IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Deployment of Improved ID3 algorithm with Havrda and Charvat entropy for Employees performance evaluation

[1*]Kirandeep, [2]Neena Madan

[1,2]Dept. of Computer Science and Engineering Guru Nanak Dev University, Regional Campus Jalandhar (Punjab), India

[*]*Corresponding Author: deepkiran527@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 12/Oct/2018, Published: 30/Nov/2018

***Abstract*—**In industrial sector, employee performance evaluation has paramount importance which will be used for predicting the performance of an employee or a number of employees by considering various aspects like employee's task type, job skills, working quality, interpersonal relationship, creativity, adherence to policy, productivity, attendance, performance etc. Data mining consists of different techniques used to complete our objectives. Using this particular research, the industry superiors will have the ability to predict the performance of employees in an organization. The decision tree method is used on the database of an employee or a number of employees in order to analyze an employee data to make a prediction. Data mining is a tool which allows us to manage the data in a superior way.

***Keywords*—** Havrda and Charvat entropy, Improved ID3 algorithm, Information gain

I. INTRODUCTION

Data mining is often known as a knowledge discovery procedure. We use data mining techniques in order to identify the patterns and relationships among the various types of data. Data mining consists of combinations of machine learning and visualization techniques for identification of the patterns. Data mining consist of various techniques which can be used to complete a goal. From whole of the data, we are required to find the data which will give more information and which will help to predict the performance. It is used for the analysis of data from various perspectives and then summarization of the useful information which can be used to increase revenues or to cut the costs. In simple terms, we can say that the process of data mining is to find meaningful information from the data [4]. Data mining is widely used in the area of telecommunication industry, education, business section, medical, fraud detection, finance and marketing.
It is supported by:

- Availability of data and data mining techniques
- Affordable processing power
- Inexpensive data storage
- Ability to generate useful information

II. THE IMPROVED ID3 ALGORITHM

As a measure of the effectiveness of an attribute in the classification of the training data, the so-called information gain was used [2]. Thus to improve the efficiency of decision tree as compared to the ID3 algorithm and to reduce the time of execution, the ImprovedID3 algorithm with our proposed technique of Havrda and charvat entropy and the is used .

Algorithm:

1.Calculate the entropy of every attribute by using the dataset.
2.Split the whole dataset into various subsets using the attribute for which entropy is minimum and highest information gain.
3.Make a decision tree root node containing that attribute.
4.Recurse on the subsets by the use of remaining attributes only.

The improved ID3 algorithm based on information gain has used. So, the problem presented in traditional ID3 problem is overcome. The more optimal tree can be obtained.

Advantages of Improved ID3 Algorithm:

1. It can be used to build a more concise decision tree in a shorter running time than ID3 algorithm.
2. Higher predictive accuracy than that of the ID3 algorithm.
3. No need to search the whole data set to create a tree.
4. It can handle missing values in dataset easily.

Disadvantages of Improved ID3 Algorithm:

1. Accuracy depends on the type of dataset that we are using.
2. Quite unstable because a small change in data can change the results to great extents [3].

© 2018, IJCSE All Rights Reserved

Havrda and Charvat entropy:
The results obtained from ID3 algorithm with Shannon Entropy makes the decision making process time consuming and complex as we have to elaborate all the attributes with complex and lengthy calculations till we get the final decision tree.
Let P = (p1, p2 …pn) be a probability distribution, p denotes the probability mass function of X and α is its inherent parameter.
Then Havrda and Charvat gave the entropy measure by formula shown under

$$h(p) = \frac{1}{1-\alpha}\left(\sum_{i=1}^{n} X_i^{\alpha} - 1\right)$$

This formula calculates Entropy. To avoid deduced solution in decision tree making process, Havrda and Charvat entropy based ImprovedID3 algorithm is proposed which gives good solution in reasonable time with effective and efficient outcomes and such algorithm can give short and fast decision for supply of goods in company [5].

Information Gain: The decision tree is usually built through top down approach. ID3 selects the splitting attribute with the highest information gain and where gain is defined as difference between how much information is needed after the split. It is a statistical property which measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected based on entropy [6].

The motive is to find the feature that best splits the target class into the purest possible children nodes - pure nodes with only one class. This measure of purity is called information. It generally represents the expected amount of information that is required to specify how a new instance of an attribute should be classified.

The formula used for this purpose is:

$$G(D, S) = H(D) - \sum P(Di)H(Di)$$

III. APPLICATION OF DECISION TREES ON PERFORMANCE OF EMPLOYEES

A decision tree is a decision support tool which uses a tree-like graph or model of decisions and their possible consequences, including the outcomes. It is a simplest method to display an algorithm. Decision trees are nowadays used in research, especially for the analysis of decision, and to support identifying a strategy that is most likely to reach a goal [2]. To construct the decision tree, we use following method:

1. Select a variable of training samples as nodes, and then create a branch to every possible value of the variables. In accordance with that the training sample set is divided into several sub-sets.
2. Do the same method to each branch, when the node of all the training samples belongs to the same classification, or no remaining attributes can be used to further divide, or the branch does not have samples, and then stop splitting the node branching and make it a leaf node.

IV. PROBLEM FORMULATION

• In existing systems, ID3 algorithm with information gain and entropy played an important role but the problem was that in ID3 Algorithm, every attribute has the binary value domain (i.e. positive or negative) [7]but in proposed system with Improved ID3 it is also possible that we can have some specific attributes that have multiple valued domain (i.e. high, medium, low, etc.) e.g. we have taken task type as (easy, moderate , heavy).
• ID3 was used only for the balanced datasets (i.e. 50% positive and 50% negative) & this can be used for imbalanced datasets too (i.e. positive classes can be 10% and the remaining negative classes can be 90% or vice versa) e.g. we have taken attributes in different ratios like in case of dataset of 20 instances the ratio is 11:5:4 for easy, moderate and heavy task types respectively.
• Accuracy in results is very low in case of ID3. Therefore decision tree is to be used in order to make a decision by considering an information gain and Havrda & Charvat entropy with Improved ID3 algorithm.

V. OBJECTIVES

1. To collect dataset for applying the ImprovedID3 algorithm
(Dataset is collected according to recent metrics which are used to predict employee performance)
2. To extract employee related variables (attributes) for classification to apply Havrda & Charvat entropy.
(Various attributes or metrics like task type, job skills, working quality, interpersonal relationship, creativity, adherence to policy, attendance, productivity, performance are evaluated).
3.To implement the dataset with ImprovedID3 using Havrda & Charvat entropy
(Implementation of datasets through NetBeans and WEKA is done over here).
4.To analyze the overall performance of the employees
(Analysis of the results after implementation)
5.To check the accuracy of applied algorithm with particular entropy
(Accuracy of the algorithm with Havrda & Charvat entropy is checked)

© 2018, IJCSE All Rights Reserved

6.To compare the results of ImprovedID3 algorithm (using Havrda and Charvat entropy) with the ID3 algorithm (using Shannon entropy)
7.The focus of our research is to enhance the performance of employees by evaluating various aspects.

Section IV contain the architecture and essential steps of ……..., section V explain the …….. methodology with flow chart, Section VI describes results and discussion …...., Section VII contain the recommendation of ……. and Section VIII concludes research work with future directions).

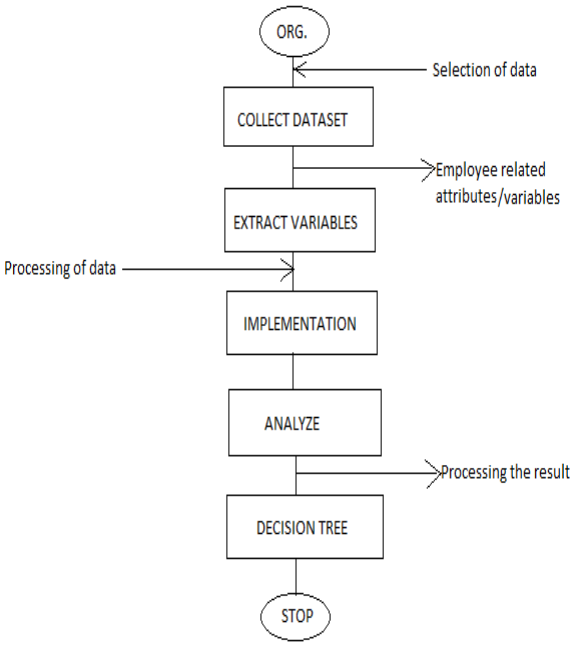

Figure1. Flowchart for the representation of objectives

VI. IMPLEMENTATION

To fulfill our objectives there are number of platforms that can be used to implement various data mining techniques like NetBeans, WEKA, MATLAB etc. In our research we use NetBeans IDE 7.4 over a data set containing 150 instances. The IDE provides comprehensive support for JDK 7 technologies and the most recent Java enhancements. We have applied some decision tree algorithms on our data set and compare their results. NetBeans IDE is a free, open source, integrated development environment (IDE) enables you to develop desktop, mobile and web applications. It supports an application development in different languages like Java, HTML, PHP and C++. The IDE provides integrated support for the complete development cycle, from project creation through debugging, profiling, deployment. The IDE runs on Windows, Linux, Mac OS X, and UNIX-based systems. In our research we have used NetBeans IDE 7.4 version.

Database Used: In our approach for performance prediction we have used 3 data sets having attributes like working task, job skills, creativity, working quality, interpersonal relationship, adherence to policy, productivity, attendance and performance. It is a sample data created by examining various sample datasets available online. The attributes defined in our dataset are according to the current aspects that are being seen nowadays in an employee. In our datasets we have used 20, 50,150 instances respectively to evaluate the performance. The data is prepared in Ms-Excel with CSV (comma separated value) format that is used to transfer large database between programs and to import and export data to office applications.

Attribute	Possible values	Description
Task type	Easy, Moderate, Heavy	The type of task performed by an employee
Job skills	Serious, Common	The practical and technical knowledge required of the job.
Creativity	Ordinary, Perfect	The extent to which an employee proposes improved work methods, suggests ideas to eliminate waste, finds new and better ways of doing things
Working quality	Good, Bad	The quality of work performed by an employee
Interpersonal relationship	Meets Expectations, Improvement Needed	The extent to which employee is willing and demonstrates the ability to cooperate, work and communicate with coworkers, supervisors, subordinates
Adherence to policy	Meets Expectations, Improvement Needed	The extent to which the employee follows company policies, procedures and follows all safety rules and regulations
Productivity	Meets Expectations, Improvement Needed	The extent to which an employee produces a significant volume of work efficiently in a specified period of time
Attendance	Perfect, Imperfect	If an employee is regular at work or not
Performance	Meets Expectations, Improvement Needed.	Employee's overall performance

© 2018, IJCSE All Rights Reserved

VII. RESULTS

WEKA is used as an external library in our research for the purpose of data mining. In this analysis, we have used classification techniques like ID3, Improved ID3 and J48 algorithms on different number of instances and the performance measurement is done on the basis of some performance metrics.
The performance of different number of employees is given as:

1. 20 instances
ID3 – 17 employees meets expectations and 3 needs improvement

Improved ID3 – 18 employees meets expectations and 2 needs improvement

2. 50 instances
ID3 – 47 employees meets expectations and 3 needs improvement

Improved ID3 – 48 employees meets expectations and 2 needs improvement

3. 150 instances
ID3 – 148 employees meets expectations and 2 needs improvement

Improved ID3 – 149 employees meets expectations and 1 needs improvement

TABLE1.1 Comparison of Detailed Accuracy

Algorithm	Instances	TP Rate	FP Rate	Precision	Recall
ID3	20	0.867	0.2	0.929	0.867
	50	0.971	0.125	0.943	0.971
	150	0.981	0	1	0.981
Improved ID3	20	0.933	0.2	0.933	0.933
	50	1	0.125	0.944	1
	150	0.99	0	1	0.99

TABLE 1.2 Comparison of accuracy and errors

Algorithm	Instances	Kappa Statistic	Mean Absolute Error	Root Mean Square Error	Accuracy
ID3	20	0.625	0.15	0.3873	85%
	50	0.8598	0.06	0.2449	94%
	150	0.9687	0.0133	0.1155	98%
Improved ID3	20	0.7333	0.1	0.3162	90%
	50	0.9049	0.04	0.2	96%
	150	0.9842	0.0067	0.0816	99.33%

On the basis of the results summarized in table 1.1 it can be concluded that Improved ID3 with Havrda and Charvat entropy gave the highest true positive rate, lowest false positive rate among different instances. So in accordance with the results of table 1.1, we can say that Improved ID3 algorithm has performed the best for huge datasets and can be useful for performance evaluation of different and large organizations.

On the basis of the results given in table 1.2, Improved ID3 with Havrda and Charvat entropy gives the highest accuracy and the lowest mean absolute error and the lowest root mean squared error when compared to ID3 with different instances. Thus, according to table 1.2, Improved ID3 gives us maximized accuracy and minimum errors that leads to enhanced performance.

Graphical Representation: At the end, the performance analyzed can be represented as

Table 2.1Tabular Comparison of Correctly classified instances

No of Instances	ID3	Improved ID3
[20]	17	18
[50]	47	48
[150]	148	149

The graphical representation of the above table can be given as:

© 2018, IJCSE All Rights Reserved

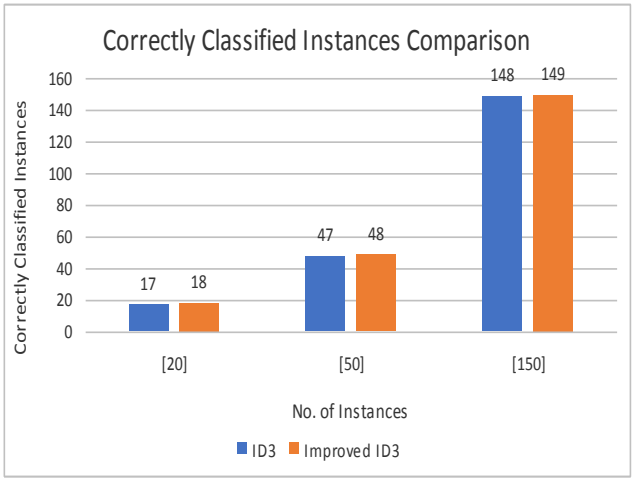

Fig.2.1 Graphical comparison of Correctly classified instances

Table 2.2 Tabular Comparison of Kappa Statistic

No of Instances	ID3	Improved ID3
[20]	0.625	0.7333
[50]	0.8598	0.9081
[150]	0.9687	0.9842

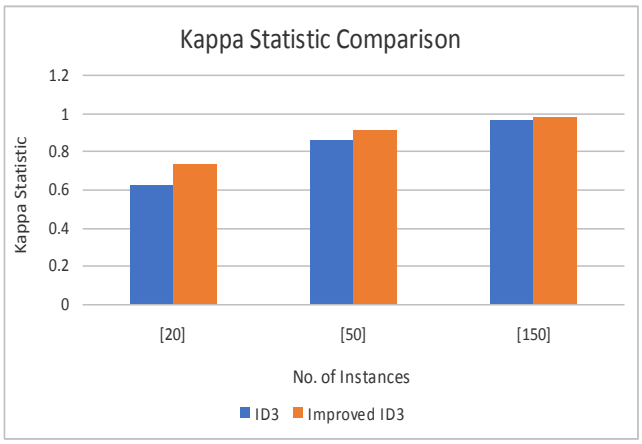

Fig. 2.2 Graphical comparison of Kappa Statistic

Table 2.3 Tabular Comparison of Mean absolute error

No of Instances	ID3	Improved ID3
[20]	0.15	0.1
[50]	0.06	0.04
[150]	0.0133	0.0067

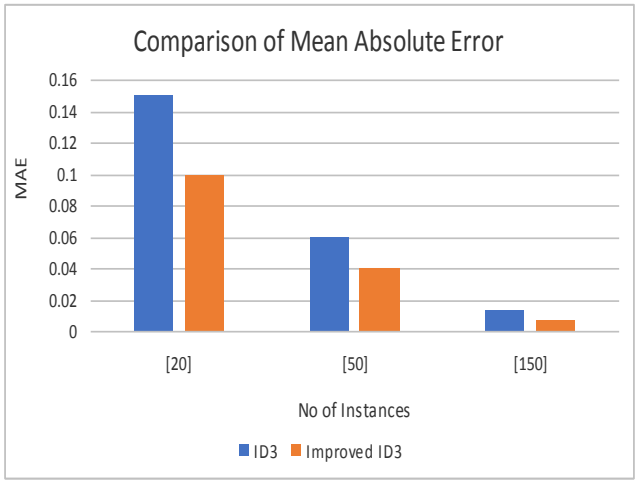

Fig. 2.3 Graphical comparison of Mean absolute error

Table 2.4 Tabular Comparison of Root mean square error

No of Instances	ID3	Improved ID3
[20]	0.3873	0.3162
[50]	0.2449	0.2
[150]	0.1155	0.0816

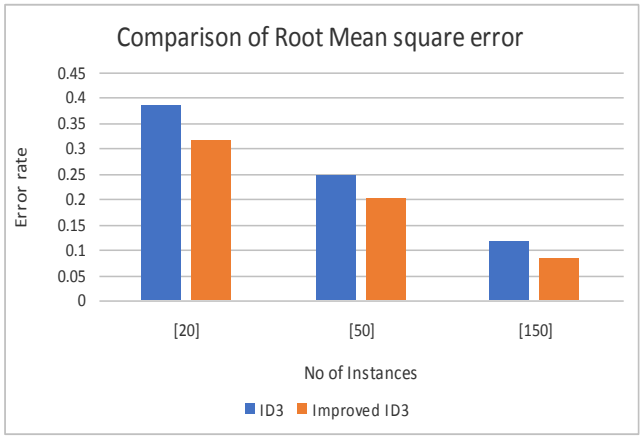

Fig. 2.4 Graphical comparison of Root mean square error

Table 2.5 Tabular Comparison of Accuracy

No of Instances	ID3	Improved ID3
[20]	85	90
[50]	94	96
[150]	98	99.33

© 2018, IJCSE All Rights Reserved

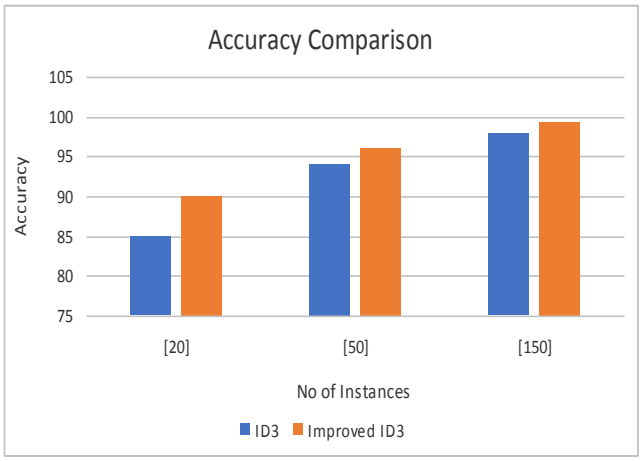

Fig. 2.5 Graphical comparison of Accuracy

Improved ID3 algorithm with Havrda and Charvat entropy for the value of alpha (α)=0.25, tree is small and less complex as compared to the use of Shannon entropy. In conclusion we can say that if we want to get a decision tree which is more effective and less complex, we can use the Havrda and Charvat entropy with this algorithm that will work best for huge datasets. The analysis of the experimental data shows that improved ID3 algorithm with Havrda and Charvat entropy when compared with ID3 algorithm has better classification accuracy and can get more reasonable results within short period of time.

VIII. CONCLUSION

In industrial sector, employee performance evaluation is an important field which will be used for predicting the performance of an employee. By considering the aspects like employee's task type, job skills, working quality, interpersonal relationship, creativity, adherence to policy, productivity, attendance, performance etc. Data mining consists of different techniques used to complete our objectives. Information that is extracted from large database is useful in decision making. By referring to the extracted information from the large databases, the processing methodologies are selected. The performance of an employee is evaluated in order to know them well and decide their position in organization.

Using this particular research, the industry superiors will have the ability to predict the employee's performance. A current performance evaluation system is required to support the recommendations for merit salary adjustments and in-grade or grade change salary increases. Data mining is a powerful analytical tool which enables the industrial institutions to improve their methods to allocate the resources and staff in an organization, and proactively manage employee's outcomes.

Conclusion of the research:

- Our algorithm has worked perfectly with highly accurate results and minimized error rate on huge datasets.
- Staffs having Heavy tasks and good quality of work, their performance is good, and most of them are actively cultivate their skills. Therefore they must be motivated to do their best.
- Staffs having Easy tasks and bad quality of work, then surely their performance is poor, and they mainly don't pay attention to cultivate their working skills, therefore such employees need a special attention.
- This evaluation helps the supervisor to find the employee's performance and taking action at right time for the profit of organization.

FUTURE SCOPE

The work has been focused on evaluation of employee's performance from various aspects and then enhancement of the performance for the organization's benefit. Our algorithm has worked perfectly with highest accuracy and minimizing error rate. There are number of ways through which we can extend our work in future. The data can be extended to have mixed characteristics of categorical and numerical data. In research it is clearly identified that our algorithm works best with Havrda and Charvat entropy for huge datasets. Thus we can use this algorithm for handling different organizations data and the work can be extended with the use of new entropy and with increased no. of attributes. Further research can be done for development and application for analysis of similar or different large datasets within short period of time.

ACKNOWLEDGEMENT

This is trivial attempt to offer my heartfelt gratitude to the various persons who have been of immense help in one way or the other for the successful completion of this dissertation. I am highly grateful to the authorities of Guru Nanak Dev University, Regional Campus, Jalandhar for providing this opportunity to carry out the dissertation work. My first and most earnest acknowledgement must go to my advisor and mentor Mrs. Neena Madan.

REFERENCES

[1] Shubham Tupe, Chetan Mahajan, Dnyaneshwar Uplenchwar, Pratik Deo," Employee Performance Evaluation System Using ID3 Algorithm", International Journal of Innovative Research in Computer & Communication Engineering, Vol. 5, Issue 2, February 2017

[2] Prof. Mr. A.M Bhadgale, Ms. Sharvari Natu, Ms. Sharvari G. Deshpande, Mr. Anirudha J. Nilegaonkar., "Implementation of Improved ID3 Algorithm Based on Association Function," Volume 114, No. 10 2017, 1-9

[3] Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, "A Modified Approach to Construct

© 2018, IJCSE All Rights Reserved

decision Tree in Data Mining Classification ", International Journal of Engineering and Innovative Technology (IJEIT) , Volume 2, Issue 1, July 2012
[4] Sartaj Sahni, "Algorithms analysis and design", Galgotia Publications Pvt. Ltd., New Delhi, 1996
[5] Nishant Mathur, Sumit Kumar, Santosh Kumar, And Rajni Jindal, "The Base Strategy For ID3 Algorithm Of Data Mining Using Havrda And Charvat Entropy Based On Decision Tree", International Journal Of Information And Electronics Engineering, Vol. 2, No. 2, March 2012
[6] L.Surya Prasanthi1, R.Kiran Kumar2, "ID3 and Its Applications in Generation of Decision Trees across Various Domains- Survey", International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015
[7] Kirandeep, Mrs. Neena Madan, "Analysis of Improved ID3 algorithm using Havrda and Charvat entropy", International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Authors Profile

Kirandeep M.Tech from Guru Nanak Dev University Jalandhar in year 2018. Currently pursuing Ph.D. Published four reaearch papers in reputed ugc approved and international journals.

Mrs. Neena Madan M.Tech from Guru Nanak Dev University Jalandhar. She is currently working as Assistant professor in GNDU,Regional Campus for Computer science and engieering department. She has published more than 40 research papers throughout her career in reputed international journals .Her research work is mainly focused on Data mining,Data analytics, Cryptography algorithms . She has 7 years of teaching experience and 3 years of Research experience.

© 2018, IJCSE All Rights Reserved