Data Mining Techniques In HR Analytics: A Review Of Domain Specific Concepts And Technicalities

Banajit Changkakati, Chayanika Das

Abstract— The review tries to deal with the usage, scope and nature of the various data mining techniques for better analysis and prediction of HR functionalities in an organization. The paper here describes with at least 30 papers mostly based on sample outside India. The author here tries to bring out few quality papers, all accumulated from known databases for review and give a vivid picture of the techniques that can be used for analyzing domain specific data. These techniques might also help the future researches to effectively predict various HR functionalities specific to region and industry.

Key words — HR analytics, Data mining, Prediction, Human Resource Management.

1 INTRODUCTION

Human Resource is one of the core competency and competitive advantage of a particular organisation. Therefore, dealing and maintaining them with utmost care is important because this capital is both qualitative as well quantitative nature. So, only stringent strategies are not sufficient, they are to be dealt with both emotion, tactics, techniques and science. In various papers, it is seen that the management perspective from the researchers end is missing or lagging behind as only technical modeling is done by the researchers on the data provided. The researches are solely on the basis of either demographic factors or domain experts knowledge of the particular organisation or industry. As management academicians few new factors can be incorporated as suggested in the models established. In recent years, a quickly growing number of research contributions aim at supporting the practical adoption of HRM data mining[Strohmeier and Piazza]. An increasing number of data mining researches have come out in the last few years. In this paper the researcher tries to compile various articles and papers that have been published along times and thus tries to find out the most accurate method for predicting various HR functions from recruiting and selection, training and development, performance management and retention of the employees in various organisations spread over various arena or industry. The researcher also tries to analyse the most accurate and user friendly techniques which the vendors can use and incorporate the model to be designed to develop commercial products to help the organisations. The paper would thus try to put a systematic angle for suggesting future scope.

2 METHOD AND FRAMEWORK

The research articles have been collected from search engines like scholar.google.com and also from online research databases (Business Source Premier, Scopus, and Science Direct). HRM data mining refers to an intersection of method and domain, respective pairs of search terms such as “data mining” and “HRM” were employed. Beyond synonyms such as “Knowledge discovery database” and “talent management”, multiple HRM sub-domains such as “retention”, “performance management”, “recruiting”, “compensation”, method-categories such as “decision trees”, “cluster analysis” etc. were used as search terms. The search was however restricted to only English language publications.

2.1 The Systematic Review of Research Papers

Chien and Chen (2008) in their paper describes about data mining techniques as discovery driven rather than assumption driven. They have opined that these techniques used in HR related areas are very rare. They proposed decision tree analysis, clustering, association techniques. In details, they proposed various algorithms for decision tree like CART, CHAID, ID3, and C4.5 and compared them. From the study, using the CHAID algorithm in the sample organisation, they were able to design various rules which in turn helped the organisation to design strategies to decrease the turnover of their employees and help the organisation. They also said that with use of more factors in hand and usage of neural network can yield better results. Yee and Chen(2009) used multi factorial evaluation model which is an application of fuzzy theory to design a performance appraisal system in a IT and Telecommunication organisation of Malaysia which in turn can segregate the poor, average and top score of performers. The architecture tends to use various aspects of performance measurement as given by the HR section too calculate an overall performance score which in turn helped in categorising and ranking the employees based on the performance. This model is supposed to be applicable for other organisations as well only by altering the aspects as relevant to the organisation and the weightages assigned to the aspects. The researcher believes that “This model follows a systematic step in determining a staff’s performance, and therefore, it creates a system of appraisal which is able to consistently produce reliable and valid results for the appraisal process. In order to
allow other companies to use this system, the aspect to be evaluated and the weightage for each of these aspects need to be define in the system beforehand.”

Jantan et al (2009) describes an architecture for talent forecasting by compiling various factors and attributes from various reviews done. They also suggested ANN, Decision trees, rough set theory as well as support vector machines techniques to be the most efficient in sequence of accuracy.

Li et al (2009) describes selection model using SVM. They proposed a weight driven SVM for the same with an accuracy rate of around 79% and believes that it provided a higher accuracy than traditional SVM model.

Kaur and Agarwal (2010) used artificial neural network and proved that prediction of company’s success and failure is very much dependent on overall HR factors. Though they used a limited set of HR factors for telecom sector in India, the accuracy rate to predict failure for the organisation was 99.99%.

Al-Radaideh and Nagi (2012) uses various data mining techniques like to predict performance of employees of basically using decision tree algorithms. the generated model was then validated by experimenting it with other organisations. They also opined that Decision tree classifiers are popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. Moreover they are easily interpreted and understood. They could also find out the strongest as well as the weakest attribute for measuring employee performance.

Fan et al illustrate (2012) the use of data mining techniques in particular machine learning clustering analysis to predict turnover of technological professionals. Here they used self-organising Map (SOM) combined with back propagation neural network to do the cluster analysis. After the analysis, the employees data were clustered into four categories ranging from low, second low, medium to high turnover tendency categories. However, they found that accuracy rate yielded was 92%whereas K means and only BPN yielded 63 and 83.5% accuracy respectively. They suggested that in such cases fuzzy algorithms and multi criteria decision making tools may yield better results.

Valle et al(2012) opines that to predict performance score of the executives of call centre using Naïves Bayesian classifier was only 83.6 % accurate. They were also of the view that using psychological attributes might augment the accuracy level or addition of more attributes may increase it.

Ancheta and the co-authors(2012) says that ‘Discovering knowledge based from previous learning makes data mining an efficient and effective tool to predict future occurrences which may affect decisions of current situations.’. they tried to predict the performance of new recruits on the basis of the rule sets designed using rule based classification algorithm.

Strohmeier et al (2013) describe their paper to be the first of its kind of reviews written on domain driven data mining. From the review the researcher reveals that though the method and technology driven researches in the said field have been contributed still the management front is lagging behind while selecting the attributes and factors to build the models by using the methods. They also say that “Despite the domain’s interdisciplinary character, the field is currently dominated by method- and technology-oriented research; managerial research rarely participates”.

Aallo et al (2013) used various decision tree algorithms like C4.5,REPTree,CART, JRip, SeeTREE, Boost SeeTREE with accuracies of 67,62,64,61,74,73,74 percent respectively. these methods were used to predict the employee retention in a higher institution of Nigeria using various demographic characteristics. The researchers designed various rules sets under various algorithms to predict employee turnover with highest accuracy rate for seeTREE and JRip algorithm.

Azar et al (2013) exhibits a case study on commercial bank to develop a selection model for the bank using data mining technique. He used various decision tree algorithms to set the rules with highest accuracy level of around 71 % with CART algorithm. The researcher opines that Cart is the most feasible tree though for better accuracy suggests fuzzy theory.

Kalugina et al (2014) describes an automated process of personnel selection which includes data crawling techniques, data conversion and personnel selection procedure. The personnel selection procedure is based on the mathematical model that includes the comparison of employers and candidates, different aggregation procedures and matching by programming language.

Shah and Ladha (2014) used fuzzy logic to develop a system based on 360 degree performance appraisal system to improve the performance of the students and teachers of academic institutions. The presented model by the researcher is an intelligent system design to identify students’ different skills in education domain. The researchers also say that, “fuzzy logic based performance appraisal allows the decision maker to introduce vagueness, uncertainty, and subjectivity into the evaluation system, which models human like decision making approach”

Shaou and Yousif (2014) uses fuzzy algorithm for designing and implementing performance measurement system among employees. The implementation of the rule sets formulated was then used to build a system using MS Access database.

Tamiheiasi and Umarani(2014) predicts employee turnover using decision tree and SEMMA technology. They advocate about decision tree to be the most flexible modelling technique and can work better with any missing values.they compare neural network and decision tree algorithms but neural network seemed to give higher accuracy rate of 60%.

Rasmussen and Ulrich (2015) shares the various challenges from management perspective about usage of HR analytics in various top ranked organisations like Google, hp etc.

Yuan et al(2016) tried to develop two networks : one on work relationship and the other on social relationship which in turn would show their implication on promotions of the employees. It was found that work-related interactions were more correlated to both promotion and resignation.it was also found out that the employee who can get more attention (higher in-degree) in the work-related network is more probable to be promoted while the employees with low values of k-shell index in the work-related network and in-degree in the social network have high risk to resign. It makes sense that peripheral employees (with low k-shell value) are more likely to leave, because they may feel worthless, under-appreciated and lonely. In turn, the central employees (with high k-shell value) though they get higher probability to be promoted, but the
classification experiments gave away negative results, therefore, it provided a novel and actionable understanding of enterprise management and suggest that even social interplays, can be helpful to reduce staffs’ turnover risk.

Gupta and Sharma (2017) explains the usage of data mining techniques like decision tree and Naive Bayesian classifier on appraisal data for performance prediction of marketers in Steel Marketing sector. Model developed during the study using data mining classification techniques demonstrates the effectiveness of these techniques in performance prediction of an employee.

Zhao et al. (2018) used machine learning algorithms starting from decision tree methods to neural networks to predict employee turnover. They conclude after their experiment that, “If there are more HR datasets available, extreme gradient boosting is recommended to use as the most reliable algorithm. It requires minimal data pre-processing, has decent predictive power, and ranks the feature importance automatically and reliably. However, due to the complexity of employee turnover prediction, one should try to find the classifier that best fits the underlying data before taking this approach.”

Attri (2018) in the project describes why an employee leaves an organisation using data mining techniques. The researcher developed four models on the attributes given by domain experts. The models were designed using random forest tree, logistic regression, SVM, Gradient boosting machine with Bayesian optimisation. Random forest tree giving the highest accuracy rate of about 85%.

Samaila et al. (2018) uses and suggests fuzzy logic model for selection of employees to make the selection biased free and judged. Though they say that further medical examination personally so as to confirm their absence related to health grounds are decreased apart from the interpretation done by the fuzzy score. This study was however validated for IT professionals’ selection in Nigeria.

Khera and Divya (2019) in their paper used machine learning algorithms to predict employee turnover among IT professionals. They tag SVM to be a efficient learned algorithm to design the predictive model for employee turnover. They found the accuracy level to be 85% for the Indian IT professionals retention prediction. They also opined that SVM should be given interest by the researchers as Neural networks to build predictive models.

Karande et al. (2019) uses diverse machine learning algorithms on dataset to predict the employee turnover. Based on the performance of the individual classification, voting is taken, using which the final classification is done. The researcher uses Multi-layer neural networks, SVM, voting and logistic regression.

2.2 Discussion of articles

Faletta (2013) in the article on HR analytics says that 51% of the global companies of fortune list believes that HR analytics provides a great input for implementation and formulation of HR strategies in their companies. The article summarizes the results of The HR Analytics Project conducted by the Organizational Intelligence Institute and Drexel University. In the article published in AIHR, predictive analytics has been of a great importance for the HR functions.

to Deloitte’s 2018 People Analytics Maturity Model, only 17% of organizations worldwide had accessible and utilized HR data. This is up from 8% in 2015, and 4% in 2014. Of this 17% in 2018, only 2% qualified as having business-integrated data, meaning they use real-time, advanced AI-aided tools to collect, integrate, and analyse data. The other 15% is able to do predictive analytics on an ad-hoc basis. As published, these predictive analytics is used by big organisation like hp, Google, Facebook, Neilson etc.

TABLE 1

<table>
<thead>
<tr>
<th>DATA MINING TECHNIQUE USED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
</tr>
<tr>
<td>Classification using Al-Radaideh and Nagi, Azar Et Al, Gupta and Sharma,</td>
</tr>
<tr>
<td>C4.5, C5.0, ID3, Random Forest tree</td>
</tr>
<tr>
<td>Clustering</td>
</tr>
<tr>
<td>Neural networks</td>
</tr>
<tr>
<td>Fuzzy theory</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

Usability and accuracy can be depicted by the following analysis:
It can be thus analysed that though the accuracy rates have been high for neural network while working in domain of HR data, yet the usability is typically lesser than fuzzy theory and decision trees. By closely introspecting the analysis the usability of fuzzy algorithm has been effectively used by the researcher to mine the data because of its accuracy rates as well as computational and technical advantages as discussed in some of the papers.

2.3 Existing Vendors dealing with HR Analytics

Pang et al (2019) have discussed the top ten vendors dealing with HR analytics software and application and their clients worldwide.

From the sources, it was also found that analytics in HR is still at the nascent stage than other domains so researches to build up the functionalities using analytics could make work of HR department easier to manage the Human capital in the organisations.

3 CONCLUSION

The intention of the author here is to put forward various insights about the data mining techniques and how they can be used to analyse datasets and effective predictions for HR specific domain and sub domain. The analysis provided with all the information now provides area for future researcher to work in this rarely touched domain. This would also in turn would enable the researchers to use their model and build softwares for the market to help the organisations. The reviews bring a systematic outline for any researcher who tends to work in this field and in which area they can take forward their research.

REFERENCES


[22] https://www.analyticsinhhr.com/blog/predictive-analytics-human-resources/