IJARIIT
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue3)
Available online at www.ijariit.com

Development of Data Mining Model for the Evaluation of Human Skill Placement in the Engineering Sector

Arun Mishra
Saroj Institute Of Technology & Management, Lucknow
er.arun2009@gmail.com

***Abstract:** Data mining is one of the widespread research areas of present time as it has got wide variety of application to help people of today's world. It is all about finding interesting hidden patterns in a huge history database. In this research work, data mining is comprehensively applicable to a domain called placement chance prediction, since taking wise career decision is so crucial for all of us for sure. A strategy to predict the overall absorption rate for every branch as well as the time it takes for all the students of a particular branch to get placed etc. are also proposed.*

The proposed method is tested on the data set provided by A.I.M.T college Lucknow and data is passed through the various data mining model, namely decision tree , neural network and navie bayes classifier on area of the application to a domain, development of classifier and future outcomes were also configured on this thesis.

At last, this research work puts forward the data mining algorithm namely C 4.5 * stat for numeric data sets which has been proved to have competent accuracy over standard benchmarking data sets called UCI datasets. It also proposes to improve the standard C 4.5 algorithm

Keywords: Data Mining, Decision Tree, UCI datasets, Patterns, C4.5* stat Algorithm.

1. INTRODUCTION

Today academic institutions and their auxiliary centers maintain a huge amount of information regarding their academic activities like student performance and placement history. Data mining can be very effectively used in analyzing such socially relevant data. In one of the existing works in literature, student retention analysis has been carried out in which, a system has been developed to find out those students who will actually finish the course and pass out from an institution. This information is useful in places, where there are many cases of students leaving an institution without completing the course creating waste of institutional resources [20]. In another work, questions like "which are the courses that are usually selected by top performers?" etc. are addressed [30].

This research work is an attempt to help the prospective students to make wise career decisions using data mining technology. Popular Data mining models like decision trees, neural networks, Naïve Bayes classifier etc. are used. The effort is taken to improve the overall performance of these models and to find an optimum model for this particular problem. A decision is made based on data like Entrance Rank, Gender (M/F), Sector (Rural/Urban), Reservation category (General, OBC (Other Backward Castes), SC/ST (Scheduled castes / tribes) and branch (Civil Engineering, Computer Science and Engineering and so on).

.

2. BACKGROUND AND RELATED WORK

Before building this model we have gone through the work done by various researchers. Romero, C., Ventura, S. and Garcia, E [1] done a lot of important work related to education data mining. It uses data mining techniques for the purpose of prediction in the education field. Nguyen et al. [2] compare the accuracy of the decision tree and Bayesian network algorithm for the prediction of performance of the undergraduate and postgraduate students. Results from this work are useful for prediction of the weak student who may fail in the exam. Affendy and Must pain [3] uses performance in various subjects to predict the CGPA of bachelor students. Al- Radaieh et al.[4] uses classification technique to improve the quality of education. Cesar et al [5] use data mining technique to develop a model which helps the student to take an academic decision. Nghe et al also provide a lot of contribution in this field. Ramaswami and Bhaskaran also develop a predictive model to evaluate achievement of a student at higher secondary level. N.S.Shah applies the various decision tree technique to categorize BBA student based on performance. Tripti Mishra et al [6] also uses classification technique to predict student's performance. In our paper, we use C4.5* algorithm to develop a model to predict placement status for a student and actually contributes to the branch wise placement of the Institution.

3. DATA COLLECTION

Data collection is a very important step of building prediction model because it is the base of all your predictions, a minor error in the data cause a blunder in the prediction. so data collected must be accurate. In this model, we collect the data related to engineering undergraduate students of the batch 2013. Data collection involves the data related to the various attributes which are considered in the model for predictions. The attributes considered in this model are branch, sector, sex, rank, students native place from which he or she belongs, student family background i.e. rich or poor, leadership ability, student's participation in the extracurricular activities.

Id	Type	Node	Parent
1	BRANCH	CS	0
2	SECTOR	RURAL	1
3	SEX	MALE	2
4	RANK	1	3
5	ACTIVITY	EXCELLENT	4
6	SECTOR	URBAN	1
7	SEX	FEMALE	6
8	RANK	2	7
9	ACTIVITY	AVERAGE	8
10	BRANCH	EC	0
11	CATEGORY	OBC	10
12	RANK	4	11
13	ACTIVITY	POOR	12
14	CATEGORY	GENERAL	10
15	RANK	6	14
16	ACTIVITY	GOOD	15

4. BUILDING PREDICTION MODEL

After the collection of desired data, we use Weka data mining tool for the development of prediction model using random tree algorithm. We simulate the random tree algorithm on the data of B.tech students of batch 2013.

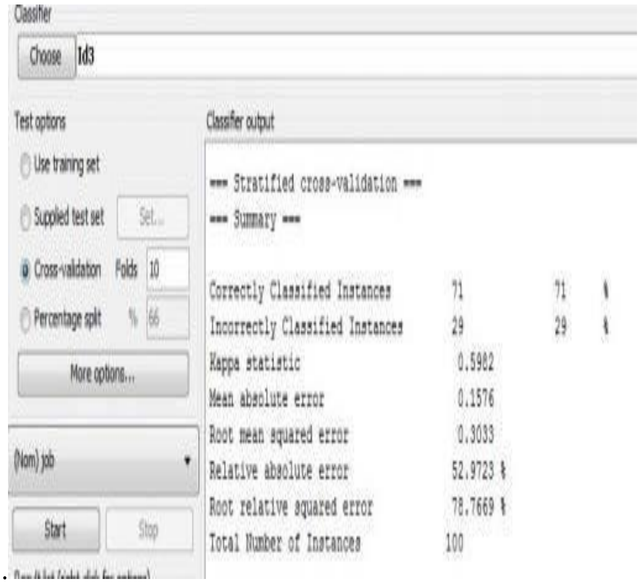
For applying the various algorithm on the given data set we need to convert excel file to arff file. For converting Excel file to Arff we use the notepad. By doing this we get the required file for processing in weak**.**

In WEKA software, a relation name is represented as
–@relation student
A list of attribute definitions
–@attribute RANK numeric

```
@attribute SECTOR {U, R}
@attribute SEX {F, M}
@attribute CATEGORY {GEN, OBC, SC, ST}
@attribute BRANCH {A, B, C, D, E, F, G, H, I, J}
@attribute CHANCE {E, P, A, G}
```

Now we apply various algorithm on the student information data set using weka tool.We apply ID3 , BayesNet , J48 ,C45* algorithms on the student information data set using weka tool

Fig 1. Model for the prediction process



Classifier: Id3

● Cross-validation Folds 10

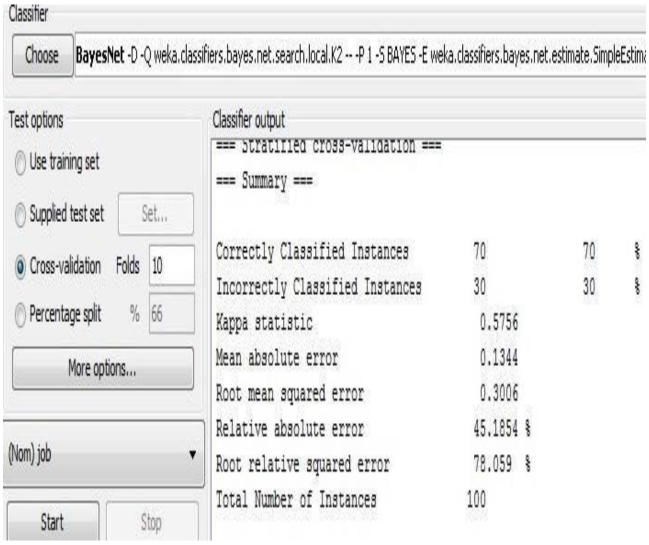
% 66

(Nom) job

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          71               71      %
Incorrectly Classified Instances        29               29      %
Kappa statistic                          0.5982
Mean absolute error                      0.1576
Root mean squared error                  0.3033
Relative absolute error                 52.9723 %
Root relative squared error             78.7669 %
Total Number of Instances              100
```

Fig1. Applying Id3 algorithm



Classifier: BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstima

● Cross-validation Folds 10

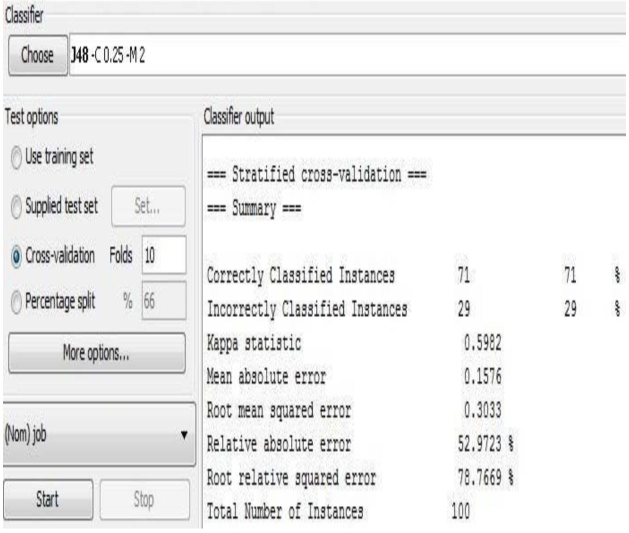
% 66

(Nom) job

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          70               70      %
Incorrectly Classified Instances        30               30      %
Kappa statistic                          0.5756
Mean absolute error                      0.1344
Root mean squared error                  0.3006
Relative absolute error                 45.1854 %
Root relative squared error             78.059  %
Total Number of Instances              100
```

Fig2. Applying BayesNet algorithm



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

(Nom) job

Start Stop

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        71               71      %
Incorrectly Classified Instances      29               29      %
Kappa statistic                        0.5982
Mean absolute error                    0.1576
Root mean squared error                0.3033
Relative absolute error               52.9723 %
Root relative squared error           78.7669 %
Total Number of Instances            100
```

Fig3. Applying J48 algorithm

5. RESULT AND ANALYSIS

By applying the various algorithms on the given dataset we find following results.As figures show after applying the various algorithm on the student information data set we find Id3 Fig1 is 71% accurate, BayesN et is 70% accurate, J48 is 71% accurate, RBF network algorithm is 65 % accurate and Random Tree algorithm is most accurate for the prediction process. i.e 73%

In conventional decision tree algorithms like C4.5, the splitting will be done based on the maximum information gain concept. But here the statistical variance is used, which is defined as follows:
In general, the population variance of a finite population of size N is given by equation (8.1)

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \qquad - (8.1)$$

Where μ is the population mean as given by equation (8.2):

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad - (8.2)$$

Here the assumption is that, if a subset of the data is having low variance then there is a chance that they converge to a particular class in a minimum number of iterations as there is minimum variation in the data for that attribute.

In Table 1, these comparisons are shown. It is observed that C 4.5*stat algorithm has competent accuracy. In statistics, to further generalize the result for a larger population, there are many hypothesis testing strategies. The statistical technique namely ANOVA is used here. In statistics, analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, for testing a hypothesis.

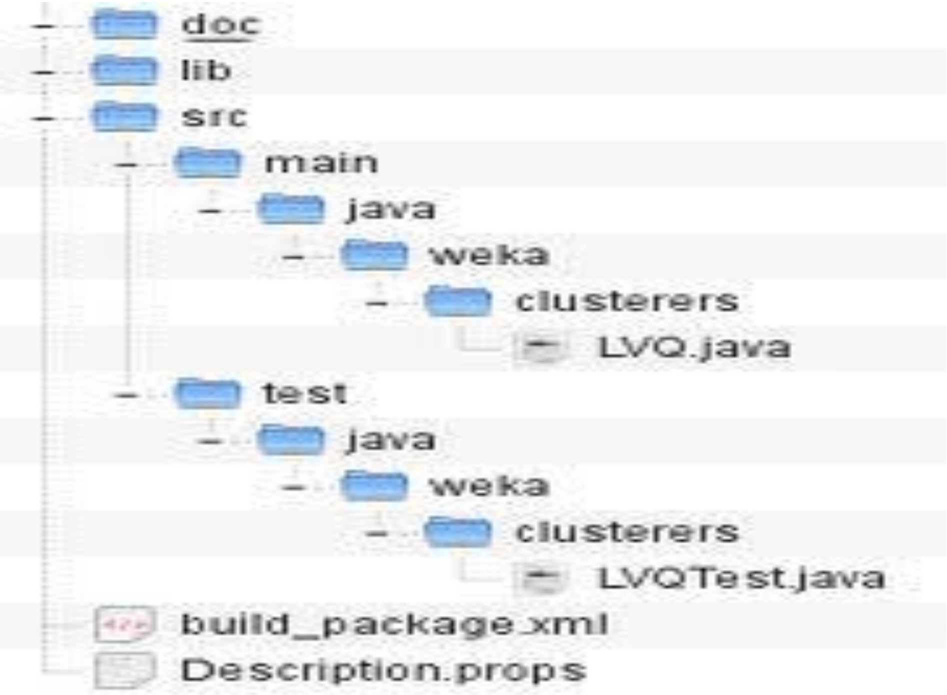
Table1: Accuracies of various decision tree algorithms on UCI Data sets

Dataset	**AD Tree**	**REP Tree**	**Random Tree**	**C 4.5*stat**
Iris	NA	96	94	95.3
Segment	NA	94.81	89.13	94.07
Diabeties	73.17	73.56	68.09	74.47
Letter	NA	NA	82.875	81.34
Breast-cancer	95.7	94.7	93.84	95.13
Glass	NA	66.82	62.61	67.28

Labor	82.45	68.42	85.96	82
Mean	83.77	82.385	82.35	84.23

The time complexity of standard decision tree algorithm is O (mn2), where m is the number of records and n is the number of attributes [72]. This is because there are total m records itself among all nodes in a particular level at a time and for computing information gain, it has to consider each of the n attributes. So at a particular level, the complexity is O (mn).

In the worst case, there will be a split corresponding to each of the n attributes. So altogether it becomes like O (mn2)in the worst case. But here as the numeric data are split based on statistical mean the number of levels in the worst case is log2m. So the time complexity becomes O (mnlog2m). So as the numbers of attributes become very high, which is common in huge data sets like bioinformatics data, this algorithm will have an edge in terms of time.



```
doc
lib
src
  main
    java
      weka
        clusterers
          LVQ.java
  test
    java
      weka
        clusterers
          LVQTest.java
build_package.xml
Description.props
```

Figure1: Integrating WEKA with Net beans IDE for developing new classifiers

C 4.5* algorithm is found to be competent in accuracy with its information gain counterpart C 4.5. Instead of information gain, this algorithm uses statistical variance. An improvement in computing time is also found, when the number of attributes of the data set increases.

CONCLUSION

Today placement is one of the critical aspect for the educational institutes. So prediction of student's placement can help them to provide assistance to improve the overall placement of the college. It also helps in the development of the system which actually suggest to students in selecting particular engineering branch which is in demand.

REFERENCES

[1] Romero, C., Ventura, S. and Garcia, E., "Data mining in Course management systems: Moodle case study and Tutorial". Computers & Education, Vol. 51, No. 1.pp.368- 384. 2008.

[2] Nguyen N. , Paul J. , and Peter H. , a comparative analysis oftechniques of predicting student performance. In the proceeding of 37th ASEE/IEEE frontiers in education.

[3] I.H. , M.Paris , L.S.Affecndy , improving prediction performance using voting technique in data mining , in world academy of science, vol 38 , 2010

[4] Al- Radaideh , Al-Shwakfa , mining student data using decision tree ,in international Arab conference on information technology , 2006

[5] Cesar V. , Javier B. , liela S. ,recommendation in higher education using data mining , educational dataminng conference 2009

[6] Tripti mishra, Dr. dharminder kumar, dr.sangeeta Gupta, mining student data for performance prediction, fourth international conference on advance computing and communication technologies, 2014

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.

[13] P.K. Chan, S.J. Stolfo, “A comparative evaluation of voting and meta- learning on partitioned data,” In Proceedings of the International Conference on Machine Learning, pp. 90-98, California, USA, 1995.

[14] N.Chawla, Bowyer, ”Designing multiple classifier systems for face tion,” In Proceedings of the Sixth International Workshop on classifier Systems, Springer, Volume 3541/2005, pp. 982- 984, 2005.

[15] Cynthia Krieger, “Neural Networks in Data Mining”, CiteSeerX, 1996.[Online].Available: http://www.cs.uml.edu/~ckrieger/user/Neural_Networks.pdf. ,pp.1-23

[16] Dan Zhu, “A hybrid approach for efficient ensembles,” Science Direct decision support systems, 48(1), pp. 480-487, 2010.

[17] T.G. Dietrich, “Ensemble methods in machine learning,” In Proceedings of the First International workshop on multiple classifier systems, pp. 1–15, Cagliari, Italy, 2000.

[18] T.G. Dietterich, “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization”, Springer Machine Learning, 40(2), pp. 139-157, 2000.

[19] S. Dudoit, J. Fridlyand, and T.P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” Journal of the American Statistical Association, 97(457), pp. 77–87, 2002.

[20] Elizabeth Murray, “Using Decision Trees to Understand Student Data,” pp.1-10.

[21] Eunju Kim, Wooju Kim, Yillbyung Lee, “Combination of multiple classifiers for the customer’s purchase behaviour prediction,” Elsevier Decision Support Systems, 34(1), pp. 167– 175, 2002.

[22] U. Fayyad, R Uthurusamy, “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, 17(3), pp. 37-54, 1996.

[23] Y. Freund, R. E. Schapire, “Experiments with a new boosting. algorithm,” In Proceedings of 13th International Conference on Machine Learning, pp. 148-156, Bari, Italy, 1996.

[24] M.Gashler, C. Giraud-Carrier, T. Martinez, “Decision tree ensemble: small heterogeneous is better than large homogeneous,” In Proceedingsof the Seventh International Conference on Machine Learning and Applications, pp. 900-905, San Diego, California, 2008.

[25] G. Giacinto and F. Roli, " Ensembles of neural network classifiers for remote sensing images", In Proceedings of the European Symposium on Intelligent Techniques, pp. 166-170, Bari, Italy, 1997.

[26] P.A. Gilson, J.A. Benediktsson, J.R. Sveinsson, “Decision fusion for the classification of urban remote sensing images,” Pattern Recognition Letters, 27 (1), pp. 294–300, 2006.

[27] M.A. Hall, “Correlation-based feature selection for machine learning,” PhD thesis, Department of Computer Science, University of Waikato, Hamilto, New Zealand, 1998.

[28] Hongjun Lu, Rudy Setiono, Huan Liu, “Effective Data Mining Using Neural Networks,” IEEE TKDE, 8(6), pp. 957-961, 1996.

[29] Jefery D Scargles, “Studies in Astronomical time series analysis,” The astrophysical journal 263(1), pp. 835-853, 1982.

[30] Jinlong Wang, Shunyao Wu, Yang Jiao, Huy Quan Vu, "Study on Student Score Based on Data Mining,” JCIT, 5(6), pp. 171-179, 2010