Performance Analysis of Software Quality using Data Mining Techniques

Rahul Gupta
Student of CSE
PCST Bhopal, India
r.gupta4124@gmail.com

Sandhya Gawhade
Assistant Professor
Dept. of CSE
PCST Bhopal , India
sandhya.gawhade@gmail.com

Dr. L. K. Vishwamitra
HOD of CSE Dept.
PCST Bhopal, India
vishwamitra@patelcollege.com

Abstract

The software quality faced a major problem of software bugs and error estimation. For the estimation of error and bugs used various data mining technique. In the series of data mining technique used clustering and classification technique. In this paper presents a survey of software quality analysis using clustering technique. Software is of high quality and highly reliable if it is error-free. Software is error-free if there is no bug present in it or it is free from bugs. Bugs are very hard to find. Software Engineering tasks are Programming, Testing, Bug Detection, Debugging and Maintenance. Data Mining Techniques are applied on software engineering tasks. Data mining techniques are used to mine software engineering data and extract the meaningful and useful information. Techniques used for mining software engineering data are matching, clustering, classification etc.

Keywords: - data mining, clustering, software quality data

.

INTRODUCTION

Cluster by nature are the collection of similar objects. Each group or cluster is homogeneous, i.e., objects belonging to the same group are similar to each other. Also, each group or cluster should be different from other clusters, i.e., objects belonging to one cluster should be different from the objects of other clusters. Clustering is the process of grouping similar objects, and this could be hard or fuzzy. In hard clustering algorithm, each element is allocated to a single cluster during its operation; however, in fuzzy clustering method, a degree of membership is assigned to each element depending on its degree of association to several other clusters. Clustering problem for unsupervised data exploration and analysis has been investigated for decades in the statistics, image retrieval, bioinformatics, data mining and machine learning fields. Basically clustering algorithms aim to divide data objects into groups so that objects in the same group are similar to one another and different from objects in other groups. Generally, clustering is identified as an unsupervised learning method which divides data objects into designated clusters based only on the information presented in the dataset without any external background knowledge and label information. Clustering is the important step for many errands in machine learning. Every algorithmic rule has its own bias attributable to the improvements of various criteria. Unsupervised machine learning is inherently an optimization task; one is trying to fit the best model to a sample of data. The terms data mining, patent mining, text mining and visualization are employed for the processing of the documents. This chapter will try to give some explanations of the terms and explain why “data mining” was chosen for the title of the study. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. To remain competitiveness among software quality field, these organizations need deep and enough knowledge for a better assessment, evaluation, planning, and decision-making. Data mining refers to extracting or mining knowledge from large amounts of databases. It is a powerful new technology with great potential to analyze important information in the data warehouse. Data mining as simple an essential step in the process of KDD (knowledge discovery in database). The various steps involve in knowledge discovery process include data selection, data cleaning, data integration, data transformation, data mining algorithm , pattern evaluation and finally knowledge presentation [1,2] . Data mining analysis trends to work up from the data and the best technique are developed with an orientation towards large volumes of data making use of as much data as possible to arrive at reliable conclusion and decision. Today there are various type of data mining available like Web mining, Sequence mining, Text mining, Temporal and Spatial data mining, Graph mining, Content

mining, Link mining. Researchers find two fundamental goals of data mining: Prediction and Description. Prediction makes use of existing variable within the databases so as to predict unknown or future values of interest, and description finding patterns describing the information and also the ulterior presentation for user interpretation. The relative emphasis of both predictive and descriptive differs with respect to the underlying application and the technique. There are many data mining technique fulfilling these objectives. A number of these are Classifications, Associations, clustering and sequential patterns [5, 8]. The essential premise of a Classification is to develop profiles of different groups. Association fined all associations, such that the presence of a set of items in a record implies the other items. Clustering segments a database in to subsets or cluster. Sequential patterns identify subject to a user specified minimum constraint. Data mining research has drawn on a number of other fields like machine learning and statistics. Review the relations of data mining with a number of the vital areas. Supervised learning: - A supervised learning is the machine learning task of inferring a function from labeled training data consists of a set of training examples. In supervised learning, every example could be a combine consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if output is discrete) or a regression function (if output is continuous). The function should predict the correct output value for any valid input object. Unsupervised learning: - In unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner square measure unlabeled, there is no error or reward signal to evaluate a possible solution. Unsupervised learning is closely associated with the problem of density estimation in statistics. However learning also encompasses many other techniques that seek to summarize and explains key features of the data. Many methods employed in unsupervised learning are based on data mining method used to p reprocess data. Approaches to unsupervised learning involve clustering (e.g. k-means algorithm, mixture models, hierarchical clustering method). The rest of paper is organized as follows. In Section II discuss clustering technique. The Section III Related work IV discusses implementation details .section V discuss performance evaluation followed by a conclusion in Section VI.

II CLUSTERING ANALYSIS

Clustering is useful technique for the discovery of knowledge distribution and patterns within the underlying data. The aim of clustering is to discover both the dense and the sparse regions in a data set. The earlier approaches do not adequately consider the fact that the data set can be too large to fit in the main memory. Clustering can be considered the most important unsupervised learning problem; so, as each problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering may be “the process of organizing objects into clusters whose members are similar in some way” [9, 5]. Clustering may be a technique of grouping data in to different groups. In order that the information in every cluster share similar trends and patterns. Clustering constitutes a significant category of data mining and a common technique for statistical data analysis used in several fields; involve pattern recognition, information retrieval, machine learning, bioinformatics, and image analysis. Cluster analysis itself is not one specific algorithmic rule, however the final task to be solved. It are often achieved by different kinds algorithms that makes an attempt to automatically partition the data space into a set of regions or clusters, to that the examples within the table are allotted, either deterministically or probability wise. The aim of the process is to identify all sets of similar examples in the data, in some optimal fashion [5]. Clustering according to similarity is a concept which appears in several disciplines. If usually similarity is available, then there are a variety of techniques for forming clusters. Another is to make set functions that measure some particular property of groups.

III Related Work

In this section discuss the related work of software quality estimation using data mining technique. Some technique discuss here.

[1] In this paper, author aims at comparing different models based on clustering techniques: k-means (KM), fuzzy c-means (FCM) and hierarchical agglomerative clustering (HAC) for building software quality estimation system. We propose quality measure of partition clustering technique (KM, FCM) in order to evaluate the results and we comparatively analyze the obtained results on two case studies. author analyzed three clustering techniques and comparatively presented the results of applying two clustering algorithms (k-means & Fuzzy c-means) and effective results can be produced by using Fuzzy c-means clustering.

[2] In this paper, applications of GA in different types of software testing are discussed. The GA is also used with fuzzy as well as in the neural networks in different types of testing. It is found that by using GA, the results and the performance of testing can be improved. Use of evolutionary algorithms for automatic test generation has been an area of interest for many researchers. Genetic Algorithm (GA) is one such form of evolutionary algorithms. Our future work will involve applying GA for regression testing in web based applications.

[3] In this paper, author gives a survey on various clustering techniques for identifying the extract class opportunities. The survey showed that there are several clustering

approaches for the identification. Among the techniques reviewed, hierarchical clustering technique identifies better extract class opportunities for performing extract class refactoring than partitioned or any other clustering algorithms.

[4] In this paper, author provides the discussion of data mining for software engineering and also provide discussion about the clustering techniques. Data mining is most efficient technique to manage large amount of data since information is highly valuable and expensive. Every technique has to solve different problems and have their own advantages and disadvantages. There is no such clustering technique and algorithm exists that is used to solve all the problems and is a best fit for all applications. As the application changes requirements also change. With this change the selection of clustering technique affected. No technique or algorithm is the readymade solution to all applications and problems. Predefined number of clusters and stopping criteria affect the accuracy and performance of clustering. Handling of noisy data, data set size, shape of the clusters all affects the clustering results.

[5] In this paper, author affirm that there are synergies to be gained by using search-based techniques within software model checking. Author will provide evidence to support this assertion in the form of existing research work and open problems that may benefit from combining Search-Based Software Engineering (SBSE) techniques and soft-ware model checking. In particular we advocate that SBSE can be used to improve the model checking process and SBSE together with model checking can be used to address common Software Engineering problems. With respect to model checking we have highlighted existing work on an EDA approach to model checking software.

[6] In this paper, author discussed the overview of strategies for data mining for secure software engineering, with the implementation of a case study of text mining for source code management tool. Data mining can be used in gathering and extracting latent security requirements, extracting algorithms and business rules from code, mining legacy applications for requirements and business rules for new projects etc. Mining algorithms for software engineering falls into four main categories: Frequent pattern mining finding commonly occurring patterns; Pattern matching finding data instances for given patterns; Clustering grouping data into clusters and Classification predicting labels of data based on already labeled data.

[10] In this paper author proposes a novel solution to adapt, configure and effectively use a topic modeling technique, namely Latent Dirichlet Allocation (LDA), to achieve better (acceptable) performance across various SE tasks. Our paper introduces a novel solution called LDA-GA, which uses Genetic Algorithms (GA) to determine a near-optimal configuration for LDA in the context of three different SE tasks: (1) traceability link recovery, (2) feature location, and (3) software artifact labeling. The results of our empirical studies demonstrate that LDA-GA is able to identify robust LDA configurations, which lead to a higher accuracy on all the datasets for these SE tasks.

IV IMPLEMENTAION DETAILS

For the evaluation of pattern of task data used k-means clustering algorithm implement in MATLAB 7.8.0 and found the similar pattern of cluster for year, month, grade and subject. The cluster found in color group. The formation of cluster gives the information of valid and invalid cluster according to cluster valid index [13]. The clustering validity criteria are classified into internal, external, and relative. The clustering work focus on the relative association of month, grade and student relative criteria is used as the validity measure. The criteria widely accepted for partitioning a data set into a number of clusters are separation of the clusters, and their compactness. Thus these criteria are obviously good candidates for checking the validity of clustering results. The process of cluster validation defines a relative validity index, for assessing the quality of partitioning for each set of the input values. The proposal formalize clustering validity index based on clusters' compactness (in terms of cluster density), and clusters' separation (combining the distance between clusters and the inter-cluster density).

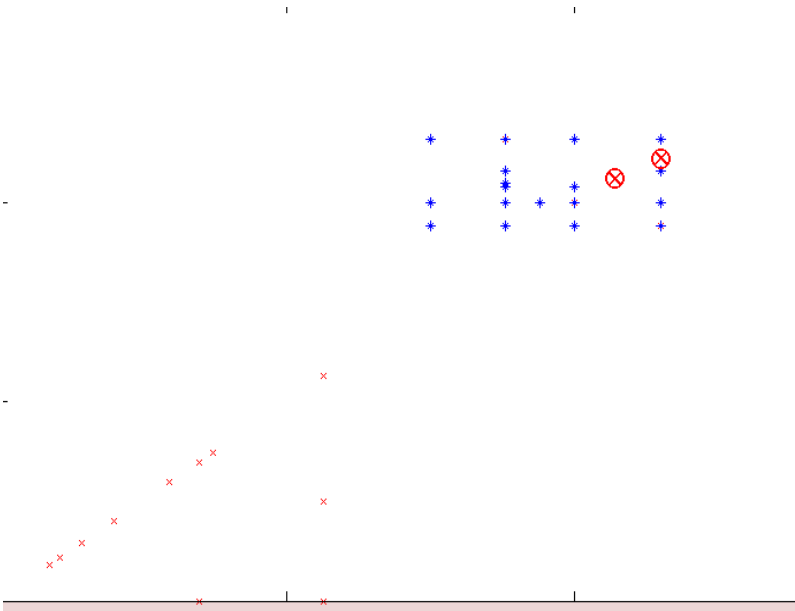

Figure 1: Gives the information about generated cluster of data point of COCOMO Data Set model.

V PERFORMANCE ANALYSIS OF DATA CLUSTER

The evaluation of clustering performance used some standard parameter such as number of valid cluster generation and number of cluster along with mean absolute

error of clustering process. The mean absolute error process induced the error rate of clustering technique. The process of clustering used some set of COCOMO data set as number of instant as row in fashion of 1000, 2000, 3000 and 4000 thousand for small data to large size of data. To test the validation of cluster each cluster property assigned the color label of data the unlabeled cluster shows that invalid cluster in the process of cluster generation[7,8]. For validity of cluster and measurement of error used some standard formula given below. In clustering the mean absolute error (MAE) is a quantity used to measure how close real or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|. \quad (1)$$

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction and y_i the true value.

Size of Data	Number of cluster	Valid cluster	Error
1000	8	6	4.607
2000	7	6	7.890
3000	8	8	2.304
4000	5	4	10.34

Table 1: Show that cluster generation of software quality data and check number of cluster according to valid cluster.

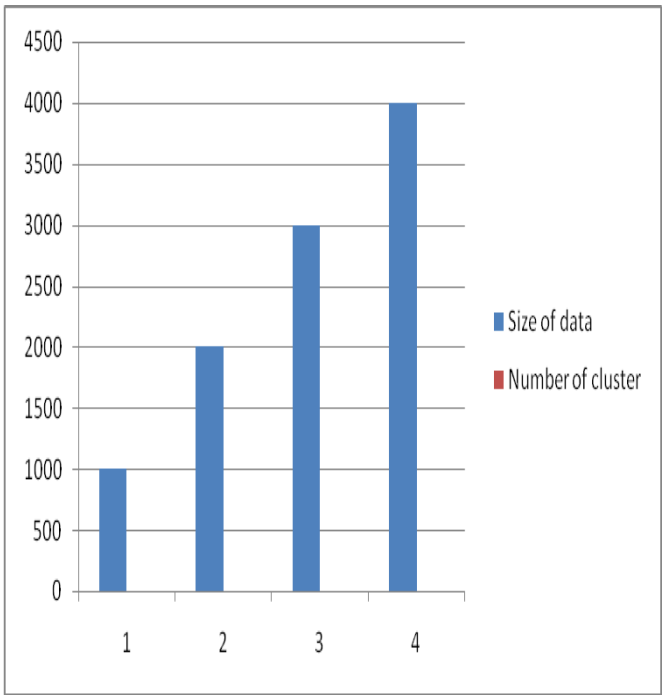

Figure 2: Shows that comparative cluster generation according to data size

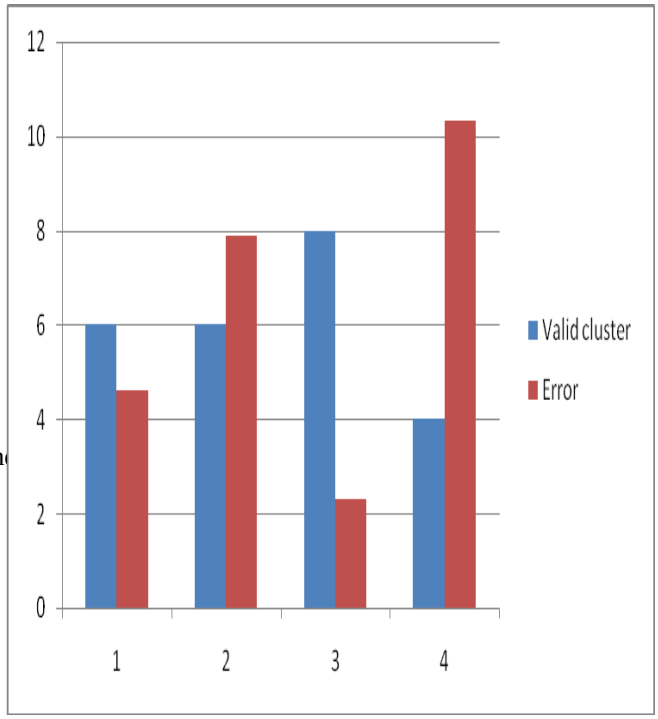

Figure 3: Shows that comparative valid cluster generation and generation of error according to cluster size

VI CONCLUSION AND FUTURE WORK

In this paper, we use of data mining technique for software quality data analysis. The process of data analysis clustering is an important tool for mining of meaning full information regarding software quality database. In this paper, author discuss task data for clustering. Also author discuss cluster generation and validation of cluster of comparative relation of two and more successive attribute such as grade and month of student. For analysis of clustering we used k-means algorithm.

REFERENCES

[1] Deepak Gupta, Vinay Kr. Goyal, Harish Mittal "Estimating of Software Quality with Clustering Techniques" Third International Conference on Advanced Computing & Communication Technologies, IEEE, 2013. Pp 20-27.

[2] Chayanika Sharma, Sangeeta Sabharwal, Ritu Sibal "A Survey on Software Testing Techniques using Genetic Algorithm" International Journal of Computer Science Issues, Vol-10, 2013. Pp 381-392.

[3] Suchithra Chandran, Bright Gee Varghese.R "A Survey On Clustering Techniques For Identification Of Extract Class Opportunities" International Journal of Research in Engineering and Technology, Vol-2, 2013. Pp 426-429.

[4] Maninderjit Kaur, Sushil Kumar Garg "Survey on Clustering Techniques in Data Mining for Software Engineering" International Journal of Advanced and Innovative Research, Vol-3, 2014. Pp 238-243.

[5] Jeremy S. Bradbury, David Kelk, Mark Green "Effectively using Search-Based Software Engineering Techniques within Model Checking and Its Applications" IEEE, 2013. Pp 67-70.

[6] A. V. Krishna Prasad, Dr. S. Rama Krishna "Data Mining for Secure Software Engineering- Source Code Management Tool Case Study" International Journal of Engineering Science and Technology, Vol-2, 2010, Pp 2667-2677.

[7] Jiawei Han, Micheline Kamber , Jian Pei " Data Mining: Concepts and Techniques" 2013.

[8] K. Kameshwaran, K. Malarvizhi "Survey on Clustering Techniques in Data Mining" IJCSIT: International Journal of Computer Science and Information Technologies, Vol-5, 2014. Pp 2272-2276.

[9] Kapila Kapoor , Geetika Kapoor "Improving Software Reliability and Productivity through Data Mining" Proceedings of the 5th National conference; INDIACom-2011,.

[10] Annibale Panichella, Bogdan Dit, Rocco Oliveto "How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms" IEEE 2104. Pp 546-555.

[11] M. Kuchaki Rafsanjani, Z. Asghari Varzaneh, and N. Emami Chukanlo "A survey of hierarchical clustering algorithms" TJMCS: The Journal of Mathematics and Computer Science, Vol-5, 2012. Pp 229-240.

[12] Anoop Kumar Jain, Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining" International Archive of Applied Sciences and Technology, Vol-3[2], 2012. Pp 68-75.

[13] Manpreet Kaur, Usvir Kaur " A Survey on Clustering Principles with K-Means clustering Algorithms Using Different Methods in Detail" International Journal of Computer Science and Mobile Computing, Vol-2, 2013. Pp 327-331.

[14] S. Revathi, Dr. T. Nalini "Performance Comparison of Various Clustering Algorithm" IJARCSSE: International Journal of Advanced Research in Computer Science and Software Engineering Vol-3, February 2013.

[15] Suma. V, Pushpavathi t.P, Ramaswamy "An Approach to Predict Software Project Success by Data Mining Clustering" international Conference on Data Mining and Computer Engineering, Bangkok (Thailand), December 2012.