Defect Data Analysis and Experiments with Defect Prediction Methods Using Topic Metrics

Prudhvi Ratna Badri Satya
A02057243
Computer Science Department
Utah State University
Logan, Utah.

***Abstract*—Software development is a complex task which requires good understanding and sound knowledge on the software system. Research Analysts have proposed different measurements in view of measurable parts of the source code entities such as techniques, classes, documents, or modules, and the social structure of a product that extend with an end goal to clarify the connections between software development and software defects. Notwithstanding, these measurements to a great extent disregard the real usefulness, i.e., the calculated concerns, of a software system framework, which are the principle technical concepts that mirror the business rationale or area of the system. The success of a software is always on par with the number of bugs that are to be found and rectified, in order to enhance the quality of the software product. The maintainers of the software system play a very crucial role in the maintaining the bug reports which contains details of a particular software failure, and detailed description on how these failures have been regenerated in the system. Defect prediction is one of the important areas that needs to be focused on in order to improvise the product quality. In this paper, I conducted the extensive data analysis and conducted experiments with the defect prediction methods using top metrics on the five publicly available defect datasets of projects namely Jdt, Equinox, Mylyn, Lucene and Pde.**

***Keywords*—Topic metrics, Defect Prediction, Correlation, Principal Component Analysis .**

I. Introduction

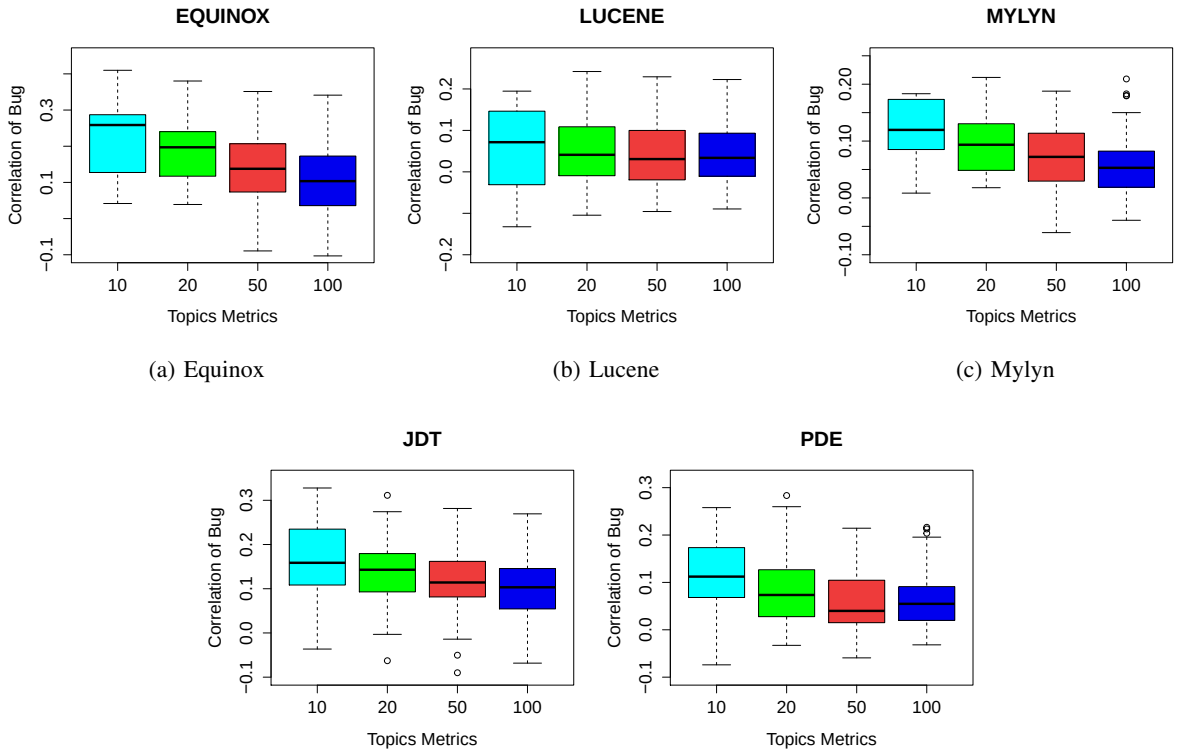
Software analytics speaks to the base segment of the software analysis that generally aims at generating findings, conclusions, and evaluations about software systems and their implementation, composition, behavior, and evolution. Software analytics uses and consolidates methodologies and strategies from statistics, prediction analysis, data mining, and scientific visualization.

Software analytics is used to explore the data in order to predict the defects in the software system which is used to improve the quality of the software. The process of finding the bugs in the software system is one of the cost and time consuming tasks. Reliability models are developed to predict the defects and the failure rates in the software. Large complex multivariate statistical models have been proposed to find a single complexity metric that will account for defects. For the purpose of this project, I have performed study on the publicly available defect data sets.

The analysis has been carried out on the defect data sets of five different projects named Jdt, Equinox, Mylyn, Lucene and Pde. The In-depth and relevant information cannot be obtained easily by scrutinizing the raw data as such without the sound support of the software analytic technologies. The information obtained by using software analytics is the most relevant information that conveys the proper understanding or knowledge towards performing the given target assignment.

In this Project, I have used linear regression model in order to predict the correlation values of 20 topic metrics of the five data sets. I have exhibited this model to apply Principal Component Analysis on different topics accordingly in order to calculate the Explanative power and the Predictive power. The explanatory power is calculated using the AIC score and the predictive power is calculated using the spearman correlation using the techniques of the cross validation. The lower the AIC values better the Explanative power and higher the rcor value better the Predictive power as per statistical analysis. The framework of a software system is nothing but a collection of software artifacts that portray a specialized perspective of various technical aspects. Those viewpoints are expected to have diverse levels of bugs, in this manner, because distinctive levels of bugs to the important software artifacts. I have used topic modeling to gauge the concerns in source code, and utilizing them as the input information for linear regression prediction models. In this paper, I focus at answering the following research questions:
1.What is the correlation of topic metrics to BUG? This question is answered by plotting as boxplots the correlation of BUG and topic metrics for the number of topics of 10, 20, 50, and 100.
2. Do topic metrics provide better and additional explanatory power and predictive power to base metrics??

The organization of the paper is as follows: Section 2 explains about the methodology. Correlation of Topic Metrics with bug in discussed in Section 2A. Section 2B discusses about the Explanatory Powers and Predictive Powers of the prediction model on the datasets considered. Results and Discussion are presented and discussed in Section 3. Finally, Section 4 contains the conclusion.



(a) Equinox (b) Lucene (c) Mylyn

(d) JDT (e) PDE

Fig. 1: Correlation of Topic metrics to Bug

II. Methodology

A. Correlation of Topic metrics with Bug

The five datasets that are analyzed in this paper are Equinox,Lucene,Mylyn, JDT and PDE . Each dataset contains topic files. These files consists of three types of metrics namely Topic metrics, Bug metrics and Base metrics.

The Topic metrics is the log transformation of the number of words assigned for each topic (from 1 to K). For example, topic5log will have 5 metrics V1-V5, each for a topic. BUG is the actual numberof post-release bugs. The Base metrics has the metrics which has the properties of the code (LOC, BF, HCM).

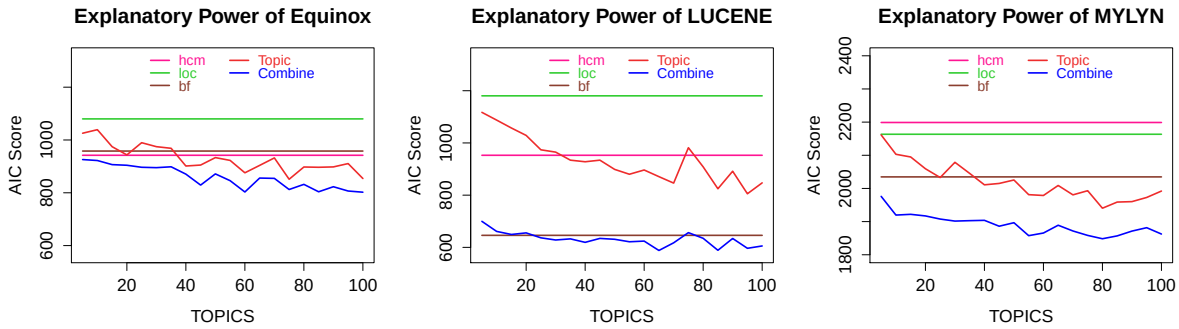
LOC: number of lines of code (measuring the code complexity) BF: number of prior bug fixes (measuring the defect history) HCM: the entropy of code changes (measuring the complexity of code changes).

The research question,*" What is the correlation of topic metrics to BUG?"* is answered by plotting and analyzing the boxplots for correlation of BUG and topic metrics for the number of topics of 10, 20, 50, and 100.

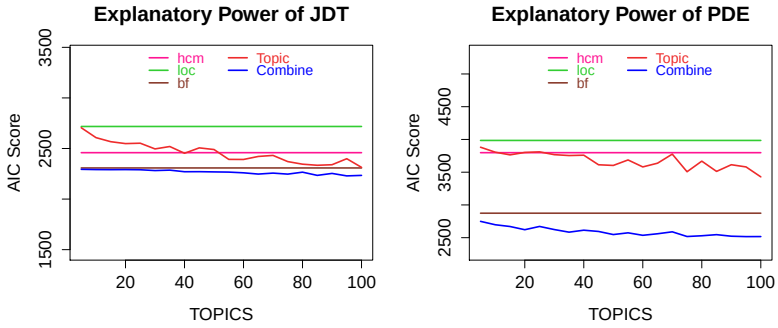
B. Explanatory and Predictive powers of the prediction model

The research question "Do topic metrics provide better and additional explanatory power and predictive power to base metrics?" is answered by plotting: Explanatory power and predictive power of 3 base metrics as baseline.The explanatory power of a prediction model is measured with AIC score. The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model. The predictive power of a prediction model is measured with Spearman correlation. For analyzing the explanatory power and predictive power , when only topic metrics are used, Plots for K = 5, 10, ... 100 topics are done. For each K, P is varied (number of selected principal components) and choose what provides the best explanatory/predictive power. Since the prediction model involves topic metrics principal component analysis is used.

Principal Component Analysis: Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much



(a) Equinox (b) Lucene (c) Mylyn



(d) JDT (e) PDE

Fig. 2: Explanatory Power of Topic Metrics

of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables.

For analyzing the explanatory power and predictive power when both topic metrics and base metrics are used, plots for K = 5, 10, ... 100 topics are plotted.The results in tables are summarized, which list the explanatory and predictive power for the best K, in addition to those for base metrics. As explanatory power and predictive power are not comparable, they are plotted in separate tables and figures and are analyzed.

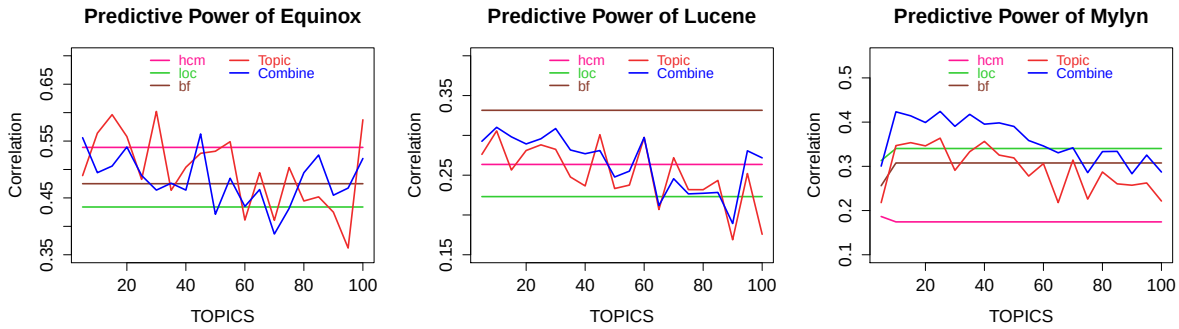
For computing the explanatory power, the AIC values are used. For each topic, you vary number of selected principal components and choose the components with low AIC values. To find the best topic, the minimum for the minimum AIC values of each topic metrics are considered and calculated. Apart from the base metrics, the combined metrics and all the three base metrics are correlated with bug using Linear Regression Model and then best metrics are analyzed. The prediction performance of the PCA components are analyzed using Linear Regression Model (LM model). We compute Spearman's correlation between the predicted PCA scores and bug metric. Such evaluation approach has been broadly used to assess the predictive power of a number of predictors. In this cross validation, for each random split, we use the training set (50 percent of the dataset) to build the regression model, and then we apply the obtained model on the validation set (50 percent of the data set).

III. Results and Discussion

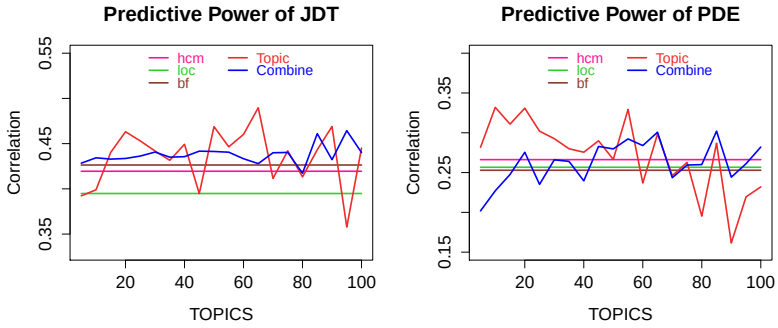
The box plots as shown in fig (1) clearly shows that there is a decrease in correlation of the topic metrics with the bugs as the number of topics increases. The median value of the metrics decreases as the number of topics increases. Even though all the metrics are present it is not mandatory that all of them provides information which is necessary.This leads us to the using of the prediction method of principal component analysis which gives us more information.

From the fig (2) it can be clearly assessed that the combine metrics has greater explanatory power when compared to other metrics as the AIC score for the combined metrics is less when compared to topic and other three base metrics.

From the fig (3) it can be stated that the combine metrics has greater predictive power when compared to other metrics such as topic and base metrics since the correlation value is higher for combined metrics.



(a) Equinox (b) Lucene (c) Mylyn



(d) JDT (e) PDE

Fig. 3: Predictive Power of Topic Metrics

Table I shows the predictive powers for the best k values of all data sets.
From Table II we can know that the combine metrics have minimum AIC score when compared to other metrics.

Dataset	LOC	bf	hcm	K (Topic)	Corr (Topic)	K (Combine)	corr (Combine)
Equinox	0.434	0.475	0.538	30	0.602	45	0.562
Lucene	0.223	0.331	0.263	10	0.305	10	0.310
Mylyn	0.340	0.307	0.186	25	0.363	25	0.424
JDT	0.394	0.426	0.419	65	0.489	95	0.464
PDE	0.256	0.252	0.266	10	0.331	85	0.302

TABLE I: Predictive powers for the best K values of all datasets

IV. Conclusion

In this paper, I have discussed Explanatory Powers and based on these results i have calculated the Predictive powers of all the metrics on the datasets. To analyze the efficiency of the predictive powers i have cross validated the results and concluded that combine metrics has more explanatory and predictive powers.

Dataset	LOC	bf	hcm	K (Topic)	AIC (Topic)	K (Combine)	AIC (Combine)
Equinox	1079.86	958.06	942.07	75	851.23	100	802.21
Lucene	1180.45	646.04	952.73	95	805.27	65	588.08
Mylyn	2163.37	2034.80	2198.88	80	1940.35	80	1848.32
JDT	2717.72	2307.71	2458.73	100	2313.59	95	2229.21
PDE	3984.95	2873.84	3797.97	100	3428.47	95	2516.21

TABLE II: Explanatory powers for the best K values of all datasets

Acknowledgements

I sincerely thank Dr.Tung Nguyen for his valuable support through out this project. I also acknowledge other students in this class for their extended help in this project.

References

[1] Tse-Hsun Chen, Stephen W. Thomas, Meiyappan Nagappan, Ahmed E. Hassan: Explaining Software Defects Using Topic Models

[2] T Patrick Knab,Martin Pinzger, Abraham Bernstein: Predicting defect densities in source code files with decision tree learners.