A Predictive Student Performance Analytics Scheme using Auto-Adjust Apriori Algorithm

Himanshu Maniar
Shri C.U.Shah College of Commerce,
Management and Computer Education
Wadhwan, Gujarat, 363035 India

S. O. Khanna, PhD
DEAN SEAS
Rai University, Dholka,
Ahmedabad, Gujarat, 382260,
India

ABSTRACT

Every academic organization needs to analyze student performance to find its overall strengths and weaknesses. At the same time, analysis helps to find out strengths and weaknesses of students along with their interests and dislikes. Any large organization with a large number of students has a large amount of result data. This data needs to be processed to find information related to student's performance. This paper presents Auto Adjust Apriori based student's results analysis scheme to predicate student's future performance. In any course, certain courses are interrelated with each other. Using this scheme, students and teachers can able to find which subjects will be more difficult in future based on student's performance in current subjects. The scheme has been implemented under .Net technology.

Keywords

Data Mining, Apriori Algorithm, DIKW

1. INTRODUCTION

Data mining is a set of techniques to find hidden information from the data. In the era of computerization, every organization, firm and individual maintains data. An academic organization maintains students' attendance, performance, subjects' data. A bank maintains customer, account, loan data. An Internet Service Provider maintains log records of all their Internet customers. A Cellular company maintains data of users and their usage. From a scientist to a student, From World Wide Web to a small business, Data is maintained everywhere. Data is indeed an important factor of computerization but even more important is to retrieve some meaningful information out of it. Just like food ingredients need to be cooked to get a dish, Data needs to be processed to find meaningful information. Every computerization is directly or indirectly based on DIKW pyramid as shown in Figure 1. In DIKW pyramid, information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge. The pyramid shown in Figure 1 is self-explanatory [1][2][3].

For any academic organization, it is necessary to analyze student's performance to identify the strengths and weaknesses of the faculties as well as of the students. This paper uses Apriori algorithm to predicate student's future performance. The main goal of this system is to alert students for those subjects which might be difficult for them well in advance. Every course has a certain dependencies in the subjects. Such dependencies are useful to predicate student's performance. Relating to the DIKW Pyramid, Data is the student's performance so far. Information is the detail of student's weak subjects as per performance or as per interest. Knowledge is the detail of finding ways to improve student's performance and wisdom is the logic which students gain to get succeed in making the predication fail [1][2][3].

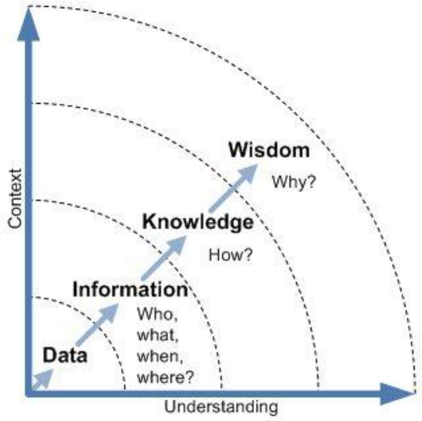


Figure 1.DIKW Pyramid

This paper is describing the system up to the transformation of data to the information. Section II describes data mining concepts, Section III describes Apriori algorithm and Section VI describes how Apriori algorithm is used for student's performance analysis [1][2][3].

2. DATA MINING CONCEPTS

2.1 Data Mining

A huge amount of data is available in the computerized world. This data has no use until converted into information. It is required to analyze this data and extract useful, hidden and unknown information from it. Data mining is the process to gain information from a specific data. Related to data mining, many other processes work together which are: Data Cleaning, Integration, Transformation, Pattern Reorganization etc [2][3].

Data mining is used today by almost every organization. With the growing computerization, the level of data mining is increasing day by day rapidly. Some of the most widely applications where data mining is used are listed below[3].

1. Financial Analytics related to Stock market, Bank transactions, Loans etc.

2. Customer Analytics related to buying patterns, recent trends, influence of other companies, satisfaction survey etc.

3. Security Analytics related to analyzing computers for possible intrusions, viruses etc.

4. Web Analytics related to users' pattern of visiting websites, page rank algorithms etc.

5. Emotional and Social Analytics related to process people's thoughts and to find physiological issues.

The main goal of this paper is to introduce data mining for the analysis of students' performance. Many researchers have designed algorithms for the same purpose and many academic organizations have adopted them too. The main focus given so far is to analyze results to decide what was happened in past examinations. This paper focuses on what was happened in past examinations and how it will affect future examinations. This system is applicable at individual level as every student has a different future than rest of the students. But the same system could be used along with data warehouse's cube concept for higher dimensional view. Data mining has various sub branches like association rule mining, classification, clustering etc. This system uses association rule mining based Apriori algorithm.

2.2 Association Rule Mining:-

Association Rule Mining could be best explained with an example of Market Basket Analysis. Lets take an example of a shopping center. By analyzing customers' buying patterns, it could be noticed that those customers who buy breads, mostly buy butter also. This information could be used by the owner for advertisement or sales purpose. Shopping center could arrange items near to each other or they announce sale related schemes over a bunch of related items. Association rule mining is all about finding such association among a set of data items.

Association rule is of the form X → Y where X and Y are item sets having no items in common. Lets take X = {bread, butter} and Y = {Jam}. The Association rule X → Y describes that the customer buying Bread and butter is more likely to buy jam also.

An Association rule has two rule evaluation metrics: Support (S) and Confidence (C).

Support is percentage of transactions containing both X and Y which is ratio of total number of transactions containing both X and Y by total number of transactions. Confidence is percentage of transactions X that also contains Y. Confidence is ratio of total number of transactions containing both X and Y by total number of transactions containing X. Those association rules which satisfy minimum support and minimum confidence are called strong association rules.

3. APRIORI ALGORITHM AND LINEAR REGRESSION

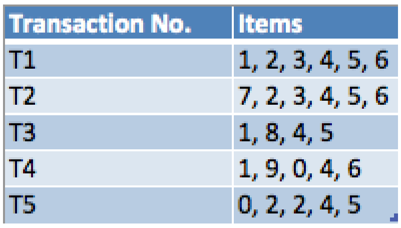
3.1 Apriori

Apriori algorithm can be used for mining frequent item sets to build association rules. Apriori has an iterative structure also called level wise searching. Here k – item sets are used to find k+1 item sets at next level. First, the set of frequent 1-itemsets is found by tracing the available database to find those items whose count satisfies minimum support. This set is called L1. L1 set is used to find L2 – The set of frequent item sets having 2 numbers of items. L2 is used to find L3. The process continues until no more frequent LK is possible. To find any Li, full trace of database is needed[4][5].

If an item set I belonging to set of item sets at level K has not count satisfying minimum support , it is not a frequent item set. So for being a frequent item set, its probability must be greater than or equal to minimum support. It is nature that if I is not a frequent item set, no super set of it could be a frequent item set. This property is called antimonotinicty and used to terminate iteration too[6][7][8].

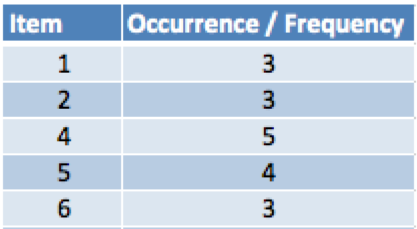
Example

Lets understand Apriori algorithm with an example. Here is a list of transactions listing the lists of items (0,1,2,3,4,5,6,7,8,9) associated with each of them. Support is 3.



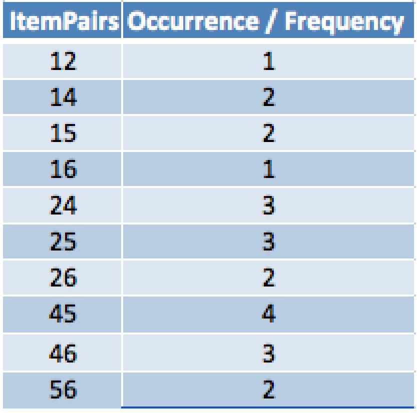
Transaction No.	Items
T1	1, 2, 3, 4, 5, 6
T2	7, 2, 3, 4, 5, 6
T3	1, 8, 4, 5
T4	1, 9, 0, 4, 6
T5	0, 2, 2, 4, 5

Step 1:- Count frequency of each of the items. Minimum support required is 3. So the list of frequent items in item set L1 is as below.



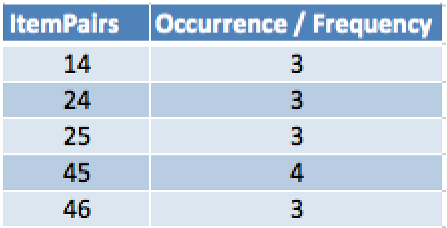
Item	Occurrence / Frequency
1	3
2	3
4	5
5	4
6	3

Step 2:- Based on antimonotinicty property, only those items which are in L1 could be in one of the item sets of L2. So in next step, for every possible pair of items in L1, frequency is counted. The possible item pairs are as below.



ItemPairs	Occurrence / Frequency
12	1
14	2
15	2
16	1
24	3
25	3
26	2
45	4
46	3
56	2

Step 3:- After discarding those item pairs for which frequency is less than minimum support, we get next level of frequent items sets which is L2.



ItemPairs	Occurrence / Frequency
14	3
24	3
25	3
45	4
46	3

Step 4:- The next step calculates L3 by making triplet combinations from the L2 and counting frequencies for deciding those item sets which are having count greater than or equal to support. Here item sets 124 and 125 have count 1 and 0 respectively. At the same time, 245 and 256 have count 3 and 2 respectively so only 245 will be in L3.

Step 5:- As there is no item set in L4, the process could be terminated by announcing frequent item sets as below.

- Frequent Item sets of Size 1:L1:- 1, 2, 4, 5, 6
- Frequent Item sets of Size 2:L2:- 1 4, 2 4, 2 5, 4 5, 4 6
- Frequent Item sets of Size 3:L3:- 2 4 5

3.2 Linear Regression

Linear regression can be used to predicate value of an unknown variable from the known value of another variable. Both variables must be related statistically. For example the relationship between height and weight, relationship between age and maturity could be analyzed using regression analysis. , if X and Y are variables related to each other statistically, then linear regression can be used to predict value of Y given value of X or vice versa[7][8].

In linear regression, one variable is independent and one variable is dependent. We need to predicate value of a dependent variable from value of an independent variable [7].

4. STUDENT RESULT ANALYTICS SCHEME

4.1 Frequent Subject Sets Identification

Apriori algorithm is used to find out the frequent subject sets. A subject set is a set of subjects which are interrelated with each other. Lets say: Set S1 = {C Programming, C++ Programming} considering C Programming in current semester and C++ Programming in next semester. Set S1 indicates that those students who fail in C Programming subject are more likely to be failed in C++ Programming also.

Apriori algorithm, based on dynamic and auto adjusted minimum support threshold value gives a list of frequent subject sets. These subject sets could be used to identify related subjects as per results dependencies. Student can use this information to identify which current semester subject could affect most in next semester subject. If a student dislikes a subject X, he can find out the other subject Y in which almost equally backlogs occurred in past. Student can imagine that if he will not able to clear X in this semester, there is a possibility that he will not able to clear Y in next semester also.

The main goal here is to find such subject sets with automation based on past history with dynamic support calculation. We have taken following parameters into consideration:

N is the total number of students

M is the total number of passed students

N-M is the total number of students who have backlogs.

S is total number of subjects in distinct backlogs list.

In above situation, we will have a list of backlogs(student wise) which will be of size N-M records.

Apriori algorithm needs a minimum support threshold value for calculation of frequent item sets. Here we used auto adjust apriori algorithm to calculate minimum support threshold as per the data.

T1 is set to Percentage of failure which is dynamically calculated using N and M by formula,

$T1 = 100 * (N-M) / N.$ (1)

T2 is set to smoothing factor based on total number of distinct subjects in list of backlogs. The more subjects in backlogs, more smoothing should be provided.

$T2 = S * [(LOG(N-M) * T1) / 100]$ (2)

Minimum Support Threshold Min_Sup_Threshold is calculated as per below formula.

$Min\_Sup\_Threshold = T1 – T2$ (3)

Following table shows some of the Min_Sup_Threshold values for given input parameters.

Table-1 Support Calculation

SR	N	N-M	S	T1	T2	Min_Sup_Threshold
1	45	0	5	0	NAN	NAN
2	45	5	5	11.11	0.4	10.71
3	45	10	5	22.22	1.1	21.12
4	45	15	5	33.33	1.95	31.38
5	45	20	5	44.44	2.9	41.54
6	45	25	5	55.56	3.9	51.66
7	45	30	5	66.67	4.9	61.77
8	45	35	5	77.78	6	71.78
9	45	40	5	88.89	7.1	81.79
10	45	45	5	100	8.25	91.75
11	45	0	7	0	NAN	NAN
12	45	5	7	11.11	0.56	10.55
13	45	10	7	22.22	1.54	20.68
14	45	15	7	33.33	2.73	30.6
15	45	20	7	44.44	4.06	40.38
16	45	25	7	55.56	5.46	50.1
17	45	30	7	66.67	6.86	59.81
18	45	35	7	77.78	8.4	69.38
19	45	40	7	88.89	9.94	78.95
20	45	45	7	100	11.55	88.45

Below is the list of observations which are sufficient to discuss strength of this dynamic minimum threshold of support calculation.

Table 1.2 Observations

Case Sr.	Observation
1 11	Here no student is having backlog in any of the subjects. In such case, there is no need of this system at all. Here Min_Sup_Threshold will be an NaN – Representing Not a Number Case.
2 to 10 and 11 to 19	These are the cases where some students have backlogs. As the number of students having backlogs increases, total number of backlog records in database increases. We can see easily that as per table 1.1, values of Min_Sup_Threshold are also increasing as number of students having backlogs is increased. The general logic is satisfied statistically here too.
10 20	These are cases completely opposite to cases 1 and 11 where all students are having backlog. Min_Sup_Threshold is highest in such cases.
2 12	The only difference between case 2 and case 12 is the total number of distinct subjects involved in backlogs. Case 2 has 5 subjects while case 12 has 7 subjects. It is natural logic that if the number of possible subjects is large, the backlog distribution will see more combinations. In such case, even if the total number of records is same, Min_Sup_Threshold should be lower to compensate increase in number of subjects. This is clearly visible for all such cases.

5. CONCLUSION

The system has been tested over a real time data of a batch of BCA. At the same time, the system was tried to formed using intuition and prediction of faculties. Each faculty was assigned task to find list of frequent subject sets based on their expertise without considering any of the past results. The faculties could identify 5 most frequent subject sets on average per faculty. The system could identify 7 most frequent subject sets on average per batch. The reason behind improved performance is faculties think from technical point of view only. This system works from technical point of view as well as other parameters which affected overall performance.

6. REFERENCES

[1] Xingquan Zhu, Ian Davidson, “Knowledge Discovery and Data Mining: Challenges and Realities”, ISBN 978- 1-59904-252, Hershey, New York, 2007.

[2] H. Johan, B. Bart and V. Jan, “Using Rule Extraction to Improve the Comprehensibility of Predictive Models”. In Open Access publication from Katholieke Universiteit Leuven, pp.1-56, 2006.

[3] Venkatadri.M and Lokanatha C. Reddy ,“A comparative study on decision tree classification algorithm in data mining” , International Journal Of Computer Applications In Engineering ,Technology And Sciences (IJCAETS), Vol.- 2 ,no.- 2 , pp. 24- 29 , Sept 2010.

[4] Wanjun Yu, Xiaochun Wang and Fangyi Wang, Erkang Wang, Bowen Chen, “The Research of Improved Apriori Algorithm for Mining Association Rules” 2008 11th IEEE International Conference on Communication Technology Proceedings, 978-1- 4244-2251-7/08/$25.00 ©2008 IEEE

[5] Shuo Yang, “Research and Application of Improved Apriori Algorithm to Electronic Commerce” 2012 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science, 978-0-7695-4818-0/12 $26.00 © 2012 IEEE DOI 10.1109/DCABES.2012.51

[6] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang, “An Improved Apriori-based Algorithm for Association Rules Mining” Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 978-0-7695-3735-1/09 $25.00 © 2009 IEEE DOI10.1109/FSKD.2009.193.

[7] Yanfei Zhou, Wanggen Wan, Junwei Liu, Long Cai, “Mining Association Rules Based on an Improved Apriori Algorithm”, 978- 1-4244-585 8- 5/10/$26.00 ©2010 IEEE.

[8] Wei-min ma, zhu-ping liu, “two revised algorithms based on apriori for mining association rules”, Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunm