Predicting Student’s Learning Behavior Prior to University Admission

Manjula V.
Research Scholar,
Department of ISE,
Jain University,Bangalore,India

A. N. Nandakumar, PhD
Professor,
New Horizon College of Engg,
Bangalore ,India

Raunak Mahesh
UG Student
Department of ISE
SET-JU, Bangalore, India

ABSTRACT

Generation of a raw data set incorporating co-related attributes, providing an insight into a student’s personality and academic performance will be our primary agenda. Subsequently, the records in the data set will be grouped into different clusters. Post clustering, each cluster will be assigned a class label considering the overall student performance in that cluster. At this stage, the raw data set is segregated into training and testing data sets. A data model can now be developed as a result of a learning algorithm which will be implemented on the training data set. Succeeding, the developed data model will be evaluated based on accuracy using the testing data set. Finally, the data model would be invoked from MATLAB for predicting a student’s performance (given all the attributes).

Keywords

Educational Data mining, EM luster, Filtered Clustered, SimpleKMeans, classification

1. INTRODUCTION

A “cause-effect” relationship is satisfied if a cause or a set of causes results in an effect. Student performance would be a great example to illustrate a cause-effect relationship. To begin with, analysis of the student performance (effect) leads us to an exhaustive set of possible causes.

Based on the resulting effects they bear, this set of causes can be broadly classified into ‘direct’ and ‘indirect’ causes. In-depth understanding of each of these causes will surely enable us to predict all possible outcomes with regard to a student’s performance.

Direct causes are the causes that are zeroed on immediately upon receiving the performance report of a student. The professor, course and the institution are the most common examples of this category of causes. For example, consider the performance report of a student who has not fared well in an examination. Spontaneously, parents conclude that either the complexity of the course or the accent of the professor or the teaching in that institution as the possible cause which has resulted in such a performance. This hasty conclusion by the parents is true only for a miniscule number of cases. Furthermore, false findings such as these do tarnish the reputation of the institution.

Indirect causes are the causes that go mostly unnoticed during the performance report scrutiny by the parents. The various roles and responsibilities carried out by the student along with his/her personality are the best illustrations for indirect causes. Consider the following example of the analysis of a good performance report of a student. Parents usually tend to praise the above mentioned direct causes and totally neglect the indirect causes. This leads to under-performance of even the good performers; reducing the duration of the role of a friend that a student plays, for most part of the day, implies that he will spend less time with his/her friends and utilize the same towards the betterment of his performance. Understanding the extent of student’s involvement in these roles and responsibilities will go a long way in facilitating a better understanding of his/her performance. Our work will based towards deriving all possible indirect causes and trying to figure out the degree to which these causes will be affecting the performance of a student.

Measuring of academic performance of student is a challenging task because student’s performance is based on different factors such as their understanding levels, capacity to learn, ability to perform well in exam and so on. So the scope of the research is always is there to find out what are the factors that affect performance of the students [1]

The main aim of the educational institutes is to provide quality education to its students and to improve the quality of managerial decisions [6].

2. BACK GROUND AND RELATED WORK

Identification of the attributes is a vital and initial part of data set design. The numerous attributes incorporated within our proposed system would comprise of the set of both the direct and indirect causes along with various academic performance related statistics such as SSC, HSC grades in addition to the other features which would provide basic details regarding a student’s demography [5]. Additionally, certain details such as library book usage pattern, attendance etcetera too can be incorporated as attributes. Hence, a data set described by ‘m’ attributes and ‘n’ records would be our raw data set.

The records in the raw data set will then undergo data pre-processing following which, the various records will then be subjected to clustering. In this step, records having different attribute values will be grouped under various clusters wherein each cluster will be mapped onto a particular class label. The data set at this stage will be split into two smaller data sets – the training data set and the testing data set. The training data set will be used for machine learning along with learning algorithm. Machine learning will be carried out using the WEKA tool which has its own default learning algorithms [2]. The learning algorithms likely to be used are the default classifiers available in WEKA tool. Post classification, the data model would be have been developed. This data model will be capable of predicting a student’s performance given all his/her attributes.

The final step before deploying the data model for prediction is to test the developed data model using the testing data set. At this juncture, the data model is primarily evaluated based on its percentage of accuracy. After the desired level of accuracy has been obtained by the data model, it will then be deployed for real-time prediction of the performance of a student prior to his/her admission to the institution. The main end-user for our proposed system will be the admissions officer, who would be given a easy-to-use graphical user interface (GUI) environment like MATLAB through which he will have to input the attributes of the student who wishes to seek admission to the institution. Post the classification of the input attributes, the predicted performance of the student will be displayed to the admissions officer who can then make a decision, considering the quality of the student, whether to grant or deny the admission for that student[3][4]. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide proper advising [7],[8].

3. PROPOSED WORK

Often, parents and guardians fail to visualize all the causes that result in their wards performance. Additionally, the incorrect causes identified by the parents' results in bringing disrepute to the professor, course and the institution. On the other hand, admitting students who can possibly bring disrepute to the institution through their under-performance can result in a potential degrade of the quality of the students trying to seek admission and may even result in potentially disastrous consequences for the future of the institution.

It is not just the institution which may suffer due to the existing ways and means of interpreting a student's performance. Improper interpretation of the performance could even possibly end a student's career even before it starts. For example, if the parents of an under-performing student keep pin-pointing to the direct causes, such as the professor, course or the institution, as the possible causes resulting in the student's under-performance whilst in reality an unnoticed indirect cause such as drug addiction may have had an effect on his/her performance all this while. This also results in improper behavior of the student.

4. RESULTS AND DISCUSSION

The primary and the most crucial step of any data mining task is the collection or gathering of data. Owing to the security and legal restrictions that real – time data of the students who were enrolled in a university possessed, we were not given the access to it. Hence, we had to find an alternative way of recreating or simulating real – time student data. We did so with the help of various free online data set generators which helped us generate our own data set which closely resembled real – time student data. This data set contains thirty three thoughtfully chosen attributes for a thousand records or instances.

The details of each of these thirty three attributes are described below.

1. School medium - the language of instruction in the school where the student studied which was either Kannada or English.
2. pumedium - the language of instruction in the pre – university course where the student studied which was either Kannada or English.
3. nativeplace - place from where a student hails from which could either be a rural or an urban area.
4. gender – the student's gender.
5. familysize – the size of the student's family which could be categorized as either greater than three or less than and equal to three.
6. parentstatus – the student's parents cohabitation status.
7. motherseducation - the student's mother's education (numeric: 1 - none, 2 - primary education (4th grade), 3 – 5th to 9th grade, 4 – secondary education or 5 – higher education).
8. fatherseducation - the student's father's education (numeric: 1 - none, 2 - primary education (4th grade), 3 – 5th to 9th grade, 4 – secondary education or 5 – higher education).
9. mothersjob – the student's mother's occupation (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other").
10. fathersjob – the student's father's occupation (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other").
11. reasontochoose - reason to choose this university (nominal: close to "home", school "reputation", "course" preference or "friends" or "other").
12. traveltime - home to university travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, 4 - >1 hour, or 5 -> 2 hours).
13. weeklystudytime – the number of hours in a week that the student utilizes for studying (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, 4 - >10 hours, or 5 -> 20 hours).
14. Gaps in education -years of gaps in education owing to failure in studies(binary: yes / no).
15. Special classes -extra educational support by means of additional coaching (binary: yes or no).
16. Family support – the support extended by the student's family towards his / her studies (binary: yes or no).
17. Extra activities – the involvement of a student in extra-curricular activities (binary: yes or no).
18. higherstudies – the student's willingness to pursue higher education (binary: yes or no).
19. internetaccess – the time spent by the student on the internet for studies (binary: yes or no).
20. relationships – the student's involvement in a relationship with the opposite sex (binary: yes or no).
21. familybond – the quality of a student's family relationships (numeric: from 1 - very bad to 5 - excellent).
22. friendshangout - the time spent by a student in hanging out with friends (numeric: from 1 - very low to 5 - very high).
23. dailyalcohol – the amount of alcohol consumed on weekdays by a student on a daily basis (numeric: from 1 - very low to 5 - very high).
24. weeklyalcohol - the amount of alcohol consumed on weekends by a student (numeric: from 1 - very low to 5 - very high).
25. currenthealth – the current health condition of the student (numeric: from 1 - very bad to 5 - very good).
26. 7thstdgrades – the seventh standard CGPA of the student (numeric: from 0 to 10).

27. 10thstdgrades – the tenth standard CGPA of the student (numeric: from 0 to 10).

28. 12thstdgrades - the twelfth standard CGPA of the student (numeric: from 0 to 10).

29. fathersincomestatus – the income of the student's father (numeric: from 1 - very low to 5 - very high).

30. mothersincomestatus - the income of the student's mother (numeric: from 1 - very low to 5 - very high).

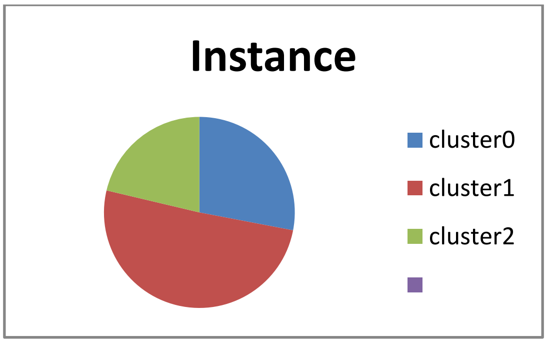
31. tvviewinghours – the time spent by the student in front of the television (numeric: from 1 - very low to 5 - very high).

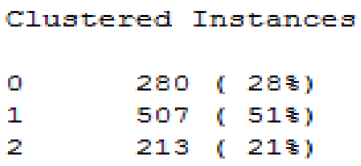
32. socialnetworkingactivenss – the engagement of the student in social networking (numeric: from 1 - very low to 5 - very high).

33. drugabuse – the student being subjected to or indulged in drug abuse (numeric: from 1 - never to 5 - very high).

The records in the data set needed to be clustered in order to be assigned class labels. Thus, we carried out the process of clustering using all the readily available clusters in Weka machine learning toolkit. The detailed description of the clustering process is provided below.

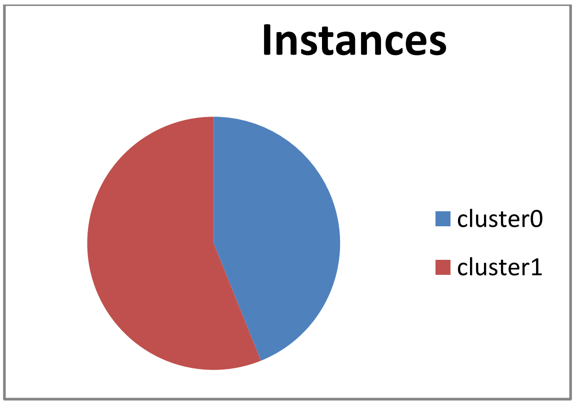
1. EM cluster :

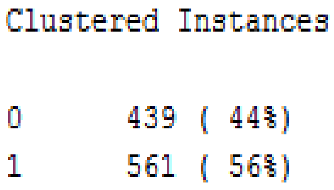




```
Clustered Instances

0       280 ( 28%)
1       507 ( 51%)
2       213 ( 21%)
```

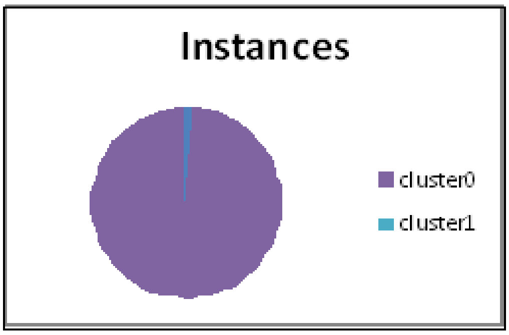
2. FilteredClusterer : (Using 'SimpleKMeans' clusterer and 'AllFilter' option)

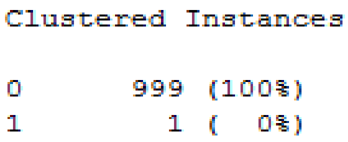




```
Clustered Instances

0       439 ( 44%)
1       561 ( 56%)
```

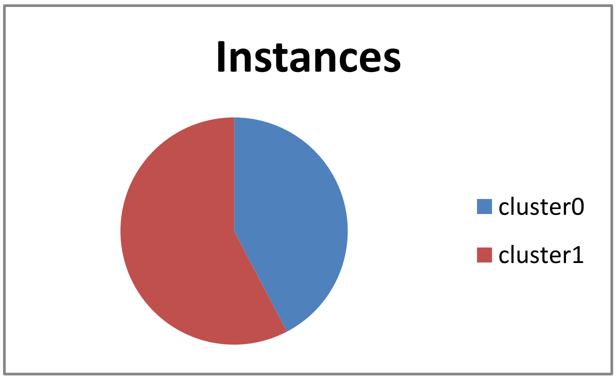
3. Hierarchical Clusterer : (Default cluster size used is 2)

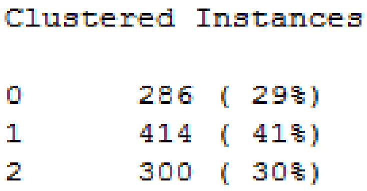




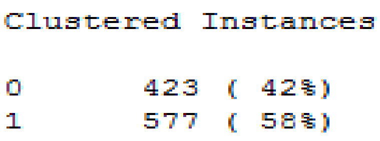
```
Clustered Instances

0       999 (100%)
1         1 (  0%)
```

4. MakeDensityBasedClusterer : (Using'SimpleKMeans'clusterer and '1.0E-6'minStdDev)



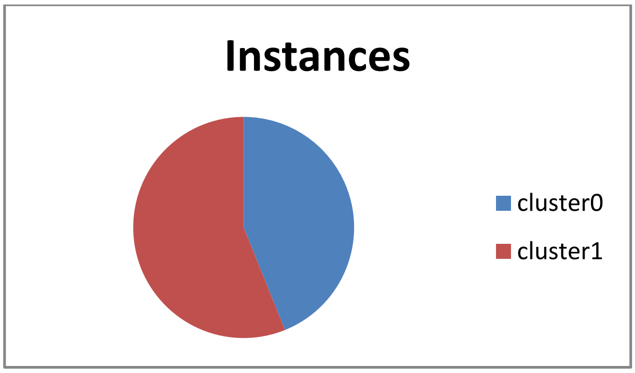


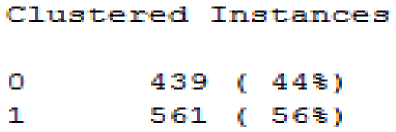
```
Clustered Instances

0       286 ( 29%)
1       414 ( 41%)
2       300 ( 30%)
```



```
Clustered Instances

0       423 ( 42%)
1       577 ( 58%)
```

5. SimpleKMeans:(Default cluster size used is 2)

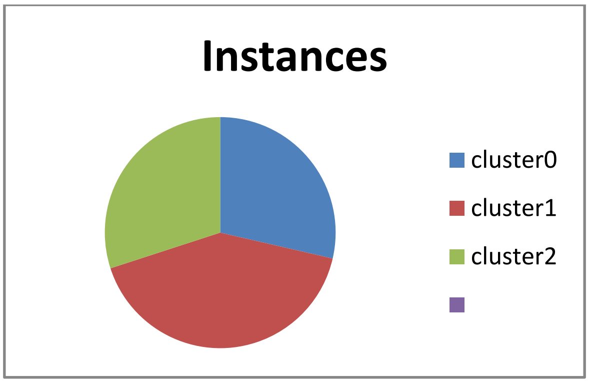




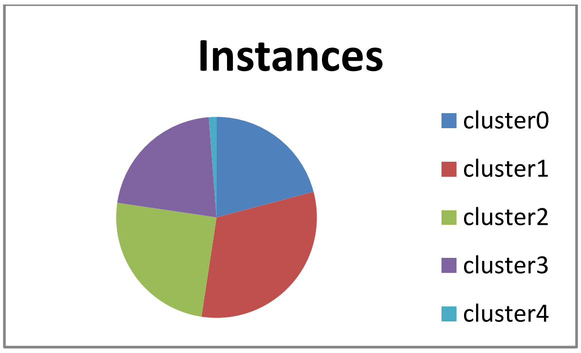
```
Clustered Instances

0       439 ( 44%)
1       561 ( 56%)
```

After performing clustering initially, it was observed that SimpleKMeans cluster gave an equal percentage of split records. Hence, for various numbers of clusters the SimpleKMeans cluster was run and the results of which are described below.

- When the number of clusters was increased to three, the clustered instances were :



- When the number of clusters was further increased to five, the clustered instances were :



Further, the cluster assignments were visualized and the clustered that was decided upon was SimpleK Means with the number of clusters being decided as two. After clustering the training data set, the class labels were assigned to the data set and then it was subjected to task of classification in order to build a data model prior to running the test data set on the data model.

5. CONCLUSION

Summarizing the work carried out in the regard of developing a student data model to predict whether a student seeking admission to a university needs to be admitted or not has its own fair share of positives and limitations. The limitations were however a point of discussion prior to taking up this work. It was then decided that every proposed work has it share of limitations but until the positive outcomes outweigh the limitations the work is liable to a successful completion. In this regard we have put across the forseen limitations of our work.

- The use of certain attributes in our data set was unnecessary as was concluded post the completion of our work.
- Not all aspects affecting a student's performance were considered as potential attributes in the dataset.
- The details as provided by the student tends to be unvalidated and inaccurate. This was considered as a major limitation.

Our work does in no way signify a potential impossibility to any related work that could be carried out upon ours. Hence, we have come up with the set of future work that would increase and enhance the scope of our project.

- The work carried out by us could serve as a potential foundation work for people wanting to pursue a Doctor of Philosophy (Ph.D) degree in educational data mining.
- The tools used by us could be worked around and different or equivalent tools could be accomplished to achieve the same set of objectives.
- We do hope to see a full scale real - time implementation of our work in universities.

6. REFERENCES

[1] Hijazi and Naïve, "Factors Affecting Students Performance" Bangladesh e-Journal of Sociology, Volume 3. Number 1, January 2006.

[2] Weka, University of Waikato, New Zealand, http//www.cs.waikato.ac.nz/ml/weka/.

[3] R. Kohavi and F. Provost, Glossary of Terms, in Spec. Issue on Apps of Machine Learning and the KDD Process, Machine Learning Journal, 30, pp. 271-274. Kluwer. 1998.

[4] I.H. Witten and E. Frank, Data Mining Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann. 2000.

[5] K. Prasad Rao and M.V.P Chandrashekar Rao, Predicting Learning Behavior of Students Using Classification Techniques. Volume 139-No 7,April 2016.

[6] P.V. Praveen Sundar ,A Comparative study for Predicting Students Academic Performance. Volume 3, Feb 2013.

[7] Baradwaj, B, and Pal,S. 'Mining Educational data to Analyze Student's perfrormance.Vol 2, no.6, 2011

[8] Dr. T.N Manjunath and Ravindra S [2012] Realistic Analysis of Data ware housing and Datamining Application in Education Domain. International Journal of Machine learning and computing Vol.2 No.4 August 2012