Study on Software Process Metrics using Data Mining Tool -A Rough Set Theory Approach

V. Jeyabalaraja
Research Scholar,
Dept of Statistics,
Loyola College, Chennai-34

T. Edwin Prabakaran
Associate Professor,
Dept of Statistics,
Loyola College, Chennai-34

ABSTRACT
Software industries are optimizing their resources to obtain the best quality software from minimum cost through identifying the potential resource for the assignments. The process matrices' are supported to increase the efficiency of the project in the development scenario. The process metrics functionalities are determined as per the development of code in terms of line of code or the size of the code which is developed by the developers without the error or the minimal development error. This paper aimed to identify the competence of the developers in their selected skill set relevant to the assigned tasks. It provides the developers those who are worked in the minimum skill set components and identified as a weak set of employees through equivalence algorithm of the rough set theory. The functional attributes are observed and analyzed to find out the maximal error process which leads to the identification employees set those are made more modification with low level expert knowledge in the working area or project. The observation, analysis and the experimental procedures using Quick reduct are presented in this paper.

Keywords
Quick reduct Algorithm, Rough set theory Approach, Software Developers Efficiency

1. INTRODUCTION
Software development system streamlined through the observation its valued measures such as project metrics, process metrics and product metrics.[5] Software metrics are the units to measure the attributes of the software process and its developmental activities. They can be classified into three categories such as product metrics, process metrics, and project metrics. *Product metrics* describe the characteristics of the product such as size, complexity, design features, performance, and quality level. *Process metrics* are used to improve software development and maintenance. *Project metrics* describe the project characteristics and execution [4].

Process and product metrics are playing vital role in the success of software delivery to their clients. Process metrics help in improving the quality of different system component & comparisons between existing systems . The metrics such as Line of Code (LOC), Token Count (TC) and Functional Process Count (FPC) measured and evaluated to determine the quality and sustainability of the software.

The software development outcomes are differing based on developers skill and their contribution level in the assigned tasks. The contribution of the skill will reflect on the error of the developers while developing the software. The developer's skills are expected to process and determine their efficiency to assign the fourth coming assignment according to the predicated skill set.

2. SCOPE AND OBJECTIVES
Data mining preprocess aims to produce the quality mining result in descriptive and predictive analysis. The software developer's process metrics data analysis and the generation of equivalence set attempted identify the potential developers. The paper attempted to obtain the relationship of software developer's efficiency as per the observed data using Quick reduct – rough set theory approach. The quick reduct –rough set theory approach explained below

3. METHODOLOGY
Data mining techniques are used in different application to analysis and predict the data for decision support system. From the observed process metrics factors, the potential variables are identified using reduction techniques [2] [6]. The predictive approach re used all the interdisciplinary sectors. The software industry developer's skill set potential identification and the predication of developer suitability attempted using data mining approach. The data set reduction carried out for various applications using rough set theory [1].The analysis process is achieved similar result from the minimum selected data through dimensionality reduction process [7].The developers minimal modification shows their efficiency on their development skill set[8].This error identification and the prevention process of developmental error will increase the software quality and its usage[9]. The process of reduction and convention of rough set approach for the reduction approach on software process metrics data is explained below.

Quick Reduct - Rough set theory in software development processes arrived the minimal error on the software efficiency. the developers contribution are arrived the following steps

i. The software developers involvement of each forms are observed

ii. The observed functional process are construed as a associate set of a individual developers in a form .

iii. Dev_Phase_FP = { F1,F2,F3…Fn}

iv. Where Dev_Phase_FP is Developers Functional Point in Forms (n =26)

v. The each form and functional points are notified according to the usage .

i. Construction of Associate Relationship

The each functional point and its relationship are constructed as a matrix represented in a associative relation process

Ass_rel = { E, M, F,FPi-n,Erc,Ers }

Where

E- Developer's Identification

M – Module

F- Form

FP – Functional Point utilization (1 – FP used , 0 – FP not used)

Erc – Error details based on LOC

Ers – Error Details based on Size

ii. Combinational Rule Generation

The combinational rule generated based on the occurrence of the Functional point. In this process, the developer's developed forms and the functional points are combined in all possible combination. The knowledge repository created based on the Combinational Rule

CR1 = Rel { (E1, F1,FP1) V (E2, F2,FP2)… (Em, Fm,FPm)

Where m is number of variables are selected for the rule generation. In this process m =5

iii. Equivalence Set Generation (ESG)

As per the constructed Combinational rule associative Unit matrix values, the occurrence equivalence set and its functional points are processed .The results are observed to determine the developer efficiency based on the modification carried out by the developer in the assigned task.

iv. Rough Set - Quick Reduct Algorithm(QRA) for Minimal/Maximum Error

The equivalence set fetched and its pair values of two different rules results are compared. The developers generated error and their performance is derived with five attributes set of all combinational functional points of each form.

The minimal error values are reflects the quality of the software; the derived approach implemented using rough set theory. The core concept of the Rough set theory explained below

3.1 Quick –Reduct: Rough sets theory

Rough sets theory was introduced by Z. Pawlak [10] as a mathematical tool for data analysis. It does not need external parameter to analyze and make conclusions about the datasets. Rough sets offer many opportunities for developing many Knowledge Discovery methods using partition properties and discern ability matrix Rough sets have many applications in KDD among them, feature selection, data reduction, and discretization . Rough sets can be used to find subsets of relevant (indispensable) features. Combining rough sets theory with a known classifier yields a wrapper feature selection method since it uses the class label information to create the indiscernibility relation. It provides a mathematical tool that can be used to find out all possible feature subsets. Analyzing the result of all possible feature sub set the same result could be achieved without using all the attributes involved in the process. This attribute reduction aid to determine the factor which influence the result using quick reduct algorithm.

Rough set concept can be defined quite generally by means of topological operations, interior and closure, called approximations.

The given a set of objects U called the universe and an indiscernibility relation $R \subseteq U \times U$, representing our lack of knowledge about elements of U. For the sake of simplicity assume that R is an equivalence relation.

The equivalence class of R determined by element x will be denoted by R(x). The indiscernibility relation in certain sense describes our lack of skill about the universe. Equivalence classes of the indiscernibility relation, called granules generated by R, represent elementary portion of skill to perceive due to R.

Rough set approach shows clear connection between functional points and the occurred error. The above concept mapped to the developers approach and the influencing factor is determined. The above concept mapped to the developer's contribution with minimal error. Quick reduct approach is described below

QR *(C, D)*

C, the set of all conditional features;

D, the set of decision features.

(1) $R \leftarrow \{\}$

(2) do

(3) $T \leftarrow R$

(4) $\forall x \in (C - R)$

(5) if $\gamma_{R \cup \{x\}}(D) > \gamma_{r}(D)$

(6) $T \leftarrow R \cup \{x\}$

(7) $R \leftarrow T$

(8) until $\gamma_{r}(D) = \gamma_{c}(D)$

(9) return *R*

The initial relationship set declared as a NULL set. Based on the generated rule the occurrence of the index values are presented and the relation ship set is generated using above presented pseudo code. The implementation process data collected from real time development environment explained below.

4. DATA COLLECTION

This research work is adopted the scientific research methods and measure the real time development project along with the components observation. The data collection process carried out around six-month observation in the development process alone. There are 27 software developers observed including the team managers and the project leader. The observation process summarized below

Duration	:	Nov-2009 to April 2010
Number of developers	:	28 in which 3 team leaders and one project manager
Observed Modules	:	24
Number of Form	:	1420
Number of functional Components	:	27

In the observation, each form and the existing components of the forms as well as its code are evaluated .each form size of

code, original production; modification and the occurred size of modification are tabulated.

The decision process on the quality factors depends on the developer's efficiency and the used components in the form and the modules. If the software developers are expert in the corresponding technology and the requirement process is defined well then error /modification is minimized .As per the observation , the total development scenario in the IDE is compressed with 27 functional process components. Therefore, the data-mining algorithm is adopted to redact the variable to identify the key variable which determines the success of the development process.

5. ANALYSIS OF RESULT AND INTERPRETATION

The collected data set is preprocessed and the Equivalence set is generated according to the generated rule.

The existence functional point rule of Button Element (3), Submit Button(4) , Reset Button(5), Text Filed (6) Select Element(11) presented below.

Table 1.Calculation of Equivalence set

Index	3	4	5	6	7
1	0	0	1	0	1
2	1	1	1	0	1
3	1	0	1	0	1
4	0	0	1	1	1
5	1	1	0	0	1
7	0	1	1	0	1
9	1	0	0	0	1

Equ_set 1= Non Existence of FP (3) U Non Existence of FP (4) U Existence of FP (5) U Non Existence of FP (6) U Non Existence of FP (11) is represented as Rel2 (0,0,1,0,0) ∩ E Where E is the identification of developer. The rule is generated 41 set values as a equivalence set from the observed data and presented as a first row of below table 1.

Equ_set2 = Existence of FP (3) U Existence of FP (4) U Existence of FP (5) U Non Existence of FP (6) U Non Existence of FP (11) is represented as

Rel2 (1,1,1,0,0) ∩ E Where E is the identification of developer. The rule is generated 41 set values as a equivalence set from the observed data and presented as a second row of the above table 1.

Where 3,4,5,6 and 7 are index values of functional points. The index 1 is the representation of the equal set of generated rule pattern of (0,0,1,0,1).for all the values the representation set is presented as a table below.

From the equivalence relationship table , the representation table is constructed with the representation of any member value of equal set.

Representation Set

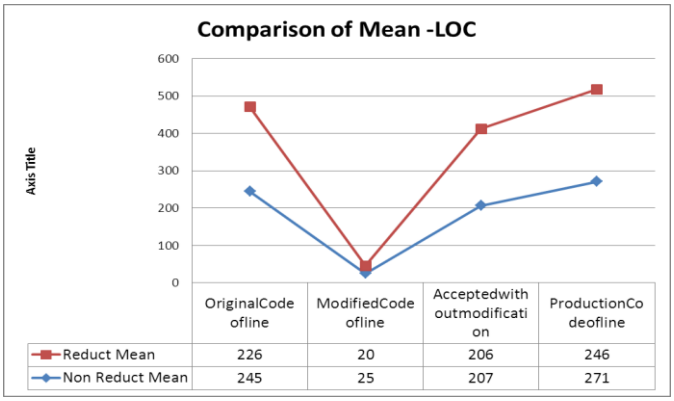
Index	3	4	5	6	11	Equal Set
1	0	0	1	0	0	{1,79,86,93,128,166,192,261,282, 310,334,347,352,451,452,525,577, 590,654,656,693,715,738,759,776, 801,871,915,1069,1094,1100,1128 ,1133,1154,1208,1270,1327,1350, 1382,1389,1403} - 41
2	1	1	1	0	0	{2,45,118,133,176,177,212,297,32 2,343,353,360,388,416,529,548,58 1,621,756,813,891,928,983,1025,1 051,1052,1072,1093,1106,1204,12 51,1345,1371,1418} - 34

34567	89101112	1314151617	1819202122	2324252627
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
7	7	6	6	7
9	8	7	7	8
10	9	8	8	9
12	11	10	10	12
13	12	11	12	14
14	13	12	13	15
18	15	13	14	16
20	18	14	15	17
21	19	15	16	18
27	20	17	17	19
46	22	19	19	24
728	24	20	20	25
	27	22	22	26
	31	28	23	27
	32	29	24	32
	33	31	25	33
	38	34	27	35
	39	36	30	37
	40	49	32	40
	45	56	35	41
	46	65	37	43
	53	78	39	49
	56	106	41	51
	64	112	52	52
	74	133	59	59
	110	148	61	65
	139	155	87	73

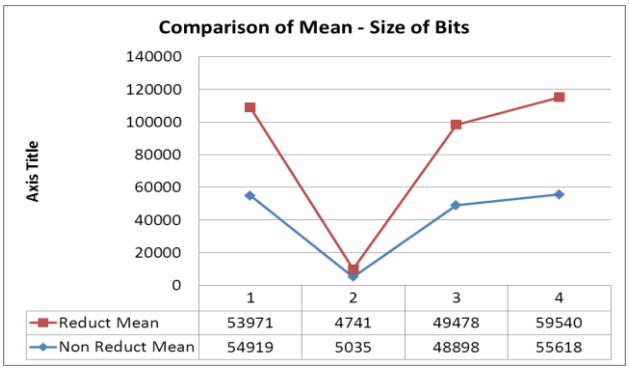
In the above table, from the functional point pairs are repeated. The repeated values are removed and the unique representation values are generated.

Resulted Reduct Set	
1	33
2	34
3	35
4	36
5	37
6	38
7	39
8	40
9	41
10	45
11	46
12	49
13	51
14	52
15	53
16	56
17	59
18	61
19	64
20	65
21	73
22	74
23	78
24	87
25	106
26	110
27	112
28	133
29	139
30	148
31	155
32	728

The resulted set contains 64 set of values. The 1420 set of values are reappeared with 64 set values. Based on the process of unique data set vales the error and the modification are evaluated and their results are discussed.



As per the evaluation of reducted set mean and the non-reducted data set vales, the mean value of original code , modified code a, accepted without modification and the production of LOC is directly reflected based on the above figure. The same attributes of quality measures are presented as a graph below . The graph represents that the change in the original data set and the reduced data set produced the similar result given below.



6. CONCLUSION

The rough set approach implemented on the software process metrics data set and arrived the 64 records out of 1420 records using equivalent set. The process time is reduced enormously while compare to process the reduced data set with normal raw data set. The reduction has attempted and succeed from 1420 tuples to 64 tuples through Rough Set - Quick Reduct algorithm .The obtained results are direct proportionate to the non reduct data set and the reduct data set as per the Mean and Median of the processed LOC and Size of bits of original production, modification, accepted without modification and production after the modification. Therefore the reduction process ensures the result with minimum set of equivalence and effective values.

7. REFERENCES

[1] A.E. Hassanien, Z. Suraj, D. Slezak, and P. Lingras, Rough Computing: Theories, Technologies, and Applications, New York: Information Science Reference, 2008.

[2] A.Skowron, Z. Pawlak, J. Komorowski, and L. Polkowski, "A rough set perspective on data and knowledge," in Handbook of Data Mining and Knowledge Discovery,Oxford: Oxford University Press, 2002, pp. 134–149.

[3] Butalia, A., Dhore, M.L. and Tewani, G.(2008), Applications of Rough Sets in the Field of Data Mining,

Emerging Trends in Engineering and Technology, ICETET '08. First International Conference, 498- 503.

[4] Cem Kaner, and Walter P. Bond , (2004),Software Engineering Metrics:What do they measure and how do we know?, 10th International Software Metrics Symposium.

[5] Everald E. Mills (1988) Software Metrics, Software Engineering Institute, Institute Carnegie Mellon University , Seattle, Washington 98122.

[6] J. Han, X. Hu, and T.-Y. Lin, "Feature subset selection based on relative dependency between attributes," in Proc. of the 4th International Conf. on Rough sets and Current Trend in Computing, Uppsala, 2004, pp. 176–185. .

[7] K. Thankavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory," A Review, Applied Soft Computing, vol. 9, no. 1, pp. 1–12, 2009.

[8] Kan, Stephen H., Jerry Parrish, Diane Manlove "In-Process Metrics for Software Testing", IBM ,Systems Journal, Vol 40, No.1, February 2001.

[9] V.R. Basili and B.R. Perricone, "Software Errors and Complexity," Comm. ACM, vol. 27, pp. 42-52, 1984.

[10] Z. Pawlak, "Rough set approach to knowledge-based decision support," European Journal of Operational Research, vol. 99, no. 1, pp. 48–57, 1997

8. AUTHOR'S PROFILE

V.Jeyabalaraja secured B.Sc., Master of Computer Applications and M.Phil in the field of Computer Science. He is working as a faculty in the Department of Computer Applications, Velammal Engineering College, Chennai-66. His research interest includes Software Engineering, Data Mining and Statistical Applications.

Dr. T.Edwin Prabakaran is working as a Associate Professor in the Department of Statistics, Loyola College, Chennai-34. His areas of research include Data mining and Statistical Applications.