Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/106504, December 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Predicting Human Productivity for an Organization

J. Karthick* and Mohemmed Yousuf

Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore –560067, Karnataka, India; karthik.mj@mvjce.edu.in

Abstract

Objective: The main objective of the paper is to evaluate the performance level of an employee using Predictive Analytics. Human capital is a major concern for an organization as they want to hire most qualified one who will perform well. This human resource can be used to find the future of the organization. **Method:** The advanced branch of data engineering is Predictive Analytics is used for this purpose. Generally, these analytics predicts some occurrence or probability based on data. The future occurrence or events, were predicted by analyzing the historical data. **Finding:** With the help of the rules generated by the decision tree classifier, performance of an employee were found by testing with the attributes. This paper concentrates on gathering information about the employees from the organizational database, based on the analysis of historical data generates an decision tree, validating the attributes of the employee with decision tree. With the latest prediction algorithm, we will predict employees' performance more efficiently than the existing system.

Keywords: Data Mining, Decision Tree, Employee Performance, Prediction, Regression

1. Introduction

Almost all the Business Organizations were putting a lot of effort for making proper plan for selecting skilled employees. After Recruiting the employees, organizations becomes more concerned and creating evaluation strategies to measure the performances of the employees and attempt to retain the best performers for a longtime. To extract a useful information from a huge data sets analytical tools such as data mining techniques were used. Data mining is nothing but an emerging technology with immense potential information system. Data mining is defined as the process of digging or mining useful information's which may contain anomalies patterns, trends, associations, and significant structures from a huge unknown data sets. Data mining indulges different set of techniques for mining useful and contextual information from a raw data. It has many tasks such as association rule mining, clustering, classification and prediction. The advanced branch of data engineering is Predictive analytics. Analytics predicts some occurrence or probability based on data. Some of the Statistical techniques from data mining, predictive modeling and machine learning were adopted by predictive analytics. The prediction about future can be done by analyzing the present and historical data's. An supervised learning technique based on the classification technique which classifies the data item into a predefined class label. In data mining, this technique is used to build an classification model is built based upon the input data set. The classification techniques used to build an common model to predict the future based on data trends.

2. Data Mining and its Techniques

Data mining tasks were divided into two namely predictive and descriptive. Predictive branching involves analyzing classification, time series and regression.

2.1 Classification

Separate groups can be classified by partitioning the data based on the classified rules. A predefined class labels were taken as a training data set to be given as a input for classification. A model has been constructed based on the class label by exploring the training data set and insists new models to be created for the future unlabeled records. The segregation based on well-known class field is called as supervised

*Author for correspondence

learning. It uses Decision trees, genetic algorithms, statistical model as some of its classification models.

2.2 Regression

The relationships among variables is estimated by an statistical process known as regression. While focusing on the relationship between dependent and independent variables(or 'predictors') many techniques were used for modeling and analyzing several variables.

3. Related Study

The classification model based on the attributes which affects the job performance were developed by Qasem A. Al-Radaideh. Three data sets have been taken by Qasem A. Al-Radaideh and Eman Al Nagi [2012] and three decision trees have been built for each data sets for these experiments. On each tested class the gain ratio is calculated to represent the weight of effectiveness of each attribute and the ordering of the tree node is specified accordingly.ID3 and C4.5 are the two versions of decision tree and Naïve Bayesian classifier are the techniques used for calculating gain ratio. The accuracy for each experiment is evaluated with 10-fold cross validation. Tremendous effect on employee performance for several factors has been found out for all the data sets by repeating the same procedure and conclusion is given. Job title is one of the most effective factor. About 20 job titles were studied, the effectiveness of the results of factor job title is not that much clear but it can be related to the complexity and responsibility of the job related to the title. The employees motivation and performance will be affected in a positive way, sometimes when burdened with high responsibilities. V. Kalaivani, Mr. Elamparithi M [2014] have used different classification algorithms to predict the employee performance. The data set was obtained by questionnaires. The questionnaire was filled by 217employees. Based on the data collected from an institution experiments were conducted with Bagging , C4.5 and Rotation Forest algorithms. The results were given below.

Table 1. % for Cross-Validation 10 folds

Algorithm used	Training set % accuracy
C4.5	84.79%
Bagging	75.57%
Rotation forest	100.00%

Table 2. % for Training Set

Algorithm used	10–folds Cross validation % accuracy
C4.5	41.47%
Bagging	45.62%
Rotation forest	51.46%

The Rotation forest algorithm had produced better performance having an maximum accuracy value of 51.46% for cross validation and 100% for training data set test option than the other three algorithms.

Therefore, it was concluded that for predicting the employee's performance Rotation forest algorithm is more efficient than the other two algorithms[2, 3]

Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas[2014] conducted similar experiments for students to analyse students' performance using classification algorithms[4]. The students' academic performance is influenced by various factors like parents' education, locality, economic status, attendance, gender and result from the different school students.260 samples were taken for the implementation.

Classification algorithm was implemented by the classify panel to estimate the dataset for accuracy of resulting predictive model, and to visualize the model. The decision tree classifier C4.5 (J48), Random Forest, Neural Network (Multilayer Perceptron) and Lazy based classifier (IB1) Rule based classifier (Decision Table) were enforced in WEKA[2, 3]. Under the "Test options", the 10 fold cross validation is chosen. The section presents the results generated from the study. The attributes were ranked in order of its importance using information gain and gain ratio measures. The ranking of each Attribute evaluators was done using ranker search method. From the above set of rules an inescapable conclusion emerges the attendance is considerably related with student performance. From the rule set it was found that parent education, locality, gender, economic status and different factors are of high potential variable that have an effect on students' performance for getting good performance in examination result[4].

4. Implementation

4.1 One R

"One Rule", abbreviated as OneR, an simple, yet but accurate classification algorithm used for generating one rule for each predictor in the data, and selects a rule as "one rule" with smallest total error. A frequency table for each predictor against the target has been constructed for creating a rule for each predictor. OneR produces very

simple and slightly less accurate rules for human interpretation than that of state-of-art classification algorithms.

Steps:

- Identify predictors,
- Make an rule for each value of that predictor
- For each appearance of the target class counting to be done.
- Most frequent class to be found.
- For a predictor having a value to a class a rule is assigned.
- For each predictor total error rules were calculated.
- A predictor with smallest error is chosen.

4.2 Elastic Net

Elastic net is a regularized regression method which linearly combines L1, L2 penalties of lasso and ridge methods, by statistics, and, in particular, in the fitting of linear or logistic regression models. Based on the penalty function, Limitations of the LASSO method were overcame by Elastic net method as

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

Several limitations while using this penalty function. In this case, "large p, small n" case(few examples with high dimensional data), the LASSO saturates by selecting utmost n variables. LASSO tends to select only one variable from the group by ignoring the others, when there is a group of highly correlated variables. A quadratic part to the penalty ($||\beta||^2$), which when used alone is ridge regression is added by the elastic net , to overcome these limitations. The elastic net method estimates were defined by

$$\beta\hat{} = \arg(\min \beta||y - X\beta||^2 + \lambda_2 ||\beta||^2 + \lambda_1 ||\beta||_1)$$

A unique minimum loss function, which is strictly convex has been made by the quadratic penalty.

The LASSO and ridge regression is included by the elastic net method, each of them is a special case where $\lambda 1 = \lambda$, $\lambda 2 = 0$ or $\lambda = \lambda$. A two stage procedure was implemented meanwhile, by the naïve version of elastic method: initially it finds the regression coefficients for each fixed and then a LASSO type shrinkage is done. Naïve elastic method estimation leads to poor predictions and increased bias due to twice the amount of shrinkage. The coefficients of the naïve version of elastic method was rescaled by the author by multiplying the estimated coefficients by $(1+ \lambda_2)$ to improve the prediction performance.

4.3 The Process

4.3.1 Data Collection and Refinement

The type of organization under study is a manufacturing organisation and the personnel under study are blue collared employees who are governed by the respective employee union. The entire implementation is divided into phases. In order to apply any of the algorithms we need the data set to do so. Thus the first step in implementation is the data collection .In this project the data was given by the respective HR of the organisation in a raw format, for some of the data it was required to personally inspect and collect data. The data set consists of 11 attributes for 3 years 2012, 2013, 2014, this means 3 individual data sets, this was so done to analyse the performance over the 3 years and finally we combine the data sets of all 3 year into a single dataset called master, this was done to compare the efficiency between the algorithms. The data finally refined was converted csv format which includes 3 individual year data set and one master dataset that is a combination of all 3.

4.3.2 Inputting to WEKA

The master data set is inputted to WEKA, if necessary it is filtered to support the type of algorithms, it is later

Table 3. Format of the Data Set

Attribute	Possible Values.
Gender	Male, Female
Marital Status	Married, Not Married
Age Range	a, b, c, d, e *
Number Of Kids	0, 1, 2, 3
Specilisation	SSLC,ITI-Elect, ITI-Turner, etc. totally 21 distinct values
Degree	SSLC, ITI, B.Com etc., totally 9 distinct values.
Experience Range	a, b, c, d, e, f *
Rank	1, 2, 3, 4, 5, 6, 7
Service Period Range	a,b,c,d,e *
Salary Range	a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p *
Performance	below average, average, good, very good, exceptional.

*Age Range – Over 40, 35-40, 25-29, 30-34, under 25.

*Experience Range – Over 20, 15-20, from 3 to 5, from 1 to2, from 10 to 14, from 2 to 3.

*Service Period Range – Over 20, 15-20, from 3 to 5, from 1 to 2, from 10 to 14.

*Salary Range – values withheld as per norms.

classified, for this we choose the algorithms of our choice , classification is done both on training the data set as well as 10 cross fold validation method. We implemented each of the algorithms available in WEKA to the master data set but for comparison purpose we have taken only a few that have given maximum efficiency on both training and 10 cross fold validation with default 66% split.

Of all the algorithms it was found that One R algorithm gives the maximum efficiency with figures of 70.0361% for training and 69.6751% for 10 cross fold validation for the performance attribute. The efficiency figures for the rest is given in the conclusion section. Thus compare our proposed method Elastic Net to One R algorithm.

4.3.3 Modifying the Data Set to Cope with Elastic Net

Elastic net being a regression techniques involves statistical modules and hence can be applied on only numeric data , thus the entire data set master was converted to numbers for example the gender attribute has 2 values namely female and male , female was taken 0 and male as 1.In this way all the attributes were converted.

5. Result and Discussion

The modified master file was inputted to WEKA and we choose Elastic Net listed under function of WEKA classifiers, we run it on the data set for both training and 10 cross fold validation, the results of Elastic Net algorithm are not in terms of efficiency % rather as correlation coefficient. We can say that higher the correlation better the efficiency. The figures we obtained for training and 10 cross fold validation are 0.3297 and 0.2502 respectively for the performance attribute. Elastic Net also provides an impact value between -1 and +1 for each of the other attributes, where -1 means inverse relation and +1 means direct relation and a value of 0 means no relation. This impact factor is compared to that generated by GainRatioAttribute Eval method by having One R as base algorithm, and the results are in form of table.

Since the efficiency of both One R and Elastic Net algorithms are not of the same unit we compare them by a common factor of Root Relative Squared Error. Here the algorithm that has the least error is the better. Here Elastic Net proved to be better with a lesser error rate for both training and ten cross fold validation. The results are in graph form under the comparison study section.

6. Conclusion and Future Enhancement

The employee data set can be studied upon and can be analyzed by various classifiers in WEKA. We have studied various papers and chosen J48 and Rotation Forest algorithms to be the best among those. The very same algorithms have given us better results on data set.

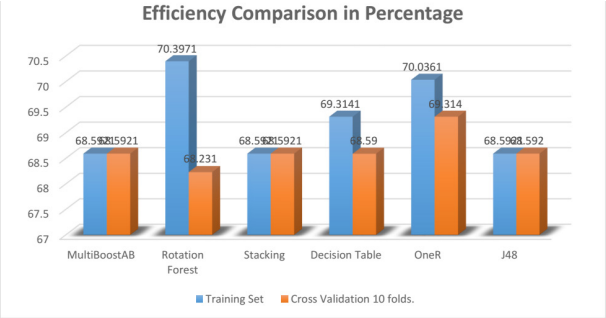

Figure 1. Efficiency comparison in %

Table 4. Comparison between Elastic Net and One R by Impact Value

Elastic Net			One R		
Impact	**Attribute**	**Rank**	**Impact**	**Attribute**	**Rank**
0.041	Salary Range	1	0.07017	Salary Range	1
0.018	Degree	2	0.06456	Total Experience range	2
0.013	Total Experience range	3	0.06416	Degree	3
0.009	Service Period Range	4	0.05806	Service Period Range	4
0.007	Rank	5	0.05722	Specialisation	5
0.006	Number of Kids	6	0.04765	Age Range	6
-0.005	Specialisation	7	0.03522	Rank	7
0.005	Marital Status	8	0.02258	Marital Status	8
-0.004	Age Range	9	0.00818	Gender	9
-0.002	Gender	10	0	Number of Kids	10

Considering J48 and Rotation Forest mentioned in our base papers we do a time study with One

R and the results as depicted above reveal that One R takes the least time to build 0 seconds. Thus WEKA has One R as the best solution to the master data set. Hence we decided to compare our implementation of the Elastic Net with the One R algorithm.

6.1 Comparison Study

The prime focus is on the two algorithms namely One R and Elastic net. While the efficiency of One R can be evaluated by correctly classified instances in percentage that of Elastic Net is in terms of correlation coefficient. This is because Elastic Net being a regression technique is a statistical model, and hence can be applied to data of numeric type.

Elastic net takes all the attributes other than the one evaluated into consideration. It overcomes the limitation of LASSO where only highly related attributes are considered for impact calculation. Elastic Net uses correlation factor to express its efficiency higher the correlation better the efficiency. Since the efficiency of the two algorithms cannot measured by the same unit. We consider the Root Relative Squared error. The comparison is as follows:

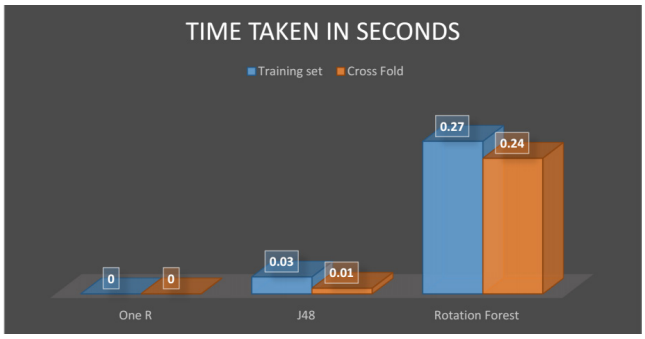

Figure 2. Time Taken in Seconds

Table 5. Comparison between Elastic Net and OneR by performance

	One R		Elastic Net	
Data set	Training Set (%)	Cross Validation 10 Folds (%)	Training Set (Correlation Coefficient)	Cross Validation 10 Folds (Correlation Coefficient)
Master	70.0361	69.6751	0.3297	0.2502
2012	75.2941	70.5882	0.3797	0.1772
2013	58.5235	49.4118	0	–0.2967
2014	83.1776	78.5047	0.3274	0.1369

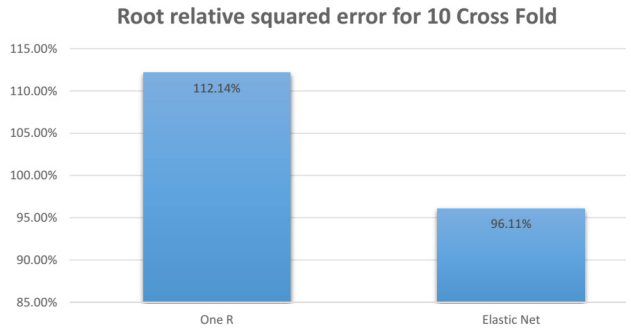

Figure 3. Root relative squared error

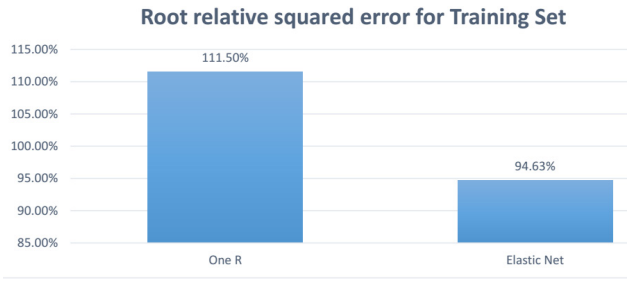

Figure 4. Root relative squarefor 10 cross fold

Thus results reveal that Elastic net has lesser root relative squared error when compared to One R on both Training sets and 10 Cross Validation .Although One R has proved better for attributes that contain a range of values, its high root relative squared error exposes its limitations. If your goal is analyze data with nominal or string attributes, and to support a range of values for attributes One R is your choice. Thus to analyze data sets with minimal attributes Elastic Net is on the upper hand, since it gives an impact value for each of the attributes. Since manual intervention in essential for data preprocessing in Elastic Net and since the efficiency is not directly in terms of percentage we find it hard to relate layman terms. Thus we recommend One R for Traditional Blue collared based organizations for data set similar to ours naturally due to its efficiency and its ease of use. Thus, we conclude that if the data pattern remains the same for the given data set, then the performance is predicted to be "GOOD". If there are variations in the data set, then the performance can vary accordingly.

6.2 Future Enhancement

The idea of this project was imbibed from the latest developments in regression techniques. With more future advancements and better algorithms, their usage to the

data set may impact the result. More number of data instances and a better refined data set may impact result.

7. References

1. Emin Kahya .The effects of job characteristics and working conditions on job performance.
2. Kalaivani V,.Elamparithi M. An Efficient Classification Algorithms for Employee Performance Prediction.
3. Mohammed Hussain K, Sheik Abdul Kadher P. On A Review of Factors and Data Mining Techniques for Employee Attrition and Retention in Industries".
4. Mythili M.S, Mohamed Shanavas A.R. An Analysis of students' performance using classification algorithms.
5. Osman M. Karatepe. The effects of family support and work engagement on organizationally valued job outcomes.
6. Qasem A, Al-Radaideh, Eman Al Nagi .Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance.
7. Thushel Jayaweera1. On Impact of Work Environmental Factors on Job Performance, Mediating Role of Work Motivation: A Study of Hotel Sector in England.