JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.7, Special Issue.2, Jan 2019 E-ISSN: 2347-2693

HR Management Using Big Data Analytics

S. Chitra[1*], P. Srivaramangai[2]

[1,2]Dept of Computer Science, Marudupandiyar College, Thanjavur, India

[*]*Corresponding Author: chitrasathish1979@gmail.com Tel.: +00-12345-54321*

Available online at: www.ijcseonline.org

Abstract— In any organization's talent management is becoming an increasingly crucial method of approaching HR functions. Talent management can be defined as an outcome to ensure the right person is in the right job. Human talent prediction is the objective of this study. Due to that reason, classification and prediction in data mining which is commonly used in many areas can also be implemented in this study. There are various classification techniques in data mining such as Decision tree, Neural networks, Genetic algorithms, Support vector machines, Rough set theory, Fuzzy set approach. This research has been made by applying decision tree classification algorithms to the employee's performance prediction. Decision tree is among the popular classification technique which generates a tree and a set of rules, representing the model of different classes, from a given data set. Some of the decision tree algorithms are ID3, C5.0, Bagging, Random Forest, Rotation forest, CART and CHAID. In this paper give the overview of C4.5 algorithms.

Keywords— HR Analytics, Talent, Prediction, Decision Tree, Algorithm, C4.5, Classification, Data Mining, Big Data**.**

I. INTRODUCTION

Data is the collection of values and variables related in some sense and differing in some other sense. In recent years the sizes of databases have increased rapidly. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data [1]. Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated. Data mining is the process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories. A widely accepted formal definition of data mining is given subsequently. According to this definition, data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data [2]. Data mining uncovers interesting patterns and relationships hidden in a large volume of raw data. Big Data is a new term used to identify the datasets that are of large size and have grater complexity [3]. So we cannot store, manage and analyze them with our current methodologies or data mining software tools.

Big Data mining is the capability of extracting useful information from these large datasets or streams of data which were not possible before due to its volume, variety, and velocity. The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge. Data Mining is analyzing the data from different perspectives and summarizing it into useful information that can be used for business solutions and predicting the future trends. Mining the information helps organizations to make knowledge driven decisions. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules [4]. It applies many computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining extract only required patterns from the database in a short time span. Based on the type of patterns to be mined, data mining tasks can be classified into summarization, classification, clustering, association and trends analysis [5].

Big Data is now part of every sector and function of the global economy [6]. Big Data a collection of datasets is so large and complex that is beyond the ability of typical database software tools to capture, store, manage and process the data within a tolerable elapsed time. Big Data is unstructured data that exceeds the processing complexity of conventional database systems. The data is too big, moves too fast, or doesn't fit the rule restricting behavior of our database architectures. This information comes from multiple, distinct, independent sources with complex and evolving relationships in a Big Data which is keep on growing day by day. There are three main challenges in Big

Data which are data accessing and arithmetic computing procedures, semantics and domain knowledge for different Big Data applications and the difficulties raised by Big Data volumes, distributed data distribution and by complex and dynamic characteristics.

Nowadays, there are many areas, such as in finance, medical, marketing, stock market, telecommunication, manufacturing, health care and customer relationship, which have adapted data mining techniques [7, 8]. However, the used of classification techniques in data mining approach does not attract much attention among people in Human Resource (HR) field [9]. HR data provides a rich resource for knowledge discovery for decision support system development. In addition, today's organization has to struggle effectively in terms of cost, quality, service or innovation. The success of these tasks depends on having enough right people with the right skills, deployed in the appropriate locations at the appropriate point of time which is known as talent management. Managing an organization talent has become one of the challenges to the HR professionals. This task involves a lot of managerial decisions in order to decide the right person for the right job at the right time. Sometimes, these types of decisions are very uncertain and difficult; and it depends on various factors such as human experience, knowledge, preference and judgment. The process to identify an existing talent in organization is among the top talent management challenges and becomes a never ending issue [10].

II. BIG DATA IN HR

Big data refers to massive and exponentially growing amounts of employee, customer, and transactional data available in organizations. In the case of HR, organizations have huge amounts of talent or people-related data (e.g., skills, performance ratings, age, tenure, safety record, sales performance, educational background, manager, prior roles, etc.) which can be used to better understand the organization's current composition, performance, and risk to improve the development of employees, products, and services. Big Data in HR sets to evaluate and improve practices including talent acquisition, development, retention, and overall organizational performance [11]. This involves integrating and analyzing internal metrics, external benchmarks, social media data, and government data to deliver a more informed solution to the business problem facing your organization.

New tools and technology is needed because big data is so big, fast changing and potentially unstructured. With these tools, HR organizations are able to perform analytics and forecasting to make smarter and more accurate decisions, better measure efficiencies and identify management "blind spots" to answer important questions regarding workforce productivity, the impact of training programs on enterprise performance, predictors of workforce attrition, and how to identify potential leaders. The ability to capture and analyze big data has enabled many companies to both increase revenues by better understanding and more accurately targeting customers and cut costs through improved business processes. The biggest problem for HR people managing talent has been a lack of numbers and a lack of data to put on the table in business discussions. With our people analytics offerings, Cornerstone Insights and Cornerstone Planning, we removed the barriers for organizations who want to answer important talent management questions but aren't quite sure where to start. We are helping them to take the power of big data from insights to action.

HR ANLAYTICS

HR analytics, also called talent or people analytics, is the application of considerable data mining and business analytics techniques to talent data. Analytics that measure performance and efficiencies that matter to HR only [12]. Examples include: time to fill a job requisition, number of people trained, number of people with certain competencies, last year's attrition, estimated attrition for next year, estimated number of candidates to have in the pipeline based on estimated attrition, which source provides the best candidates, compliance reporting, diversity reporting. The goal of human resources analytics is to provide an organization with insights for effectively managing employees so that business goals can be reached quickly and efficiently. HR analytics does not only deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve the processes. The challenge of human resources analytics is to identify what data should be captured and how to use the data to model and predict capabilities so the organization gets an optimal return on investment on its human capital. Cornerstone Analytics offers a suite of people analytics products that apply sophisticated data science and the most refined machine learning system for talent management to help organizations harness the power of real-time talent data and more efficiently and effectively manage their people. The Cornerstone Analytics suite includes, Cornerstone Reporting, Cornerstone View, Cornerstone Insights and Cornerstone Planning.

III. DECISION TREE IN BIG DATA

Decision Tree [DT] is ideal to use as the filter to handle the large amount of data. DT is a basic way of classification can have satisfactory efficiency and accuracy of those datasets. Decision Tree algorithm is good at tuning between precision which can be trained very fast and provide sound results on those classification data [13].

Big Data are now rapidly expanding in all domains with the fast development of networking and increase in the data storage and collection capacity. The instances are divided into a set of discrete valued set of properties, known as various features of the data. For example, classifying a received email as “spam” or “not spam” could be based on analyzing characteristics of the email such as origin IP address, the number of emails received from the same origin, the subject line, the email address itself, the content of the body of the email, etc. All these features will contribute to a final value which will allow the algorithm to classify the email. It is logical that the more number of examples of spam and non-spam emails the Machine Learning system goes through, the better will be its prediction for the next unknown email.

Decision Tree learning is reasonably fast and accurate. The approach is to learn on large data sets is to parallelize the process of learning by utilizing Decision Trees [14]. It is straightforward to reduce a Decision Tree to rules. The strategy follow here is to break a large data set into n partitions then learn a DT on each of the n partitions in parallel. A DT becomes bigger on each of n processors independently. After that they must be combined in such a way that the Decision Tree remains individual tree, for this approach Decision Tree can used Meta-learning. Meta-learning is the process by which learners become aware of and increasingly in control of habits of perception, inquiry, learning, and growth.

Now other aspect of creating final DT is pruning the tree which removes the nodes that do not provides accuracy in classification results in reduced size tree. Pruning is likely to be very important for large training set which will produce large trees. There are a number of methods to prune a Decision Tree. In C4.5 an approach called pessimistic pruning is quite fast and has been shown to provide trees that perform adequately.

Big Data challenges are growing day by day; traditional Decision Tree algorithms have multiple limitations. First, building a Decision Tree is a very time consuming when the available dataset is extremely huge. Second, although parallel computing clusters can be leveraged in Decision Tree based classification algorithms, the strategy of data distribution should be optimized so that required data for building one node is localized and meanwhile the communication cost is minimized. To overcome these limitations, distributed C4.5 algorithm is used. When available dataset is extremely huge then C4.5 algorithm performs well in short time and it is robust in nature as well as simple to understand [15].

As figure 1 shows, it’s a very simple decision tree. A, B, C represents different attributes separately in one data set. And each branch like a1,a2,b1,b2,c1,c2 represents the value of split attribute. Leaf node 1,2,3,4 represents the class of decision attribute in each sample set.

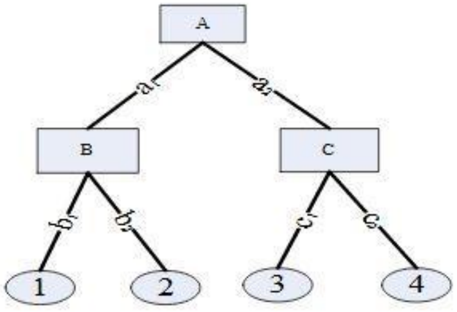


Figure 1: Decision tree structure

There are mainly two steps in decision tree, build a decision tree and do pruning. The thought of building decision tree is called CLS. We have a data set S, the attributes set is A, decision attributes set is D, the whole process is as follow:
(1) Make S be the root node, if all data in S belongs to the same class, turn node to leaf node.
(2) Otherwise choose one attribute a∈A and divide nodes according to different values of attribute a. S has the number of m lower layer nodes, branches represent the situation of different values of a.
(3) Induct step 1 and step 2 for m branch nodes.
(4) If attributes in one node belong to same class or there is no node to divide, then stop.

The most two important things in decision tree are: 1 how to decide best split node? 2 when to stop splitting? Because real data can’t be pure. There must be data attributes miss, data inaccurate, noise these situations, which will result in overfit. Overfit may lower the accuracy of the classification and prediction of decision tree and increase the complexity of tree structure. So after building a tree, we also need to pruning.

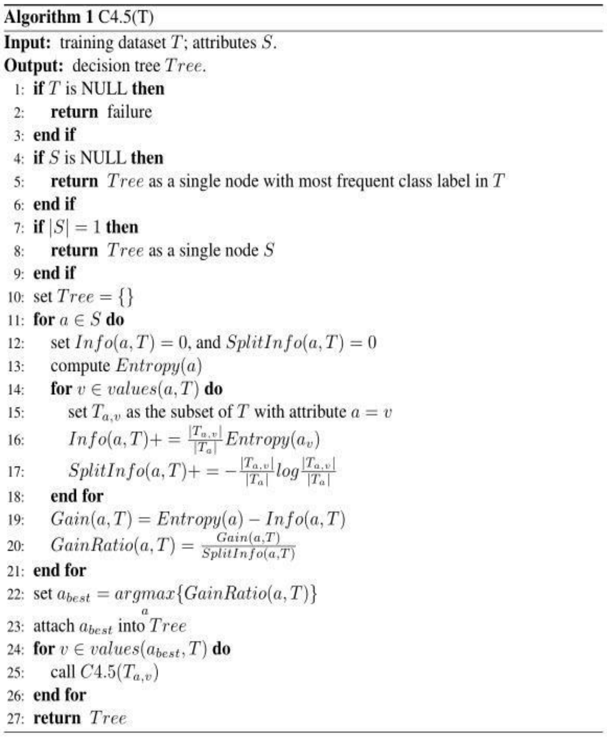
3.1 C4.5 ALGORITHM

The C4.5 technique is one of the decision tree family that can produce both decision tree and rule-sets; and construct a tree for the purpose of improving prediction accuracy [16]. Besides that, C4.5 models are easy to understand as the rules derived have a very straightforward interpretation. The C4.5/C5.0/J48 classifier is among the most popular and powerful decision tree classifier [16]. C5.0 and J48 are the improved version of C4.5 and IDS algorithms. WEKA classifier package has its own version known as J48. J48 is an optimized implementation of C4.5 rev. 8, and C4.5 first grows an initial tree using the divide-and-conquer algorithm.

A decision tree is a classifier which conducts recursive partition over the instance space. A typical decision tree is composed of internal nodes, edges and leaf nodes. Each *internal node* is called *decision node* representing a test on an attribute or a subset of attributes, and each edge is labeled with a specific value or range of value of the input attributes. In this way, internal nodes associated with their edges split the instance space into two or more partitions. Each *leaf node* is a terminal node of the tree with a *class label*.

The general process of building a decision tree is as follows. Given a set of training data, apply a measurement function onto all attributes to find a best splitting attribute. Once the splitting attribute is determined, the instance space is partitioned into several parts. Within each partition, if all training instances belong to one single class, the algorithm terminates. Otherwise, the splitting process will be recursively performed until the whole partition is assigned to the same class. Once a decision tree is built, classification rules can be easily generated, which can be used for classification of new instances with unknown class labels.

C4.5 [17] is a standard algorithm for inducing classification rules in the form of decision tree. As an extension of ID3 [17], the default criteria of choosing splitting attributes in C4.5 is *information gain ratio*. Instead of using information gain as that in ID3, information gain ratio avoids the bias of selecting attributes with many values.



```
Algorithm 1 C4.5(T)
Input: training dataset $T$; attributes $S$.
Output: decision tree $Tree$.
if $T$ is NULL then
    return failure
end if
if $S$ is NULL then
    return $Tree$ as a single node with most frequent class label in $T$
end if
if $|S| = 1$ then
    return $Tree$ as a single node $S$
end if
set $Tree = \{\}$
for $a \in S$ do
    set $Info(a,T) = 0$, and $SplitInfo(a,T) = 0$
    compute $Entropy(a)$
    for $v \in values(a,T)$ do
        set $T_{a,v}$ as the subset of $T$ with attribute $a = v$
        $Info(a,T) += \frac{|T_{a,v}|}{|T_a|} Entropy(a_v)$
        $SplitInfo(a,T) += -\frac{|T_{a,v}|}{|T_a|} log \frac{|T_{a,v}|}{|T_a|}$
    end for
    $Gain(a,T) = Entropy(a) - Info(a,T)$
    $GainRatio(a,T) = \frac{Gain(a,T)}{SplitInfo(a,T)}$
end for
set $a_{best} = \underset{a}{argmax}\{GainRatio(a,T)\}$
attach $a_{best}$ into $Tree$
for $v \in values(a_{best},T)$ do
    call $C4.5(T_{a,v})$
end for
return $Tree$
```

Figure 2: C4.5 algorithm

A. Construction

Some premises guide this algorithm, such as the following [18]:

- if all cases are of the same class, the tree is a leaf and so the leaf is returned labeled with this class;
- For each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Also calculate the gain in information that would result from a test on the attribute (based on the probabilities of each case with a particular value for the attribute being of a particular class);
- Depending on the current selection criterion, find the best attribute to branch on.

B. Counting gain

This process uses the "Entropy", i.e. a measure of the disorder of the data. The Entropy of$\vec{y}$ is calculated by

$$\textbf{Entropy}\ (\vec{y}) = -\sum_{j=1}^{n} \frac{y_j \vee}{\vec{y}\vee} \textbf{ log } \frac{\vec{y}\vee}{y_j \vee -}$$

iterating over all possible values of $\vec{y}$. The conditional Entropy is

$$\textbf{Entropy (j | } \vec{y}\textbf{) = } \frac{\vec{y}\vee}{y_j \vee -} \textbf{ log } \frac{\vec{y}\vee}{y_j \vee -}$$

and finally, we define Gain by

$$\textbf{Gain (}\vec{y}\textbf{), j) = Entropy (}\vec{y}\textbf{)} - \textbf{Entropy(j |}\vec{y}\textbf{)))}$$

The aim is to maximize the Gain, dividing by over- all entropy due to split argument $\vec{y}$ by value j.
Where splitinfo is,

$$\textbf{Splitinfo (p , test) =}\sum_{j=1}^{n} p`(\frac{j}{p}) \textbf{ X log (p`(}\frac{j}{p}\textbf{))}$$

P' (j/p) is the proportion of elements present at the position p, taking the value of j-th test. Note that, unlike the entropy, the foregoing definition is independent of the distribution of examples inside the different classes.

C. Pruning

This is an important step to the result because of the outliers. All data sets contain a little subset of instances that are not well-defined, and differ from the other ones on its neighborhood. After the complete creation of the tree, that must classify all the instances in the training set, it is pruned. This is to reduce classification errors, caused by specialization in the training set; this is done to make the tree more general.

IV. CONCLUSION

Classification is the one of the hottest topics in the area of data mining. The research activities on this topic is reviewed hence, the study guides the researches to get an idea about the recent advancements with Classification. This paper has described the significance of the study on the use of data mining classification techniques for employees' performance prediction. However, there should be more data mining classification techniques applied to the different problem domains in HR field of research to broaden the horizon of academic and practice work on data mining in HR. In this paper, we studied C4.5 classifier has a great potential for performance prediction.

REFERENCES

[1]. More about "*Big Data"* Online Available From: http://en.wikipedia.org/wiki/Big_data

[2].https://www.youtube.com/watch?v=Pq3OyQOl3E Hilbert & López 2011

[3]. Bright Planet's Blog(2012), *"Structured vs. Unstructured Data",* Online Available from: https://www.brightplanet.com/2012/06/structure d-vs-unstructured-data/

[4]. *A Quick Guide to Structured and Unstructured Data*(2014) Online Available from: http://smartdatacollective.com/michelenemschoff /206391/quick-guide-structured-andunstructured- data

[5]. R. Thoran(2012), "*10 emerging technologies for Big Data"* Online Available from: http://www.techrepublic.com/blog/big-dataanalytics/ 10-emerging-technologies-for-big-data/

[6]. *MapReduce* Online Available from: http://en.wikipedia.org/wiki/MapReduce

[7*]. IBM's report on What is MapReduce* Online Available from: http://www- 01.ibm.com/software /data/infosphere/hadoop/mapreduce/ SAS Report on Hadoop Online Available from: http://www.sas.com/en_us/insights/bigdata/ hadoop.html

[8]. Ranjan, J., *"Data Mining Techniques for better decisions in Human Resource Management Systems"*. International Journal of Business Information Systems, 2008. 3(5): p. 464-481.

[9]. DeNisi, A.S. and R.W. Griffin, "*Human Resource Management*". 2005, New York: Houghton Mifflin Company.

[10]. *A TP Track Research Report Talent Management: "A State of the Art"*. 2005, Tower Perrin HR Services.

[11]. Tso, G.K.F. and K.K.W. Yau, *"Predicting electricity energy comsumption : A comparison of regression analysis, decision tree and nerural networks"*. Energy, 2007. 32: p. 1761 - 1768.

[12]. Han, J. and M. Kamber, *"Data Mining : Concepts and Techniques"*. 2006, San Francisco: Morgan Kaufmann Publisher.

[13]. J. Ranjan, "Data Mining Techniques for better decisions in Human Resource Management Systems,"*International Journal of Business Information Systems,* vol. 3, pp. 464-481, 2008.

[14]. C. F. Chien and L. F. Chen, "*Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry,*"*Expert Systems and Applications,* vol. 34, pp. 380-290, 2008.

[15]. A TP Track Research Report *"Talent Management: A State of the Art,"* Tower Perrin HR Services 2005.

[16]. G. K. F. Tso and K. K. W. Yau, *"Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks,"Energy,* vol. 32, pp. 1761-1768, 2007.

[17]. I. Becerra-Fernandez, S. H. Zanakis, and S. Walczak, *"Knowledge discovery techniques for predicting country investment risk,"Computers & Industrial Engineering,* vol. 43, pp. 787-800, 2002.

[18]. P. R. Kumar and V. Ravi, *"Bankruptcy prediction in banks and firms via statistical and intelligent techniques : A review,"European Journal of Operational Research,* vol. 180, pp. 1-28, 2007.

Authors Profile

Dr.P.Srivaramangai received her Ph.D Degree from Mother Teresa University, Kodaikanal in the year 2012. She received her M.Phil Degree from Manonmaniam University, Tirunelveli in the year 2003. She received his M.C.A Degree from Bharathidasan University, Trichy in the year 1996. She is working as Associate-Professor, PG and Research Department of Computer Science, Marudupandiyar College of Arts & Science, Thanjavur, Tamilnadu, India. She has above 30 years of experience in academic field. She published 25 papers in National & International journals so far. Her areas of interest include Computer Networks, Internet of Thing, Grid Computing, Cloud Computing and Mobile Computing.

Ms S.Chithra is pursuing her Ph.D from Marudupandiyar College, Thanjavur, affiliated to Bharathidasan University, Trichirapalli in part time and she is working as an Asst.Prof in the Dept. Of computer Sciecne at Srimad Andavan College of Arts and Science, Srirangam , Trichy. Her areas of Research and interest includes Data Mining, Data science, Big Data, Predictive Analytics. She has presented and publisher more than five papers in various Ntional and Internal Journals.