IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Enhancing The Prediction of Absenteeism By Decision Cluster Based Rule Generation

S. Adaekalavan

Department of Computer Science,J.J. College of Arts and Science (Autonomous),Pudukkottai, India

Corresponding Author: kingsmakers@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.326330 | Available online at: www.ijcseonline.org

Accepted: 17/May/2019, Published: 31/May/2019

***Abstract*—**The Data analysis inspires many applications in the field of Computing. It may be a phase either in design or on-line operation. The procedures of Data analysis are dichotomized as exploratory and confirmatory. Irrespective of these two types, the primary component for both procedures is grouping or classification. It can be done based on either (i) goodness-of-fit to a postulated model or (ii) natural groupings (clustering) revealed through analysis. Clustering is a process of partitioning a set of data or objects into a set of meaningful sub-classes, called clusters based on similarity. Obliviously, clustering has its own impact in solving complex real world problems. This paper addresses the impact of clustering algorithms for one such problem i.e. for the prediction of absenteeism at work place. The proposed method will draw predictions about absenteeism at work place by decision cluster based rule generation.

***Keywords*—**Computing, Data Analysis, Clustering, Classification

I. INTRODUCTION

The present scenario of technological upcoming, raised the level of knowledge to handle huge volume of data. People are being aware of the technological trends in their respective fields which obliviously contribute to the tremendous growth of data. The primary source is Internet. Since data are created in each and every aspect of Internet. The collected data should be used properly. Data Mining is primary phase of Knowledge Discovery in Database (KDD) process which contains examination of data and extraction of useful knowledge from the raw data [10]. Data Mining employs different techniques to handle this data and extract useful knowledge. The knowledge thus obtained can be used for future Decision Making. Cluster Analysis is one among the essential strategies used to mine the database [5]. Clustering is usually utilized as a solitary instrument for understanding the problem domain or applied to investigate the domain, preparation of information, pre-process the data for different purpose.

Clustering is a process of grouping a set of given data into a set of meaningful sub-classes, called clusters. Clustering [1] is the unsupervised classification of patterns which may be observations, data items, or feature vectors etc into groups. Thus the Cluster analysis divides data into meaningful or useful groups called clusters. The problem of clustering has been addressed in many contexts by the researchers belonging to different disciplines. This reveals the wide range of demand for including clustering as an inevitable step in groping the data while performing analysis. Irrespective of the difficulties faced in the definition of clusters, its assumptions and context, clustering has its own impact in solving difficult problems. Clustering algorithms are broadly categorized as Hierarchical, Partitioned and Density based [2].

This research paper is organized as follows. Section II of this paper elaborates the background study about Clustering algorithms. Section III discusses about the proposed architecture and presents the test bed for the proposed work to predict absenteeism at work place. Section IV describes the experimental study and depicts the performance of the proposed work. Section V concludes this research work with summary of the research findings and suggests some future research directions.

II. REVIEW OF LITERATURE

The objective of clustering is to group the data into meaningful clusters. The group thus formed will consists of objects that are similar to each other and different or unrelated to the objects in the other groups. The clustering is said to be more distinct if the similarity within a group and the difference between the groups are greater [7].

The expected outcome of every cluster analysis should have a crisp grouping of data into non-overlapping groups. As an understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining [3].

The resulting clusters should depict the "natural" structure of the data. The clustering is being applied for varying applications either as a primary functionality or as a preprocessing phase. For example, the applications such as finding the similar protein structures, identifying the earthquake prone locations, grouping the web documents for searching purpose, etc are domains where clustering is done for the prime objective. In some other cases like image or data compression, pattern matching, and recommender systems clustering is used as preliminary phase [9].

This research work is motivated by many research works. Some are described in this section. The [6] research paper presented a mechanism to predict absenteeism using artificial neural networks. The authors use rough set based approach to reduce the number attributes. The author also illustrated that the computational cost is reduced due to the attribute reduction.

The authors [8] proposed a Data Mining approach to predict forest fire. The SVM based method is used in this work and with the help of only four direct attributes, the authors can able to predict about the size of forest fire.

In [4], the authors analyzed the reason for absenteeism in work in this paper. The productivity in any organization is affected by the absenteeism to work. They analyzed the UCI dataset using the Naive Bayes, Decision Tree and Multilayer Perceptron and concluded that they offer better results when compared to the other algorithms.

The authors of [12] proposed GBA – a Gradient boosting algorithm to deal with plenty of data and to make a prediction with high prediction power. The algorithm is used for prediction of status and investigated its properties in the Student criteria of a sample taken from Bschool during the admission period.

In [13] the research proposed a methodology to improve the clustering accuracy using feature extraction model. Here the authors used K-means clustering to improve the clustering accuracy on Document term matrix (DTM). This proposed methodology used in this research work also uses the K-means algorithm.

In [14], the researcher used a combination of data mining methods to expand the accuracy rate to recognise the content of benzene in air. The combination of random forest and J48 are used in this research work.

III. PROPOSED METHODOLOGY

In this research work clustering is used to predict the absenteeism in work. The proposed architecture is shown in the figure below. The architecture consists of two phases namely the training phase and the testing phase. In the training phase initially the decision clusters are formed and the rules for the concerned clusters are generated. The model for prediction is formed in the training phase. The predictions are drawn using the model obtained in testing phase.

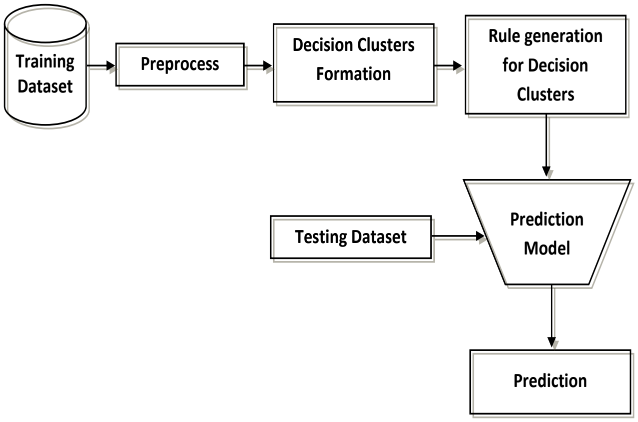


Fig. 1. Proposed Architecture

A. Training Phase

During the training phase, the given dataset is preprocessed to remove the unknown and missing values. The preprocessed dataset is spilt into categories to form the Decision clusters. The process of splitting the dataset into categories is done by considering the attribute absenteeism in hours. The initial clustering will result in only two decision clusters, not absent and absent. The absent dataset can be further grouped using K-Means clustering [5].

The factors affecting the absenteeism in the decision clusters vary with respect to each other. Predications will be more accurate if the rules are generated by considering this observation. So, rules are generated for each decision clusters. The predication model is then formed based on the knowledge gained.

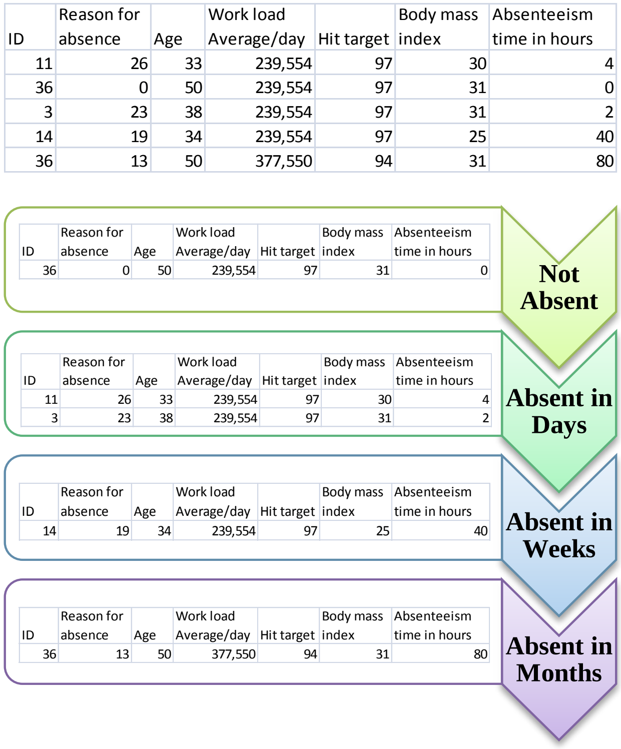
B. Testing Phase

In the testing phase, the given input is checked against the decision clusters. Once the clusters are identified the rules for the corresponding cluster is applied to predict the result. Here not only the possibility of absent is predicted but also the time span of absenteeism is also predicted using the proposed method.

IV. EXPERIMENTAL STUDY

The Absenteeism data set [11] is downloaded from the UCI data repository for the proposed work. The database was created with the records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. This dataset consists of 21 attributes and 740 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. The reason for absence is expressed as International Code of Diseases (ICD) code.

In the training phase the decision clusters are formed. As a result four decision clusters are formed namely not absent, absent in days, absent in weeks and absent in months. The cluster forming is depicted in the following figure 2. Based on the clusters, rules are generated.



ID	Reason for absence	Age	Work load Average/day	Hit target	Body mass index	Absenteeism time in hours
11	26	33	239,554	97	30	4
36	0	50	239,554	97	31	0
3	23	38	239,554	97	31	2
14	19	34	239,554	97	25	40
36	13	50	377,550	94	31	80

ID	Reason for absence	Age	Work load Average/day	Hit target	Body mass index	Absenteeism time in hours
36	0	50	239,554	97	31	0

ID	Reason for absence	Age	Work load Average/day	Hit target	Body mass index	Absenteeism time in hours
11	26	33	239,554	97	30	4
3	23	38	239,554	97	31	2

ID	Reason for absence	Age	Work load Average/day	Hit target	Body mass index	Absenteeism time in hours
14	19	34	239,554	97	25	40

ID	Reason for absence	Age	Work load Average/day	Hit target	Body mass index	Absenteeism time in hours
36	13	50	377,550	94	31	80

Fig. 2. Spliting process of dataset

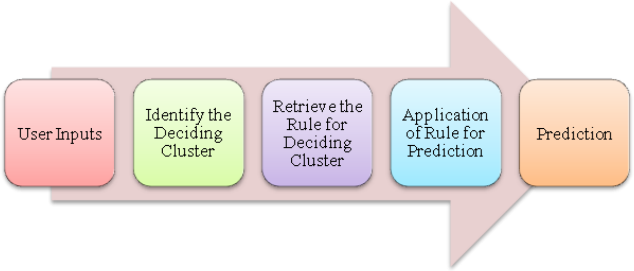

User Inputs
Identify the Deciding Cluster
Retrieve the Rule for Deciding Cluster
Application of Rule for Prediction
Prediction

Fig. 3. Workflow of Prediction Process

The work flow of the prediction process is depicted in the following figure 3. Since the attributes does not contain null values, the preprocessing stage is not needed for this dataset.

After the clustering, rule generation and prediction, the performance of the proposed method need to be analyzed. The results of the proposed methodology are analyzed using confusion matrix. The confusion matrix is shown in the following figure 4.

		Classifier Predication	
		Positive	Negative
Actual Value	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Fig. 4. Confusion Matrix

The metrics used to analyze the performance of the proposed system are accuracy, misclassification rate, precision, recall and F1 score. The formula for calculating the metrics are given below.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Misclassification\ Rate = \frac{(FP+FN)}{(TP+TN+FP+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

$$F1\ Score = \frac{2}{(\frac{1}{Recall}+\frac{1}{Precision})}$$

The performance of prediction algorithms are measured in terms of the metrics mentioned above. The algorithms is said to offer better prediction if the accuracy is higher and misclassification rate is lower. The proposed method is compared with some of the existing prediction algorithms such as Random Forest, Naive Bayes and Decision Tree algorithms and the results are tabulated in Table I.

TABLE I. PERFORMANCE ANALYSIS

Metric	Random Forest	Naïve Bayes	Decision Tree	Proposed Work
Accuracy	85.405	81.081	91.216	95.676
Misclassification Rate	14.595	18.919	8.784	4.324
Precision	0.852	0.809	0.911	0.966
Recall	0.992	0.988	0.995	0.988
F1 Score	0.917	0.889	0.951	0.977

The misclassification rate, precision, recall and F1 Score of the existing and proposed methods are shown in the following figures 5 to 9. The performance analysis chart in Fig. 5

shows that the proposed method is having better accuracy (95.676%) when compared to the others. Similarly the misclassification rate in Fig. 6 of the proposed method is also very low (4.324%) than the other existing methods.

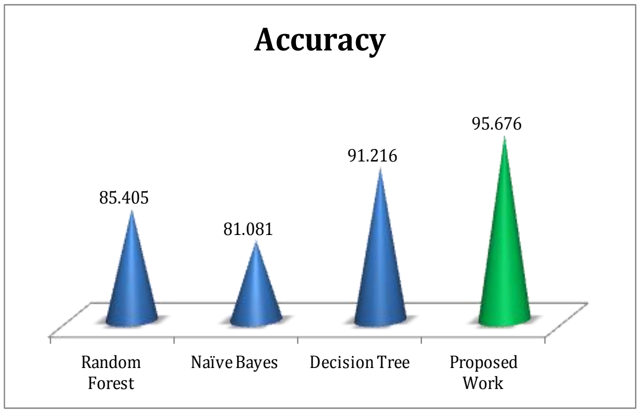


Fig. 5. Performance Analysis in terms of Accuracy

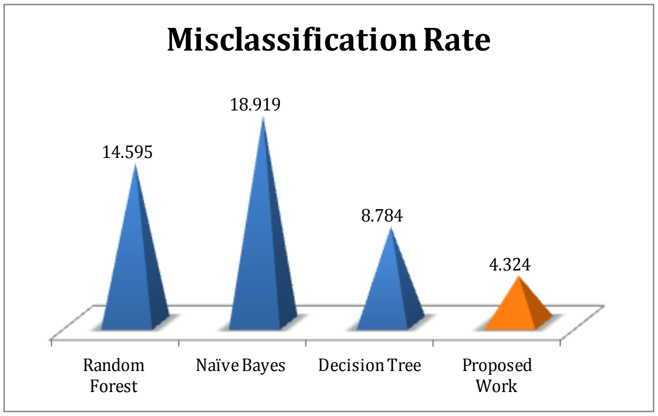


Fig. 6. Performance Analysis in terms of Misclassification Rate

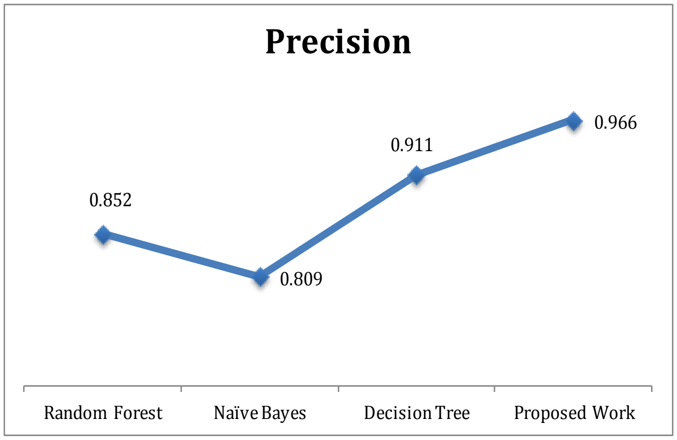


Fig. 7. Performance Analysis in terms of Precision

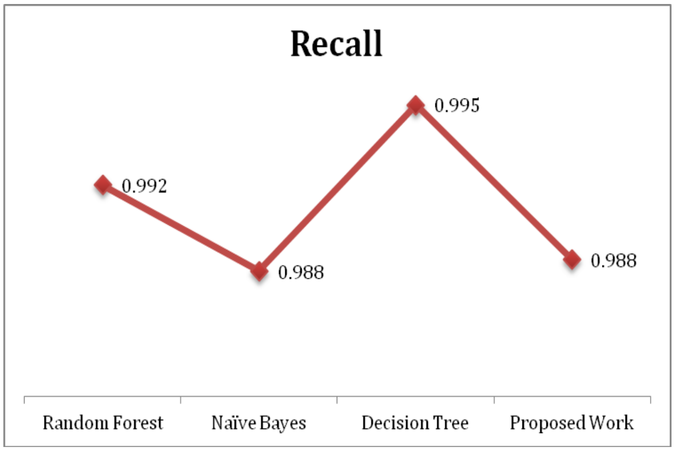


Fig. 8. Performance Analysis in terms of Recall

Fig. 9.

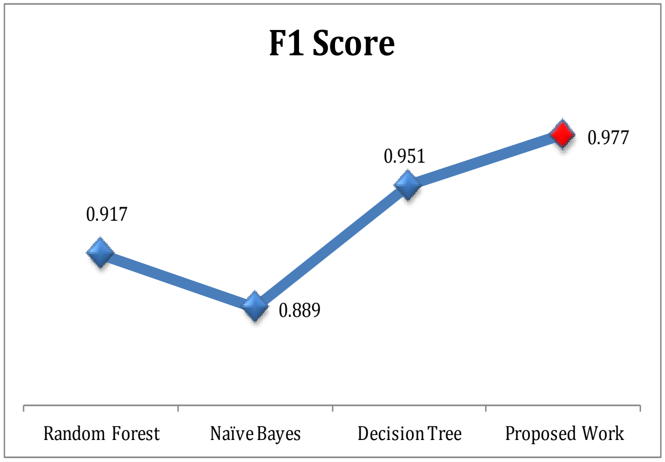


Fig. 10. Performance Analysis in terms of F1 Score

From the above figures, it is explicitly visible that, the proposed method is having very good performance. This work can also be implemented for similar domains. The knowledge about decision cluster categorization need to be varied according to the dataset being used.

V. CONCLUSION

This research paper has proposed a methodology to predict the absenteeism at work using K-means clustering algorithm. The proposed system not only predicts the absenteeism but also identifies the category of absenteeism i.e. whether the employee belongs to any one of the groups namely, not Absent, Absent in days, absent in months or absent in years. Based upon the results, it is possible to take preliminary steps to address the absenteeism issue. The proposed method shows better results when compared to the existing methods such as Naive Bayes, Decision Tree and Random forest. Thus the proposed "Enhancing the prediction of absenteeism by Decision Cluster based Rule generation" can be applied to similar domains.

.

REFERENCES

[1] A K Jain, M N Murty, P J Flynn, "Data Clustering : A Review", ACM Computing Surveys (CSUR) Journal, Volume 31 Issue 3, Pages 264-323, Sept. 1999.

[2] Richard C. Dubes and Anil K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.

[3] Gasparetti, F, "Modeling user interests from web browsing activities", Data Mining and Knowledge Discovery, Springer, Volume 31, Issue 2, pp 502–547, March 2017.

[4] Gayathri.T, "Data mining of Absentee data to increase productivity", International Journal of Engineering and Techniques - Volume 4 Issue 3, pp. 478- 480, ISSN: 2395-1303, , May 2018.

[5] Shivangi Bhardwaj, "Data Mining Clustering Techniques - A Review", International Journal of Computer Science and Mobile Computing, Vol.6 Issue.5, pg. 183-186, ISSN 2320–088X, May- 2017.

[6] Ricardo Pinto Ferreira et al., "Artificial Neural Network And Their Application In The Prediction Of Absenteeism At Work", International Journal of Recent Scientific Research, Vol. 9, Issue, 1(G), pp. 23332-23334, January, 2018.

[7] Saroj et al, "Study on Various Clustering Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , pp. 3031-3033, ISSN : 0975-9646, 2015.

[8] Cortez, Paulo & Morais, A. " A Data Mining Approach to Predict Forest Fires using Meteorological Data", 2007

[9] Gopinath Ganapathy and K.Arunesh, "Models for Recommender Systems in Web Usage Mining Based on User Ratings", Proceedings of the World Congress on Engineering 2011 Vol I, July 6 - 8, ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online), 2011.

[10] Pragati Shrivastava, Hitesh Gupta, “A Review of Density-Based clustering in Spatial Data,” *IJACR*, vol. 2, pp. 200-202, September 2012.

[11] Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C., “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, 2012.

[12] A.Deepa , E. Chandra Blessie, “Input Analysis for Accreditation Prediction in Higher Education Sector by Using Gradient Boosting Algorithm”, Int. J. Sci. Res. in Network Security and Communication, Vol.6(3), pp. 23-27, E-ISSN: 2321-3256, Jun 2018.

[13] T.SenthilSelvi , R.Parimala, “Improving Clustering Accuracy using Feature Extraction Method”, Int. J. Sci. Res. in Computer Science and Engineering, Vol-6(2), pp. 15-19 , E-ISSN: 2320-7639, April 2018.

[14] Gagandeep Kaur , Harmanpreet Kaur, “Ensemble based J48 and random forest based C6H6 air pollution detection”, Int. J. Sci. Res. in Computer Science and Engineering, Vol-6(2), pp 41-50, E-ISSN: 2320-7639, April 2018.

AUTHORS PROFILE

Dr. S. Adaekalavan, is Currently working as Assistant Professor of Computer Science, JJ College of Arts and Science (Autonomous), Pudukkottai. He got his Ph.D. degree from Periyar University, Salem. He has published more than 10 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it’s also available online. His main research work focuses on clustering Network Security, Data Mining, IoT and Computational Intelligence based education. He has 13 years of teaching experience and 5 years of Research Experience.