International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015
ISSN: 2395-3470
www.ijseas.com

A Study on Classifier Performance Using Machine Learning Algorithm

[1]R.Yogatharani, [2]R.Bhaskaran

[1] Head Department of CS, SW, M.Sc (I.T) & M.C.A, Parvathy's Arts & Science College, Tamilnadu 624 004, India.

[2] School of Mathematics, Madurai Kamaraj University, Tamilnadu 625 021, India

Keywords: Data mining, Classification, Prediction.

Abstract:

Data mining is the process of analyzing data from large data sets on different perspectives and extract meaningful and useful information that can be used to acquire insight into the data. Data mining is used today in diversified applications by researchers and educational institutions for gaining knowledge. In the case of educational institutions and researchers working on problems related to education, they concentrate on the performance analysis of the students with reference to a particular aspect or develop a model that would help the institutions for better performance. In our study we have utilized data mining techniques to focus on Arts and Science College in both rural and urban area around Madurai, to evaluate the Academic performance of the students as well as their performance towards placement. By means of a questionnaire we collected data from 1000 undergraduate students from 15 Arts and Science Colleges. These 1000 datasets were evaluated by twenty classifiers of machine learning algorithm. From these predictive based classifiers we observe that students at the colleges in the rural area need to develop more communication skill as it is very important in connection with their placement for jobs.

I.INTRODUCTION:

As in the past many students after under graduation prefer job and only a few go in for higher education. It may be noted that many organizations prefer to conduct campus interviews, walk-in or job fair to select candidates for the required jobs. Consequently students get better opportunities. In spite of such good opportunities being provided many students fail to get selected. Subject knowledge, computational skill, communication skill etc. may be attributed to such failure. Now a day, most of the educational institutions are focusing on pass percentage of the students and thereby they inculcate only the bookish knowledge and the students do not get an opportunity to think and develop good skills. So there is no innovation. We have made an attempt to study the aspects related to the failure. In this connection we selected 15 Arts and Science colleges at random in and around Madurai and obtained through a

questionnaire data from 1000 students. We have 1000 instances and nine attributes namely Gender, Medium of Education, Level of Communication, Location of Educational Institution, Type of School, (Public or Private) Percentage of pass in tenth, twelfth, and in undergraduate course, Non academic activities. We have used classifiers and evaluators with two different search methods like Bf and Rk. Study reveals that the skill set of the students depends on the location of their educational institutions. We have used WEKA, the open source program for all the computations carried out and the results shown are in the format of WEKA. The study also shows that there is a vast difference in communication skill and the way of presentation between the students studying in rural and urban.

II.RELATED STUDIES:

It is very natural to use data mining techniques to study the difference in the characteristics between rural and urban students. The problem is quite universal. Hannaway, J. and Talbert, J.E., in their paper (1993) Bringing context into effective school research: Urban-suburban differences have shown that there is a difference in the characteristics of urban and rural areas that account for most of the differences in the performance of students. Moore, E. J., Baum, E. L., and Glasgow, R. B. (1984) have studied Economic factors influencing educational attainment and aspirations of farm youth from both rural and urban. Vandernberghe, V. and Robin, S. (2004) have compared methods in evaluating the effectiveness of private education across countries. In exploring the relationships between school location (urban vs. rural) and students' occupational and educational aspirations J**.** David McCracken and Jeff David T. Barcinas, (1991) have observed considerable differences between urban and rural schools and similar differences between urban and rural students. More specifically interns of occupational aspirations. Investigating, using six learning strategies, the overall characteristics of the rural and urban high school students' learning strategy selection and use, the results obtained by Yanfeng Hu ,Sep 2009 indicate that the urban students, compared to the rural ones, do better in using all the strategies. Through T- test analysis no evident distinction exists between the two groups when using metacognitive, cognitive, affective and memory strategies. A comprehensive study conducted by Lin Siew Eng, Abdul Rashid Mohamed and Muhammad Javed, 2013 go to show that there is no significant difference between the performance of male and female students and the students of public and private schools, whereas there was a significant difference between the performance of urban and rural students. H. Jeraltin Vency and E.Ramganesh, Nov 2013 undertook a study of finding the language proficiency of post graduate students, the results convey the need for in depth revision in sowing the skills of English language among college students.

III.DATA MINING STRATEGIES:

The main idea of data mining is to extract useful information from that is hidden in large data sets. Several algorithms are used for this purpose to extract nuggets of knowledge from large set of data. These algorithms go towards classification, Association, Clustering etc. In this paper we use Classification algorithm as it helps in the prediction of the output of data according to the given input. For this prediction, the algorithm processes the data in a training set

which contains the given set of attributes and instances. The overview of Machine Learning is depicted in figure 3.1

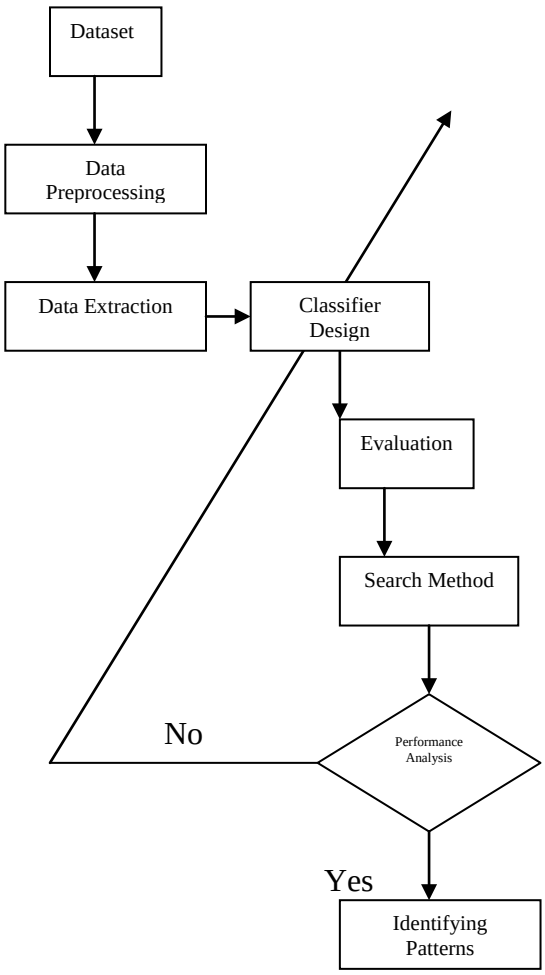

Figure 3.1Overview of Machine Learning

IV.ATTRIBUTE SELECTION:

Attribute selection is a process of selecting the best subset of attributes from original data set; the selected subset is determined and validated according to the goal. The attributes name and their description is given below:

Table 4.1 Students related Variables

Variable Name	Description
1.GEN	Student's gender
2.MOE	Medium of education
3.LOC	Level of communication
4.TOS	Type of School
5.TP	Tenth Percentage
6.TWP	Twelfth Percentage
7.UGP	Under Graduate Percentage
8.NAP	Non Academic Performance
9. LOEI	Location of Educational Institutions

In the above mentioned table the field Location of Education Institution (LOEI) and Level of Communication (LOC) are the class labels. In LOEI there are two variables namely r and u, where r stands for rural area and u stands for urban area. In class label LOC there are five variables namely vd - very much dissatisfied, vs - very much satisfied, ss - somewhat satisfied,
n - neutral and sd - somewhat dissatisfied. These class variables are used to compare and predict the level of communication between the rural and urban area students.

V. CLASSIFICATION ALGORITHM:

In this study, data is analyzed by using twenty different classifiers namely J48, DT, BN, NB, MLP, RBFnet, RSS, LB, MCC, HP, CR, JRip, DTNB, BFT, NBT, ID3, RT, REPT, RF and SC. These classifiers are built from the training set made up of database associated with class labels. Further the classification rules are applied to the dataset by using seven different evaluators namely Cfs, Gar, Lsa, Ing, Chi, Css and Rae with two different search methods Bf and Rk.

Phase 1: Evaluating the dataset by using LOEI as class label

International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015
ISSN: 2395-3470
www.ijseas.com

Phase 2: Evaluating the dataset by using LOC as class label

Building Classifier Model for Phase I:

When the class label LOEI is subjected to machine learning algorithm it selects Level of Communication as a locally predictive attribute. The outcome for J48 pruned tree for LOEI is given below.

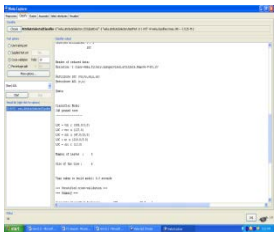

Fig 5.1 J48-CFS-BF Classifier

J48 pruned tree

LOC = a: vd: r (595.0/3.0)
LOC = b: vs: u (127.0)
LOC = c: ss: u (47.0/15.0)
LOC = d: n: u (218.0/5.0)
LOC = e: sd: r (13.0)

According to the evaluation of J48 pruned tree, in rural background the ratio of very much dissatisfied level of communication is found to be larger than the somewhat dissatisfied level of communication. But in urban background the ratio of neutral level of communication is larger than the ratio of very much satisfied level of communication which is comparatively higher than somewhat satisfied level of communication.

Building Classifier Model for Phase II:

Here the class label LOC selects two locally predictive variables namely Location of Education Institutions and Medium of Education; the output for J48 pruned tree for LOC is given below.

J48 pruned tree

MOE = t
| LOEI = r: vd (623.0/33.0)
| LOEI = u: n (252.0/39.0)
MOE = e
| LOEI = r: vd (2.0)
| LOEI = u: vs (123.0)

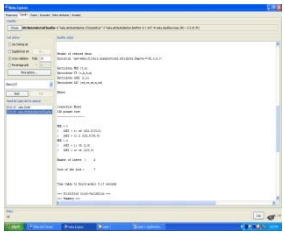

Fig 5.2J48-CFS-BF Classifier for class LOC

From this classifier model we observe that Medium of Education is closely related to Location of Education Institution, to predict Level of Communication. In urban area the ratio of Level of Communication is high in English medium while comparatively there is very low Level of Communication in Tamil medium in the rural area.

VI. PREDICTION:

Data mining is a “Deviation Analysis”, in that prediction is a “supervised learning Task”. In prediction, dataset is used directly to identify the class value. This analysis defines the potential of data and predicts future behavior. However, only good data can produce good prediction. As a result of prediction the statistical values of Accuracy, F1-Measure and ROC of data set is evaluated.

Evaluating the Accuracy of Classifier:

International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015
ISSN: 2395-3470
www.ijseas.com

According to the classifier, the accuracy predicts the correctness of class label .While the accuracy of predictor refers how well a given predictor can guess the value of predicted attribute for a new data. Accuracy is defined by

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP is the number of true positive cases, TN is the number of true negative cases, FP is the number of false positive cases and FN is the number of false negative cases.

Table: 6.1 Classifier Vs Accuracy

CLASSIFIER NAME	ACCURACY
1. J48-CFS-BF	97.7
2. DT-CFS-BF	97.7
3. BN-GAR-RK	96.1
4. NB-LSA-RK	62.5
5. MLP- LSA-RK	55.1
6. RBFNet –ING-RK	96.5
7. RSS-CHI-RK	92.8
8. LB- CSS-BF	62.5
9. MCC- CSS-BF	62.5
10. HP- GAR-RK	75.2
11. CR-CFS-BF	96.4
12. JRip-CSS-BF	97.7
13. DTNB-GAR-RK	97.7
14. BFT -GAR-RK	97.6
15. NBT-CFS-BF	97.7
16. ID3-CFS-BF	97.7
17. RT-CHI-RK	96.6
18. REPT- CHI-RK	97.7
19. RF-RAE-RK	97.1
20. SC-CFS-BF	97.7

According to the table 6.1 these classifiers J48, DT, JRIP, DTNB, NBT, ID3, REPT, SC produce high accuracy percentage (97.7%), but MLP classifier (55.1%) produces low accuracy percentage. The output of accuracy is depicted in the form of chart and graph.

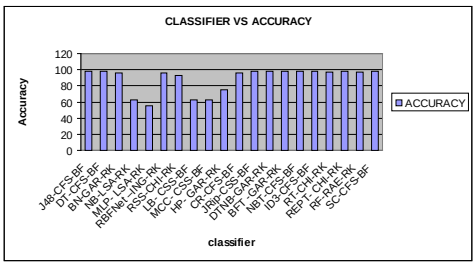

Fig 6.1 Chart for Classifier Vs Accuracy

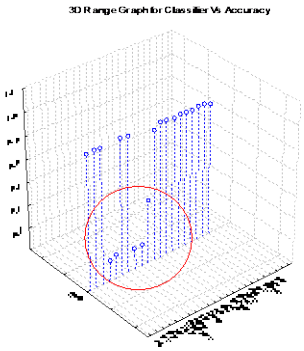

Fig 6.2:3D range Graph for Classifier Vs Accuracy

F1-Measure:

F1-Measure is the combination of precision and recall .whereas in precision the retrieved document is relevant but in recall relevant document is retrieved.

$$\text{F1-measure} = \frac{2*(\text{Precision}*\text{Recall})}{(\text{Precision} + \text{Recall})}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Table: 6.2 Classifier Vs F1-Measure

CLASSIFIER NAME	F1-MEASURE
1. J48-CFS-BF	0.977
2. DT-CFS-BF	0.977
3. BN-GAR-RK	0.961
4. NB-LSA-RK	0.481
5. MLP- LSA-RK	0.539
6. RBFNet-ING-RK	0.965
7. RSS-CHI-RK	0.927
8. LB- CSS-BF	0.481

9. MCC- CSS-BF	0.481
10. HP- GAR-RK	0.711
11. CR-CFS-BF	0.964
12. JRip-CSS-BF	0.977
13. DTNB-GAR-RK	0.977
14. BFT -GAR-RK	0.976
15. NBT-CFS-BF	0.977
16. ID3-CFS-BF	0.977
17. RT-CHI-RK	0.966
18. REPT- CHI-RK	0.977
19. RF-RAE-RK	0.971
20. SC-CFS-BF	0.977

From the table 6.1 & 6.2 we can understand these classifiers J48, DT, JRIP, DTNB, NBT, ID3,REPT, SC produce good result for both accuracy and f1-measure but three classifiers namely NB, LB and MCC produce low f1-measure (0.481).
The output of F1-Measure is given below

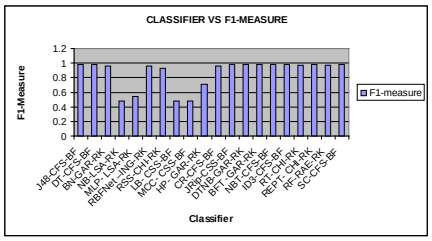

Fig: 6.3 Chart for Classifier Vs F1-Measure

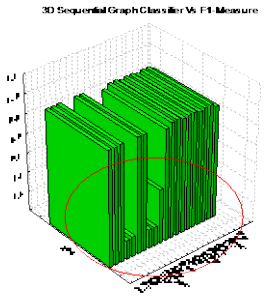

Fig: 6.4 3D ranges Graph for Classifier Vs F1-Measure

ROC:

It illustrates the performance of a binary classifier system as its discrimination threshold is varied. In statistics Roc is a graphical plot this curve is created by plotting the true positive rate against the false positive rate at various threshold settings, where

TP (True Positive) is recall and
FP (False Positive) is fall-out.

P_1 (T) = Threshold parameter belongs to the class.
P_0 (T) = Threshold Parameter not belonging to the class

$$FPR(T) = \int_{T}^{\infty} P_0(T)\, dT$$

In False Positive Rate according to the varying parameter threshold parameter do not belong to the class

$$TPR(T) = \int_{T}^{\infty} P_1(T)\, dT$$

In True Positive Rate according to the varying parameter the threshold parameter belong to the class.
ROC is parametrically TPR (T) versus FPR (T) with T as varying parameter

Table: 6.3 Classifier Vs ROC

CLASSIFIER NAME	ROC
1. J48-CFS-BF	0.987
2. DT-CFS-BF	0.987
3. BN-GAR-RK	0.99
4. NB-LSA-RK	0.495
5. MLP- LSA-RK	0.501
6. RBFNet-ING-RK	0.99
7. RSS-CHI-RK	0.989
8. LB- CSS-BF	0.495
9. MCC- CSS-BF	0.495
10. HP- GAR-RK	0.676
11. CR-CFS-BF	0.967
12. JRip-CSS-BF	0.987
13. DTNB-GAR-RK	0.989
14. BFT -GAR-RK	0.981
15. NBT-CFS-BF	0.987
16. ID3-CFS-BF	0.987
17. RT-CHI-RK	0.979

18. REPT- CHI-RK	0.987
19. RF-RAE-RK	0.988
20. SC-CFS-BF	0.97

Here the peak value is 0.99 produced by two classifiers namely BN and RBFNet but NB, LB and MCC produce very low ROC value (i.e.,) 0.495 .By comparing table 6.2 & 6.3 we can understand NB,LB and MCC classifiers produce low output for both ROC and F1 measure. The graphical representation for ROC is given below.

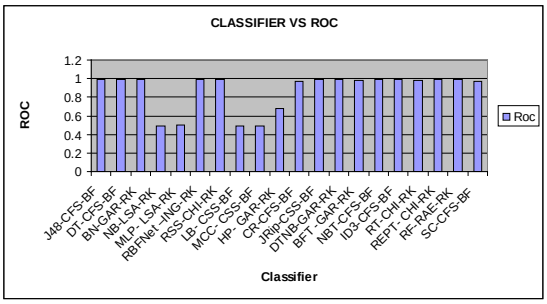

Fig: 6.5 Chart for Classifier Vs ROC

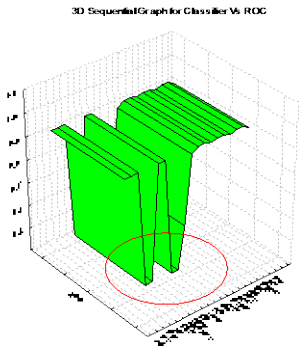

Fig: 6.6 3D Sequential Graph for Classifier Vs ROC

VII.CONCLUSION

Classification is one of the significant functions of data mining which accurately predicts the target class for each case in the data. In our study, we have taken various classification methods and compared the results of various algorithms on the basis of Accuracy, F1-Measure and ROC. According to the table 6.1, 6.2 and 6.3,we observe that J48, DT, JRIP, DTNB, NBT, ID3, REPT, SC produces high Accuracy and F1-Measure values of 97.7% and 0.977,while if we consider ROC 0.99 is the highest value produced by BN and RBFNet algorithm. However from the output of the classification algorithm it follows that rural background students have very low level of communication skill when compared with urban background students. Hence this work suggests that the education Institutions, to focus more on communication skill to get appropriate placement for their students.

References:

[1] Lin Siew Eng, Abdul, Rashid Mohamed and Muhammad Javed., 2013 .Analysis of Student's Competency in Listening Comprehension of the English Language at Pakistani Secondary School Level, PP.331-341.

[2] J. David McCracken and Jeff David T. Barcinas., 1991, Differences Between Rural and Urban Schools, Student Characteristics, and Student Aspirations in Ohio., Vol. 7, No.2, pp. 29-40

[3] Yanfeng Hu , Sep 2009, A Survey on the English Learning Strategy of the Rural High School Students and Urban High School Students,Vol.2,No.3, pp.232-236.

[4] H. Jeraltin Vency and E. Ramganesh , Nov 2013, Is Language Proficiency Taken Care of at Higher Education Level? Need for Self Efficacy of Post Graduate Students., Vol. 4, No. 6, pp. 1176-1183.

[5] Cosby, A. & Picou, J. S. (1973). Structural models and occupational aspirations: Black and white variations among deep-south adolescents. Journal of Vocational Behavior, 3, pp.1-14.

[6] Han, Jiawei and Micheline Kamber. 2001. Data Mining: Concepts and Techniques. Morgan Kaufman Publishers.

[7] Hannaway, J. and Talbert, J.E., 1993. Bringing context into effective school research: Urban-suburban differences.

Educational Administration Quarterly 29, pp. 164–186.
[8] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.
[9] A. Merceron and K. Yacef, "Educational Data Mining: a Case Study,"In *C. Looi; G.* McCalla; B. Bredeweg; J. Breuker, editor, Proceedings of the 12th international Conference on Artificial Intelligence in Education AIED, pp. 467–474. Amsterdam, IOS Press, 2005
[10] Ming-Syan Chen, Jiawei Han and Philip S. Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, Dec. 1996.
[11] Moore, E. J., Baum, E. L., & Glasgow, R. B. (1984, April). Economic factors influencing educational attainment and aspirations of farm youth. Washington, DC: Economic Research Service, Resource Development Division. (ERIC Document Reproduction Service Document No. ED 015797), pp.1-43.
[12] Opdenakker, M. C. and Van Damme, J., 2006. Differences between secondary schools: A study about school context, group composition, school practice, and school effects with pecial attention to public and Catholic schools and types of schools. School Effectiveness and School Improvement 17(1), pp. 87-117.
[13] Stevans, L. K. and Sessions, D. N., 2000. Private/public school choice and student performance revisited. Education Economics 8 (2), pp. 169-184.
[14] Vandernberghe, V. and Robin, S., 2004. Evaluating the effectiveness of private education across countries: A comparison of methods. Labor Economics 11(4), pp. 487-506
[15] I. H. Witten and E. Frank, "Data mining: Practical Machine Learning Tools and Techniques," 2nd ed., Morgan-Kaufman Series of Data Management Systems San Francisco Elsevier, 2005.
[16] Young, D.J., 1998. Rural and urban differences in student achievement in science and mathematics: A multilevel analysis, School Effectiveness and School Improvement 9(4), pp. 386-418.

1. R.Yogatharani is Head Department of CS, SW, M.Sc. (I.T) and M.C.A, Parvathy's Arts and Science College, TamilNadu, India. She obtained her M.Sc. (CS & I.T) from Madurai Kamaraj University, and M.Phil in Computer science from Madurai Kamaraj University. She is a research Scholar in Madurai Kamaraj University.
2. Dr.R.Bhaskaran received his M.Sc. (Mathematics) from IIT, Chennai, India (1973) and obtained his Ph.D. in Mathematics from the University of Madras (1980). He then joined the School of Mathematics, Madurai Kamaraj University as Lecturer and retired in June 2012 as Professor. His research interests include Non-Archimedean analysis, Image Processing, Data Mining and Software development for learning Mathematics.