Application and Assessment of Classification Techniques on Programmers' Performance in Software Industry

Sangita Gupta[1*], Suma V.[2]

[1] Research Scholar , Jain University, Bangalore, India.
[2] Dayananda Sagar Institutions, Bangalore, India.

Corresponding author. Tel.: 917204194550; email: sgjain.res@gmail.com
Manuscript submitted January 10, 2015; accepted July 8, 2015.
doi: 10.17706/jsw.10.9.1096-1103

Abstract: Software companies seek to develop and maintain high quality products along with increasing reliability and maintainability. Therefore, companies look into every aspect for project success. Software companies try to improve efficiency in process by using different techniques. With best hardware, operating system and methodologies, still there have been disastrous software failures. It is found that most software failures are due to lack of focus on people working on the process. Human aspect of software engineering has thus become an emerging stream and it is identified as core factor for project success. Therefore, human aspect needs deep investigation in this ever changing field of software development. One of the strategies is to process and analyze previous data of software companies to predict future failures. Data mining techniques have the ability to uncover hidden patterns in large databases. Software companies can build models that predict with a high degree of accuracy the attributes required in human aspect for success. Through these predictive data mining models, companies can effectively address issues ranging from selection, retention to development and effectively manage the team members. The purpose of this paper is to use different data mining algorithms on project personnel data and compare the accuracy of these algorithms.

Key words: Data mining, classification, software personnel performance, comparison.

1. Introduction

Human aspect of software engineering has emerged as one of the core area for maintaining quality standards in software industry [1]. Software is not only a product but in recent years more of service which needs to be agile and innovative. This can be achieved by having the best project personnel. Therefore it is utmost important to look into people aspect of the company. Data related to project personnel have been extracted from resumes, interviews, training and this data is then processed and analyzed to enhance software quality. This dormant data about project personnel can be a base for providing knowledge for project enhancement. Predicting employee performance is an important issue in software industry. Performance is based on various diverse technical and non technical attributes. Some of the technical attributes taken are programming skills, domain skills, training, reasoning skills, aptitude tests score, college tier, level of education and experience. Data mining techniques have proved to be a promising tool to extract knowledge and reveal patterns from the databases related to project personnel. Data mining is also discovering knowledge through a combination of classification, clustering and association methods [2]. Prediction with high accuracy is beneficial for identifying correctly those attributes which are required for good programmers. Classification algorithm in data mining is one of the best methods for both predictive

and descriptive analysis [3]. In this study various classification techniques are applied on project personnel database and thereby assessment is made on accuracy and results thus obtained are put forth from various data mining techniques using WEKA environment.

The paper is further organized as follows. Section 2 gives insight of related work. Section 3 and 4 elaborates the research methodology and research work. Section 5 gives the performance analysis and section 6 gives results. Section 7 concludes the paper.

2. Related Work

There has been constant improvement and revolution in software engineering for effective management of software developed. Practitioners have found the impact of various major elements of software development on quality of software. Boehm considered the quality of staff the most important influence on products when constructing the Constructive Cost Model [1]. Right people have a high positive impact whereas wrong people have a very high negative impact. Authors in [1] have strongly pointed out that software professional performance is a strong factor affecting software cost, quality and success. There is a link between project success and the personnel performance factor and this missing link of software engineering that is the human component needs a deeper investigation. In this knowledge era, technology has proven to be a tool for improvement in many diverse fields like education, medicine, sales, manufacturing industry etc [4]. Data mining and parametric analysis is applied by many researchers. Authors in [5] have used data mining for software quality estimation by investigating previous data from similar project. They further did a comparative study of various accuracy parameters of data mining techniques. They have showed that data mining techniques can be effectively used to reduce development time by mining data for reuse process. Authors in [6] have made a comparative study of various classification techniques and provided a review about the methods. Data mining has improved and enhanced sales and productivity in diverse industries. Authors in [7] have used and compared data mining classification methods on data related to customer relation management and extracted meaningful information thereby to enhance sales. Authors in [8] too have used data mining techniques effectively for extracting information in code reuse and thereby effective code management for next project from data available of previous similar projects. Authors in [9] have used data mining classification techniques like ID3, CART, and C4.5 etc. and brought forth important attributes required in project personnel for project success. They have shown that assumptions of experience and academic performance which has been followed by many software companies did not result in project success. Data Mining revealed that other skill factors like programming and reasoning skills contributed to good performance. However, mining software project data poses several challenges, requiring various algorithms to effectively mine sequences, graphs and text from such data [10]. Using well established data mining techniques and comparing them will give knowledge about data and data mining methods [11]. Detailed study on data and data mining techniques will help practitioners to explore the potential of data and techniques. Deep investigation on data mining will assist for better management of projects and will result in high quality software systems that are delivered within cost and time limits.

3. Research Methodology

The research framework was thus constructed to explore the relationships between personnel profiles and performance at work using data mining techniques and further evaluate the classification model. The methodology uses data mining methods to extract knowledge from the processed database containing data related to project personnel and extract parametric information for better decisions and thereby enhance software process. This study applies the classification technique for predicting performance and further

reveals the most accurate classifier. It has mainly concentrated on accuracy of classifiers and a comparative study of various classification algorithms is has been done.

1) Defining Objectives: The problem along with limitations and constraints need to be specified firstly. It is also important to specify and understand the data mining methods for applying on the available data. This study is limited to software projects which are web based application and more of service oriented rather than product oriented. Also it is limited to classification techniques which can deal with discrete values.
2) Data collection and preparation: Collecting and preparing the right data is the basis of data mining . Data was collected from various databases, internal assessment and project managers' feedback. The data available was converted to discrete values and various techniques in WEKA support this type of data and therefore the research work was carried successfully in WEKA toolkit.
3) Model application and evaluation: There are numerous classification methods in WEKA environment. The data was subject to classifiers. The classifiers produced trees depicting the importance of attributes and results related to accuracy of the algorithm subjective to the data set. Thereby the accuracy parameters of models were reviewed and compared.
4) Interpretation and knowledge extraction: Later data mining results should be interpreted and assessed by domain experts in order to justify the meaning of extracted knowledge.
5) Using discovered knowledge: The discovered knowledge can be the basis for decision support to generate human management and selection strategies. It can also be used to improve related activities. Furthermore, since the empirical models derived from data mining have life cycle and thus need to be reviewed periodically to maintain its validity.

4. Research Work

Data collected from the industries are pre-processed and subject to classification methods using Waikato Environment for Knowledge Analysis (WEKA) data mining tool [9].WEKA with 10-fold cross validation for deriving results by various classification algorithms. 10- fold cross validation is the best option to judging the accuracy of classifier since it repeats the process ten times with the data for giving the results. The other options like split 66% are also available for fast computing. The attributes taken into consideration is listed in Table 1.

Table 1. Description of Attributes

Attribute	Description	Possible Values
GPA	Institution aggregate	First > 7.5 Second >6.5 & <7.5 Third >5.0 & <6.5
DS	Domain skills	poor ,Average, Good
PS	programming skills	Poor, average, good
GP	General Proficiency	Yes, No
CS	Communication skills	Poor, average, good
TE	Time efficient	Yes, No
RS	Reasoning skills	First > 7.5 Second >6.5 & <7.5 Third >5.0 & <6.5
CT	College Tier	Govt, aided, semi aided
EX	Experience	Low - less than 2 years, medium-2-5 years, high- above 5 years
DO	Degree Obtained	Graduate, Post Graduate
Target class	Predicted output of Performance	Good, average, poor

Data was collected from multiple companies in Bangalore. The various profile attributes predictors were processed without gender, social and economic discrimination.

GPA- General Percentile Assessment was obtained from the personal database of employee. It is mapped into

good (for >7.5), average (for <7.5 an >6.5) and poor for (<6.5) according to expert opinion.

DS-Domain Skills was taken after training and denotes the domain skills on that particular platform of the personnel.

PS- Programming skill. This attribute values were again taken from training and placement department. It was further categorized into discrete classes as good, average and poor based on the marks in the scale of 10 like GPA was done.

GP-General Proficiency denotes the overall rating of the employee in various domains.

CS-Communication Skills were rating was also available in many companies.

TE-time efficiency was obtained from project team lead for his team members.

RS-Reasoning skill. During selection company takes various placement assessments. This variable was one of the assessments, which is categorized similar to GPA.

TE-Time Efficiency. It was obtained from the project data through project leaders. It was obtained in the form of YES and NO.

Experience-This attribute was again taken from personal database and indicates total work experience. It is has three categories. High for above 5 years, medium for 2-5 years and low for less than 2 years.

Education(DO-Degree Obtained) This attribute is related to personal database. It indicated whether the project personnel are a graduate or post graduate.

College tier-This variable denotes type of college from which the candidate has passed out. In India there are basically three types of colleges. Government colleges in which generally the best students study and which gives high quality education. Following it are aided colleges. Apart from them there are private colleges. Therefore all three categories were considered to find its impact on performance.

Target Class-Performance. This is the target or output class. The value for performance variable is acquired from project team leaders in terms of good, average or poor, which is based on the quality of software developed. The company has an evaluation system every month and the consolidated results for the tenure of the project were considered.

The classification models applied on the data are listed in Table 2. Table 2 also briefs about the differences and similarity of the various methods [13].

Table 2. Description of Classification Techniques

Algorithm	**Proposed by**	**Dealing with data types**	**Speed of classification**	**Knowledge extraction of data from classification**
Id3	Quilan	discrete and continuous	Excellent	Excellent
J48 pruned tree	Quilan	discrete and continuous	Excellent	Good
RandomTree		discrete and continuous	Excellent	Excellent
CART Decision Tree	Breiman	discrete and continuous	Excellent	Good
Naivebayessimple	Baye	only discrete	Excellent	Excellent

5. Performance Measures

First of all the classification methods applied gave an insight into the attributes. It revealed to important attributes. After application, there is a need to assess the accuracy of various methods.

For comparing the various techniques accuracy, precision, recall, F-measure, ROC Area (Receiver operating Characteristic Area), RMSE (root mean Square Error) and MAE (mean Absolute Error) have been

taken into consideration.[12]. Accordingly, parameters such as True positive (TP) rate which is be predicted to positive and is actual positive or is the proportion of samples are classified as class x and truly have class x. False positive (FP) which is be predicted to positive but is actual not or those tuples are classified as class x but belong to a different class. True negative (TN) which is predicted to be negative and is actual negative and false negative (FP) is predicted to be negative but is actual not. Accuracy, precision and recall are indicated by equation 1, 2 and 3 respectively. In cross-validation, it needs to decide on a fixed number of folds or partitions of the data. Then the data is split into equal partitions. One portion is taken as training set and other as test set. 10-fold cross validation option is used in all the classification methods.

These classification algorithms produced outputs giving information related to data and classifier model in form of trees and rules. The trees and rules provide insight to dominating attributes. However, it is also important to compare the various algorithms for final knowledge deployment. The algorithms provided information on various parameters of the classifier like accuracy, recall, mean absolute error, root mean square error, f-measure etc [12].

Accuracy is one of the parameters to determine the accuracy of a classifier. This measure indicates that what percentage of the total test set records correctly classified. Equation 1. shows the calculation of accuracy.

$$\text{accuracy} = (TN + TP)/(TN + FN + TP + FP) \tag{1}$$

Precision in classification is called Positive Predictive Value. Precision is the proportion of the examples which truly have class x or the number of items correctly labeled as belonging to the positive class divided by the total number of elements labeled as belonging to the positive class [2]. A classifier with precision 1 is the most accurate one. The precision equation is indicated in equation 2.

$$\text{precision} = TP/(TP + FP) \tag{2}$$

Recall is called sensitivity since it gives the true positive rate. Recall is the number of true positives divided by the total number of elements that actually belong to the positive class or the sum of true positives and false negatives (FN), which are items which were not labeled as belonging to the positive class but should have been. The recall can be calculated as equation 3.

$$\text{Recall} = TP/(TP + FN) \tag{3}$$

Precision score of 1.0 for a class means that every item labeled as belonging to class does indeed belong to that class but it does not take in account the number of items from that particular class that were not labeled correctly. Recall of 1.0 means that every item from class was labeled as belonging to class but does not take into account how many other items were incorrectly also labeled as belonging to class.

Mean absolute error (MAE) is the average of the difference between predicted and actual value in all test cases or it is the average prediction error. The formula for calculating MAE is given in equation 4.

$$MAE = 1/n \sum_{i=1}^{n} |e_i| \tag{4}$$

where $e_i = (a_i - p_i)$ which is difference between actual value and predicted value.

The final parameter taken for comparison is Relative Mean Square Error (RMSE) that is used to show. RMSE as mentioned in equation 5.

$$RMSE = \sqrt{(1/n \sum_{i=1}^{n} e_i{}^2)} \qquad RMSE = \sqrt{(1/n \sum_{i=1}^{n} e_i{}^2)} \tag{5}$$

The f-measure of classification algorithm on the data is computed as stated in equation 6.

$$F - Measure = 2 \times Precision \times Recall/(Precision + Recall) \tag{6}$$

ROC Area- Receiver operating Characteristic Area is useful to visualize and compare between classifiers.

It shows the trade-off between true positive rate and false positive rate of classifier [2]. It is a measure of accuracy of the model. A model with ROC Area 1 will be fully accurate. The following section shows the results obtained of all the defined parameters.

6. Results

Various classifier algorithms were analyzed in WEKA toolkit and the results of ID3 has been shown in Fig. 1. The result infers about the important attributes. Skills have proved to be root attribute. Also college tier has impact in performance. People from government colleges and with just graduate degree performed well. However there is no significant impact of academic performance or communication skills on performance. Fig. 2 is the tree generated by Random tree algorithm. It is also having programming skills at root node followed by reasoning skills, college tier and experience. Based on the results obtained by data mining methods companies can derive rules for dynamic selection and retention of people working on the projects.

Results of accuracy related parameters like TP, Recall, precision, MAE and RMSE are shown in Table III. These values give a clear indication about accuracy of the classifier and easy comparison of various classifiers on dataset.

```
ps = good
|  GP = Yes: good
|  GP = No: av
|  GP = yes: good
|  GP = no: null
ps = av
|  rs = good
|  |  GP = Yes: good
|  |  GP = No: av
|  |  GP = yes
|  |  |  college tier = govt: av
|  |  |  college tier = aided
|  |  |  |  DS = good: av
|  |  |  |  DS = av: good
|  |  |  |  DS = poor: null
|  |  |  college tier = private: av
|  |  GP = no: null
|  rs = av: av
|  rs = poor
|  |  college tier = govt: good
|  |  college tier = aided
|  |  |  degree = ug: poor
|  |  |  degree = pg: av
|  |  college tier = private: poor
ps = poor: poor
```

Fig. 1. ID3 tree.

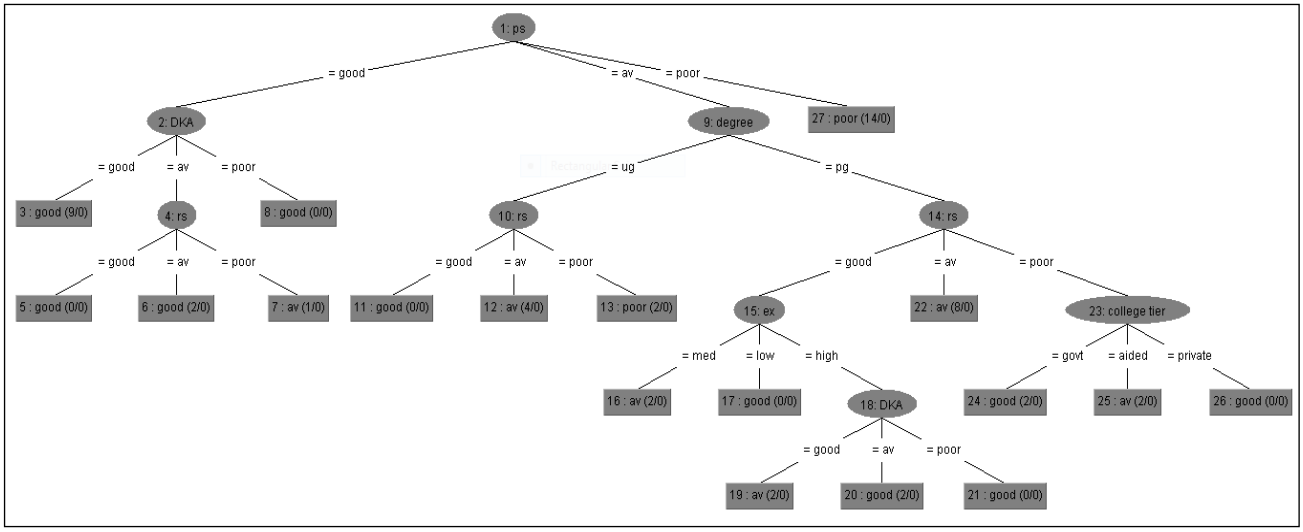


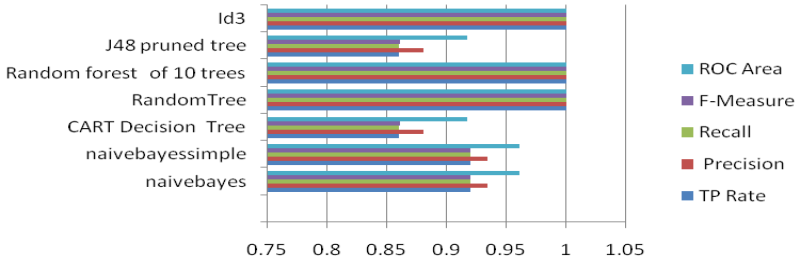
Fig. 2. Tree generated by random tree.

Table 4 gives the bar graph of various classifier and its parameters for better visualization of output results. Though most classifier gave consistent results, but ID3 and random forest gave more accurate results comparatively. Table III and IV infers about the accuracy of classifier and reveals that ID3 and random forest have proved to be most accurate classifier with less error and FP rate.

Table 3 Accuracy Measurements of Classifiers

Classifier parameters								
Classifier	TP Rate	Precision	Recall	F-Measure	ROC Area	MAE	RMSE	FP Rate
naivebayes	0.92	0.934	0.92	0.92	0.961	0.1046	0.2253	0.049
naivebayessimple	0.92	0.934	0.92	0.92	0.961	0.1046	0.2253	0.049
CART Decision Tree	0.86	0.88	0.86	0.861	0.917	0.1533	0.2769	0.082
RandomTree	1	1	1	1	1	0	0.0000	0
Random forest of 10 trees	1	1	1	1	1	0	0.1043	0
J48 pruned tree	0.86	0.88	0.86	0.861	0.917	0.1533	0.2769	0.082
Id3	1	1	1	1	1	0	0.0000	0

Table 4. Visualization of Accuracy Parameters



7. Conclusions

In this paper, we not only apply data mining methods for the prediction of performance of the project personnel but also reveal the most promising classifier. This study has discovered the performance related attributes and also shown the accuracy of the various methods applied on the data. Approach like this helps in the decision making capability of the developers of the company prior to starting and during the project. This study has given insight into accuracy of various techniques. ID3 and Random Forest proved to be more accurate than other classifiers. However the results were consistent in all classifiers. This paper has given statistical validity of classification methods as applied to performance analysis of project personnel.

More intensive study can be done for various types of projects and discover the patterns of human performance for diverse projects across various companies. Thereby the companies can ensure quality of the product developed by recruiting and maintaining capable people.

References

[1] Barry, B., June, V., & Bernard, W. (2007). Fifth workshop on software quality. *Proceedings of the 29th International Conference on Software Engineering* (pp. 131-132).

[2] Han, J. W., & Micheline, K. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann Publishers.

[3] Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning, 42(3)*, 203 –231.

[4] Liao, S. H. (2003). Knowledge management technologies and applications — literature review from

1995 to 2002. *Expert Systems with Applications, 25 (2)*, 155–164.

[5] Ngai, E. W. T., Li, X., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36 (2)*, 2592- 2602.

[6] Thair, N. P. (2009). Survey of classification techniques in data mining. *Proceedings of the International Multi Conference of Engineers and Computer Scientists*.

[7] Prasad, A. V. K., & Krishna, S. R. (2010). Data mining for secure software engineering — Source code management tool case study. *International Journal of Engineering Science and Technology, 2(7)*, 2667-2677.

[8] Ajay, P., & Ashoka, M. A. (2012). Application of data mining techniques for software reuse. *Procedia Technology, 4*, 384 – 389.

[9] Sangita, G., & Suma, V. (2014). Prediction of human performance capability during software development using classification. *Advances in Intelligent Systems and Computing,* Springer, *249*, 475-483.

[10] Usama, F., Gregory, P. S., & Padhraic, S. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39(11)*, 27–34.

[11] Fu, Y. J. (1997). Data mining: Tasks, techniques and applications. *IEEE Potentials, 16(4)*, 18–20.

[12] Quinlan, J. R. (1986). Induction of decision tree. *Machine Learning*, *1(1)*, 81–106.

[13] Ian H. W., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools And Techniques* (2nd ed.). Morgan Kaufmann, San Francisco.

Sangita Gupta is a research scholar of Jain University Bangalore. She has done her B.Sc, M.Sc and M.Phil in computer science. Her specialization is data mining, software engineering, pattern recognition, RDBMS. She has been awarded by best paper award twice for two papers during her research journey.

Suma V. is a dean, research and industry incubation center in Dayananda Sagar Institution, Bangalore, India. She specializes in software engineering, UML, cloud computing and has many ongoing projects related to it. Dr. Suma has more than 100 papers in her credit.