

EMPLOYEE PERFORMANCE AND LEAVE MANAGEMENT USING DATA MINING TECHNIQUE

¹Himanshu Kumar Singh, ²Kshitij Vishnavat, ³R. Srinivasan

^{1,2,3}CSE, SRM University, Chennai, India

¹himanshukumar_singh@srmuniv.edu.in, ²kshtijsinghvishnava@gmail.com,

³srinivisan.r@ktr.srmuniv.ac.in

Abstract: Almost every company has a performance management system or conduct a performance analysis of its employees. However there is little evidence that this existing system is efficient for evaluating the performance of the worker in a big company. The aim of this paper is set references and guidelines for performance analysis of employees for a typical sales company. Our system takes in employee data such as the attendance, pay rate, revenue, expense, gender, and remark to evaluate the performance of the employees of the company. Our system would be using Bayesian Classification algorithm for evaluating the performance of the employees.

Keywords: Data-mining, Employee performance analysis, Data mining Classification.

1. Introduction

A tremendous growth in the available data as a result of the computerization of the every aspect of organization has immensely to the development of the intelligent decision making technologies. A young but promising of these kind of technology is Data Mining which is the process of analyzing data from different perspectives and summarizing it into useful information. Data Mining aims at developing knowledge from available data and could further be used for improving the process as the whole.

There are many classification technique in Data Mining such as Decision Tree, Neural network, Rough Set Theory, Bayesian theory and Fuzzy logic. In this we have considered the Bayesian theory for the purpose of classification. Bayesian

Classification is less complex as compared to the other classification but does effect the accuracy of the system due its less complexity.

Data mining stand out due to its wide array of techniques from the different domains such as statistics, artificial intelligence, machine learning, database systems and visualization. These serve as groundwork for its application to employee performance analysis for a sales company. Data Mining has gained popularity due its tools with potential to identify trends within data

and turn them into knowledge that could serve as the strong basis for the analysis.

Our proposed project is typically helpful for big sales company or the companies which consider criteria such as revenue, pay rate as its main criteria. We are taking criteria other than net revenue like attendance, last appraisal rating for analysis of the performance of the employees. In order to include the behavioral component into account we have included remark component, which is typically given by HR managers of the employee, who judge by interacting with employees and observing the behavior of them. All of the above criteria is used to calculate the performance score of the employee and amount of leave which could be granted to the employee.

We basically use the training data with the five component in it to train the system to construct the Bayesian Table. Based on which a score is calculated which further being used to generate the performance rating of each employee. Using Bayesian Classification algorithm we are able to calculate the score of the employee. The leave to a specific employee granted depending on the performance score of the employee. In the end we finally try to illustrate the performance of the specific employee graphically in comparison to other employees. This graphical depiction typically illustrate the performance of an employee in comparison to other employees with the help of the bar chart.

2. Related work

Several studies have used Data Mining for extracting rules and predicting certain behavior in several areas science, information technology, management, education, biology and Medicine. Data mining tools [1] are being used in order to analyze large amount of data generated by large organizations and drawing fruitful conclusion and inferences. An overview of the Data Mining system and some of its application is given in [2].

[3] used the decision tree algorithm for predicting the performance of the employee using the data of the public development institute of Kenya. It has been

noticed from that the performance of the employee is being highly affected by the experience, age, academic qualification, professional training, gender, marital status and previous appraisal scores.[4] have used the clustering and Navies Bayes on educational data for the purpose of classification of the student under various categories.

[5] used the data mining tools to predict the employee's performance based upon diverse factors like personal, social skill,punctuality,dependability,Interpersonalrelations,task,responsibility,work standards, actual performance, Psychological and other environment variables. It has basically used decision tree algorithm in implementing it. [6] provides us with the references and set of case studies for management to better understand the process of performance management within a particular organization. Based on the three companies analyzed it provides recommendations to managers for how they can work within the boundaries of their system to conduct more effective performance reviews.

[7] adopted the CRISP-DM data mining methodology for building the classification model. It is also being used for predicting the performance of the employees of the company. Decision tree is the main data mining tool used to build the classification model, where several classification rules were generated. [8] Used a Naive Bayes classifier to predict job performance in a call center with the aim of knowing what levels of the attributes are indicative of individuals who perform well. By using operational records, they predicted future performance of sales agents, achieving satisfactory results.

[9] Developed a Data Mining framework based on decision tree and association rules to generate useful rules for personnel selection. This framework can be used to develop an effective personnel selection mechanism to find the talents who are the most suitable to their own organizations. Their results show that specific recruitment and human resource management strategies were created.[10] Used rule-based classification Data Mining technique to extract knowledge significant for predicting training needs of newly-hired faculty members in order to devise the necessary development programs. They used the Cross Industry Standard Process for Data Mining (CRISP-DM) in discovering significant models needed for predictive analysis and demonstrated the required professional trainings to prepare faculty members to perform their tasks effectively.

[11] Built models that used classification algorithm like decision trees and Naïve Bayes to rank the applicants for a job profile based on their resume and social media presence. Theirs is a match making system where the companies will be given a list of ranked candidates using information retrieval technique like two way relevance ranking.

[12]Proposed a potential Data Mining technique for talent forecasting by identifying potential talent using past experience knowledge. [13] Show how the potential human talent can be predicted using a decision tree classifier with the pattern of talent performance identified through the classification process. They used decision tree C4.5 classification algorithm to generate the classification rules for human talent performance records.

[14]Suggested a model for talent management that can be used as a decision support tool and performs several different type of analysis. This model uses supervised as well as unsupervised techniques. The authors have further tested the accuracy of prediction using four different classification algorithms and have achieved significant accuracy in the classification results. [15]Also built similar model but with data warehousing perspective of the data. [16]Applied Data Mining in the context of higher educational system to discover useful knowledge.

The abundance of data has attracted Data Mining research towards the domain of Human Resource Management. The review by [17] shows that HRM constitutes a noteworthy new domain of Data Mining research that is dominated by method and technology-oriented work. However, there is need for specific domain requirements, such as performance evaluation, or compliance with legal standards. [18] Reviewed HR applications and talent management, the use of Data Mining technique in HR and proposed the potential HR system architecture for talent forecasting.[19] Showed the ability of data mining in improving the quality of the decision-making process in HRM Systems by showing how to discover and extract useful patterns from large data sets in order to find observable patterns in HR.

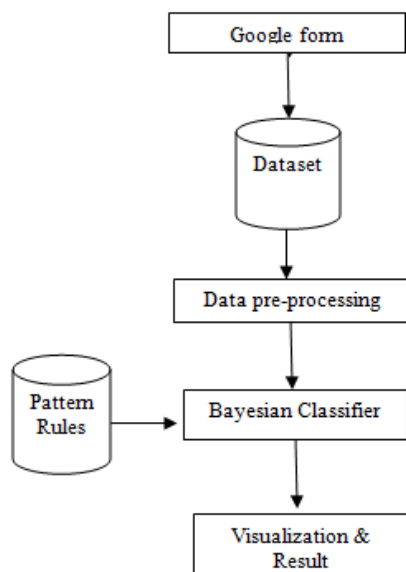
Each of the work cited above has primarily used classification supervised learning approach or clustering – unsupervised learning for constructing Data Mining models in HRM.

3. Proposed System

A. Architecture Design

The architecture diagram for predicting the performance of the employees is schematically represented in the flowchart given below. Here the HR manager of the company would be giving input in the Google form. The input data would be stored in the excel sheet which would act as the dataset for our project. Our system can analyze at about 800 employees performance at one time. The next section is the data preprocessing. In this unwanted data such as @, ~, ` etc are being removed from the remark section of the dataset. Only the necessary data are being send for further process. In the data pre-processing section another which is being added is the spam detection. If

the data in the datasets are being repeated then our system detect the repetition of the data in the data in the dataset. Our system then detect the number of male and female employees simultaneously. The training data in the form of pattern rules has already been stored in the system. The Bayesian classifier takes the training data for classifying the employees in the dataset. The employees are being rated on the scale of 6. We would basically be using Bayesian classification algorithm for the purpose of classifying the the employee from the dataset.



B. Dataset

Dataset basically consist of the employees data which is being used for calculating the score of the employees on the scale of 6. The employee data include :-

1. Revenue
2. Pay Rate
3. Attendance
4. Expense
5. Employee Remark

Based on the above attribute the performance score of each of the employee is calculated and leave is being granted to each of the employee accordingly.

C. Data Understanding

The data for training the model was collected through the extraction and organization of the employee performance data for one of the reputed sales company in the country. The data collected was for one years and included 430 employee's performance appraisal records, described by 5 parameters. These parameters basically represented the skills set required to be

possessed by an employee. This data was divided into two datasets. The first

Dataset of 30 records was used for training and testing, while rest other dataset was used for validating the model.

D. Algorithm used

The algorithm used in our model is Navies Bayes Algorithm. The Bayes Classifier is used for classifying the new dataset based on the training data. In machine learning, **Naïve Bayes classifiers** are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c|x) = P(x|c) * P(c) / P(x)$$

$$P(c/x) = P(x1/c) * P(x2/c) * \dots * P(xn/c) * P(c)$$

Where,

- $P(c|x)$ is the posterior probability of class (target) given predictor (attribute).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

Our employee performance management system is based on this above equation. Here we would be calculating the performance of the employees on the scale of 6 by calculating the probability of a employee lying on a particular scale out of six, which is the probability of the employee lying on a particular scale. We would be having known data such as Revenue, Pay Rate, Attendance, Expense, and Employee Performance Score is calculated based on maximum probability of all the given scale.

E. Classification Model

Our classification model aims at assigning points to each of the employee based on the training data available to us. We would also be assigning the leave to the employees based on the calculated score. Our pre-processor would first remove the tuples in the dataset which have been repeated. The unwanted elements in the dataset such as character such as @,* etc would be removed from the dataset. Our model would also remove the unwanted columns in the dataset other than Revenue, Pay Rate, Attendance, Expense and the performance Remark. The rest of the other columns would not be considered while detection of the data. Our model would be having different criteria for calculating the performance of the male and female employees. Our system front end would be implemented with the C# and back end would be implemented SQL server.

The training module consist of set of 30 performance remarks and each of the performance remark in the training module would consist of set of 4 possible combination of revenue, Pay Rate, Attendance, Expense, points scale out of 6. The possible remark is given by the HR manager of the company. The remark of employee performance could be like the good, excellent, brilliant, need improvement, bad, employee should be fired etc. This way the 30 possible remark could be adjusted with each remark adjusted with the 4 possible combination of the revenue, pay rate, attendance etc. This training module could be used to construct the Bayesian Table with the above mentioned 5 features. Whenever we come across a new tuple of revenue, pay rate, Attendance, Expense, then we compare our data with the available Bayesian table to arrive at the point score of 6. Based on the point score we are able to calculate the leaves which are needed to be granted to each of the employees of the company. For any particular tuple we are basically calculating the probability of score of any particular employee lying in each of the scale of the points from 1 to 6 by comparing with the Bayesian table. The maximum probability among all of the possible scores from 1 to 6 will be assigned to any particular employee. The leave is given based to the possible score. For e.g. the employee having score of 1 be assigned 1 day leave, with the score of 2 be assigned 2 days of leave, with the score of 3 be assigned 3 days of leave and just like that for rest of the 3 point scores. A possible tuple for the analysis would be like:-

Employee ID-1103024456
Employee Name-Mr. Brown
Age-32
Date of Hire-10/27/2008
Department -Admin Office
Position-Accountant
Revenue-\$75000
Pay Rate-\$100000

Expense-\$12000

Attendance-56%

Based on the above tuple the point score of 3 is calculated and the leave of 3 days is being assigned to that specific person. In case the possible score in the decimal arises then the nearest whole number of days of leave would be assigned to the specific person. For e.g.: if the point score of 2.6 arises, then 3 days of leave would be assigned to that specific person.

First of all the class attribute is calculated. Here the class attribute is the point scored calculated. The class attribute probability is calculated by dividing the number of times a particular class has occurred in the training data by total number of element which occur in the training data. Here there are 6 possible classes in the training data. The 6 classes include 1 to 6 point scored. These points are assigned according to the performance of the particular employee. So accordingly the class probability are being calculated. For e.g.: the class probability of point score 1 is being calculated as number of times 1 has occurred in the training data divided by the total number of the element which are there in the dataset. Similarly the class attribute for rest of the other class attribute are being calculated.

The next part include dividing the various attribute features such as Revenue, Pay Rate, Attendance, and Expense into various ranges so that probability within certain ranges could easily be calculated. For e.g.: let us take revenue. The revenue could easily be divided into certain range. If we are having revenue from \$0 to \$10000. So we can easily divide the revenue into ranges of \$5000. So the revenue distribution would be like \$0 to \$5000, \$5000 to \$10000 and so on. So we basically calculate the number of individual lying within the certain range of revenue for the individual lying within the training data. This data is further used for the probability calculation. Similarly the range is calculated and the number of individual within the certain range is calculated for the rest of the attribute like Pay rate, Attendance, Expense. These data are further used for construction of Bayesian probability distribution table. Similarly as like the probability distribution of the class attribute, the probability distribution of the performance remark is also calculated. The above data calculated is used further for the analysis. Further the probability distribution of the attribute range is being calculated in the model.

The probability of the gender is also being calculated. This calculation is being done by dividing the number of employees in the particular gender by the total number of employees in the training data. Next step involve calculating the probability of various attribute such as revenue, attendance etc within certain ranges. Finally we come up with the Bayes Distribution table with the probability of the each five attribute within certain ranges. The range is being set at the discretion of the developer of the employee management system.

Now let us take the element from the dataset for understanding the performance analysis of the employee. We would consider the tuple with the employee name Mr. Brown for analysis. We have illustrated the tuple with name of Mr. Brown above. So for convenience we name the employee tuple with the name of Mr. Brown as **T**. We would represent the point scored by the employee which is the class attribute as **Q**. So for the purpose of analysis the final class probability which would be calculated would be like, with the class be like the point scored by each employee as:-

$$P(Q/T) = P(T/Q) * P(Q)/P(T)$$

The class attribute **Q** range from 1 to 6. The calculation of $P(T/Q)$ take place by multiplication of various probability of the attribute given the class attribute. This calculation is being done by calculating the probabilities from the Bayesian Table. For e.g.: let us take the class attribute being point score 1 for tuple **T**. So we would calculate the probability of each attribute like revenue, expense, attendance etc each of which in a given range. Each of the probability would be calculated given that the point scored by the person is 1 from the Bayes Distribution Table. Below equation represent the tuple probability given the class 1:-

$$P(T/Q1) = P(T's \text{ revenue}/Q1) * P(T's \text{ attendance}/Q1) * P(T's \text{ Pay rate}/Q1) * P(T's \text{ Expense}/Q1) * P(T's \text{ Remark}/Q1) * P(T's \text{ gender}/Q1)$$

Here **Q1** represent the class attribute which is the point score of 1. And each of the revenue, attendance etc represent the probability within the certain range. By multiplication of above $P(T/Q)$ with $P(Q)$ we arrive at the likelihood of the tuple lying within the particular class. So the calculation of the equation $P(T/Q1) * P(Q1)$ would give us the likelihood of the tuple having the pointed scored equal to 1. Similarly the likelihood of the tuple **T** for every possible class, with the point scored from 1 to 6 is being calculated. So when we calculated the all possible likelihood for the tuple **T**, we sum up all the possible tuple to arrive the estimated value for the tuple **T** with is being given by the $P(T)$. Since we have already calculated the likelihood by multiplication of $P(T/Q)$ and $P(Q)$, so when we divide the likelihood by estimated value $P(T)$ we get probability of the point scored by the employee given the particular employee in the dataset with the tuple **T** which is given by $P(Q/T)$. For any particular tuple **T**, we calculate the $P(Q/T)$ for all the possible class from 1 to 6. The class value for any tuple is given by the class for which the $P(Q/T)$ is maximum. Here the class value is the point scored by the employee. So in this way the point scored by the employee given the other five attributes is being calculated. The leave is finally calculated by the point scored by the employee. An employee which is having the higher point score is being given the more number of leave in comparison to the other employees. So in this way amount of leave is being assigned to any particular employee. This model

is implemented for 400 employees in the dataset and finally we represent the performance of the employees virtually with the help of the bar chart. The success of the employee performance management system depend upon the quality of the training data. The better the quality of the training data the better is the performance of the system. So in this way our classification model is being finally implemented.

4. Evaluation

Overall the analysis of the employee performance system give us the accuracy of about 65%. It means that this model give us the correct analysis of the employee performance 13 times for a dataset of 20 data. The analysis was performed for about 400 element dataset of employees. 30 dataset were used as the training data. Accuracy of the system could be increased by increasing the tuples in the training data. The accuracy of the existing model could decrease if the analysis of the system is performed with more than 400 employees. The analysis of the system accuracy is done at the discretion of the developer and accuracy calculation of the system could vary depending on the analysis of by any third person. But the range accuracy analyzed by any third person could vary in the maximum range of 5%. It has to be noted that with the increase in the attribute for analysis the accuracy of the system decreased simultaneously in proportion. It is also noteworthy that the attributes which increases the accuracy of the analysis system are revenue and attendance, which was analyzed based on individual analysis of five attributes with the class attribute.

5. Conclusion and Future Prospect

This paper focused on the possibility of building a classification model for predicting employee performance. We used 5 attributes for analysis of the performance of the employees. These 5 attribute were used to calculate the performance score. The classifier used in the analysis was the Navies Bayes Classifier. We used the 30 item dataset for training our system. Here our class attribute was point scored which was calculated by our system based on the training data. Our system was implemented with the performance analysis of dataset of 400 employees. The accuracy of our system was found out to be 65%. With the increasing in the training data the system accuracy could be increased at least by 5%. In future with the training data of about 300 dataset and more attribute along with 5 attribute which has been used in our model, we can built a real life system which could predict the employee performance management system which could predict the performance of the employee with accuracy of about 95%. A better system than our existing model could be implemented with the algorithm like decision tree algorithm which could

work on more complex employee data and with better accuracy. Our system is helpful for big companies which judges its employees on regular basis based on the revenue, pay rate, attendance, employee as an expense to company and behavior of the employee.

References

- [1] Witten, I., H., Frank, E., (2005). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, Elsevier Inc.
- [2] Mishra, P., Padhy, N., Panigrahi, R. (2013). The survey of Data Mining applications and feature scope, Asian Journal of Computer Science & Information Technology, 2(4), pp 67 – 77.
- [3] John M. Kirimi, Christopher A. Moturi (2016). Application of Data Mining Classification in Employee Performance Prediction, International Journal of Computer Applications (0975 – 8887) Volume 146 – No.7, July 2016
- [4] Ashish Dutt, Maizatul, Akmar Ismail (2016), A systematic review on Educational Data Mining. IEEE (2016).
- [5] Rohit Kotalwar, Raj Chavan, Sunny Gandhi, Vinayak Parmar (2014), Data mining: Evaluating Performance of Employee's using Classification Algorithm Based on Decision Tree. IRACST – Engineering Science and Technology: An International Journal (ESTIJ), ISSN: 2250-3498 Vol.4, No. 2, April 2014.
- [6] Tom Shott, Chris Imondi, Ryan Fedie (2012), Evaluation of Performance Management Systems for Knowledge Workers, 2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies.
- [7] Qasem A. Al-Radaideh, Eman Al Nagi (2012), Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.
- [8] Valle, M.A., Varas, S., Ruz, G.A., (2012). Job performance prediction in a call center using a Naive Bayes classifier, Expert Systems with Applications, 39(11), pp 9939–9945.
- [9] hein, C.F., Chen, L.F., (2008). Data Mining to improve personnel selection and enhance human capital: A case study in high technology industry, Expert Systems with Applications, 34(1), pp 280–290.
- [10] Ancheta, R.A., Cabauatan, R.J.M., Lorena, B.T.T., Rabago, W., (2012). Predicting faculty development trainings and performance using rule-based classification algorithm, Asian Journal of Computer Science and Information Technology 2: 7, pp 203 – 209.
- [11] Sarda, V., Sakaria, P., Sindhu Nair, S., (2014). Relevance Ranking Algorithm for Job Portals, International Journal of Current Engineering and Technology, 4(5), pp 3157 – 3160.
- [12] Jantan, H., Hamdan, A. R., Othman, Z. A. (2011). Towards applying Data Mining techniques for talent Management. International Conference on Computer Engineering and Applications IPCSIT Vol. 2.
- [13] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm, International Journal on Computer Science and Engineering, 2(08-2010), pp 2526-2534.
- [14] Hamidah J., Abdul R.H., Zulaiha A.O., (2009). Knowledge Discovery Techniques for Talent Forecasting in Human Resource Applications, World Academy of Science, Engineering and Technology, 3.
- [15] Zhao, X. (2008). A Study of Performance Evaluation of HRM: Based on Data Mining, FITME 2008, International Seminar on Future Information Technology and Management Engineering.
- [16] Delavari, N., Phon-Amnuaisuk S., (2008). Data Mining Application in Higher Learning Institutions, Informatics in Education, 7(1), pp. 31–54.
- [17] Strohmeier, S., Piazza, F., (2013). Domain driven Data Mining in human resource management: A review of current research, Expert Systems with Applications, 40(7), pp 2410–2420.
- [18] Jantan, H., Hamdan, A. R., Othman, Z. A. (2009). Knowledge discovery techniques for talent forecasting in human resource application. World Academy of Science, Engineering and Technology, Penang, Malaysia, pp 803- 811.
- [19] Ranjan, J., Goyal, D.P., S I Ahson, S.I., (2008). Data mining techniques for better decisions in human resource management systems, International Journal of Business Information Systems 3(5) pp 464 – 481.
- [20] T.Padmapriya and V.Saminadan, "Utility based Vertical Handoff Decision Model for LTE-A networks", International Journal of Computer Science and Information Security, ISSN 1947-5500, vol.14, no.11, November 2016.
- [21] S.V.Manikanthan & T.Padmapriya "Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g", International Journal of Pure and Applied Mathematics, ISSN NO:1314-3395, Vol-115, Issue -8, Sep 2017.

