IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Enhancing test case reduction by k-means algorithm and elbow method

A. Pandey[1*], A. K. Malviya[2]

[1] Dept. of CSE, K.N.I.T., Sultanpur, India
[2] Dept. of CSE, K.N.I.T., Sultanpur, India

[*]*Corresponding Author: ankp67@gmail.com, Tel.: +91-9839419533*

Available online at: www.ijcseonline.org

Accepted: 11/Jun/2018, Published: 30/Jun/2018

***Abstract*—** Software testing plays an indispensable part in the software development process. A huge number of test cases are required to be tested to improve the quality of the software which is a tedious and time-consuming process. In this paper we aim to minimize the number of test cases by eliminating redundant test cases and thereby assisting us in reducing the time consumed in testing huge number of test cases. We have used the popular data mining k-means algorithm along with an elbow method to reduce the number of test cases required to be tested. Experimental result presents better clustering accuracy and significant elimination of redundant test cases by using the proposed approach.

***Keywords*—** Testing; data mining; test case reduction; test case minimization; test suite reduction; test suite minimization; cluster.

I. INTRODUCTION

Software testing is the process of detecting errors that help programmers improve the quality of the software and minimize the cost associated with these defects. It can also be stated as the process used to measure the information related to quality of the product or service under test. A substantial amount of time and effort in software development is spent in various forms of testing. Besides, assuring that the software works as intended the testers also ensure that there are no unintended consequences of using the software. Test cases in large number are already being built automatically using the automated test case generation. Though these huge numbers of test cases are generated at a faster rate but in an industry, programs containing thousand lines of code are used and each test case may take considerable amount of time to execute. As a result, testing of thousands of automatically generated test cases may take several days to execute completely. These automatically generated test cases also contain many redundant test cases and lot of time is wasted in testing these redundant test cases.

In our approach we have used data mining technique to identify and eliminate the redundant test cases which in turn reduce the time spent in testing the automatically generated test cases. Our rest of the paper is organized as follows. Section II discusses the work done in the past in the area of test case reduction. Further, in section III we discuss about the k-means algorithm along with the elbow method a technique to find an accurate estimate of k. Section IV represents the research methodology proposed for identification and elimination of redundant test cases. Section V shows the experimental results produced by implementing the proposed research methodology. In section VI we discuss the future prospects and conclude our work.

II. RELATED WORK

Several approaches have been proposed in the past to reduce the number of test cases. In [1] the author uses the heuristic approach to choose a representative set of test cases with equivalent code coverage as the original test suite and thus minimize the number of test cases required for testing. The author has illustrated the technique that only involves relation between the testing requirements and test cases fulfilling the requirements and is thereby not dependent on the test methodology used. Data flow testing methodology is used to demonstrate the technique developed by the author.

Similarly, in [2] the authors were of the opinion that there are two special kinds of test cases namely the essential test cases and the 1-to-1 redundant test cases. They presented the GRE heuristic algorithm that uses the following procedures alternatively until all the requirements are covered: (1) to discover all the essential test cases, (2) remove the redundant test cases (3) the test cases that cover the maximum number of unsatisfied requirements are identified are determined.

In [3] the author has proposed a new methodology for test suite reduction by optimizing the test suite requirements using graph contraction. Various empirical studies were conducted [4], [5] to compare the sizes of the reduced test

suite using the test-suite reduction techniques proposed in the past to provide the guidelines to choose the appropriate test suite reduction technique.

In the recent studies, [6] the author has presented a mining system to undergo a more desirable knowledge of the test cases to produce and use more effective test cases. The knowledge mining system proposed by the author accepts the test suite as input which is mined by attribute selection and application of clustering techniques. The output of the system results in a reduced test suite.

The author in [7] focuses on the capability of the reduced test cases to detect faults in comparison to original test cases and has used fault detection capability as the measure of suite quality. The system proposed by the author comprises the selection and evaluation of metrics, clustering of test cases and selection of test cases.

The author in [8] aims to reduce the cost of testing by reducing the number of test cases using the data mining techniques that provide an aid in detecting redundant test cases incorporated by the automatic test case generator. The author uses a program with two input variables and generates random values using the automatic test case generator. Then the author applies k-means algorithm to these automatically generated random values to form clusters. Randomly sample from each cluster is picked up and stored in a file. These samples of clusters are used to test the coverage of the test cases. To obtain optimal coverage the process is repeated for different values of k.

The author in [9] used the data mining classifier technique to reduce the number of test cases. The author generates the test cases for the program under test and builds a dataset based on some attributes like test cases input, output, and test case coverage details. J48 and Naïve base classifier algorithms are applied on the dataset to extract the results that emphasis the use of data mining classifier technique in removing the redundant test cases.

Classification of functional and non-functional requirements is done using the software requirement specification to generate test cases which are then reduced using mining techniques by the author in [10].

In [11] the author uses the Density based clustering technique to group identical data objects based on density and reduce the number of test cases. Test cases are generated using selenium software which is then loaded in Weka for application of DBSCAN algorithm and filters for generation of desired results. The author in the literature [12] has used k means algorithm to cluster the independent paths of the program to minimize the test case.

However, it has been observed that the dataset are not properly grouped by the clustering technique, thereby are less efficient in identifying the redundant test cases. In order to identify and remove adequate redundant test cases, proper clustering of the dataset is important. In this paper we have proposed a methodology that incorporates elbow method to estimate the correct value of k along with k-means algorithm to efficiently cluster the test cases.

III. BACKGROUND

A. Data mining

Data mining is an interdisciplinary subfield of computer science that involves computational process to extract knowledge from any large set of raw data by identifying patterns and establishing relationships thereby helps in solving problems through data analysis [12]. Different methods like association rule, classification and clustering are available for mining different kinds of data [13, 14, 15, 16]. Clustering is an important unsupervised learning problem that aims to segregate data points with similar characteristics and assign them to clusters. There are various clustering algorithm available but algorithms that are used popularly are k-means, fuzzy c-means, hierarchical and DBSCAN.

B. K-means algorithm

The k-means algorithm is an unsupervised learning algorithm that is used to classify data into certain number of clusters. K-means starts by randomly defining k centroids that represent initial group centroids. Now, the algorithm works in repetition to perform the next two steps.

Step 1. Assign each object to the group that has the closest centroid, using the standard Euclidean distance.

Step 2. After all the objects have been assigned, the centroid of each of the k clusters is recalculated by finding the mean of all the values belonging to a cluster.

The above steps are repeated until the centroids no longer move and convergence is achieved.

Euclidean distance (1) is used to figure out the root of square difference between two points or co-ordinates of pair of objects.

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \quad (1)$$

K-means algorithm main objective is to minimize squared error function represented by (2).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \quad (2)$$

Where, J is the objective function, k is the number of clusters, n is the number of cases and c is the centroid for cluster j.

C. Elbow method

The oldest method for determining the true number of clusters in a data set is inelegantly called the elbow method [17]. The blueprint of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 2 to 10), and for each value of k calculate the sum of squared errors (SSE). Then a line chart of the SSE for each value of k is plotted. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The purpose is to seek for a small SSE, but SSE tends to depreciate with increasing value of k. So our goal is to choose a small value of k that still has a low SSE, usually represented by the elbow.

IV. METHODOLOGY

The approach used by us mainly consists of four steps. Firstly, the source program is selected and test cases for the same are generated. Then in the next step we prepare the dataset using the test cases generated for the source program. In the third step we apply the clustering algorithm to the dataset prepared in the previous step and use the elbow method to predict the accurate value of k. Finally, the clustered results are saved and appropriate filters are applied to eliminate redundant test cases. Figure 1 represents the proposed methodology in detail.

A. Selection of source program and test case generation

We have used the famous triangle problem [18] to demonstrate the application of our proposed methodology. The program accepts three variables as input and returns the type of triangle formed as output. In order to execute the program we have used an IDE for Java called Eclipse SDK 4.6.3. It is the most widely used IDE that provides a base workspace and can be easily customized due to its extensible plug-in system. The details of the test cases generated are shown in Table 1. A set of 500 test data was randomly generated and was executed using Junit 4.12 a unit testing framework for Java.

B. Preparation of Dataset

Using the outputs generated by the Junit we prepared our dataset containing the test id, input arguments supplied to the source program and expected output for each test case. A sample of the data set is shown in the Figure 2. This dataset contains 500 test cases input and expected output values. The attributes used in the dataset are described in the Table 2.

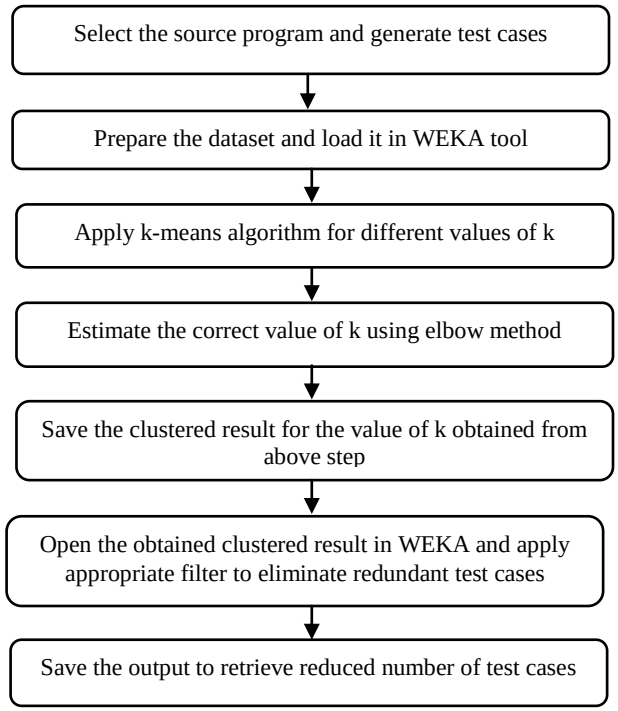

Table 1. Description of test cases generated for Triangle Problem

Si. No.	Description	No. of Test Cases
1.	Not a Triangle	76
2.	Equilateral	75
3.	Isosceles	59
4.	Scalene	89
5.	Length of the first side is invalid	75
6.	Length of the second side is invalid	57
7.	Length of the third side is invalid	69
Total		500

File Edit Format View Help

TestId	Input a	Input b	Input c	Expected Output
test0001	174	4	88	Not a triangle
test0002	78	81	166	Not a triangle
test0003	175	4	53	Not a triangle
test0004	104	190	68	Not a triangle
test0005	63	100	191	Not a triangle
test0006	91	0	113	Length of second side is invalid
test0007	76	75	10	Scalene
test0008	155	41	10	Not a triangle
test0009	175	98	88	Scalene
test0010	145	152	109	Scalene
test0011	0	4	32	Length of first side is invalid
test0012	53	52	78	Scalene
test0013	0	100	100	Length of first side is invalid
test0014	98	2	35	Not a triangle
test0015	1	1	78	Not a triangle

Figure 2. Sample dataset

C. Applying k-means algorithm and elbow method

Next, the above prepared dataset is converted to csv format and loaded into WEKA [19] for clustering. We employed k-means clustering algorithm on our dataset for different values of k (2, 3, and so on) and plotted the graph for each value of k against the sum of squared errors generated by Weka as shown in Fig. 3. Using the graph we are able to identify the suitable value of k, here 7 as an elbow is formed at this value of k. The cluster file formed by applying k-means (for k=7) is saved in the arff format.

D. Identification and elimination of redundant test cases

The arff file obtained from the above step is loaded again into WEKA and suitable filters are applied for identification and elimination of redundant test cases. In our approach we have used unsupervised filter that applies sampling algorithm to produce a random subsample of the dataset. After applying the filter we obtain reduced number of test cases detailed in the Table 3.

Table 2. Dataset attributes

Attribute Name	Attribute Description	Data Type
TestId	A sequential identifier for a test case	String
Input a	Fist input to a test case	Numeric
Input b	Second input to a test case	Numeric
Input c	Third input to a test case	Numeric
Expected Output	The expected output of a test case	String

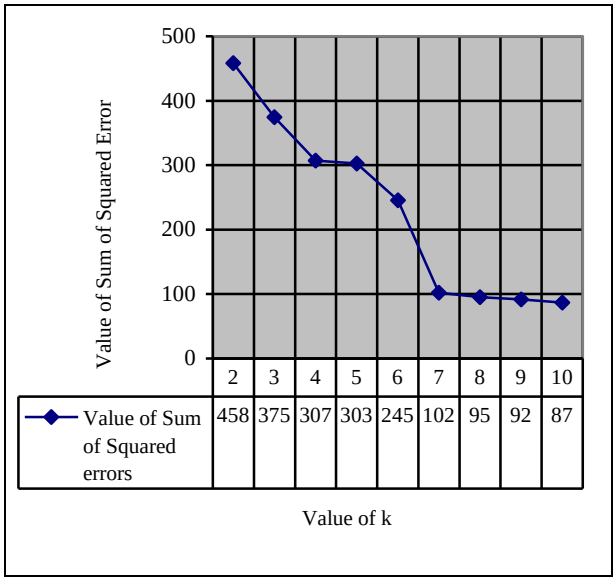

Figure 3. Graph for Elbow method

Table 3. Description of reduced test cases generated for Triangle Problem

Si. No.	Description	Reduced number of Test Cases
1.	Not a Triangle	19
2.	Equilateral	14
3.	Isosceles	10
4.	Scalene	20
5.	Length of the first side is invalid	19
6.	Length of the second side is invalid	5
7.	Length of the third side is invalid	13
Total		100

V. RESULTS AND DISCUSSION

The result evaluation is done by comparing the number of test cases generated originally to the reduced number of test cases as shown in Figure 4. Table 4 shows percentage of correctly classified instances and incorrectly classified instances along with weighted average of F-measure for each value of k used in our experiment.

From the above table we can see that for k=7 we have 100% correctly classified instances which is the highest in comparison to results obtained from other values of k. Also, the weighted average of F-measure reaches its perfect value 1 for k=7.

With the help of clustering technique the number of test cases required to test the program are reduced. This in turn will reduce the amount of time required to test programs with large number of lines used in industry and also the cost.

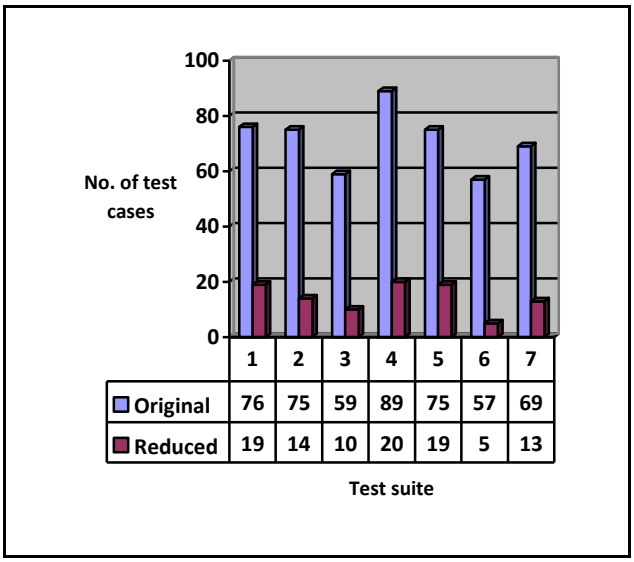

Figure 4. Graph showing original number of test cases vs reduced number of test cases

Table 4. Percentage of Correctly classified instances, incorrectly classified instances and Weighted average of F-Measure

Value of K	Correctly classified instances	Incorrectly classified instances	Weighted average of F-measure
2	97.8%	2.2%	0.978
3	98%	2%	0.980
4	95.2%	4.8%	0.952
5	96.6%	3.4%	0.966
6	98.6%	1.4%	0.986
7	100%	0%	1.000
8	99.6%	0.4%	0.996
9	98.6%	1.4%	0.986
10	99.2%	0.8%	0.992

VI. CONCLUSION AND FUTURE SCOPE

In this paper we present a mining approach, k-means along with elbow method to correctly estimate the value of k in k-means algorithm and their applicability in reducing the number of test cases by eliminating the redundant test cases. In future, we would like to include more parameters like cyclomatic complexity, branch coverage and method coverage in our dataset to enhance the effectiveness of our approach. Also, we would like to extend our work on large programs.

REFERENCES

[1] M. J. Harrold, R. Gupta., & M. L. Soffa, "*A methodology for controlling the size of a test suite*", ACM Transactions on Software Engineering and Methodology (TOSEM), Vol. 2, No. 3, pp. 270-285, 1998.

[2] T. Y. Chen and M. F. Lau, "*A new heuristic for test suite reduction*", Information and Software Technology, Vol. 40, No. 5, pp. 347-354, 1998.

[3] Zhenyu Chen, Baowen Xu, Xiaofang Zhang, and Changhai Nie "*A novel approach for test suite reduction based on requirement relation contraction*", In Proceedings of the 2008 ACM symposium on Applied computing (SAC '08). ACM, New York, NY, USA, pp. 390-394, 2008.

[4] H. Zhong, L. Zhang, and H. Mei, "*An experimental study of four typical test suite reduction techniques*", Information and Software Technology, Vol. 50, No. 6, pp. 534–546, 2008.

[5] T. Chen and M. Lau, "*A simulation study on some heuristics for test suite reduction*", Information and Software Technology, Vol. 40, No. 13, pp. 777–787,1998.

[6] L. Ramesh, "*Knowledge Mining of Test Case System*", International Journal on Computer Science and Engineering, Vol. 2, No. 1, pp. 69-73, 2009.

[7] L. Rameesh and G.V. Uma, "*An Efficient Reduction Method For Test Cases*", International Journal of Engineering Science and Technology, Vol. 2, No. 11, pp. 6611-6616, 2010.

[8] K. Muthyala, & R. Naidu, "*A novel approach to test suite reduction using data mining*", Indian Journal of Computer Science and Engineering, Vol. 2, No. 3, pp. 500-505, 2011.

[9] A. Saifan, "*Test case reduction using data mining classifier techniques*", Jouranal of Software, Vol. 11, No. 7, pp. 656-663, 2016.

[10] L. Ramesh, & G. V. Uma, "*Reliable Mining of Automatically Generated Test Cases from Software Requirements Specification (SRS)*", International Journal of Computer Science (IJCSI), Vol. 7, No. 3, pp. 87-91, 2010.

[11] R. Chauhan, P. Batra, & S. Chaudhary, "*An Efficient Approach for Test Suite Reduction using Density based Clustering Technique*", International Journal of Computer Applications, Vol. 97, No.11, pp. 1-4, 2014.

[12] B. Subashini, D. JeyaMala, "*Reduction of Test Cases Using Clustering Technique*", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, No. 3, pp. 1992-1996, 2014.

[13] L. Ramesh, G. V. Uma, "*UML Generated Test Case Mining Using ISA*", International Conference on Machine Learning and Computing, IPCSIT, Vol. 3, pp. 188-192, 2011.

[14] A. K. Upadhyay, A. K. Misra, "*Prioritizing Test Suites Using Clustering Approach in Software Testing*", International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue-4, pp. 222-226, 2012.

[15] Marie Fernandes , "*Data Mining: A Comparative Study of its Various Techniques and its Process*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.19-23, 2017.

[16] R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra, "*A Review: Design and Development of Novel Techniques for Clustering and Classification of Data*", International Journal of Scientific Research in Computer Science and Engineering, Vol.06, Issue.01, pp.19-22, 2018.

[17] T. M. Kodinariya, P. R. Makwana, "*Review on determining number of Cluster in K-Means Clustering*", International Journal of Advance Research in Computer Science and Management Studies, Vol. 1, No. 6, pp. 90-95, 2013.

[18] L. C. Briand, Y. Labiche and Z. Bawar, "*Using Machine Learning to Refine Black-Box Test Specifications and Test Suites*", *2008 The Eighth International Conference on Quality Software*, Oxford, pp. 135-144, 2008.

[19] I. H. Witten and E. Frank, "*Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann Publisher, San Francisco, 2005.

Authors Profile

Mr. Ankit Pandey pursued B.Tech (Computer Science & Engineering) from Amity University, U.P., India in 2012. He is currently pursuing M.Tech. (Computer Science & Engineering) from K.N.I.T., Sultanpur, India. His main research work focuses on Software Engineering and Data Mining. He has 3 years of software industry experience.

Dr. A. K. Malviya is currently working as Professor in Department of Computer Science & Engineering, K.N.I.T., Sultanpur, India. He has published more than 50 research papers in reputed international journals and National/ International conference proceedings. His main research work focuses on Software Engineering, Web Engineering and Data Mining. He has approx 20 years of teaching experience.

