International Journal of Engineering and Techniques - Volume 4 Issue 3, May 2018

RESEARCH ARTICLE OPEN ACCESS

Data mining of Absentee data to increase productivity

Gayathri.T[1]

Assistant Professor, Department of Computer science, New Horizon college of Engineering, Bangalore

Abstract:

Productivity of an organization reduces when employees are absent for prolonged duration. This could be avoided if the employee's absentia is understood in the early period and a substitute is assigned. This paper is a research to create classification model to predict whether an employee would be absent for short or long duration. Various classification models are studied and Multilayer perceptron in found to be the best suited model for this purpose. Absenteeism record of Courier company from UCI data repository is used for this study.

***Keywords*- Classification, absent employee, Multilayer Perceptron, Naive Bayes, J48, Data mining.**

I. INTRODUCTION

Data mining provides us a forum for predicting, analysing and grouping different problem statement of various genres without a subject matter expert. Data mining has its importance in many fields such as forecasting weather, predicting customers purchase pattern, spam detection, disease diagnosis, robotics and numerous other fields. There are four steps in the process of Data mining.Data collection, Data pre-processing, machine learning and Data visualisation. Data collection involves understanding the domain, collecting the data. Data pre-processing is where the collected data is cleaned and transformed before Machine learning. Machine learning is the process where the system learns the data. From the learned knowledge it predicts, associates classifies and clusters the data. For this purpose, various algorithms are used.The information in raw form can be difficult for the user to understand. So this data is visualized in graphical and projected form using data visualization tools.

A domain which requires the interference of data mining is Human Resource management. It is difficult to analyse the pattern in human behaviour. Machine learning helps us to identify hidden, interesting pattern in human behaviour. Productivity of an organization improves when a staff provides dedicated full time work. When an employee is absent for a short duration, it is possible for him/her to finish the pending task. If a person is absent for a prolonged duration, it would be difficult for him/her to finish task. This would also lead to decreased productivity.

This paper is an attempt to study the various algorithms to classify absenteeism dataset based on the duration of absenteeism using WEKA [1].

For this purpose UCI dataset – Absenteeism at work data set is used [2]. The dataset was collected from the records of courier company from 2007-2010. Andrea Martiniano, Ricardo Pinto Ferreira and Renato Jose Sassi created the dataset in 2012.

II. LITERATURE SURVEY

Absenteeism dataset is used in [3] to analyse the reason for absence from work. The productivity in industry and organisations can decrease because of employee absence. Employee might be irregular to work due to multiple reasons including, sickness, depression, unhappiness in work etc. this purpose uses an Artificial Neural Network to predict absentees.

The future work of [3] suggests that a model can be used to find out whether an employee would be absent for a week or a month. This paper classifies the employee record to find if he would be absent for hours, day, week or month.

III. DATA COLLECTION

Absenteeism data set is downloaded from UCI data repository. This dataset was used in the prediction of absentee in an organization in [3]. The same data set is used in this paper.

The dataset contains 21 attributes and 741 instances. The attributes present in the dataset are: ID, Reason for absence, Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Hit target, Disciplinary failure, Education, Son, Social drinker, Social smoker, Pet, Weight, Height, Body mass index, Absenteeism time in hours

IV. DATA PRE-PROCESSING

Through history the importance of data pre-processing in undermined. Data pre-processing can be data cleaning or data transformation. When the data is collected, data is unstructured. To get meaning from this unstructured data, they are converted to rows and columns. The row signify the data sample and column signify varied information on the samples. Even though the data is structured, it is not advisable to feed them directly to algorithm. [4] suggests the use of data pre-processing to improve machine learning.Classification and clustering accuracy is predominantly dependent on the proper representation of data. Data pre-processing involves data cleaning, data transformation, data reduction, oversampling data anddata selection.

Absenteeism dataset does not have a class, as it is used to create a prediction model. When a person is working for an organization, hisproductivity is affected when he is absent for a prolonged duration. Absent duration is an important attribute it has value between 0 and 120 hours. To use this dataset in classification model, the class attribute should be of nominal data. Further the problem statement is to predict prolonged absenteeism. So the numeric data of 0-120 is transformed as follows.

TABLE I
CONVERSION OF NUMERIC DATA OF ABSENTEEISM IN HOURS FIELD

Numeric Value	Nominal value
0	NOT ABSENT
1-16	DAYS
17-56	WEEK
>56	MONTH

Discretization was also tried on the field. The problem with discretization was the number of bins had to specified. The correct prediction of the number of bins is not always accurate. When discretization was applied the nominal value is a range values like 0-17.5 hours. It is more meaningful to consider the 0 hours as a separate nominal value as it signifies no absent.

V. MACHINE LEARNING

Machine learning is a method of getting information or knowledge analysis the various patterns and rules in data [5]. Many machine learning algorithms are available, their correctness vary from application to application [6]. For any application it is important to apply few machine learning algorithms to find out the best suited model. Machine learning algorithms can be grouped under Bayes, Rule Based, Neural network and Decision tree.

A. Naïve Bayes

Naïve Bayes is conditional probabilistic model. Naïve Bayes assumes that each field contribute to the classification of data independently [7]. Each instance is considered as a vector. The probability of occurrence of class if found by multiplying its prior probability and likelihood divided by the evidence.

B. Decision Tree

Decision tree is arrived at by finding the optimum way to arrange the various nodes. Each node is a filter which has a rule based on one of the attributes and splits the record for further process.[8]

C. Multilayer Perceptron

Multilayer perceptron contains large number of nodes called as neurons, joined together so that they for input layer hidden layer and output layer. The instances are supplied though the input layer, bias and weight are added at the hidden layer and supplies the class in output layer [9].

VI. EXPERIMENTATION

The dataset was downloaded from UCI repository it contained 21 attributes and 741 instances. The dataset does not contain class information. The dataset is pre-processed so that Absenteeism in hours act as a class. This pre-processing is done such that no absenteeism is considered as a separate class, days of absentia, weeks of absentia and month of absentia are considered as separate class. According to [10] J48, Multiplayer perceptron and Naïve Bayes are algorithms that provide predominantly good result. So the same was studied for absenteeism in employees.

TABLE III
ACCURACY RESULTS FOR CLASSIFICATION MODEL FOR ABSENTEEISM RECORDS

Algorithm	**Accuracy**	**Root mean squared error**
J48	93.5135	0.1754
Multilayer Perceptron	97.5676	0.0969
Naïve Bayes	89.3243	0.2053

VII. CONCLUSIONS

Absenteeism dataset provide better results with Multilayer perceptron. Even though J48 provides an accuracy of 93%, it cannot be used for this application because. J48 classifies all the instance as DAYS ABENT. J48 for this algorithm suffers because of imbalance in dataset. Multilayer perceptron works well giving an accuracy of 97% and it classify all the four classes with minimum error.

After preprocessing and Multilayer perceptron this model can be used in the finding the employee who would be in prolonged absentia.

ACKNOWLEDGMENT

I thank my college New Horizon college of engineering for providing support and tools for this research. I thank Head of Department, Dr.B.Rajalakshmi for her support and guidance.

REFERENCES

1. *G Holmes, A Donkin, IH Witten, "WEKA: a machine learning workbench", Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems, pp. 357-361, 1994*
2. https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work#
3. *Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.*
4. *D. H. Deshmukh, T. Ghorpade, and P. Padiya, "Improving classification using preprocessing and machine learning algorithms on nslkdddataset," in Communication, Information & Computing Technology(ICCICT), 2015 International Conference on. IEEE, 2015, pp. 1–6*
5. *Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.*
6. *Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1.1 (1997): 67-82.*
7. *Lewis, David D. "Naive (Bayes) at forty: The independence assumption in information retrieval." European conference on machine learning. Springer, Berlin, Heidelberg, 1998.*
8. *Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993*
9. *Goodman, Rodney M., and Zheng Zeng. "A learning algorithm for multi-layer perceptrons with hard-limiting threshold units." Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop. IEEE, 1994.*
10. *Borges, Lucas Rodrigues. "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection." Group 1.369 (1989).*