Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 1 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm

By Ananya Sarker, S.M. Shamim, Dr. Md. Shahiduz Zama
& Md. Mustafizur Rahman
Mawlana Bhashani Science and Technology University

Abstract- Employee is the key element of the organization. The success or failure of an organization depends on the employee performance. Hybrid procedure based on Data Clustering and Decision Tree of Data mining method may be used by the authority to predict the employees' performance for the next year. This paper presents how data clustering method can be applied for evaluating the employee's performance as well in decision making process. Different performance evaluation factors like personality, punctuality, tact oral expression etc has been studied. The result of this paper predicts the number of employee those are selected for promotion or designation and discharged according to their performance. This study help to find out the inefficient employee, magnitude of inefficiency and aids to eliminate inefficiencies with a relatively easy to employ framework.

Keywords: *data clustering, data mining, decision tree, employee's performance, prediction.*

GJCST-C Classification: *I.5.3, H.3.3*

Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm

Ananya Sarker [α], S.M. Shamim [σ], Dr. Md. Shahiduz Zama [ρ] & Md. Mustafizur Rahman [ω]

Abstract- **Employee is the key element of the organization. The success or failure of an organization depends on the employee performance. Hybrid procedure based on Data Clustering and Decision Tree of Data mining method may be used by the authority to predict the employees' performance for the next year. This paper presents how data clustering method can be applied for evaluating the employee's performance as well in decision making process. Different performance evaluation factors like personality, punctuality, tact oral expression etc has been studied. The result of this paper predicts the number of employee those are selected for promotion or designation and discharged according to their performance. This study help to find out the inefficient employee, magnitude of inefficiency and aids to eliminate inefficiencies with a relatively easy to employ framework.**

Keywords: data clustering, data mining, decision tree, employee's performance, prediction.

I. Introduction

Most of the organizations or companies have a formal performance evaluation system in which employee job performance is graded on a regular basis, usually once or twice a year. A good performance evaluation system can prominently benefit an organization. It helps employee behavior toward organizational aims by permitting employees know what is expected for them, and it yields information for making employment decisions, such as those regarding pay raises, promotion or releases. Developing and implementing an effective system is no easy task [1]. An Employee can improve their performance by way of monitoring the progression of their performance [2]. Machine learning algorithms i.e. clustering algorithm and decision tree of data mining technique can be used to find out the key characteristics of future prediction of an organization. Clustering is a method to group data into classes with identical characteristics in which the similarity of intra-class is maximized or minimized [3]. This method is most widely used procedure for future prediction.

The main objective of this paper isto partition Employees into homogeneous group according to their characteristics and abilities using clustering. This study makes use of cluster analysis to segment employees in to groups according to their performance. Decision tree has been used for making meaningful decision for the Employee. Based on the employee's performance results possible to take decision whether advanced training, talent enrichment or further qualification required or not. These applications also help administrative staff to enhance the quality of the organizations.

The structure of this paper is as follows. Section 2, the technical background described. In Section 3, Test Bed Setup has been explained. Section 4, evaluates the results after the experiment. Section 5, concludes the paper with future work.

II. Background

Data Clustering is unsupervised and arithmetic data analysis procedure. Cluster analysis is used to segment a large set of data into subsets called clusters. Each cluster is a collection of data objects that are similar to one another place within the same cluster but are dissimilar to objects place other clusters. It is used to classify the same data into a homogeneous group. It's also used to operate on a large data-set to discover hidden pattern and relationship which helps to make decision quickly and efficiently. Data clustering and decision tree algorithm [4] has been used to evaluate the employee performance. Firstly, apply K-means clustering for separating Employees performance into four clusters which is Excellent, Good, Average and Poor according their Performance. Then apply Decision tree Algorithm for predicting next year Performance.

a) K-Means Clustering

K-means clustering [5] is a type of unsupervised learning, which is used in unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method

Author α ρ: Department of Computer Science and Engineering, Rajshahi University of Engineering & Technology.
e-mail: ictshamim@yahoo.com
Author σ ω: Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University.

produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function. Given a set of data or documents ($x_1, x_2, \ldots, x_i$), where each data point is an *M*-dimensional real vector, the objective of the algorithm is to partition *n* documents into *k* clusters ($k \leq n$) with minimizing an objective function, which may be expressed as,

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} || x_i^j - c_j ||^2$$

Where J is the object function, k number of cluster, n number of cases, and $||x_i^j - c_j||^2$ is a chosen distance measure between a data point x_i^j and the cluster center c_j, [6, 7]. The algorithm and flow-chart of K-means clustering is given below.

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters

- Take first k instances or
- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters.

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.
The flow chart of the k-means algorithm is given below.

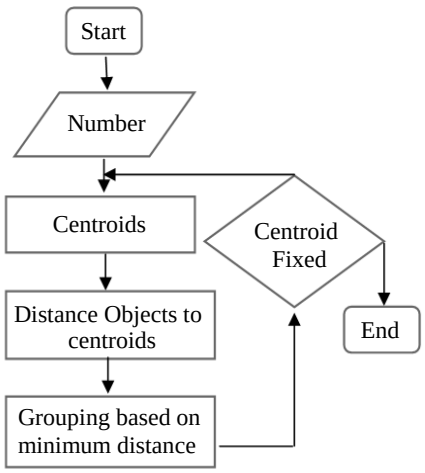


Fig. 1: Flowchart of K-Means Clustering

b) Decision Tree

Data mining [8]consists a set of techniques that can be used to extract relevant and interesting knowledge from data. It has several tasks such as association rule mining, classification and prediction and clustering. All classification methods are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build prototypes that are used to guess future data developments. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner [9].

Decision Tree generates a decision tree from the given training data. It is one of the most used techniques, since it creates the decision tree from the data given using simple equations depending mainly on calculation of the gain ratio, which gives automatically some sort of weights to attributes used, and the researcher can implicitly recognize the most effective attributes on the predicted target. As a result of this technique, a decision tree would be built with classification rules generated from it. Here Decision tree is used for predict Employee Performance.

III. Test Bed Setup

Employee evaluations are an important part of maintaining a motivated and skilled workforce. Every company maintains a confidential report form for measuring the quality of an employee throughout the year. The assessment is comparison with other stead of the same grade, should be recorded by putting their initials in the appropriate. For experiment confidential form has been used which are nineteen categories. These are Intelligence & Mentalalertness, Initiative & Drive, Oral Expression, Written Expression, Ability to plan & organize work, Ability to supervise work, Quality of Work, Perseverance & Devotion to duty, Capacity to guide & train subordinates, Attitude towards superiors, Ability to work with others, Tact, Moral integrity, General sense of responsibility, Responsibility in financial matters, Personality, Public relations, Punctuality and Observance of security measures. The rating scale is the user input to the organization.

Table 1: Performance Rating Scale

Very good	5
Good	4
Average	3
Below Average	2
Poor	1

IV. Result and Analysis

The main objective of this paper is to cluster/group employee according to their performance using K-means clustering and decision tree algorithm. Four years data have been collected from an organization employee's database which consist 100 samples of data.

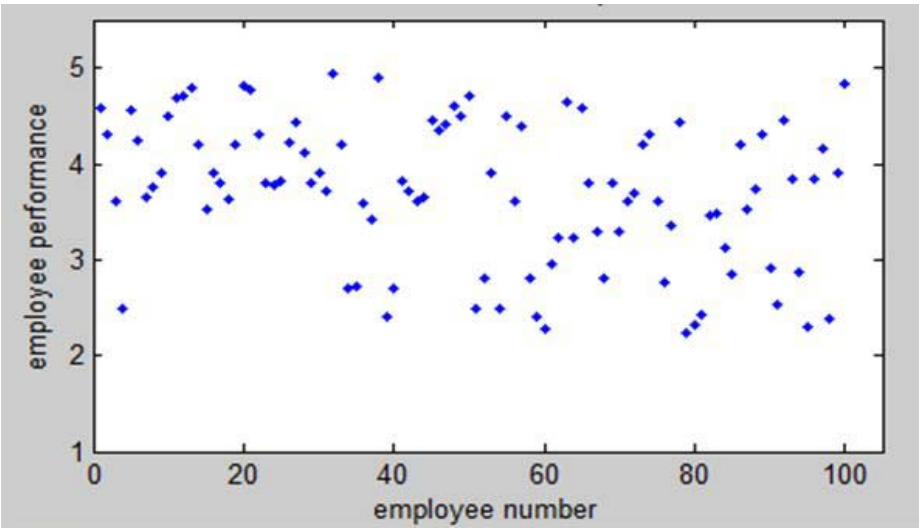


Fig. 2: Data without Clustering

By applying K-means clustering algorithm on the training data four group Excellent, Good, Medium and Poor has created according to employee's previous year performance. Figure 2 shows the initial values before clustering. One is excellent, second is good, third is average and the last one is poor. Figure 3 shows the 4 cluster such as Excellent, good, medium and poor based on their performance throughout 1st year.

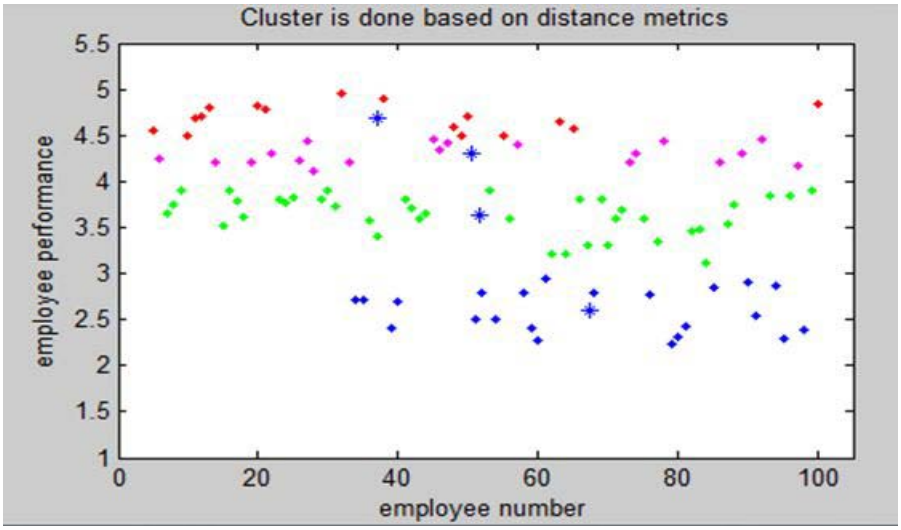


Fig. 3: Clustered Employee into Four Different Regions in 1st Year

Figure 4 shows the clustering result after applying K-Means clustering on 2nd year database of Employee.

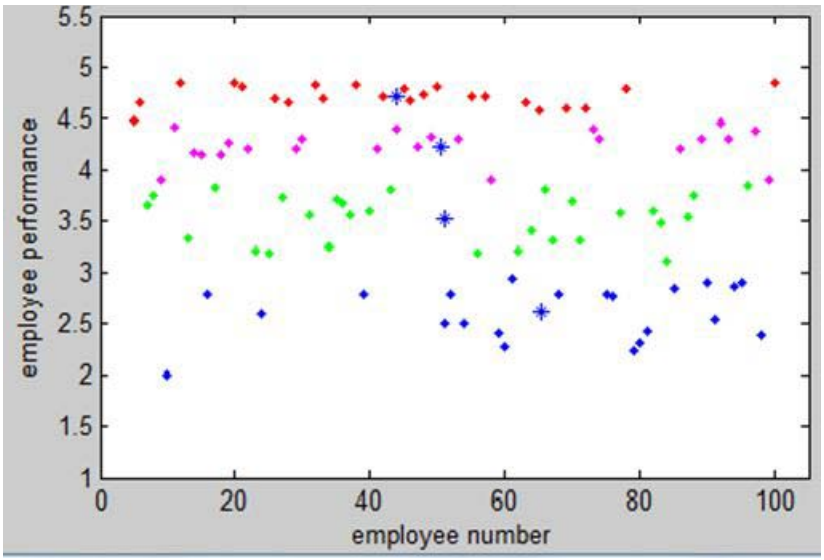


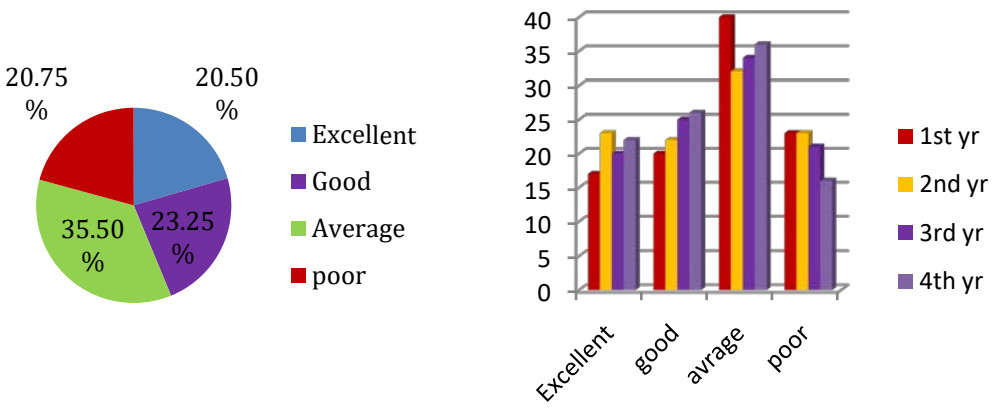
Fig. 4: Cluster Employee into Four Different Regions in 2nd Year

In the similar way 3rd and 4th year performance of 100 employees has been clustered. After clustering calculates the percentage of different cluster around 4 year get the following result.

Table 2: Employee Performance Clustered into 4 Categories

Cluser	Range of Category	No of Employee in different cluster				Percentage
		1st yr	2nd yr	3rd yr	4th yr	
Excellent	Quality>=4.5	17	23	20	22	20.50%
Good	4.0<=Quality<4.5	20	22	25	26	23.25%
Average	3.0<=Quality<4.0	40	32	34	36	35.50%
Poor	Quality<3.0	23	23	21	16	20.75%

Now the Employees four group has been created. One is excellent, second is good, third is average and the last one is poor. Graphical representation of these four categories is given below.



(a). Percentage curve of 4 year Performance (b). Categorical curve of 4 year Performance

Fig. 5: Performance Analysis Graph

Using decision tree algorithms first calculate the entropy from which Information Gain (IG) has to estimate. There are 4 attributes which are 1st year, 2nd year, 3rd year and 4th year Performance of Employee. Maximum Information Gainattribute will be the root node. 1st year attribute has highest Information Gain which is the root node and the 4th year attribute has the lowest Information Gain. Figure 6 shows the decision tree based on some rules.

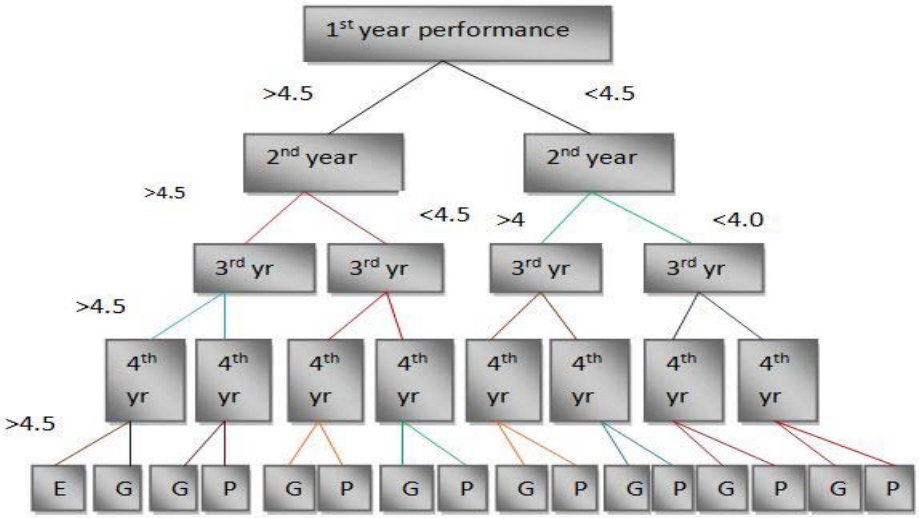


Fig. 6: Decision Tree Based on Rules.

The following rules have been generated after calculating the entropy and information gain:

- If 1st, 2nd, 3rd and 4th year Performance is greater than 4.5 then the final result Excellent.
- If 1st year Performance is greater than 4.5 but 2nd, 3rd and 4th year Performance is greater than 4.0 and less than 4.5 then the final result will be good.
- If 1st year and 2nd year performance is greater than 4.0, but less than 4.5 and 3rd year and 4th year performance is greater than 3.0, but less than 4.0 then the final result Average.
- If 1st, 2nd, 3rd and 4th year Performance is less than 3.0 then the final result will be poor.

From the generated rules administrative stuff or authorities can easily predict he next year performance. Moreover, from this prediction authority can take necessary steps for the qualified and inexpert Employees. The authority may take the following action based on the performance-

- If the Employee is excellent then he/she will get the opportunity of promotion.
- If the Employee's Performance is good he/she need not to take special care but involved themselves with some development activities.
- If the Employee's Performance is medium he/she need some training and try to understand the weak area of him/her, then always prepared for improving themselves.
- If the Employee's Performance is Poor he/she will be selected for expulsion.

Several rules have been created based on four year performance of an organization's employee. If generated rules apply to predict 5th year performance based on previous years, prediction result will be nearly closed to the actual performance.

V. Conclusions

This paper presents an overview of k-means clustering algorithm and decision tree algorithm. K-means clustering algorithm is a common way to define classes of jobs. Here we apply K-means clustering algorithm for partitioning Employee into different cluster based on their quality of Performance. We use decision tree algorithm for classify Employee easily and take appropriate decision quickly. Several actions can be taken in this circumstance to avoid any danger related to hiring poorly performed employee. Future work involves more proper data from several companies. When the appropriate model is generated, these algorithms could be developed for predicting performance of employees in any kind of organization.

References Références Referencias

1. Hameed, A., &Waheed, A.,"Employee Development and it's Affect on Employee Performance a Conceptual Framework". *International journal of business and social science*, *2*(13), 2011.
2. Azar, A., Sebt, M. V., Ahmadi, P., & Rajaeian, A.,"A Model for Personnel Selection with A Data Mining Approach: A Case Study in A Commercial Bank: Original Research". *SA Journal of Human Resource Management*, *11*(1), 1-10, 2013.
3. Varghese, B. M., Unnikrishnan, A., Sciencist, G., Kochi, N. P. O. L., & Kochi, "Clustering Student Data

to Characterize Performance Patterns". *Int. J. Adv. Comput. Sci. Appl*, 138-140, 2010.

4. Lakshmi, T. M., Martin, A., Begum, R. M., & Venkatesan, V. P. "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data". *International Journal of Modern Education and Computer Science*, *5*(5), 18, 2013.
5. Shovon, M., Islam, H., & Haque, M.,"An Approach of Improving Students Academic Performance by using k Means Clustering Algorithm and Decision Tree". *arXiv preprint arXiv:1211.6340,*2012.
6. Nazeer, K. A., & Sebastian, M. P., "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm". In *Proceedings of the World Congress on Engineering* (Vol. 1, pp. 1-3), 2009.
7. Oyelade, O. J., Oladipupo, O. O.,& Obagbuwa, I. C.,"Application of k Means Clustering algorithm for prediction of Students Academic Performance". *arXiv preprint arXiv:1002.2425*, 2010.
8. Al-Radaideh, Q. A., & Al Nagi, E.,"Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance". *International Journal of Advanced Computer Science and Applications*, *3*(2), 2012.
9. Priyama, A., Abhijeeta, R. G., Ratheeb, A., & Srivastavab, S. "Comparative analysis of decision tree classification algorithms". *International Journal of Current Engineering and Technology*, *3*(2), 334-337, 2013.