International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 2 Issue 1 January 2015

Using Data Mining Techniques for Performance Analysis of Software Quality

Nilesh Patidar
RITS, Bhopal
Patidarnilesh1@gmail.com

Abstract: - The software quality moon-faced a serious problem of software bugs and error estimation. For the estimation of error and bugs used numerous data processing techniques. Within the series of knowledge mining technique used cluster and classification technique. During this paper presents a survey of software quality analysis victimization cluster technique. Software is of top quality and extremely reliable if it's error-free. Software is error-free if there's no bug present in it or it's free from bugs. Bugs are terribly exhausting to search out. Software Engineering tasks are Programming, Testing, Bug Detection, Debugging and Maintenance. Data mining Techniques are applied on software engineering tasks. Data mining techniques are accustomed mine software engineering knowledge and extract the significant and helpful data. Techniques used for mining software engineering knowledge are matching, clustering, classification etc.

Keywords: - data mining, clustering, software quality data.

1. INTRODUCTION

Cluster naturally is the gathering of comparable objects. Every cluster or cluster is homogeneous, i.e., objects belonging to a similar cluster are just like one another. Also, every cluster or cluster ought to diverge from alternative clusters, i.e., objects belonging to at least one cluster ought to diverge from the objects of alternative clusters. Clustering is that the method of grouping similar objects, and this might be exhausting or fuzzy. In exhausting bunch algorithmic program, every component is allotted to one cluster throughout its operation; but, in fuzzy clustering methodology, a degree of membership is appointed to every component betting on its degree of association to many alternative clusters. Clustering drawback for unattended knowledge exploration and analysis has been investigated for many years within the statistics, image retrieval, bioinformatics, data processing and machine learning fields. Primarily clustering algorithms aim to divide knowledge objects into teams in order that objects within the same cluster are just like one another and completely different from objects in other teams. Generally, clustering is known as an unsupervised learning technique that divides information objects into selected clusters based only on data conferred within the dataset with none external information and label information. Clustering is that the vital step for several errands in machine learning. Each algorithmic rule has its own bias because of the enhancements of assorted criteria. Unsupervised machine learning is inherently associate improvement task; one is attempting to suit the most effective model to a sample of knowledge. The terms data processing, patent mining, text mining and image are utilized for the process of the documents. This chapter can try and provide some explanations of the terms and justify why "data mining" was chosen for the title of the study. Data processing is that the analysis of

(often large) empiric information sets to seek out unsuspected relationships and to summarize the information in novel ways in which are each understandable and helpful to the information owner. Clustering could be a division of knowledge into teams of comparable objects. Representing the data by fewer clusters essentially loses sure fine details, however achieves simplification. It models information by its clusters. to stay competitiveness among software quality field, these organizations want deep and enough information for a stronger assessment, evaluation, planning, and decision-making. Data processing refers to extracting or mining information from massive amounts of databases. It's a strong new technology with nice potential to research necessary info within the information warehouse. Data mining as easy a vital step within the process of KDD (knowledge discovery in database). the assorted steps involve in information discovery method include information selection, data cleaning, data integration, data transformation, data processing algorithm , pattern analysis and eventually information presentation [1,2] . Data processing analysis trends to figure up from the information and therefore the best technique are developed with an orientation towards massive volumes of information creating use of the maximum amount data as possible to hit reliable conclusion and decision. nowadays there are numerous type of data mining accessible like web mining, Sequence mining, Text mining, Temporal and spatial data mining, Graph mining, Content mining, Link mining. Researchers find two elementary goals of data mining: Prediction and description. Prediction makes use of existing variable inside the databases therefore on predict unknown or future values of interest, and description finding patterns describing the knowledge and additionally the ulterior presentation for user interpretation. The relative emphasis of each predictive and descriptive differs with reference to the underlying application and the technique. There are units several data mining technique fulfilling these objectives. a number of those are Classifications, Associations, cluster and successive patterns [5, 8]. The essential premise of a Classification is to develop profiles of various groups. Association fined all associations; such the presence of a collection of items during a record implies the other items. Cluster segments a database in to subsets or cluster. Sequential patterns identify subject to a user nominal minimum constraint. Data mining analysis has drawn on variety of alternative fields like machine learning and statistics. Review the relations of knowledge mining with variety of the important areas. Supervised learning: - A supervised learning is that the machine learning task of inferring a function from labeled training knowledge consists of a group of training examples. In supervised learning, each example may be a mix consisting of associate degree input object (typically a vector) and a desired output price (also known as the supervisory signal). A learning algorithmic program analyzes the training knowledge and produces an inferred function, that is called a classifier (if output is discrete) or a regression function (if output is continuous). The function ought to predict the right output value for any valid input object. Unsupervised learning: - In unsupervised learning refers to the matter of attempting to seek out hidden structure in unlabeled knowledge. Since the examples given to the learner unit of measurement unlabeled, there's no error or reward signal to judge a potential solution. Unsupervised learning is closely related to the matter of density estimation in statistics. But learning also encompasses several alternative techniques that look for to summarize and explains key features of the data. Several strategies utilized in unsupervised learning are supported data mining methodology wont to p reprocess information. Approaches to unsupervised learning involve clump

(e.g. k-means algorithm, mixture models, hierarchical clustering method). the remainder 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 analysis followed by a conclusion in Section VI.

2. CLUSTERING ANALYSIS

Clustering is beneficial technique for the invention of knowledge distribution and patterns inside the underlying data. The aim of clustering is to find each the dense and therefore the distributed regions during a knowledge set. the earlier approaches don't adequately think about the very fact that the information set is large to suit within the main memory. Clustering is thought-about the foremost vital unsupervised learning problem; therefore, as every downside of this sort, it deals with finding a structure during a assortment of unlabeled knowledge. A loose definition of clustering could also be "the method of organizing objects into clusters whose members are similar in some way" [9, 5]. Clustering is also a method of grouping information in to totally different groups. So as that the knowledge in each cluster share similar trends and patterns. Clustering constitutes a big class of information mining and a standard technique for statistical data analysis utilized in many fields; involve pattern recognition, data retrieval, machine learning, bioinformatics, and image analysis. Cluster analysis itself isn't one specific algorithm, but the ultimate task to be solved. It are typically achieved by totally different kinds algorithms that creates an endeavor to mechanically partition the information space into a collection of regions or clusters, thereto the examples inside the table are assigned, either deterministically or likelihood wise. The aim of the method is to spot all sets of comparable examples within the information, in some optimum fashion [5]. Clustering consistent with similarity may be an idea that seems in many disciplines. If sometimes similarity is on the market, then there are a range of techniques for forming clusters. Another is to create set functions that live some explicit property of groups.

3. Related Work

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

[1] During this paper, author aims at comparing completely different models supported clustering techniques: k-means (KM), fuzzy c-means (FCM) and hierarchical agglomerative clustering (HAC) for building software system quality estimation system. We have a tendency to propose quality live of partition clustering technique (KM, FCM) so as to judge the results and that we relatively analyze the obtained results on two case studies. Author analyzed three clustering techniques and relatively conferred the results of applying two clustering algorithms (k-means & Fuzzy c-means) and effective results may be created by victimization Fuzzy c-means clustering.

[2] During this paper, applications of GA in numerous varieties of software testing are mentioned. The GA is additionally used with fuzzy as well as within the neural networks in numerous varieties of testing. It's found that by victimization GA, the results and also the performance of testing may be improved. Use of evolutionary algorithms for automatic test generation has been a vicinity of interest for several researchers. Genetic algorithm (GA) is one such variety of evolutionary algorithms. Our future work can involve applying GA for regression testing in internet primarily based applications.

[3] During this paper, author gives a survey on varied clump techniques for characteristic the extract category opportunities. The survey showed that there are many clustering approaches for the identification. Among the techniques reviewed, graded clustering technique identifies higher extract class opportunities for activity extract class refactoring than partitioned off or the other clustering algorithms.

[4] During this paper, author provides the discussion of knowledge mining for software engineering and conjointly provides discussion concerning the clustering techniques. data processing is most effective technique to manage great amount of knowledge} since information is very valuable and expensive . Each technique should solve totally different issues and have their own blessings and downsides. There's no such clustering technique and algorithm exists that's wont to solve all the issues and could be a best fit for all applications. Because the application modifications needs also change. With this variation the choice of clustering technique affected. No technique or algorithm is that the readymade answer to all or any applications and issues. Predefined variety of clusters and stopping criteria have an effect on the accuracy and performance of clustering. Handling of noisy information, information set size, form of the clusters all affects the clustering results.

[5] During this paper, author affirm that there are synergies to be gained by mistreatment search-based techniques among software model checking. Author can offer proof to support this assertion within the sort of existing analysis work and open issues that will have the benefit of combining Search-Based software Engineering (SBSE) techniques and soft-ware model checking. Specially we have a tendency to advocate that SBSE may be accustomed improve the model checking method and SBSE in conjunction with model checking may be accustomed address common software Engineering problems. With reference to model checking we've highlighted existing work on an EDA approach to model checking software.

[6] During this paper, author mentioned the summary of ways for data processing for secure software engineering, with the implementation of a case study of text mining for source code management tool. data processing may be employed in gathering and extracting latent security needs, extracting algorithms and business rules from code, mining inheritance applications for needs and business rules for brand spanking new comes etc. Mining algorithms for software engineering falls into four main categories: Frequent pattern mining finding ordinarily occurring patterns; Pattern matching finding data instances for given patterns; cluster grouping information into clusters and Classification predicting labels of information based on already tagged data.

[7] During this paper author proposes a unique solution to adapt, configure and effectively use a topic modeling technique, specifically Latent Dirichlet Allocation (LDA), to realize higher (acceptable) performance across numerous SE tasks. Our paper introduces a unique answer known as LDA-GA that uses Genetic Algorithms (GA) to work out a near-optimal configuration for LDA within the context of three totally 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 ready to spot robust LDA configurations that result in a better accuracy on all the datasets for these SE tasks.

4. **Implementation details**

For the analysis of pattern of task data used k-means clustering algorithm implement in MATLAB7.8.0 and located the similar pattern of cluster for year, month, grade and subject. The cluster found in color cluster. The formation of cluster provides the data of valid and invalid cluster in line with cluster valid index [13]. The bunch validity criteria are classified into internal, external, and relative. The bunch work concentrate on the relative association of month, grade and student relative criteria is employed because the validity measure. The standards wide accepted for partitioning information set into variety of clusters are separation of the clusters, and their compactness. There for these criteria are clearly sensible candidates for checking the validity of clustering results. The method of cluster validation defines a relative validity index, for assessing the standard of partitioning for every set of the input values. The proposal formalize clustering validity index supported clusters' compactness (in terms of cluster density), and clusters' separation (combining the space between clusters and therefore the inter-cluster density).

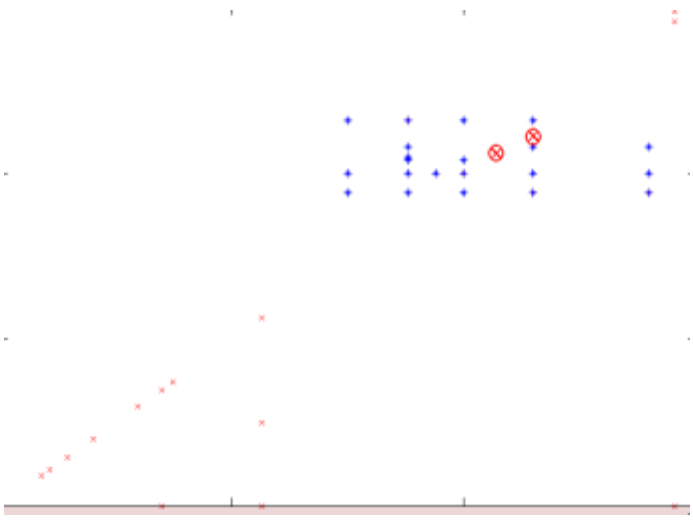

Figure 1: Gives the information about generated cluster of data point of COCOMO Data Set model.

5. PERFORMANCE ANALYSIS OF DATA CLUSTER

The analysis of clustering performance used some customary parameter like variety of valid cluster generation and variety of cluster in conjunction with mean absolute error of clustering method. The mean absolute error method induced the error rate of clustering technique. the method of clustering used some set of COCOMO information set as variety of instant as row in fashion of one thousand, 2000, 3000 and 4000 thousand for tiny information to massive size of information. To check the validation of cluster every cluster property assigned the color label of information the untagged cluster shows that invalid cluster within the method of cluster generation [7, 8]. For validity of cluster and menstruation of error used some customary formula given below. In clustering the mean absolute error (MAE) could be a amount used to measure how close real or predictions are to the ultimate 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.

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

Size of Data	Number cluster	Valid cluster	Error
2000	16	12	9.214
4000	14	12	15.78
6000	16	16	4.608
8000	10	8	20.68

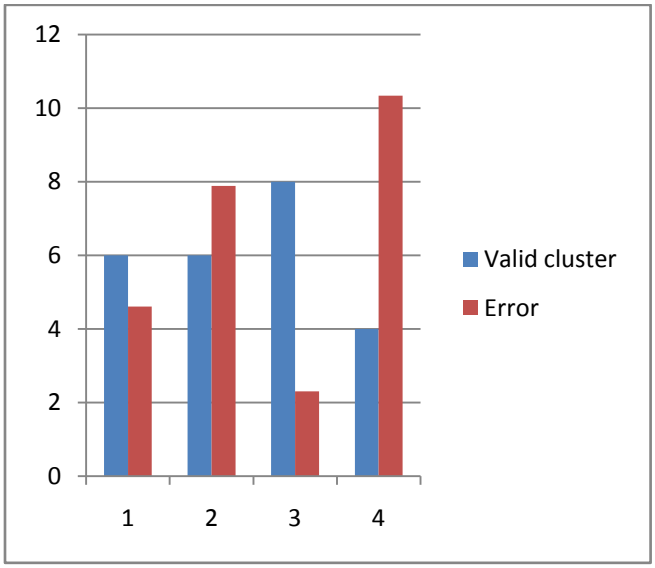

Figure 2: Shows that comparative valid cluster generation and generation of error according to cluster size

6. Conclusion and future work

In this paper, we tend to use data mining technique for software system quality data analysis. The method of {information} analysis clustering is an important tool for mining of which means full information concerning software quality database. During this paper, author discusses task information for clustering. additionally author discuss cluster generation and validation of cluster of comparative relation of two and a lot of successive attribute like grade and month of student. For analysis of clustering we tend to used k-means algorithm.

REFERENCES

1. [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. [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. [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. [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. [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. [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. [7] Jiawei Han, Micheline Kamber , Jian Pei " Data Mining: Concepts and Techniques" 2013.
8. [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. [9] Kapila Kapoor , Geetika Kapoor "Improving Software Reliability and Productivity through Data Mining" Proceedings of the 5th National conference; INDIACom-2011,.
10. [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. [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. [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. [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. [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. [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.