Performance Analysis using Bayesian Classification- Case Study of Software Industry

Sangita Gupta
Jain University Bangalore

Suma V
Dayanand Sagar Instituitions

ABSTRACT
There has been rapid improvement in the ability to construct software systems , firstly by developing reliable hardware and second by developing effective process oriented methodologies. However there has been a lack of emphasis in people component of software engineering. This paper examines the human component and brings out interesting patterns for enhancing human performance. IT companies are constantly reaching out to find ways for software success and quality and innovations for having a competitive edge over other companies. It has been found that human aspects seem to dominate the most expensive project failures. Therefore Human Aspect of Software Engineering has emerged as a new dimension in software engineering in recent years. IT industry have a large set of data having precious information related to project personnel. Data mining technique has the ability to discover knowledge from this unexplored data. In this paper, data mining techniques particularly Bayesian classification method is applied on the data related to team members of a software company to help an organization to find the right project personnel who will contribute largely to project success by performing well. This study will help the organization to reduce the failure ratio of software to a significant level and enhance the quality of the software by deploying the right person at the very start of all software process.

Keywords
Software project personnel, performance, data mining, Bayesian classification

1. INTRODUCTION
Human aspect of software engineering has become one of the main concerns in software companies to achieve quality objectives. Also software companies have to struggle effectively in terms of cost, quality, service or innovation. The success of these tasks depends on having right people with the right skills. Software development has been discussed as existence of mutually cohesive triangle of software, process and people [13]. For all process of software development , people are the driving force and therefore there is a need to focus on human aspect of software engineering. Right people have been identified as success factor for most industries [16]. Human talent is considered as the capability of any individual to make a significant difference to the current and future performance of the organization and thereby enhance success and fulfill quality objectives [19]. Software companies are now paying attention to select the right talent who can perform consistently throughout all generic framework activities and execute the process properly. A proper system to evaluate the human performance still remains unexplored and not much related work has been done to analyze the human talent. They are attempting to build a proper evaluating system for the same using various methods like expert system and other decision support systems. Various companies have different yardstick of deploying project personnel which they are following without proper validations. Therefore they failed to achieve the quality standards and gain profits [11]. Companies started scrutinizing the past data to find the reasons. In any organization, past experience plays a key role in improvement and management of the new project. Effective project management depends on how well this experience is captured and organized to enable learning and reuse[10]. There is enormous data stored in software industry databases which can be reused. These database have a lot of knowledge embedded within which can be extracted by various methods like statistical, numerical analysis and data mining techniques [14]. Further rules can be extracted and a expert system for performance prediction can be derived [12]. These methods can be applied for decision making for the betterment of the projects by providing a parametric analytical approach which were earlier running on assumptions [15]. To derive knowledge by means of data mining , there are various techniques such as: classification, clustering, association, decision tree, neural network etc . Classification is the most familiar and popular data mining technique. Predictive classification is a method of classifying an attribute or a set of attributes or features into one of a set of possible classes based on historical data[2]. Predictive classification is a supervised machine learning method . This study aims to analyze previous project data and predict the performance of a new candidate with similar attributes in a new project. The properties used to classify the performance are called attributes of the project personnel. The classes of performance have been derived by software project managers into good , average and poor performance which are discrete values. The classification technique applied to the training set is Bayesian classification. The implementation of the model produced information about high potential attributes of team members during software development. Based on the findings , software industry can refocus on human talent factors. In this connection, the objectives of the present investigation were framed so as to derive a mathematical method for performance analysis:

The further sections will deal with related work of data mining in various domains including software engineering and human resource in other industry, Bayes theory, research methodology with application of Bayesian classification model to the data and thereafter the conclusion.

2. RELATED WORK
The increasing demand of software has led to continual research in the areas of quality assurance and effective project management with the new dimension of the science called software engineering[18] . Along with quality of product , software companies are also looking for innovations in their products for having a competitive edge over other companies.

Therefore they are looking for methods for decision making in various aspects of software engineering and management . Data mining has been used for many aspects of software projects like defect prediction and thereby prevention, test analysis and improvement, code optimization etc[4]. Data mining results in decision through methods and not assumptions. In [6] authors have conducted an empirical study for selection criteria for software industry by introducing a knowledge based decision tree algorithm. It was found that though academic performance was given importance by most of software company , talents like programming and reasoning skills contributed largely to performance in software companies. Depending on few selected attributes related to employee, the model could predict their performance. Some of these attributes which were considered by authors of [3] were personal characteristics, educational and professional attributes. As a result for their study, they found that employee performance is highly affected by education degree and experience.

The authors in [5] worked for finding factors that affect the performance at work place. They mainly studied on work conditions along with position of the employee. They used various data mining methods in WEKA environment. WEKA tool supports most data mining methods. By applying various data mining techniques they found that factors such as position of the employee in the company, working conditions and environment affected the performance of the employee. In most of the studies done earlier academic performance did not yield clear relationship with the performance. In this study too Bayesian classification showed that other talent factors contributed largely to work performance rather than academic performance. The further sections will deal with theory and application of bayes classification.

3. BAYESIAN CLASSIFICATION

Predictive modeling is the process by which a model is created to try to best predict the probability of an outcome. This model is chosen on the basis of detection theory by calculating the probability of an outcome given a set amount of input data [9]. Classification is a predictive data mining technique which makes prediction using historical data. Predictive models can predict class membership of a variable given known values of other variables. Classification maps data into predefined groups or classes. It is referred to as supervised learning because the classes are determined before hand by examining the data by expert or many experts of that domain[17].

Bayes classification in pattern recognition and data mining methods has been developed based on Bayes rule of conditional probability. Bayes classification is statistical classifier. Bayes rule is a technique to estimate the likelihood of a property given the set of data as input also called evidence or measurement made on attributes[1]. The approach is called "naïve" or class conditional independence. Naïve Bayes classification is both a descriptive and a predictive type of algorithm[8]. The probabilities are descriptive since it represents the dependencies among subset of attributes. Further the result is used to predict the class membership for a target tuple with certain values of the attributes. Therefore it is predictive too. The naïve Bayes approach is simpler to use because it works in small or large training data set with accuracy.

The Bayesian decision making refers to choosing the most likely class, given the value of the features or attributes. The probability of class membership are calculated from the bayes' theorem. Bayes theorem is explained below:

If the tuple X is denoted by vector(x1-----xd) and class of Ci ,given the probability p(Ci) and P (X| Ci) which denotes the prior probability that the random sample is a member of class Ci and P(X/ Ci) is the conditional probability of obtaining attribute values X given the sample is from Ci. Our goal is to estimate the probability that a sample tuple belongs to class Ci , given that it has attribute values X . The probability is denoted by P(Ci |X) which can be calculated according to (3.1) as stated by bayes theorem.

If there are k classes and d attributes then probability of the attribute vector is denoted by equation 3.1.

$$P(X_1,.....X_D)=$$

$$\sum_{j=1}^{k} P(Cj)P(x1,.....xd|Cj)..... \quad (3.1)$$

which can be computed assuming the naive assumption that each attribute is independent within the class by equation 3.2

$$P(x_1,.....x_d|C_j) = P(x_1|C_j)*P(x_2|C_j).....P(x_d|C_j).....(3.2)$$

Then by bayes theorem the conditional probability that a tuple with attributes values x1,x2....xd belongs to class Ci is denoted by equation 3.3

$$P(C_i|x_1,....,x_d) = \frac{P(Ci)P(x1|Ci).....P(Xd|Ci)}{\sum_{j=1}^{k} P(Cj)P(x1|Cj)....P(xd|Cj)}.----------(3.3)$$

The Derivation of bayes' classification is thus written as mentioned below:

D : Set of tuples

Each Tuple is an 'd' dimensional attribute vector

X : (x1,x2,x3,…. xd)

Let there be 'k' Classes : C1,C2,C3…Ck

Naïve Bayes classifier derives and predicts X belongs to Class Ci if

P (Ci/X) > P(Cj/X) for 1<= j <= k , j <> i

Maximum Posteriori Hypothesis is given by equation 3.4

P(Ci /X) = P(X/Ci) P(Ci) / P(X) .....(3.4)

Bayes classification aims to Maximize P(X/Ci)* P(Ci) as P(X) is constant.

With many attributes, it is computationally expensive to evaluate P(X/Ci). therefore with Naïve Assumption of "class conditional independence" the computations are made. The final derived equation with the naive assumption of class independence is given by equation 3.5 and 3.6.

$$P(X/Ci) = \prod_{k=1}^{d} P(Xk/Ci) \quad .....(3.5)$$

P(X/Ci) =

P(x1 /Ci) * P(x2 /Ci) *.. P(xd/Ci).....(3.6)

In this study X is the attribute vector of the project personnel having d attributes with values x1, x2,--xd. There are three classes: C1- good, C2- average and C3-poor for performance.

This study aims at finding through Bayesian classification, attribute value X which gives highest probability in the respective classes as per equation 3.6.

Further sections will deal with details about data, attributes and application of Bayesian model on the data.

4. RESEARCH METHODOLOGY

The objective of this study is to find a suitable technique for making the best decision for deploying the right candidate for a particular software project . the project is short duration and web based application which demands innovations by project personnel for competing with similar applications by other software companies.

Problem statement

The hypothesis can be stated as

Hypothesis- There is a clear and directly proportional relationship between project personnel profile and project success thereby contributing to company profits.

Constraints of the hypothesis- This study can be applied to projects which deal with C/C++ or similar programming language.

The research method or technique used is Bayesian classification to find the high potential attributes for good performance. The steps involved in developing the study is shown in Fig 1.

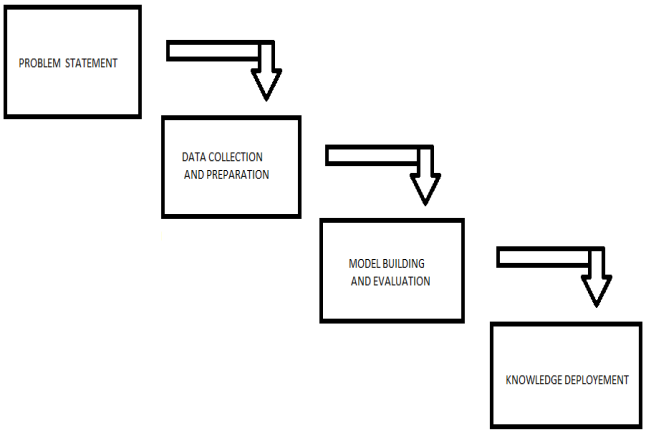

Fig 1-Research Framework

After defining the problem objectives, the next task was gathering and pre processing the data.

Data preparation

The next task to accomplish the problem statement was to collect data . Data was obtained from a software company for a project of tenure 2 years with a team of 40 members.

Data came from various sources like personal data of employee, questionnaire given to project leaders as well as project personnel,human resource department and training and placement cell. Many attributes related to personal, educational and family background were taken during data collection. But important attributes without gender, social and economic discrimination were taken for building the model in this study. Human related attributes are very complicated and thus those attributes were taken on which quantitative approach could be employed in practice. The list of attributes with description is given in Table 1.

TABLE 1. Software Project Personnel Attributes

Attribute	Description	Discrete Attribute Values
GPA	General percentile aggregate	good, average ,poor
DKA	Domain knowledge assessment	Poor ,Average, Good
PS	programming skills	Poor, average, good
GP	General Proficiency	Poor ,Average, Good
CS	Communication skills	Poor, average, good
TE	Time efficient	yes, No
RS	Reasoning skills	Poor ,Average, Good
P	Performance	Poor ,Average, Good

For this investigation, the possible values for some of the variables were defined as per company yardsticks by their experts into good , average and poor.

Sample data tuples are shown in Table 2

Table 2: Subset of Training Data Set

S.No.	GPA	DKA	PS	TE	CS	RS	P
1	Good	Good	Good	Yes	Good	Good	Good
2	Good	Good	Average	No	Good	Good	Good
3	Good	Good	Average	No	Average	Average	Average
4	Good	Average	Good	Yes	Good	Good	Average
5	Good	Average	Average	Yes	Good	Good	Average
6	Good	Poor	Poor	No	Average	Poor	Poor

The values were clustered and preprocessed accordingly to get good and equally partition set of data so that computations can be done easily by applying our technique. However with continuous values of attributes bayes' classification can be applied by using gaussian distribution. Computations are lengthy but simple[9]. For continuous valued attribute ,a Gaussian distribution with a mean μ and standard deviation σ is taken. Further it is applied to bayes' theorem as in equation 4.1 and 4.2

$$g(x,\mu,\sigma)= \frac{1}{\sqrt{2\pi\sigma}} \quad e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{---------(4.1)}$$

$$P(x_k|\, C_i)=g(x_k, \mu_{Ci}, \sigma_{Ci}) \qquad \text{----------(4.2)}$$

where μci is the mean , σci, is the standard deviation, of the values of attribute xk for training tuples of class Ci.

The data has been converted into a matrix format. For output class i.e C1(good) there were 13 tuples, 13 for C2(average) and 14 for C3(poor). Therefore the probability P(Ci) is approximately same for the training set which makes computations more simpler. Table 3 shows the output from processed data after applying Bayesian classification. The high probability values are highlighted.

Table 3- Output of Bayesian Model

Attribute	Attribute Values	Probability P(X|Ci)		
		C1 Good	C2 Average	C3 Poor
GPA	Good	0.3076923	0.1538462	0.1428571
	Average	0.4615385	0.3846154	0.5000000
	Poor	0.2307692	0.3846154	0.5000000
PS	Good	0.9230769	0.0769231	0.0000000
	Average	0.0769231	0.9230769	0.0714286
	Poor	0.0000000	0.076923077	0.8571429
DKA	Good	0.3846154	0.3076923	0.0714286
	Average	0.5384615	0.6153846	0.0714286
	Poor	0.0769231	0.0769231	0.6428500
CS	Good	0.3846154	0.2307692	0.3571429
	Average	0.5384615	0.6153846	0.1428571
	Poor	0.0769231	0.1538462	0.4285714
RS	Good	0.6923077	0.2307692	0.0714286
	Average	0.3076923	0.4615385	0.2142857
	Poor	0.0000000	0.3076923	0.7142857
TE	Yes	0.6923077	**0.7692308**	0.5000000
	No	0.3076923	0.2307692	0.5000000

Implementing bayes' classification and results

Given a training set the naïve Bayes algorithm first estimates the prior probability P (Cj) for each class by counting how often each class occurs in the training data. Out of 40 records, there were 13 for class good, 13 for average and 13 for poor. Therefore

$P(C_1$=good)= 13/40=.325

$P(C_2$=average)=13/40= .325

$P(C_3$=Poor)= 14/40 =.35.

In this training set P(Cj) are approximately same for each class. For each attribute value, xi can be counted to determine P (xi).Similarly the probability P (xi | Cj) can be estimated by counting how often each value occurs in the class in the training data divided by number of instances for Cj for that attribute. Therefore for a attribute GPA , P(xgood| C good) = 4/13 .Similarly for all attributes the data has been segregated accordingly for application of bayes classification as shown in table 3. Table 4 shows high probabilities attributes obtained from the model.

The results obtained from bayes classification method are shown in table 4.

The result clearly depicts the classification of attributes with a particular value with performance classes. Table 4 infers the classification results by depicting the high potential attributes. For example if a candidate is having PS good then the probability of his performing good is high. Similarly if his PS is average then probability of his performance to be average is very high. If the attribute value for PS is poor then there is high probability that his performance is poor. The next attribute to show high probability is RS. If RS is good then there is high probability that performance is good and if RS is poor then performance is having high probability to be poor. After RS , DKA showed that if DKA is average then it can be predicted that performance will be good or average. GPA and CS did not show any impact on performance. However TE showed mixed performance. When TE was YES, performance could be good or average. However a personnel who was not time efficient did not perform well. The highest potential attributes for project personnel were PS and RS for a software industry developing their products in C/C++. Table 4 shows the attribute vector X which has yielded maximum probability. It shows that if PS and RS are good and Time Efficient Yes, and with other attributes having average values too it gives very high probability for good performance. Therefore PS and RS have proved to be dominating attributes for a project personnel working in a software company which produces products in C/C++ platform.

Table 4- High Probability Attributes

High Potential Attributes	
Attribute	Value
PS	Good
DKA	Average
CS	Average
RS	Good
TE	Yes

With the above results, this study has accomplished the application of Bayesian classification with the objective of finding the talent matrix related to good performance and thereby enhance project quality. Thus, the useful information or patterns which the classifier has extracted can be further summarized into rules for decision support system. The discovered knowledge from bayesian classification can be the basis for human management in software companies. It can be used to improve human resource management activities of a software industry.

Table 5- Visualization of classification results

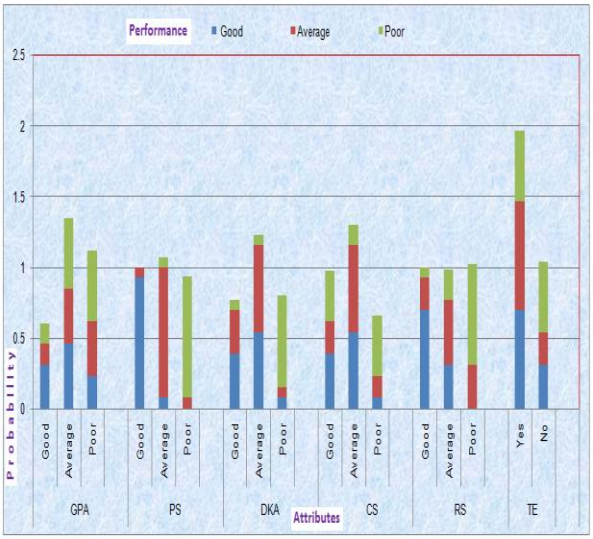

Table 5 depicts the visualization of bayesian classification on human talent. Clearly it shows that PS is a attribute which shows high correlation with performance. For a attribute value within a class say good , the table should show maximum probability for that class . For example if the attribute value is good for PS, the outcome should be such to have a bar with more of blue color. A attribute which shows mixed color bar depicts that it does not have any significant impact on the performance. GPA has shown mixed results therefore it is an attribute of no significance.

5. CONCLUSIONS

The aim of the paper was to investigate the predictive capabilities of bayesian classification system for performance analysis of software team members. For quality software products , software companies are looking for accurate and parametric methods for most activities during development. Data mining have given interesting patterns for most process like code optimization, quality assurance etc. This study has used, Bayesian classification method to predict the performance of the project member on the basis of few personal attributes. Historical data was taken from the previous project and applied to data mining technique for identifying those attributes in a project personnel that will contribute towards good software quality and increase the company profitability factors. The classification model clearly showed the attributes which had high probability for good performance. This study will help the software company to improve the development of software by choosing the right talent at the very start of the project. Furthermore, this study can also be applied to other jobs and other industry with other attributes be it discrete or continuous to find the right talent to enhance human capital. This study has shown that software success greatly depends on certain skills of the project member rather than marks obtained in institutions. Though academic performance may be low or average but innovative mind , good programming concepts and aptitude have got significant impact on performance. The proposed data mining approach for performance analysis of project personnel will help the software company to improve the software process and thereby decrease the failure ratio of software.

Software companies which were following the old criteria of good academic scores and experience have achieved better results by following the human centric rules derived in this paper. They could attain much higher quality along with innovations in their products by deploying the right skill according to talent matrix at the very start of the project.

Further study can be done by collecting other possible input variables such as experience, passion for working in that field , type of school and college , position in the company, working condition etc. Taking such factors and seeing the impact of these on a performance can be future scope of this study.

6. REFERENCES

[1] C. Eklan," Boosting and naive Bayes learning", Technical Report Computer Science , University of San Diego, pp97-557,Sept 1997.

[2] Earl Gose ,Richard Johnsonbaugh and Steve Jost. "Pattern Recognition and Image analysis". Prentice Hall PTR, 1996

[3] C. F. Chien and L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," Expert Systems with Applications, vol. 34, pp. 280-290, 2008

[4] Suma.V, Pushpavathi T.P, and Ramaswamy. V, "An Approach to Predict Software Project Success by Data Mining Clustering", International Conference on Data Mining and Computer Engineering (ICDMCE'2012), pp. 185-190.

[5] Hamidah Jantan et al, "Human Talent Prediction in HRM using C4.5 Classification Algorithm", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, pp. 2526-2534

[6] Sangita Gupta, Suma V, "Empirical study on selection of team members for software project- A data mining approach", International Journal of Computer Science and Informatics, ISSN (PRINT): 2231 –5292, Volume 3, Issue 2, 2013, pp 97-102.

[7] Reiter, Ashley. 1995. Writing a research paper in mathematics web.mit.edu/jrickert/www/mathadvice.html

[8] V. Barnett, T. Lewis, " Outliers in Statistical Data", John Wiley and Sons. 1994.

[9] Hogg, R.V., and Tanis, E.A.," Probability and Statistical Inference", Macmillian Publishing Co., Inc., 2nd Edition.1983.

[10] Campbell, J.P. " Modeling the performance prediction problem",Iindustrial and organizational psychology.pp 687-731,1990

[11] Cooper, D., & Robertson I.T.,"Recruitment and selection: A framework for success", London: Thompson.(2002).

[12] R. S.Hooper , T. P Galvin,R. A Kilmer, J. Liebowitz, (1998). "Use of an expert system in a personnel selection process", Expert Systems with Applications, Vol 14,Issue 4, pg 425–432.

[13] Suma V, T. R. Gopalakrishnan Nair, "Four-Step Approach Model of Inspection (FAMI) for Effective Defect Management in Software Development",arXiv preprint arXiv:1209.6466 (2012).

[14] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge Information System,2008.

[15] Beckers, A. M., & Bsat, M. Z.. " A DSS classification model for research in Human Resource Information Systems.", Information Systems Management , Vol 19, issue 3, pg 41–50,2002.

[16] A. S. Chang, & Leu, S.S., "Data mining model for identifying project profitability variables," International Journal of Project Management, vol. 24, pp. 199-206, 2006.

[17] D. Heckerman, Geiger and Chickering," Learning Bayesian Networks: The combination of Knowledge and Statistical Data", Machine Learning, Vol20,pp197-243,1995.

[18] M. Jackson," Systems Development", Prentice Hall.2000.

[19] L. M. Hough and F. L. Oswald, "Personnel selection: Looking toward the future—Remembering the past," Annual rev. Psychology, vol. 51, pp.631–664, 2000.

.