Human Resource Utilization Analysis and Recommendations Using Software Analytics

Subhash Ajmani, Satya Sai Prakash K., and D. S. U. M. Prasad

Abstract—Software organizations induct employees belonging to diverse backgrounds with varied work experience. Resources work on different projects with varied complexity. We propose a formalism to identify the significant parameters that contribute to evaluate the resource utilization and contribution in the projects. Proposed formalism was tested with a sample project data consisting of 23 employees, 11 measured parameters across 10 product releases. Parameters such as individual effort, product level experience, defect percentile, complexity level, and manager rating have been captured for each release. Once data is pre-processed, principal component analysis (PCA) and K-means clustering were used to assess the relative association of employees based on experience to complexity. The PCA analysis helped in identifying two parameters i.e., ‘product level experience’ and ‘defect percentile’ that captures the major variation in the data. In addition, k-means provided 3, 5, and 9 clusters for further investigation. This analysis helped in identifying the impact of various parameters such as product level experience, project complexity etc. on resource utilization. Based on the present analysis recommendations could be made for future resource allocation and areas of improvement.

Index Terms—Human resource utilization analysis, K-means clustering, principal component analysis, software analytics.

I. INTRODUCTION

Employees can be the most valuable and possibly, the most costly asset to any corporate organization. Research shows that industries using data-driven decision-making and predictive analytics are more viable and have higher returns than industries that don’t [1]. Because of this, the most ambitious companies are engaged to obtain more data on their employees, but then what type of data has to be collected is a big question? Moreover gathering this information truly can be delicate. Nevertheless, by collecting and using data in a way that appreciate, motivate, recognize and reward employees could possibly make a great influence on organization growth.

Software projects are people-intensive and require employees with different skills to work in different geographical locations to solve diverse problems [2]. Hence, various types of data obviously exist in the software industry such as employee data, location data, project data, compensation data etc. Analytics techniques were used on these data sources to help understand influence of these data type on the progress of software projects.

Software analytics is to employ data-driven methods to allow software experts to do data examination and study in order to find insightful and actionable information for finishing numerous jobs around software systems, software users, and software development process [3].

Human resources analysis offers an organization with understandings for efficiently dealing with employees so that organizational goals can be reached swiftly and proficiently. The more you know about your employees, the more you can appreciate their professional performance. Additionally, organizations can expect to gather real understanding into their employees and intensely increase their chances of growing productivity in the place of work. The task of human resources analysis is to find what data should be captured and how to use the data to model and predict capabilities so the organization gets maximally benefited by its human capital [4].

As software organizations induct employees belonging to diverse backgrounds with varied work experience, resources work on different projects with varied complexity. Assigning employees to the best-fitted tasks and human resource allocation has become a crucial part in efficient software project planning. Hence, there is a growing need for developing effective tools for analysis of the employee data. The present work proposes a method to identify the significant parameters that contribute to evaluate the resource utilization and contribution in the projects. This would help in making recommendations for future resource allocation and areas of improvement. Also the proposed method can be utilized in general to analyze any set of parameters of interest and could aid in making decision on variety of organizational problems.

II. PRIOR ART SEARCH

There are reports where in K-means clustering is applied to group jobs with similar value to the organization into pay grades, personal management system, and performance of human resource management on performances of the firms etc. [5]-[7]. Also literature reports application of Principal Component Analysis in job evaluation, performance management, and project management etc. [8]-[10].

To the best of our knowledge not much literature is available which reports human resource analysis similar to the perspective of the present work.

III. APPROACH

In the present work for data analysis two methods principal component analysis (PCA) and K-means clustering were utilized. The PCA offers two main advantages as compared to...
other methods like self-organizing maps (SOM) or multidimensional scaling (Sammon mapping) i.e. PCA is fast and provide dimensionality reduction along with the visualization of the data points in 2 dimensional or 3 dimensional graphs. Herein, K-means clustering was preferred over hierarchical clustering. Since, in general, K-means clustering is faster as compared to hierarchical clustering and also K-means produce tighter clusters than hierarchical clustering. An employee data collection and analysis workflow pertaining to the present work is depicted in Fig. 1. Following subsections will define data collection, preprocessing and methodology used for data analysis in the present work.

A. Dataset Collection

The data set is collected from a client data source. Basically it is derived from multiple repositories like Bugzilla, SVN etc. The employee and project attributes are heterogeneous and complex in nature. Also it is observed that there were many missing values in the data, which makes it sparse and incomplete in nature.

![Data Analysis](image)

**Fig. 1.** An employee data collection and analysis workflow.

B. Data Preprocessing

The project data consists of 23 employees (E1-E23) as 11 parameters collected across 10 product releases. These collected parameters with their definition are following: Defect percentile (Def_p): Number of defects fixed by developer

Project complexity (Com_p): overall complexity of bug or defect

Individual effort (Efforts_p): Number of hours spent to fix total number of defects

Relevant experience (Rel_p): Relevant experience

Product level experience (P_exp_p): Product level experience

Number of files change (NFC): Number of files change/added

Average file complexity (AFC): Complexity of the handled files

Function complexity (FC): Complexity of the handled functions

Class complexity (CC): Complexity of the handled classes

Delta functional point change (DFPC): Number of functions changed/added

Manager rating (MR): Rating given by the Manager/Supervisor.

At first parameters were normalized between 0 and 1. Then average of 11 parameters for each employee was calculated using the corresponding parameters based on their contribution across various product releases.

C. Methodology

In the present work two well established statistical methods were used to analyze the obtained employee data using R language version 3.2.1 platform. These methods were Principal Component Analysis for dimensionality reduction and K-means clustering to analyze the hidden clustering pattern in the data. A brief description of these two statistical methods is given below:

1) Principal component analysis (PCA)

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into (smaller) number of uncorrelated variables called as principal components [11]. The first principal component captures the maximal possible variation in the data as possible, in a particular direction. And each following component accounts for maximal possible variation of the remaining variability in the data.

In the present study principal components (as shown below equations 1-4) were obtained as linear combinations of the four variables viz. individual effort (V_i), product level experience (V_j), defect percentile (V_p) and complexity level (V_c) that were chosen for further analysis

\[
PC1 = w_{11}V_1 + w_{21}V_2 + w_{31}V_3 + w_{41}V_4
\]

\[
PC2 = w_{12}V_1 + w_{22}V_2 + w_{32}V_3 + w_{42}V_4
\]

\[
PC3 = w_{13}V_1 + w_{23}V_2 + w_{33}V_3 + w_{43}V_4
\]

\[
PC4 = w_{14}V_1 + w_{24}V_2 + w_{34}V_3 + w_{44}V_4
\]

Where \(w_j\) represents the weight/coefficient of the \(j^{th}\) variable (V_j) towards the \(i^{th}\) principal component (PC_i)

2) K-means clustering

K-means is a simplest unsupervised learning procedure that answer the popular clustering problem [12]. It is a non-hierarchical clustering method that can be utilized for cluster analysis of huge amount of data. The aim of K-means is to minimize the total sum of the squared distance of every point to its corresponding cluster centroid.

IV. RESULTS AND DISCUSSION

To begin the analysis the main challenge was how to characterize employees with respect to their contribution in different product releases. So data preprocessing was done by normalizing data between 0 and 1, followed by taking average of selected parameter for each employee based on their contribution across various product releases.

At first an attempt was made to understand each of the 11 collected parameters with respect to their impact on evaluation of resource utilization and contribution in the projects. Preliminary data analysis was carried out by performing correlation and univariate statistical analysis such as mean, variance etc. Later, four variables viz. individual effort, product level experience, defect percentile and project

185
complexity were chosen for further analysis.

The `princomp` function in R was utilized for performing PCA analysis with the default parameters i.e. using data mean centering scaling and extracting principal components on the covariance data matrix. The following R code was used for PCA analysis:

```r
ResourceData<-read.csv(file="ResourceData.csv", header=TRUE)
rownames(ResourceData)<-ResourceData$Developer
ResourceData$Developer<-NULL
ResourceData_4p <- ResourceData[,c(1,2,3,5)]
pca_4p<-princomp(ResourceData_4p)
summary(pca_4p)
```

To visualize and assess the relative association of the employees, principal component analysis (PCA) was performed on the selected four parameters. This has resulted in dimensionality reduction and revealed that the first two principal components captures major variation (~ 86%) in the data. Fig. 2 shows the scree plot of principal component analysis, which details the relative importance of each component during sequential extraction of information from the dataset.

The loadings (as shown in Table I) revealed that first and second principal component axis were predominantly correlated with ‘product level experience’ (i.e. $r = -0.912$) and ‘defect percentile’ (i.e. $r = 0.932$) respectively, showing their importance in capturing maximal variation in the data. Fig. 3 shows the scatter plot of scores projection along the first two principal moments.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Def_p</td>
<td>0.142</td>
<td>0.932</td>
<td>-0.168</td>
<td>-0.287</td>
</tr>
<tr>
<td>com_p</td>
<td>0.341</td>
<td>-0.311</td>
<td>-0.808</td>
<td>-0.367</td>
</tr>
<tr>
<td>Efforts_p</td>
<td>0.178</td>
<td>0.174</td>
<td>-0.394</td>
<td>0.884</td>
</tr>
<tr>
<td>P_exp_p</td>
<td>-0.912</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For further investigation K-means clustering method was utilized to understand hidden clustering patterns within relative association of employees. The `kmeans` function in R was used with default parameters i.e. Hartigan-Wong algorithm minimizing Euclidian distance, with different number of clusters = 3, 5, 9 and first two principal components.

Following R commands were used for K-means clustering:

```r
km <- kmeans( pca_4p$scores[,c(1,2)], centers = 3)
km <- kmeans( pca_4p$scores[,c(1,2)], centers = 5)
km <- kmeans( pca_4p$scores[,c(1,2)], centers = 9)
```

Figs. 4-6 show the relative position of the formed 3, 5 and 9 clusters respectively formed along with the distribution of employees on first two principal components. It can be noticed that the bigger cluster on the left bottom side and a small cluster on top middle of the graph (encircled in red) remains intact irrespective of the number of cluster centers. Also further increase in number of clusters i.e. from 3 to 9 happened by diffusion of left bottom cluster as shown in Figs. 4-6. This showed the significance of these formed clusters.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Low (1)</th>
<th>Moderate (2)</th>
<th>High (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Def_p</td>
<td>&lt;0.30</td>
<td>0.30 – &lt;0.60</td>
<td>&gt;0.60</td>
</tr>
<tr>
<td>com_p</td>
<td>&lt;0.35</td>
<td>0.35 – &lt;0.60</td>
<td>&gt;0.60</td>
</tr>
<tr>
<td>Efforts_p</td>
<td>&lt;0.40</td>
<td>0.40 – &lt;0.65</td>
<td>&gt;0.65</td>
</tr>
<tr>
<td>P_exp_p</td>
<td>&lt;0.35</td>
<td>0.35 – &lt;0.60</td>
<td>&gt;0.60</td>
</tr>
</tbody>
</table>

Table III provides details of the clusters formed by K-means setting number of cluster values = 3, 5, and 9.
Analysis of the 3 clusters formed by K-means evidently divides the dataset into three classes based on the product level experience (low, moderate and high). Increasing further the K-means clusters to 5 resulted in the previous clusters belonging to product level experience (moderate and high i.e. C3 = 1, 2) to be intact but further divided the cluster with employees of low product level experience into 3 sub-clusters considering project complexity and defect percentile. Setting the K-means to 9 clusters resulted in the 3 clusters (C5 = 1 to 3) from the previous five clusters (K-means = 5) to be intact, and further divided clusters C5 = 4 and 5.

Overall, this analysis revealed following important facts on resource utilization:

1) Employees with the high product level experience (clst3, clst5, clst9 = 2) were mostly (except E14) engaged in handling low project complexity and were fixing moderate number of bugs/defects and that too spending moderate time.

2) Employees with the moderate product level experience (clst3, clst5, clst9 = 1) were engaged in handling low project complexity, were fixing high number of bugs/defects and that too spending moderate time.

3) Employees with the low product level experience (clst9 = 3, 5) were mostly (except E1) engaged in handling high project complexity, were fixing high number of bugs/defects and that too spending moderate time.

The above found facts suggests following:

1) Employees with the high and moderate product level experience were not getting utilized to their best potential. This recommends that these employees should be assigned project with the high or moderate complexity. Conversely, the data analyzed here (as these employees have currently underperformed) put’s following question marks:

   a) Would they be able to take up the project with the high or moderate complexity efficiently?

   b) Why these employees were not assigned projects with moderate/high complexity?

We will be assigning required projects to some of these employees and report the finding in our future publication. Also this recommends that there should be a rational procedure for project assignment considering employees product level experiment, project complexity etc.

2) Employees with the low product level experience in general are performing as per expectation but few employees like E10 and E6 were outperforming as they have handled high project complexity, have fixed high number of bugs/defects and that too spending moderate time. These employees may be recommended for reward or promotion to encourage them.

This finding recommends that certainly there is a need to realign the task/job distribution amongst the employees. This could increase the efficiency of the project progress as well as may also lead to improve employee work satisfaction. This study provided the confidence that proposed methodology viz. principal component analysis followed by K-means clustering using relevant contributory parameters could be applied for large employee dataset to obtain meaningful clusters. The resulting clusters may perhaps guide through the areas of improvement and aid in resource allocation.

![Fig. 5. Depiction of 5 clusters formed by K-means with respect to scatter of two principal components.](image)

![Fig. 6. Depiction of 9 clusters formed by K-means with respect to scatter of two principal components.](image)

<table>
<thead>
<tr>
<th>EmpID</th>
<th>C3</th>
<th>C5</th>
<th>C9</th>
<th>P_exp_p</th>
<th>com_p</th>
<th>Def_p</th>
<th>Efforts_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>E22</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>E23</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>E3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E9</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E14</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E17</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E10</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>E21</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>E2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>E13</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E19</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E20</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E5</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E15</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E16</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>E1</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>E6</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>E4</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

TABLE III: CLUSTER MEMBERSHIP OF 23 EMPLOYEES WITH RESPECT TO K-MEANS CLUSTERING, PRODUCT LEVEL EXPERIENCE, PROJECT COMPLEXITY, DEFECT PERCENTILE AND INDIVIDUAL EFFORTS

1 – low, 2 – moderate, 3 – high
C3 – K-means 3 clusters, C5 – K-means 5 clusters, C9 – K-means 9 clusters,
E1-E23 refer to employee
V. CONCLUSIONS

A pragmatic method for employee data analysis to gain insights into efficiency estimate of resource utilization in various projects has been developed. The PCA analysis helped in identifying two parameters, i.e., ‘product level experience’ and ‘defect percentile’ that captures the major variation in the data. Furthermore, K-means clustering based on first two principal components yielded 3, 5, and 9 clusters for exploration. This analysis helped in identifying the impact of various parameters such as product level experience, project complexity on resource utilization. While examining clusters it was noticed that employees with high product level experience were dealing with low project complexity and few employees with low product level experience were dealing with high project complexity. This finding recommends that certainly there is a need to realign the task/job distribution amongst the employees. Based on this analysis recommendations for future resource allocation and areas of improvement are given.

ACKNOWLEDGMENT

We thank senior management of HCL for encouraging and allowing us to publish this work. We acknowledge the efforts of our development team without which this report would not have been a reality.

REFERENCES


Subhash Ajmani is a Sr. data scientist at HCL Technologies. He received his PhD degree from Devi Ahilya Vishwavidyalaya, Indore for his dissertation in the area of computer aided drug discovery. After his doctoral studies he held postdoctoral research fellowship at the University of Portsmouth, UK researching to develop new methodology and software to support high throughput experimentation (HTE) projects. He has 14 years of industrial research and development experience employed in pharmaceutical, chemical and software organizations. In addition, he has served as a visiting faculty in Interdisciplinary School of Scientific Computing, University of Pune. He has published numerous papers in reputed journals and international conferences. His current research interests includes knowledge discovery, machine learning, predictive analytics, text mining and analytics.

Satya Sai Prakash is a Sr. data scientist at HCL Technologies. In his earlier stints, he has worked with Sulekha.com and BharatMatrimony.com where he was instrumental in personalization and recommendation engines development for improved UXP and adoption of technologies such as Solr / Lucene for improved search, Hadoop MapR with R for data analytics. His strategic research initiatives at Excel Soft have gone long way in the areas of Semantic Learning and Mobile based Learning systems. In addition to the corporate experience (about 11 years), he has about 7 years of academic research experience where he spent his time at IIT Madras as Infosys fellow and at Amrita University as an Associate Professor and the Head of Center for Digital Earth. He also served S K R Engineering College as the Dean of School of Computing. Dr. Sai has a PhD from IIT Madras and double masters (M.Sc. (mathematics and computer science)) and M.Tech (computer science), specialization in artificial intelligence and knowledge systems from Sri Sathya Sai Institute of Higher Learning. He also has PGDM and PGDFM from IGNOU. He has delivered invited talks at State University of New York, Buffalo, USA; Johannes Kepler University, Linz, Austria; Digital Earth Conference, Taiwan; CSCI Chennai chapter and many more. He is professional member of ACM and IEEE & Senior life member of CSI and ISCA.