE computer society press Los Alamitos CA, USA.

[7]en.wikipedia.org/wiki/software bug

[8]XiTan, Xinpeng, Sen Pan, Wenyun Zhao "Assessing oftware quality by program clustering and defect prediction" 2011 18th working conference on Reverse Engineering.

[9]Deepak Gupta, VinayKr.Goyal, Harish Mittal"Analysis of clustering Techniques for software quality prediction"