System Verification (SV) Resourcing Predictive Analytics

Rahul D. Sapnar
Research Scholar,
Department of Computer Engineering,
Vishwakarma Institute of Technology,
Pune, India

Virendra D. Pawar
Assistant Professor,
Department of Computer Engineering,
Vishwakarma Institute of Technology,
Pune, India

Smideep Nair
Module Lead,
Pune, India

***Abstract* - Human Resource Management is an important area of project management. Software projects are more critical as compared to projects of other disciplines because success of software projects dependents on human resources. In software projects, Project Manager (PM) allocates resources and level resources using Resource Leveling techniques which are already implemented in various project management softwares. Predictive analytics is proving to be very useful for achieving resource estimation for the System Verification (SV) process of software development in IT company. This work is based on the historical data set collected from various sources like the company's data warehouse and Defect Management Tools for example JIRA and Quality Centre. This article provides the conceptual architecture for resource estimation to fulfill the requirement for System Verification (SV) process of software development.**

***Keywords* - Resource Management, Predictive Analytics, Machine Learning, Support Vector Machine.**

I. INTRODUCTION

Resource Allocation is a critical task but it is still governed by dated models like COCOMO and CPM. Human resource allocation to software project activities is very important because project's success depends on how resources are set out to software project activities. PERT and CPM are traditional methods and assist project managers to analyze network diagrams to allocate Human Resources to project activities. These however are not complete resource allocation techniques. As an Industry Resource management tool, predictive analytics helps to improve the quality decisions taken by decision makers address critical issues of resource management and estimation. This article presents an analytical study for resource planning.

In recent years, research has shown that there has been extensive study of various machine learning techniques and algorithms for classification and prediction analysis. Among these are Bayesian network classifier [3], k-nearest neighbor classifier [4] and decision tree [5]. These methods are conventional learning methods compared to those new approaches; however, it has simple algorithms and relatively high efficiency. For examples, there are neural network, support vector machines, fuzzy kmeans and maximum entropy models also have good results.

Predictive analytics involves technologies and methods enabling users to spot the patterns and trends in the data set having large number of independent variables [1]. According to research reports, the predictive analytics market is estimated to grow from $1.70 billion in 2013 to $5.24 billion in 2018 at a compound annual growth rate of 25.2%. Many organizations now view data and analytics as important corporate assets. Real-time decision making and resource management is part of advanced analytics used to achieve optimization functionalities. In this sense, optimization means managing, to the best of an organization's ability, the deployment of predictive analytics such that organizational objectives are achieved as opposed to just local, functional area objectives.

Create a statistical basis for resource requirement or number of people required to perform particular process like testing and development of the product. The Statistical Model is generated based on the historical data to calculate the expected number of resources (here persons) required for testing of product. Extend the statistical model to estimate resources based on value features of the product for example cost of product, revenue generated from the product etc.

The rest of the paper is as follows: The next section describes motivation of proposed system. Section 3 contains the objectives and roadmap of the system. The Framework of proposed system is given in Section 4. Finally, the experimental results are given in section 5 & conclusions are given in section 6.

II. MOTIVATION

The traditional software development teams are unable to prioritize the available resources based on their skill set. They also fail to estimate the resourcing requirements for the application according to the business parameters. The tool is an in-house reporting and data analysis tool that looks at the various measures of productivity and helps management to make predictive decisions based on future market strategy for product.

IMPACT OF THE PROBLEM

A. Imbalance of Resources v/s User

Some projects tend to the problem of being either under resourced or over resourced, leading to improper resource utilization. Lack of resource optimization is certain to result in shortages or delays in project delivery or on the other hand underutilization of resources. Without an analysis of how often resources need to be replenished, these necessities will not be found where and when needed.

B. Increase in Expenses

Cash flow problems can go so far as to result in the inability to pay employees on time. This is bound to have a negative impact on employee loyalty and retention.

C. Productivity and Timeline

Without planning, there will be no mission statement and no vision. Employees are most productive when they understand the bigger picture behind what they are doing, so productivity will decrease. There is also likely to be much wasted time, as some workers will be duplicating the work of others, while some essential tasks will be overlooked. This is all likely to result in the need for crisis management. In addition larger projects will take longer than necessary, or may never reach completion, because no one did the planning necessary

to break them down into more manageable segments.

III. OBJECTIVES

The primary objective is to generate a statistical model to study the data points available like number of defects, verified defects and customer found defects based on the historical data. Use the data points to evaluate the confidence measures for unit testing confidence and regression confidence. By using the confidence values create the data model to calculate the expected number of resources required for testing a particular Application. . It basically includes the requirements for managing the personnel data, controlling authentication and authorization mechanism, and evaluating of employees' performance.

More specifically, SV resourcing controls and manages the personal database such that any user with different role types as manager, admin, employee, and human resource will be able to manipulate their personal data.

IV. PREDICTIVE ANALYTICS FRAMEWORK FOR SV

1. Data-Set Collection

This data which is collected, along with data sourced from the DW, is then collated and displayed in the form of graphs to the users. Since users are not assigned permissions directly, but only acquire them through their role (or roles), management of individual user rights becomes a matter of simply assigning appropriate roles to the user's account, this simplifies common operations, such as adding a user, or changing a user's department.

The first step in predictive analytics process is data collection. The data is collected from resources such as data warehouse and the defect management tool. The predictive model for SV Resourcing requires features like MR count of the product, Resolved MR count, Customer found defects (CFDs) of the product and total number of defects. The month wise historical data from last five years is collected from the available sources. The feature set for every product is collected from company data warehouse. Data collected from the company data warehouse and defect management tools are generally structured data. Unexpected changes recorded about underlying data distribution over time are referred to as concept drift [6]. The data type of independent variables is integer and data type of dependent variables is text because it is categorical data.

The data set consists of 60 records describing month wise details of particular product's defect information. The feature set of independent variables are collected in 5 columns named Months, Closed MR count, New MR Count, Resolved MR Count and Count of Defects. The dependent variable is in the last column named Resources Required. The dependent variable is a categorized data into S- small, M-medium, L- large, XL- extra-large. The following Table 1 shows the format of collected data.

Table 1: Input Data Set

Month	Closed MR	New MR	Resolved MR	Tot. Defects	Resources Req.
Mar-2011	5	26	8	122	S
Apr-2011	11	10	10	395	M
May-2011	12	11	16	285	M
Jun-2011	25	40	27	258	M
Jul-2011	15	42	17	272	M
Aug-2011	39	37	43	344	M
Sep-2011	30	56	54	295	M
Oct-2011	64	36	48	335	M
Nov-2011	38	57	42	291	M
Dec-2011	47	45	28	263	M

2. Building Predictive Model

First, we assemble the data for training purposes. Because a predictive model combines data with predictive modeling techniques or tools, the data presented to such techniques determine the model's learning patterns. The predictive model is generated using machine learning approaches. Here, the technique called Support Vector Machine (SVM) is used to build the model. Support vector machine as classifier must be trained before it can be used for categorization. Training of the Support vector machine classifier is done by SVM ^{multiclass} proposed by Joachim and available at http://svmlight.joachims.org/svm_multiclass.html.The SVM library also helps to train the model by using LibSVM() class in java environment and available at https://www.csie.ntu.edu.tw/~cjlin/libsvm.html. The set of training data and a specification of the pre-defined categories are required to train the SVM[2]. Following Fig.1 shows the overall architecture of proposed system.

The data set consists of 60 records. The total data set is divided into two parts. 10% of whole data set is randomly considered as test data and remaining 90% is considered as training data. Training of a SVM classifier based on supervised learning is done by using SVM*multiclass* available at link http://svmlight.joachim.org/svm_multiclass.html. SVM*multiclass* consists of a learning module (svm_multiclass_learn) and a classification module (svm_multiclass_classify). The classification module can be used to apply the learned model to new examples. In the experiment using SVM, the input file"SVMInput.txt" contains the training data set. This file is trained using *svm_multiclass_learn* and the result in the model which is learned from the training data in train file. To make predictions on test instances, *svm_multiclass_classify* reads model file generated by SVM*multiclass*.

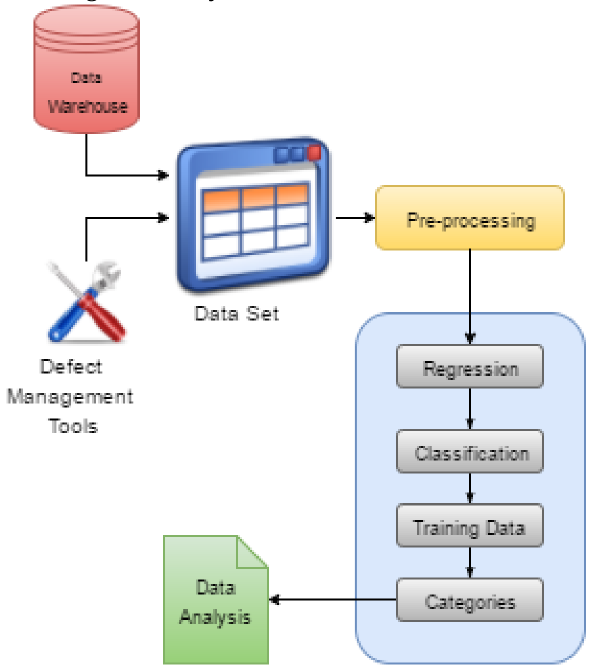

Figure 1: Proposed System Architecture

For all test records in"test.txt" the predicted classes are written to "output.txt". The first value in each line is the predicted category, and each of the following numbers is the discriminated values for each of the n categorization. The performance of SVM is largely dependent on the selection of the kernel parameters. The string *c* is a learning option to determine the trade-off training error and margin, and string *t* is a kernel option to store with the vector. In SVM*multiclass*, the value of *t* range from 0 to 4. This experiment uses four values of *t* (0, 1, 2, 3) that is combined with the value of *c* range from 0.10 – 1.0. The boundary equation of SVM is given by eq.(1).

$$f(x) = w^T x + b \qquad (1)$$

Here, *b* is the bias value, x is the input vector and *w* is its weight vector. The vector of test data set is applied on the following equation eq.(2) to get the predicted output.

$$f(\mathbf{x}) = \sum_{i}^{N} \alpha_i y_i (\mathbf{x}_i^\top \mathbf{x}) + b \qquad (2)$$

Training a classifier based on supervised learning is also done by using data mining tools like Weka 3.6. The input file *"SVMInput.txt"* is converted into *"SVMInput.arff"* as Weka tool accepts input files wile *.arff* extension. The classifier used by Weka tool is LibSVM. The parameter for the LibSVM classifier is gamma value is set to 0.25(i.e 1/no. of independent attributes)and another parameter is kernel type consists of linear function.

V. EXPERIMENTAL RESULTS

As mentioned above, the data set consists of 60 records and among that 10% data set i.e. 6 records is selected randomly as test data and remaining 54 records are used as training data. By using 90% of total data set the predictive model is generated. The test data set is then applied on the model generated by LibSVM function in weka. The accuracy of the predictive model is 80%, as shown in Fig2.

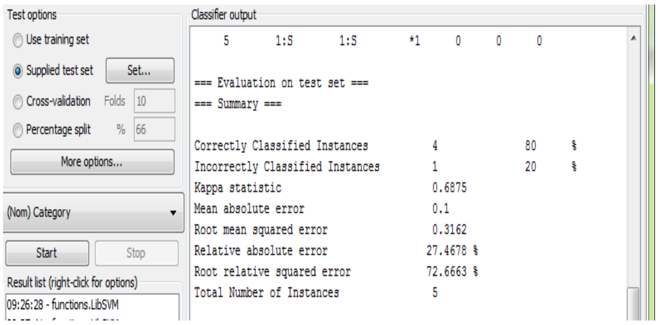

Figure 2: Result

VI. CONCLUSION

The proposed predictive model is a feasible solution to forecast the project resource requirement based on work-in-hand, quality of resources and other business factors affecting product. The whole proposed system is achievable. The system establishes the quality of statistical scale to adjust for resource quality in project resource estimations.

ACKNOWLEDGEMENT

This work has been supported by Smideep Nair. Also we would like to thank Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, for their extended help and suggestions.

REFERENCES

[1] Michael Goul, Sule Balkan, Dan Dolk, "Predictive Analytics Driven Campaign Management Support Systems", 48th Hawaii International Conference on System Sciences, pp. 4783-4791, 2015.

[2] Erlin, Unang Rio and Rahmiati, "Text Message Categorization of Collaborative Learning Skills in Online Discussion Using Support Vector Machine", International Conference on Computer, Control, Informatics and Its Applications, pp. 295-300, 2013.

[3] W. Lam and K.F. Low, "Automatic document classification based on probabilistic reasoning: Model

and performance analysis," Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol.3, pp. 2719-2723, 1997.

[4] P. Soucy and G.W. Mineau, "A simple k-NN program for text categorization," The first IEEE International Conference on Data Mining (ICDM_01), Vol. 28, pp. 647-648, 2001.

[5] L. Ma, J. Shepherd and Y. Zhang, " Enhancing text classification using synopses extraction," The fourth International Conference on Web Information System Engineering, pp. 115-124, 2003.

[6] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.(2014). "A Survey on Concept Drift Adaptation" ACM Computing Surveys 6(4), article 44.

Authors' Profile:

Rahul D. Sapnaris a student pursuing Master of Technology, in Computer Science & Engineering from Department of Computer Engineering at Vishwakarma Institute of Technology, Pune.

Virendra D. Pawar is Assistant Professor at Department of Computer Engineering inVishwakarma Institute of Technology, Pune.

Smideep Nair is a module lead for the CDTE Team with a leading technology company in Pune.