Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp. 8258-8267
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.04.765214

Algorithms For Predicting Student Dropouts With Feature Selection

Ubhaida Aslam P.hd scholar Department of computer Applications and Sciences, Sarvepalli RadhakrishnanUniversity, Bhopal, MP, INDIA. (Email id:uaslambhat@gmail.com)

Dr. Ravindra Kumar Gupta: - Associate professor, Sarvepalli Radhakrishnan University, Bhopal, MP, INDIA.

ABSTRACT

Proof that a feature is applicable has become a prerequisite for using data mining calculations successfully in real-world contexts. In order to achieve the relevant feature subsets in the writing's order and grouping goals, numerous feature selection approaches have been offered. The concepts of feature pertinence, general strategies, assessment standards, and the qualities of feature selection are presented in this work. Last but not least, the chi square test will be used in feature selection calculations to predict school dropouts. This paper's goal is to find comparable instances of usage in the data compiled from datasets so that, in the end, we may make predictions for each student based on various segment, scholastic, and point-of-view features. In conclusion, information gleaned from the study may shed light on how to support students who are in risk even more effectively. We will wrap up this work with real-world applications (such early student dropout prediction), challenges, and potential directions for future study.

KEYWORDS:- Feature Selection, Filter method, Student dropout, Data mining.

ABBREVIATIONS:- FS, Feature Selection ; FM, filter method SD, Student Dropout

1. INTRODUCTION

The abundance of data in modern datasets necessitates the progress of acute computations for the discovery of significant data. Data models are created using data mining tasks, but typically in the areas of organization, recurrence, and grouping. Preparing the datasets in advance occurs frequently for two main reasons:

1) Reducing the dataset size to carry out an increasingly effective investigation, and
2) Adapting the dataset to the selected research method [1]

There are two important approaches to handle feature selection, and the following

explanation is becoming more and more important these days. Let me continue my investigation on this, There are two significant ways to deal with feature selection.

The first is Individual Evaluation, and the second is Subset Evaluation. Positioning of the features is known as Individual Evaluation . In Individual Evaluation, the heaviness of an individual feature is doled out as indicated by its level of significance.[3]

In Subset Evaluation, up-and-comer feature subsets are developed utilizing search methodology. The general methodology for feature selection has four key strides as appeared in Figure 1.

- Feature Subset Generation
- Evaluation of Feature Subset
- Stopping Criteria
- Result Validation

  Subset creation is a heuristic endeavour in which each state selects a rival subset for evaluation in the research domain. The concept of the subset generation process is determined by two key factors.

First off, the replacement generation decides on the inquiry course by selecting the pursue beginning stage. Forward, in reverse, compound, weighted, and arbitrary approaches may be taken into consideration to select the inquiry's initial steps at each state.

Second, the feature selection technique is accountable for scan association with a certain method, such as sequential inquiry, exponential hunt [6] or irregular pursuit.

A recently created subset needs to be evaluated according to certain evaluation criteria. The integrity of the up-and-coming subset of the features has thus been determined by a number of assessment measures that have been put out in the writing. Assessment rules can be divided into two groups based on how heavily they rely on mining algorithms: free and ward measures. Subordinate measures use predetermined mining algorithms for feature selection to choose features based on the presentation of the mining algorithm applied to the chosen subset of features, while free measures make use of the fundamental attributes of the preparation data without incorporating any mining algorithms to evaluate the integrity of a feature set or feature [7].

Finally, stop standards must be decided upon in order to terminate the selection process. The process of feature selection ends with the approval method. Finally, stop standards must be decided upon in order to terminate the selection process. The process of feature selection ends with the approval method. Although it isn't a part of the feature selection process, the method for selecting features must be approved by passing numerous tests and correlations with previously decided outcomes or real-world datasets, or both.

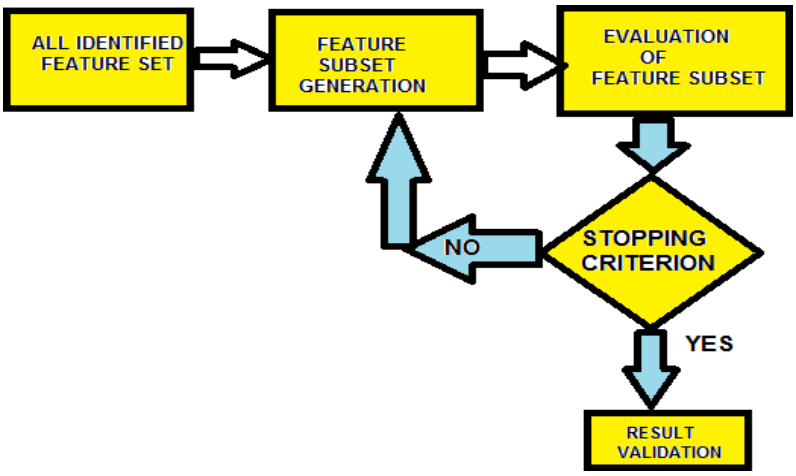


Figure 1:- Process of Feature Selection[2].

For feature selection, there are three general methods. The Filter Method first takes advantage of the general characteristics of data preparation without the need of a mining algorithm.

The Wrapper Approach also looks into the relationship between choosing the ideal feature subset and importance. The ideal feature subset tailored to the specific mining algorithm is searched for [9].

Thirdly, the Embedded Approach has a specific learning algorithm that completes feature selection throughout the preparation period.We will use the main strategy and realise its algorithm in this work to forecast student dropout using the chi square test.

2.RELATED WORK

Student dropout prediction via feature selection Although some work has been done on the feature selection algorithm, it is still a fresh commitment to the data mining field to really implement it in predicting early student dropout. Algorithms are generally a new area for the inspection.

Many of the works that have looked at feature selection algorithms are as follows:

El-Halees [4] has organised a transparent contextual analysis that makes use of academic data management to study the altered behaviour of researchers who are learning. His study aims to draw attention to how accommodating data processing would be applied in educational activities to enhance students' performance. They used a call tree, pack, and anomaly investigation to apply info mining techniques to enormous datasets of affiliation rules and grouping rules.

In order to evaluate the student's presentation, Bharadwaj and Pal [5] developed an innovative strategy that makes use of the decision tree methodology for categorising. This contextual analysis aims to identify the information that illustrates students' performance in the final semester evaluation. This investigation was very helpful in identifying students who dropped out at an earlier stage, students who required special consideration from the United Nations office, and it allowed the coach to request prior consideration from the researchers.

Significant contributions in this area have been made by Chang, Verhaegen, and Duflou (2014), Guyon and Elisseeff (2003), Kohavi and John 1997, H. M. Harb and M. A. Moustafa Scheirer, H. Liu and H. Motoda, among others.

2. METHODOLOGY:-

The theory work of the author, which depends on both necessary and ancillary data and data, has occupied this study. Anantnag, which is instructively typical and the second-evolved location in Kashmir division of the state with an education pace of 64.32%, has been selected for the study's focus area in Jammu and Kashmir, India. The information acquired from the area reveals the net enrollment percentage (GER) of pupils up to the twelfth grade, as well as information on sexual orientation and dropout rates.

Following the collection of auxiliary data and data from schools, we have focused a field review of the family unit of drop-out students for the assortment of key data. In order to gather various supplementary data and information, we visited ZONAL EDUCATION OFFICES, various higher optional schools in the Anantnag region of Jammu and Kashmir, and sought advice from various libraries. Also, data from the JKBOSE (Jammu and Kashmir State Leading Body of School Education Kashmir) website, Director school training Kashmir, and published sources on the topic of dropout were used.

Finally, dissuades more reactions were broken down using chi square test of feature selection algorithm on the data gathered to forecast if these features were critical or not. We obtained several reasons why students can't progress with higher education (why they fizzled).

4 FILTER METHODS

Machine learning uses a variety of techniques to determine whether or not our information attributes are relevant to the expected outcome. With the Filter approach, positioning tactics are used as guiding principles.

The factors are given scores using an appropriate positioning basis, and those with scores below a certain edge value are eliminated. These techniques are less expensive computationally, keep a safe distance from over fitting, but filter techniques ignore conditions between the features. As a result, the selected subset is probably not perfect, and an extra subset might be obtained. Below, usage and one of the key filter feature selection

algorithms are studied.

4.1 CHI-SQUARE TEST

Methods for choosing filter features include those that use factual techniques (such as the Chi-Squared test) to evaluate the relationship between the target variable and the data type of the information. Karl Pearson, a mathematician, is honoured by having his name attached to the Chi-Squared test and Pearson's Chi-Squared test. A measurement known as "integrity of fit" is another name for it. A chi-square (2) measurement is a test that determines how preferences compare to actual observed data (or model outcomes). The information used to calculate a chi-square measurement must be unreliable, unpolished, fundamentally unrelated, derived from independent variables, and derived from a large enough example. Chi-square tests are frequently used when assessing theories.

The Chi-Square Equation Is

$$\chi c^2 = \sum Ei(Oi - Ei)^2/Ei \qquad (i)$$

Where:

c=Degree of freedom
O=Observed value(s)
E=Expected value(s)

5. IMPLEMENTATION OF CHI SQUARE TEST ON SCHOOL DATA TO PREDICT STUDENTDROPOUT.

After acquiring optional data and data from schools, we conducted a field evaluation of the family unit of drop-out kids to collect the critical data. In the Jammu and Kashmir region of Anantnag, we visited ZONAL EDUCATION OFFICES, various higher auxiliary schools, and were given library recommendations for the collection of optional data and information. Moreover, statistics from the jkbose website, Director of School Instruction Kashmir, and published sources on the topic of dropout have also been utilised.

When the class twelfth assessment for Jammu and Kashmir state's top group of school training was held in December 2019, we received data on 10364 pupils from the locality/district of Anantnag. Of them, 8342 students have been accepted into government high schools. In the area, there are 42 government higher secondary schools.

Out of 8342 students at government schools, 4589 passed the test (55.01%), while 3753 (44.9%) students failed to do so.We selected data from two higher secondary schools, one from a rural area and the other from an urban area, for our exploration purpose.

The total number of students from these upper optionals who attended class and took the 2019 twelfth assessment is 282.

152 students passed the exam.

130 pupils dropped out or failed.

Also, they won't be able to attend school as a result (advanced education). When we asked kids, parents, teachers, and professors to describe these disappointed students, we received a variety of reactions.

The responses that were produced are listed below (Reason of failure).

- Poverty
- A teacher's
- Unfavourable conditions

However, the majority of these students, guardians, teachers, and academicians were concerned about poverty, poor teacher behaviour, and early marriage in addition to hartal/strikes, careless parenting, parental illness, and orphanages (for example the rate for reactions of poverty, negative behaviour of teacher and early marriage was more when contrasted with rest of the features). Let's have a look at a very simple dataset with only two parts. We'll check to see if the Failure Reason is connected, subordinated, or related to Gender.

Gender	**Reason of Failure**
Male	Poverty
Female	Early Marriage
Male	Poverty
Male	NEGATIVE BEHAVIOUR OF TEACHER
……	**……**

Let's do a hypothesis test, a quantifiable procedure that evaluates two assertions (speculations) and determines which is true.
Let the beginning assertion, the null hypothesis, be denoted as H0, and the exchange hypothesis, which is typically related to the first, be denoted as H1. The theories for our model are:
Gender and Reason for Failure are free features (which implies they are not related).
H1: Reasons for Failure include Cause of Failure and Gender (which implies they are

related). Let α = 0.05
Smaller values, such as those between 0.01 and 0.10, are frequently preferred, where alpha is a measure of centrality that, for instance, indicates how confident we should be in our results.

NOW WE CREATE A CONTINGENCY TABLE

A possibility table is a chart showing the recurrence of one variable in lines and another in segments that is used to investigate the link between the two parameters (otherwise called a cross arrangement or crosstab).

	Poverty	**Early Marriage**	**NEGATIVE BEHAVIOUR OFTEACHER**	**Row total**
Male Female	20 25	20 30	25 10	65 65
Column Total	45	50	35	130

Table 2:- <u>Contingency table</u>

Out of 130 candidates, the dataset reveals that 20, 20, and 25 guys, respectively, have reasons for dissatisfaction with poverty, early marriage, and teacher-bad behaviour. Consequently, 25, 30, and 10 females, respectively, are interested in poverty, early marriage, and teacher behaviour that is not positive. We refer to the values in the table as watched values.

CALCULATE EXPECTED FREQUENCY

We calculate the average recurrence mean for each cell. Expected recurrence is calculated using the formula E = (push complete * segment all out)/grand aggregate.
For the first cell, such as male poverty, the expected recurrence will be
E1 = (65 * 40)/130 = 20.

We calculate the expected frequency for the remaining cells to produce the table below, where values in brackets denote expected frequencies:

	Poverty	**Early Marriage**	**NEGATIVE BEHAVIOUR OF TEACHER**	**Row total**
Male Female	20[22.5] 25[22.5]	20[25] 30[25]	25[17.5] 10[17.5]	65 65
Column Total	45	50	35	130

CALCULATE THE CHI-SQUARE VALUE (CHI-SQUARE STATISTIC)

Chi-square value, or 2, is calculated using the following formula:

$\chi^2 = \sum Ei(Oi-Ei)2/Ei$

This is the Greek alphabet in Chi order, not the English alphabet that we are familiar with. As it initially looked, 2 is the total of the squared difference between the observed and expected frequencies divided by the expected recurrence for all cells. The estimations are shown in the following manner:

χ^2 = ((20-22.5)^2/22.5) + ((20-25)^2/25) + ((25-17.5)^2/17.5) + ((25-22.5)^2/22.5) + ((30-25)^2/25) + ((10-17.5)^2/17.5)

$\chi^2 = 0.277 + 0 + 3.214 + 0.277 + 0 + 3.214$ $\chi^2 = 6.982$

CALCULATE DEGREES OF FREEDOM

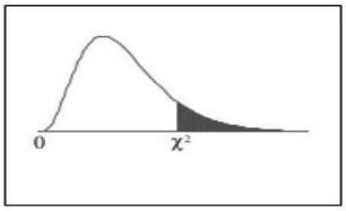
In order to calculate the degrees of opportunity, use the formula:

df= (total rows-1) * (total cols-1).

FIND P-VALUE

The distribution tables for Chi Square are visible.

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766

Table 4 :- chi square Distribution Table (which was used to calculate p value)

Using the Chi Square and degrees of opportunity esteems, find the p-value.

The degrees of chance esteem (2) on the left track with its column to the closest number to the Chi-Square value (6.982), and then verify the corresponding number in the main line to acquire the p-esteem, which is 0.025. Although there are several websites that calculate the p-value, we used the above table to determine the p value.

6. RESULTS AND DISCUSSIONS:-

Simply put, we reject the null hypothesis if our p value is less than or equal to the essentialness value and we do not reject it if our p value is greater than or equal to the centrality value.

We reject the null hypothesis, which suggests that there is a relationship between gender and the reason for disappointment, because 0.025 is not quite our essentialness threshold of 0.05.These two characteristics are therefore sufficient to identify student dropouts.

7. CONFLICT OF INTEREST:-

The authors state that there are no requirements beyond reconciliation with the remainder of the research.

8. REFERENCES:-

Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, "Order and prediction based data mining algorithms to anticipate moderate students in training segment", 3rd Int. Conf. on Recent Trends in Computing, Vol 57, 2015,pp. 500-508.

[1] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," in: C. Aggarwal (ed.),
Data Classification: Algorithms and Applications. CRC Press, 2014

[3]. A. Mueen, B. Zafar, and U. Manzoor, Demonstrating and Predicting Students' Academic Performance UsingData Mining Techniques,
International Journal of Modern Education and Computer Science, 8:36, 2016.

[4]. Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.

[5]. Bharadwaj B.K. and Pal S. "Mining Educational information/data to Analyze Students? Performance",International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69.

[6] E. Scheirer and M. Slaney, "Construction and evaluation of a powerful /robust multifeature music/speechdiscriminator," in Proc. ICASSP' 97, Apr. 1997, vol. II, pp. 1331–1334.

7] Daily News and Analysis, India, online:
http://www.dnaindia.com/india/report_rte-report-carddropout-rate-in-schools-falls_1669959 ,, April 1 (2012**)** accessed on October 30, 2012

[8] Study on Feature Selection Methods/Techniques in Educational Data Mining, M.
Ramaswami and R.Bhaskaran,Vol1,Dec,2009.

[9] "feature selection for high-dimensional data: a fast correlation-based filter solution", lei yu and huan liu.

[10] "A review of feature selectionmethods/techniques in bioinformatics", yvan saey, in aki inza and pedrolarran aga. vol. 23 no. 19 2007, pages2507–2517