International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Classifying the Category of Workers using Crowd-sourced Job Seekers Data

S. Rajathilagam, K.Kavitha

Abstract: ***Crowdsourcing refers to decomposing complex jobs to multiple tasks and solve those task with multiple workers through open call in distributed networking environment. The recruitment of employees for organization has undergone transformation from traditional method to digital domain. Online recruitment facilitates just-in-time hiring to requesters and enables the workers to compete in the global market. This paper proposed an Efficient Machine Learning Crowdsourced(EMLC) method for E-recruitment which uses Crowdsourcing method to collect resumes from the workers and details of work from the requesters. The data is collected from a private job agency through an online recruitment portal which consist of recruiters from companies and job seekers based on qualifications and experience related to their field. The data collected from recruitment portal is analyzed with Machine Learning Approach with decision tree algorithms like ID3, CART and C4.5 for better selection of efficient person to complete the job. Various performance metrics such as Accuracy, Error rate, Recall etc were used to the Crowdsourced Database to categorize the job seekers efficiently. The proposed method gives better result for online recruiting through Crowdsourcing.***

Keywords: Crowdsourcing, Online Recruitment, Machine Learning, Decision Tree.

I. INTRODUCTION

Crowdsourcing plays an alternate problem solving technique where automated computers faces difficult to solve the problem when compared to humans. Crowdsourcing works by outsourcing the work normally done in company by a group of professional employee to a large group of undefined community or crowd with help of using internet through an open call method[1]. The editors of Wired magazine Jeff Howe and Mark Robinson coined the term Crowdsourcing in 2006 in their book "The Rise of Crowdsourcing". It describes the uses of internet in developing the business by outsourcing the task and seeking opinion about various companies decision to a community of people developed through internet[2]. The advancement in web technology gives rise Crowdsourcing websites where it consists of two groups of users the requesters and the workers[3]. The requesters presents the set of task, salary paid for that task and the duration in which the task has to be completed. A worker selects his preferred task from the available list and complete the task and submit the completed task. If the submission of the completed task best suits the need of the requesters then the worker who completed the task receives a reward[4].

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
S.Rajathilagam*, Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India. Email:raji.mdu2011@gmail.com
Dr.K.Kavitha, Assistant Professor, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India. Email:kavitha.urc@gmail.com

Crowsourcing was applied in voting system like Amazon Mechanical Turk (AMT) where workers are allowed to vote from a list of answers, Information sharing system like Wikipedia, yahoo answers where workers are allowed to share their knowledge through internet and game based system like Google Image Labeler where the objective is to collect labels for the image and creative thinking based system like lego, Eyeka which uses crowdsourcing to develop their products by inviting the user to share their ideas of designing a new products[5]. They allow the participants to share their knowledge by creating a set of questions and user respond to visual questions. Recruiting of employees through online community is one of the application of crowdsourcing[6]. Most of the companies recruit their workers by recruitment agency by crowdsourcing method. These companies post jobs vacancies through online and receive the resume of skilled persons and store these records in a database[7]. They analyze the database and find out the skilled persons from the available database and link the skilled person to the company to conduct interview. Companies recruit their employees then through interview and select the right persons needed for them. Crowdsourcing algorithms like sort, top-1, top-k, select, count, join were used to analyze the database and select the suitable tasks related to the category of workers [8].

II. LITERATURE REVIEW

Norases Vesdapunt[9] proposed a hybrid machine learning method to solve the Entity Resolution problem arise in the user generated content from social media like facebook, twitter. This method identifies mismatched present in the database by generating false positives and false negatives and find the deduplicating values generated by the users. Yuchen Zhang[10] proposed a new method for solving the problem of collecting multiple labels for each item provided by the non-professional crowdsourced workers. The task associated with each category calculated by the minimum likelihood is stored in a confusion matrix and then calculated with Expectation Maximization algorithm to identify the correct labels for multiple labeled problems. Michael J. Franklin[11] in their paper proposed a new query extension operator that can overcome the problem of missing data and subjective comparison. They used CROWD as a keyword to represent the column which is crowdsourced and the value for that particular column is not present at time of data preparation. The values of the attributes with CROWD are filled with CNULL at the time of data preparation. They used to combine the crowdsourced query with traditional SQL query and formed CrowdDB and implement Random Sort algorithm on the CrowdDB and presents the results to the AMT platform.

III. PROPOSED EMLC FRAMEWORK

This paper proposed a methodology called Efficient Machine Learning Crowdsourced(EMLC) framework to extract the skilled persons to do IT jobs required by the requesters of the IT Company and is illustrated in Figure 1 give below.

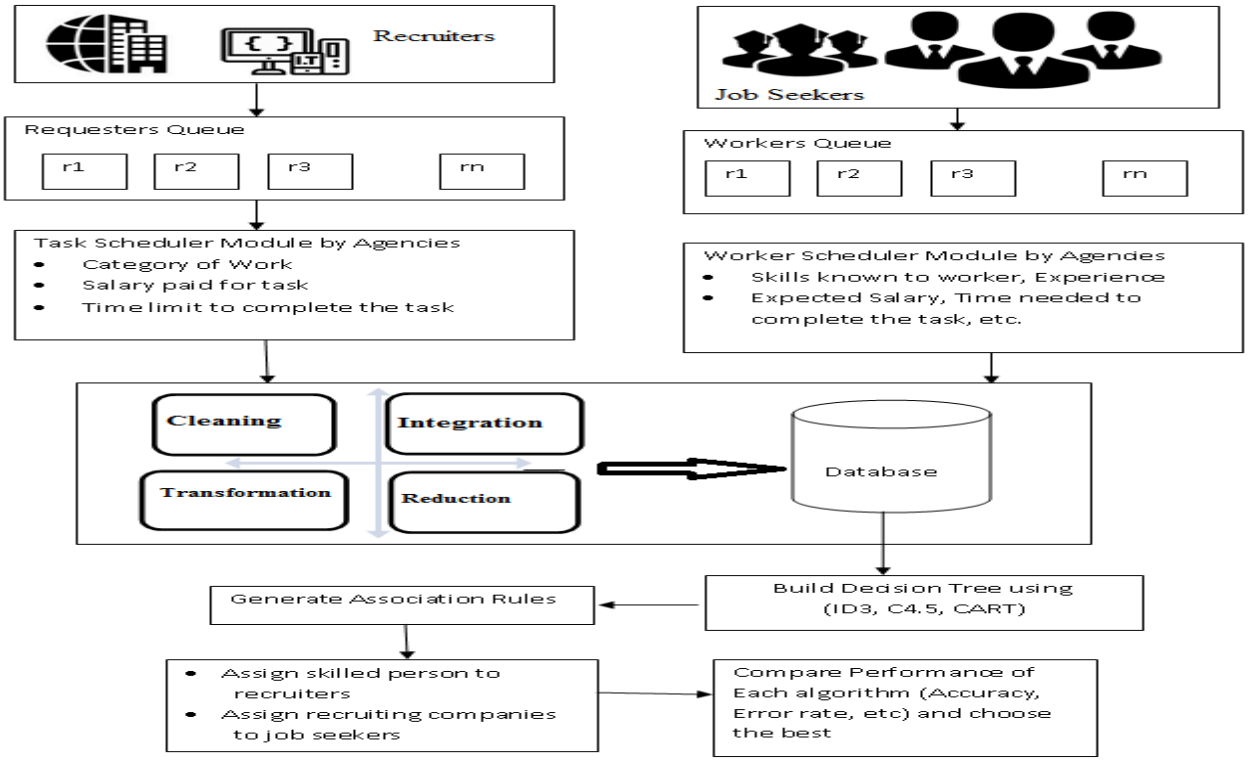

Figure1: Proposed EMLC Framework

The requesters who have work to be done by the workers were asked to register their company details, category of work, number of persons needed to do the job, location of work etc through the job portal. The registered requesters are placed in the requesters queue. The details of the job seekers such as their qualification, skills like database known, languages known, web technology known are collected through the job portal. The details collected were analyzed with query processing operators for cleansing the data entered by the recruiters and job seekers like misspelling of the values, filling the incomplete values given to the attributes and preprocessing was done and finally a Crowdsourced Database was formed in the first phase of the framework. The second phase of this framework deals with applyig Machine Learning techniques called classification to the collected data. Decision tree Induction Algorithm like ID3, CART, and C4.5 were implemented in the Crowdsourced Database and a decision tree was formed. Rule extraction was done on the generated decision tree and attributes which are highly relevant in the dataset was identified. The category of workers was categorized with the help of the generated rules and the task was allotted to the applied candidates. The third phase of the framework deals with applying various performance metrics such as accuracy, error rate, recall etc to the classifier and the classifier which best classifies the Crowsourced database was identified.

Pseudocode of The EMLC Algorithm

Input: Crowdsourced database D with n attributes, t tuples with details of jobseekers

Output: A Decision tree containing the category of workers

Case 1: Online Environment

CrowdDb_Creation(Database D, Tuples T, Attributes A)

1. Create a GUI module to get the details of the Recruiters, Job seekers
2. If all the attributes entered by the users are in correct order then insert into CrowdDb
3. Else begin
4. **Foreach** mis-spelled or incomplete values a_i of T
5. Apply query extension operator and complete the values then insert into the CrowdDb
6. End

Case 2: Offline Environment

Decision Tree_Creation(Dataset D, Dependent attribute Da, Independent Attributes Ii)

Begin

1. If D is Vacant then
2. return T as root node with most common values in T
3. Else begin
4. Crack the dataset into training and testing dataset
5. **Foreach** possible value I_i of Da begin

Size up Entropy for dependent attribute using

$$\text{Info}(D)=\sum_{i=1}^{m} p_i \log_2 p_i. \quad (1)$$

Size up Information gain for each attribute using

$$\text{InfoA}(D)=-\sum_{j=1}^{V} \frac{|D_j|}{|D|}\times\text{Info}(Dj). \quad (2)$$

Compute Gain for each attribute using

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (3)$$

end

6. **Foreach** possible value I_i of Da begin

Size up Entropy for target attribute using

$$\text{Info}(D)=\sum_{i=1}^{m} p_i \log_2 p_i.$$

Compute Information gain for each attribute using SplitInfoA(D)=-$\sum_{j=1}^{V}\frac{|Dj|}{|D|}$×Info(Dj).

Compute Gain Ratio for each attribute using Gain Ratio(A)=$\frac{Gain(A)}{SplitInfo(A)}$ (4)

end

7. **Foreach** possible value I_i of Da begin

Compute Impurity for target attribute using Impurity (D)=$\sum_{i=1}^{n}(y_i - \bar{y})^2$ (5)

Compute Gini Index for each attribute using Gini=1-$\sum_{i=1}^{n}(\frac{n_j}{n})^2$ (6)

End

8. Decision Tree was built using the maximum value obtained using the selection criterion mentioned in the steps 5,6, and 7.

9. Return rules to identify the category of workers from the crowdsourced dataset

10. Apply performance metrics to each classifier and compare the results of each classifier

End.

IV. RESULTS

The database used to store the details was first developed statically and the fields required for the attributes of the table were designed as a web form. The details of the recruiters and job seekers were collected by crowdsourcing method and the records were collected dynamically from the user. Implementing machine learning techniques like decision was implemented in Rapidminer Studio 9.2. The dataset contains 24 attributes like gender, educational qualifications, technical qualifications in software fields, conference attended, paper presented etc.

Table 1: Category of job attribute details

Database Administration	Software Developing	Software Testing	Web Designing	Total
116	172	85	127	500

The target attribute used to classify the dataset was Category of job and the following Table 1 shows the details of category of job attribute having four values which are Database Administration, Software Developing, Software Testing, Web Designing listed below.

The Entrophy is calculated using Equation (1) as shown below

Info(D)=($\frac{116}{500}\ log_2\frac{116}{500}-\frac{172}{500}\ log_2\frac{172}{500}-\frac{85}{500}\ log_2\frac{85}{500}-\frac{127}{500}\ log_2\frac{127}{500}$) =0.4890-0.5295-0.4345-0.5021

Info(D)=1.9551.

The following Table 2 describes the details of the Designation attribute which has the highest influence in deciding assigning job to the job seekers in the database.

Table 2: Designation Attributes details

Designation	Database Administrator	Software Developing	Software Testing	Web Designing
Dbadministrator	53	0	0	0
Dbadministrator/project manager	18	0	0	0
Dbadministrator/teamleader	13	0	0	0
Swdeveloper	0	138	0	0
Swtester	0	0	51	0
Swtester/teamleader	0	0	18	0
Swtester/project manager	0	0	14	0
Web designer	0	0	0	14
Web designer/teamleader	0	0	0	3
Fresher	32	34	2	110
Total	**116**	**172**	**85**	**127**

The information needed to split the attribute is calculated using Equation (2) as given below

$Info_A(D)=\frac{116}{500}[(\frac{53}{116}\ log_2\frac{53}{116})+(\frac{18}{116}\ log_2\frac{18}{116})+(\frac{13}{116}\ log_2\frac{13}{116})+(\frac{32}{116}\ log_2\frac{32}{116})]+\frac{172}{500}[(\frac{138}{172}\ log_2\frac{138}{172})+(\frac{34}{172}\ log_2\frac{34}{172})]+\frac{85}{500}[(\frac{51}{85}\ log_2\frac{51}{85})+(\frac{14}{85}\ log_2\frac{14}{85})]+\frac{127}{500}[(\frac{14}{127}\ log_2\frac{14}{127})+(\frac{3}{127}\ log_2\frac{3}{127})]$

$Info_A(D)$=(0.4173+0.9892+0.1709+0.0542)=1.6316

The gain of an attribute is calculated using Equation (3)

Gain(A)= 1.9551-1.6316=0.3235.

The Split information is calculated as follows

Split Info(A)=$\left[\left(\frac{53}{500}log_2\frac{53}{500}\right)+\left(\frac{18}{500}log_2\frac{18}{500}\right)+\left(\frac{13}{500}log_2\frac{13}{500}\right)+\left(\frac{138}{500}log_2\frac{138}{500}\right)+\left(\frac{51}{500}log_2\frac{51}{500}\right)+\left(\frac{18}{500}log_2\frac{18}{500}\right)+\left(\frac{14}{500}log_2\frac{14}{500}\right)+\left(\frac{14}{500}log_2\frac{14}{500}\right)+\left(\frac{3}{500}log_2\frac{3}{500}\right)+\left(\frac{178}{500}log_2\frac{178}{500}\right)\right]$

=(0.3523+0.1726+0.1369+0.5125+0.3359+0.1726+0.1444+0.1444+0.4428+0.53044)=2.54632.

Gain Ratio=$\frac{Gain(A)}{Split\ Infor(A)}=\frac{0.3235}{2.5463}=0.1270$

Among the 24 attributes present in the dataset Designation has the highest Gain Ration and it was chosen as the root node and the process was repeated and finally the decision tree obtained for the job seekers dataset was shown in Figure 2 below.

- designation in [fresher]
  - extracurricular in [sports]
    - 12grade in [distinction] then category = **webdesigning** (29 tuples)
    - 12grade in [first]
    - gender in [male] then category = **webdesigning** (6 tuples)
    - gender in [female] then category = **swdveloping** (6 tuples)
    - 12grade in [second] then category = **webdesigning** (2 tuples)
    - 12grade in [third]
    - uggrade in [distinction] then category = **webdesigning** (7 tuples)
    - uggrade in [first] then category = **webdesigning** (7 tuples)
  - extracurricular in [nill]
    - workshop in [nill] then category = **webdesigning** (6 tuples)
    - workshop in [one] then category = **swdveloping** (12 tuples)
    - workshop in [three] then category = **webdesigning** (5 tuples)
    - workshop in [four] then category = **webdesigning** (18 tuples)
  - extracurricular in [nss]
    - 10board in [state] then category = **webdesigning** (5 tuples)
    - 10board in [CBSE]
      - webtechnology in [html/css/js] then category = **swdveloping** (9 tuples)
      - webtechnology in [html/css/coofescript] then category = **swdveloping** (5 tuples)
  - extracurricular in [rrc] then category = **dbadministrator** (19 tuples)
  - extracurricular in [yrc]
    - 12grade in [distinction] then category = **webdesigning** (22 tuples)
    - 12grade in [first] then category = **webdesigning** (8 tuples)
    - 12grade in [second] then category = **dbadministrator** (12 tuples)
- designation in [swdeveloper] then category = **swdveloping** (138 tuples)
- designation in [swtester] then category = **swtesting** (51 tuples)
- designation in [swtester/teamleader] then category = **swtesting** (18 tuples)
- designation in [dbadministrator] then category = **dbadministrator** (53 tuples)
- designation in [dbadministrator/teamleader] then category = **dbadministrator** (13 tuples)
- designation in [dbadministrator/projectmanager] then category = **dbadministrator** (18 tuples)
- designation in [swtester/projectmanager] then category = **swtesting** (14 tuples)
- designation in [webdesigner] then category = **webdesigning** (14 tuples)
  - designation in [webdesigner/teamleader] then category = **webdesigning**(3 tuples)

Figure 2: Decision generated for Job seekers dataset

V. RULES GENERATED FROM DECISION TREE

R1: If designation is fresher and Extra Curricular Activity is Sports and 12th marks grade is distinction then Category of work is Web designing (29 tuples). R2: If designation is fresher and Extra Curricular Activity is Sports and 12th marks grade is first and gender is male then Category of work is Web designing (6 tuples). R3: If designation is fresher and Extra Curricular Activity is Sports and 12th marks grade is first and gender is female then Category of work is Web designing (6 tuples). R4: If designation is fresher and Extra Curricular Activity is Sports and 12th marks grade is second then Category of work is Web designing (2 tuples). R5: If designation is fresher and Extra Curricular Activity is Sports and 12th marks grade is third and ug mark grade is distinction then Category of work is Web designing
(7 tuples). Likewise 26 rules were generated from the decision tree. The dataset is again analyzed with C4.5, and CART algorithm and decision tree were constructed as described above and performance of each algorithm is measured as described below.

1.1 PERFORMANCE MEASURES

The performance of the decision tree is measured with the help of the confusion matrix and the confusion matrix obtained for the CrowdDb jobseekers database containing 500 records is shown below.

Table4: Confusion Matrix of C4.5 Algorithm

	Web Designing	Software Developing	Software Testing	Database Administrator	Sum
Web Designing	127	0	0	0	127
Software Developing	3	169	0	0	172
Software Testing	1	1	83	0	85
Database Administrator	1	0	0	115	116
Sum	132	170	83	115	500

The performance metrics used to measure the classifier such as Accuracy, Error rate, Recall etc were calculated from each confusion matrix obtained for the CrowdDb Job seekers database containing 500 records are listed below in Table 5.

Table5: Performance Comparison of various Decision Tree Algorithms

	ID3	CART	C4.5
Accuracy	0.984	0.948	**0.988**
Error rate	0.016	0.052	**0.012**
Recall(A)	0.9914	0.8879	**0.9914**
Recall(B)	0.9942	0.9419	**0.9826**
Recall(C)	0.9765	0.9765	**0.9765**
Recall(D)	0.9685	0.9921	**1.0000**
1-Precision(A)	0.0160	0.0000	**0.0000**
1-Precision(B)	0.0339	0.0182	**0.0059**
1-Precision(C)	0.0000	0.0000	**0.0000**
1-Precision(D)	0.0000	0.1544	**0.0000**

The same procedure is followed to analyze the performance of various decision tree algorithms like ID3, CART, C4.5 on CrowdDb database on increasing the number of records present in the database such as 1000, 2000, 3000, 4000, 5000 and the performance of those algorithms are shown in graphs.

The following Figure 3, Figure 4, Figure 5, Figure 6, Figure 7 shows the performance of the above mentioned algorithms with 1000, 2000, 3000, 4000, 5000 tuples respectively.

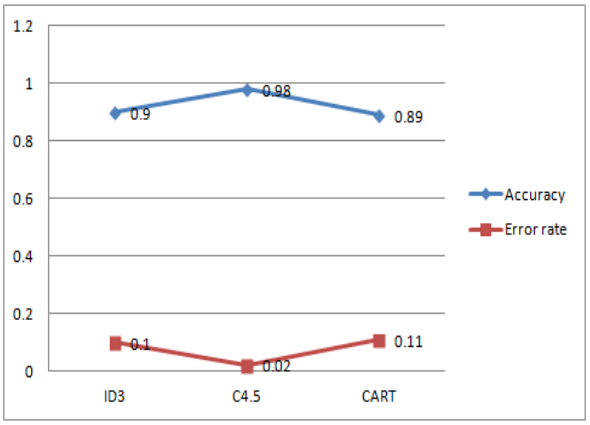

Figure 3: Performanace for 1000 records

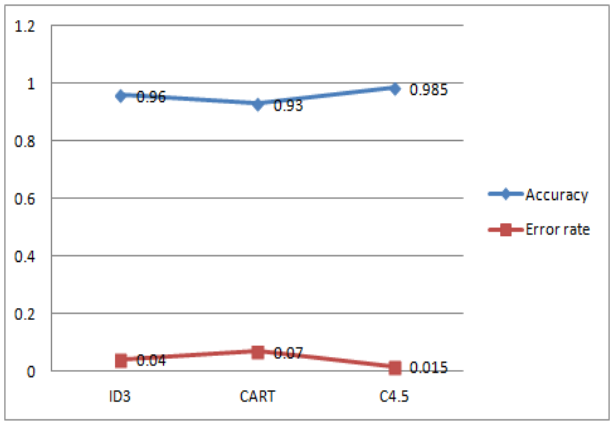

Figure 4: Performanace for 2000 records

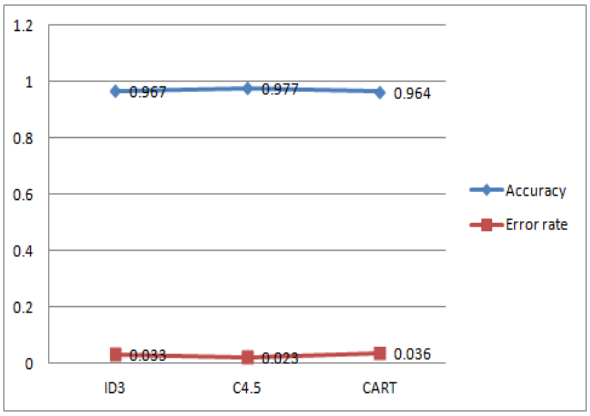

Figure 5: Performanace for 3000 records

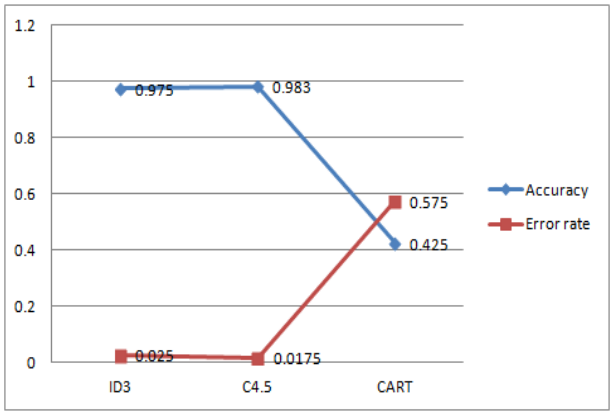

Figure 6: Performanace for 4000 records

\

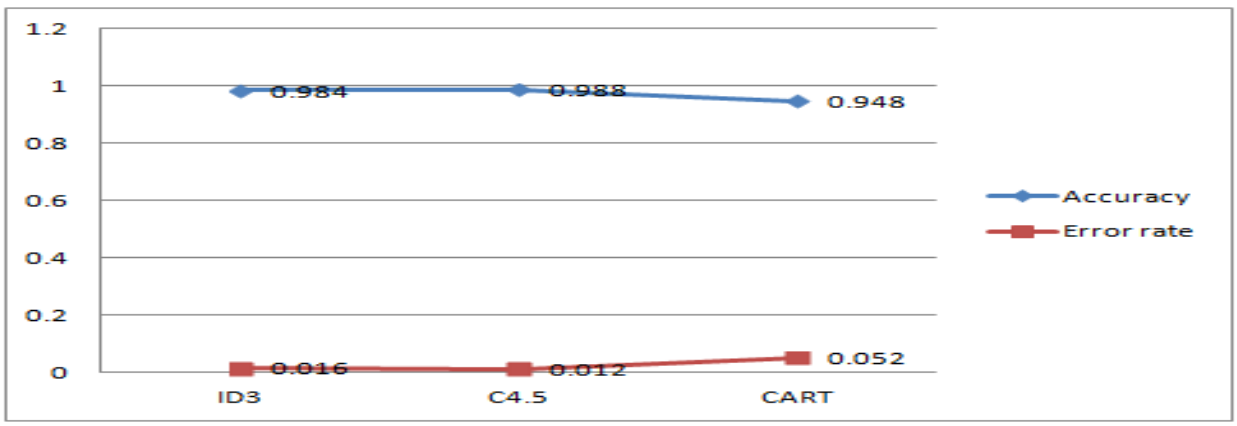

Figure 7: Performanace for 5000 records

VI. DISCUSION

From the decision tree obtained it clearly shows machine learning technique was correctly applicable to crowdsourcing platform. The category of workers was correctly identified by their working experience on the respective platform and the candidates who are fresheres were identified with the extracurricular, co-curricular, conference attended and paper presented details. The technical qualifications of the fresher play a very important role in identifying the carrier for them. The performance of the various decision tree algorithms were measured using various records in the database. It is clear from the various graphs shown above C4.5 Algorithm performs better when compared to the remaining algorithms. Applying machine learning techniques is best suitable for crowdsourcing because of its scalability. Since Crowdsourcing based attributes are filled only at runtime, machine learning is recommended to analyze for Crowdsourcing.

VII. CONCLUSION

This paper proposed an Efficient Machine Learning Algorithm to Crowdsourcing (EMLC) to analyze the data collected through outsourcing method.

The data entered by the user dynamically is treated with query extension operator and the imperfect entries are cleaned dynamically and the database was formed.
The data analyzing was done offline by decision tree induction algorithm and the performance of the C4.5, CART, ID3 were measured. This research shows better result than traditional crowdsourcing algorithms like sort algorithms by predicting exact attributes responsible for selecting the skilled employee from the job seekers database.

REFERENCES

1. Abraham, I., Alonso, O., Kandylas, V., & Slivkins, A, "Adaptive crowdsourcing algorithms for the bandit survey problem", In Conference on learning theory, June 2013, pp. 882-910.
2. Deng, X. N., & Joshi, K. D, "Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions", Journal of the Association for Information Systems, Vol 17, Issue 10, pp:648.
3. Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. CrowdDB: answering queries with crowdsourcing. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 2011, June, pp. 61-72.
4. Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., & Horton, J, "The future of crowd work. In Proceedings of the 2013 conference on Computer supported cooperative work", 2013, pp:1301-1318.
5. Kulkarni, A., Can, M., & Hartmann, B, "Collaboratively crowdsourcing workflows with turkomatic", In Proceedings of the acm 2012 conference on computer supported cooperative work, 2012, pp:1003-1012.
6. Venetis, P., Garcia-Molina, H., Huang, K., & Polyzotis, N. (2012, April). "Max algorithms in crowdsourcing environments", In Proceedings of the 21st international conference on World Wide Web, pp:989-998.
7. G. Suresh, K. Arunmozhi Arasan, S. Muthukumaran, "Perfomance Comparison of Decsion Tree Algorithms to Findout the Reason for Student's Absenteeism at the Undergraduate Level in a College for an Academic Year" International Conference on Computing and Intelligence Systems Volume: 04, Special Issue: March 2015, pp: 1235 – 1241.
8. Venkatesan, N., Arasan, K. A., & Muthukumaran, S, "An ID3 Algorithm for Performance of Decision Tree in Predicting Student's Absenteeism in an Academic Year using Categorical Datasets". Indian Journal of Science and Technology, Vol. 8, Issue - 14, pp:1-5, 2015.
9. Vesdapunt, N., Bellare, K., & Dalvi, N, "Crowdsourcing algorithms for entity resolution", Proceedings of the VLDB Endowment, Vol 7, Issue 12, 2014, pp:1071-1082.
10. Zhang, Y., Chen, X., Zhou, D., & Jordan, M. I. (2014). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In Advances in neural information processing systems, 2014, pp. 1260-1268.
11. Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. CrowdDB: answering queries with crowdsourcing. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 2011, June, pp. 61-72.
12. Vengatesan N, Arunmozhi Arasan K, Muthukumaran S, "Efficiency of Decision Tree in Predicting Student's Absenteeism in an Academic Year using C4.5 Algorithm", International Journal of Innovative Research in Computer Science and Engineering(IJIRCSE), Vol.1, Issue-3, 2015.

AUTHORS PROFILE

Rajathilagam.S Research Scholar, Dept.of.Computer Science, Mother Teresa Women's University, Kodaikanal. Her area of research interests are Data Mining, Network Security, Software Engineering. Mail Id: Raji.mdu2011@gmail.com .

Dr.K.Kavitha, Assistant Professor, Dept.of.Computer Science, Mother Teresa Women's University, Kodaikanal. Her Area of research interests are, Data Mining, Big Data Analytics.Mail Id: Kavitha.urc@gmail.com