International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Finding the Most Frequent Dropout Reasons Using Interestingness Measure

Anju Lata Gajpal[1], Umesh Kumar Pandey[2]

[1] Research Scholar, [2]Research Supervisor,

MATS School of Information Technology, MATS University Raipur CG

ABSTRACT

Data processing field address need of all dimension of human life. It has been realized that finding interesting and useful patterns from data is inevitable, for this purpose many methods developed. Interestingness measures in association rule mining find these interesting measures using statistical tool. In this paper interestingness measure were used to find the most interesting reason for dropping out from educational institution. The study will help decision maker to take necessary action to reduce the drop out from the institution.

Keywords: Interestingness measure, Dropout analysis, Association analysis.

I. INTRODUCTION

Number of undergraduate student are becoming serious challenge for our educational system. This challenge comes because of dropout among students of undergraduate.

After 1992, affiliation to private investor opened to establish private higher education institution. So that requirement of higher education human resource is provided to the nation. The result of this step was positive and private investors showed their keen interest and established higher education institution in every subject domain. In recent years India received number of private higher education institution in federal states of India. As per official website of UGC, Chhattisgarh has 9 private university and other self-financed higher education institution spreading the knowledge in this region.

Chhattisgarh also witness dropout in higher education institution. After observing this scenario, a research is conducted to identify the reason and the pattern of reason severity behind dropout.

Now a day's data mining is used to study the data of educational environment. The field which studies educational data using data mining is termed as educational data mining. Campbell and Oblinger [1] defined educational data mining as studying educational data by statistical technique and data miningthat help faculty and administrative persons to take appropriate and necessary action.

In data mining to find the interesting relation association rule mining is applied on the dataset. Interesting relations are identified by interestingness measure which includes statistical formula. Interestingness measure are of two types i.e. objective and subjective. In this paper association rule and objective interestingness measure data mining method is used for identifying the dominating dropout factor among the student of undergraduate.

II. LITERATURE REVIEW

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Association rule mining is widely used in different areas to find frequent patterns and the strengths of the patterns. Piatesky-Shapiro [2] considered statistical independence for interestingness measure. Soon interestingness measures were the hot topic and several new interestingness measures proposed for association rule. Agrawal and Shrikant [21] proposed support and confidence as interestingness measure. Hilderman and Hamilton [22], Tan et. al. [27] compared various interestingness measure in their research work. Lee et. al. [24] and Omiecinski [25] mentioned in their research work that the confidence, coherence and cosine measures gave good effect on correlation mining. Tan et. al. [27] discussed in his research work about the properties 21 objective interestingness measures.

Sanjeev Rao et. al. [8] used association rule mining for retail business. They predicted product sales trends and customer behavior in retail business.

Mahmood A Rashidi et.al. [11] used apriori algorithm to find cooccurrences of disease in patients.

Sheenu Tomas et.al. [16] proposed correlation analysis as alternative of support and confidence. At the end they concluded that correlation gives additional information about association rule.

Monika Gandhi et.al. [18] applied decision tree algorithm on medical data sets. The objective of this paper was to prepare automated system for diagnosing heart disease using machine learning.

Tinto [19] in his study expresses' that educational environment play role in student attrition.

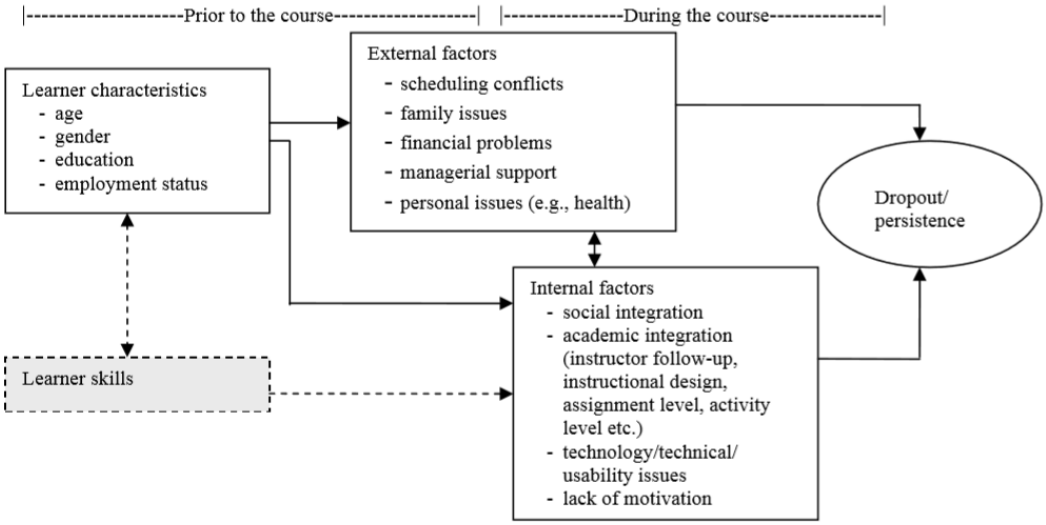


Fig 1: Model of adult dropout in online learning [28]

Rovai [14} prepared a persistent model that identify factor which affect student dropout in online learning.

Park [15] prepared a review study for identifying factors which affect dropout in nontraditional and nondegree online program. Persistent model proposed in this paper is shown in the figure.

Pandey UK and Pal S [20] organizes a study on classroom teaching language. to perform this study interestingness measure were used to find the most appropriate teaching language for classroomenvironment.

III. ASSOCIATION RULE MINING AND METHOD

Knowledge data discovery steps include selection of data, preprocessing, transformation, application of appropriate or selected data mining method and interpretation of the result. Association rule analysis is also one

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

of the among data mining techniques. In association rule mining interestingness measures are used to find the strength of the associations. The objective of association rule to find the frequent occurring itemset in the dataset which is interesting for study. It is also known as interestingness measure. Interestingness measure are grouped in two categories i.e. symmetric and asymmetric.

Symmetric measure gives same resultant irrespective of antecedent and precedent of same itemset. It means value of X→Y is similar to Y→X.

Sometimes it is needed to find the strength of the given rule X to Y and Y to X. In this situation measures are very useful. Table shows the list of symmetric and asymmetric measures.

Table 1: Symmetric Interestingness measure

Correlation
Odds Ratio
Kappa
Lift
Cosine
Leverage
Chi-squared

Table 2: Asymmetric Interestingness Measure

Confidence
J-Measure
Gini Index
Conviction
Added Value
Certainty Factor
Mutual Information

Following association tools are used for interestingness measure:

Support: Support measures that how many times an item or itemset occurs in the transactions. Higher value of support indicates that an item or itemset frequently occurs in the itemset.

$$Support = \frac{itemset\ mumber}{Total\ number\ of\ transactions}$$

Confidence: Confidence indicate that how many times support for item, whose confidence is measured, with respect to support of two items.

$$Confidence\ P(\frac{Y}{X}) = \frac{Support\,(XUY)}{Support\ (X)}$$

Lift: It is used to enhance the response performance of association rule. It is one step ahead of confidence and overcome the disadvantage of confidence. The value of lift may be either positive or negative to show the interdependency. Lift value 1 indicates that X and Y are independent. Higher value of lift indicates both itemset frequently observed in transaction.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

$$Lift = \frac{P(X \cap Y)}{P(X) * P(Y)}$$

Leverage: Leverage value is used to interpret the gap of independence. This shows difference between X and Y which appear together in data set and what would be expected if X and Y are statistically independent. Value of leverage measure varies from -1 to 1. If value of leverage is 0 then it indicates independence.

$$Leverage(X \rightarrow Y) = support(X \rightarrow Y) - support(X) * support(Y)$$

Coverage: Sometimes coverage referred as antecedent support. It covers LHS-Support. The value of coverage ranges from [0,1].

Conviction: Conviction analysis is useful for motivational measure in compare to confidence and interest. Conviction compare appearance probability between observed items. In contrast to lift it is measured directly. Higher conviction value indicates that the rule is more interesting. Conviction value range 0 to ∞.

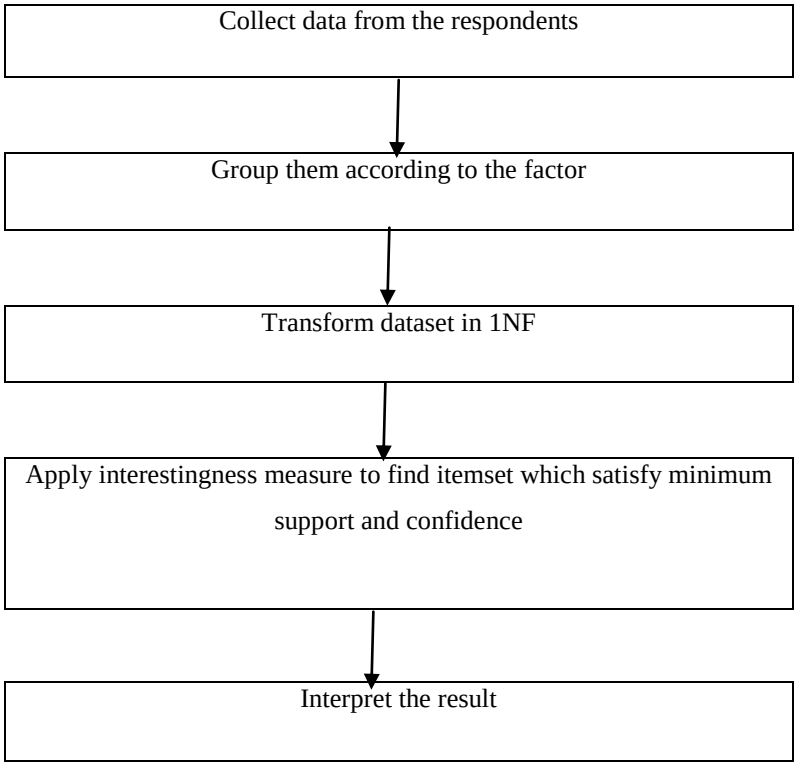
$$Conviction(X \rightarrow Y) = \frac{1 - support(Y)}{1 - confidence(X \rightarrow Y)}$$

Cosine: It measure the similarity between two itemset i.e. X and Y. if the cosine value is 1 then all transaction containing X also contain item Y. If the value of cosine is the 0 then X and Y do not appear together.

$$Cosine(X \rightarrow Y) = \frac{support(X \rightarrow Y)}{Sqrt(support(X)\,Support(Y))}$$

IV. METHODOLOGY

Through observation and various literature review related to the factor for dropout of student from higher education it has been inferred that number of factors are responsible for dropout. Dropout reasons are collected and summarized under five factors related to each other. To perform association analysis among this factor following steps are followed:



IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Fig 2: Methodology of research

V. DATA SET

Data is collected from the student living in the region of Raipur. Students are personally contacted for this purpose. Respondents information is collected on various attribute. Subset of the dataset related to dropout is used in this paper for association and interestingness analysis. Total respondent for this analysis is 172 with 298 answers. Respondents opted following causes for dropout with possible values.

Table 3: factors causing dropout with description and their possible values

Variable	Description	Possible value
FP	Family and personal Problem	Marriage, Home Sickness, Adjustment, Change in goal
FiP	Financial Problem	
IP	Institutional Problem	Campus Environment, too many rules, Over extracurricular activity, enrolled to other institution, Poor teacher learner interaction, Satisfaction from institution, Return on investment
HP	Health Problem	
CP	Course Problem	Curriculum, Language of instruction, Learning problem, Employability of program

Subset of the database is transformed into a new data structure so that association and other mining techniques can be applied. Screenshot of the transformed dataset is shown in fig 3:

Tid	Response
34	IP
35	IP
36	IP
37	IP
38	IP
39	IP
40	Fip
41	Fip
42	Fip

Fig 3: Transformed dataset of respondents.

VI. DATA ANALYSIS TOOL

All data analysis done in R language. For interestingness measure R language provide “arules” package and “apripori” function. The function has facility to decide what would be support and confidence value for finding the itemset out of possible itemset. The code used in R language are following:

```
library(arules)
library(readxl)
```

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

```
question <- read_excel("F:/gajpal/question.xlsx", sheet = "anju")
trans<-split(question$Response,question$Tid,"anjugajpal")
head(trans)
library(arules)
rules<-apriori(trans, parameter=list(support=0.1, confidence=0.2))
inspect(rules)
summary(rules)
interestMeasure(rules, c("support", "chiSquare", "confidence", "conviction","cosine", "coverage", "leverage", "lift"), trans)
```

VII. ITEMSET OF MINIMUM THRESHOLD VALUE

Dataset has 5 different causes for dropout. Total pair of itemset formed using these 5 causes are 80. Some of them are significant and some of them are insignificant or has 0 values. The itemset of minimum threshold value is selected from all the itemset. For the study purpose itemset having support level more than 10 % and confidence level more than 20% is considered as minimum threshold value to select itemset. The list of itemset which satisfies the minimum support and confidence is shown in the table.

Table 4: Itemset with minimum support threshold value

Sr. No.	Itemset	Support	Count
1	FiP	0.273	47
2	FP	0.297	51
3	CP	0.465	80
4	IP	0.587	101
5	FiP → CP	0.110	19
6	CP → FP	0.116	20
7	CP → IP	0.320	55

Table 5: Itemset with minimum confidence threshold value

Sr. No.	Itemset	Confidence	Count
1	FiP → CP	0.404	19
2	CP → FiP	0.238	19
3	FP → CP	0.392	20
4	CP → FP	0.25	20
5	CP → IP	0.688	55
6	IP → CP	0.545	55

VIII. RESULT & ANALYSIS

Table 4 shows that the itemset which satisfies minimum threshold value for support i.e. 0.1. Table 4 indicates that Financial problem (FiP), Family problem (FP), Course problem (CP), Institutional problem (IP) has significant reason for dropout of the student. Collective factor responsible for the dropout which satisfies

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

minimum support threshold value are 3 i.e. Financial problem and Course problem (FiP → CP), Course problem and Family problem (CP → FP) and Course problem and Institutional problem (CP→IP).

Table 5 shows confidence measure. Confidence measure contains 3 itemset with 6 pairs because of asymmetric result. Between Financial problem and Course problem confidence is higher for relationship Fip →CP. This means that 0.4 times financial problem occurs so does course problem. Between Family problem and Course problem confidence is high for relationship FP→CP, this means that 0.392 times family problem occurs so does course problem. Between course problem and institutional problem confidence is high for relationship CP→IP, this means that 0.688 times course problem occurs so does institutional problem.

Table 6: Asymmetric interestingness measure

Sr. No.	Itemset	Conviction	Coverage
1	FiP → CP	0.898	0.273
2	CP → FiP	0.953	0.465
3	FP → CP	0.880	0.297
4	CP → FP	0.938	0.465
5	CP → IP	1.321	0.465
6	IP → CP	1.174	0.587

Table 6 shows asymmetric interestingness measure i.e. conviction and coverage for the itemset which satisfies minimum threshold for support and confidence. Among all the itemset conviction value is high for itemset CP→IP, this means that among all the itemset CP→IP is more interesting and appear most of the time. Among all the itemset coverage value is high for the itemset IP→CP, this means that institutional problem is playing important role over course problem.

Table 7: Symmetric interestingness measures

Sr. No.	Itemset	Cosine	Leverage	lift
1	FiP → CP	0.310	-1.663061e-02	0.869
2	CP → FP	0.313	-2.163332e-02	0.843
3	CP → IP	0.612	4.664684e-02	1.171

Table 7 shows symmetric interestingness measure i.e. cosine, leverage and lift for the itemset which satisfies minimum threshold for support and confidence. Among all the itemset CP→IP has maximum cosine value, this means course problem and institutional problem occurs most of the time compare to other itemset. Leverage value is positive only for itemset CP→IP. Lift is also high for itemset CP→IP this means that probability of occurrence of course problem and institutional problem is higher than multiplication of probability of occurrence each item which form the itemset.

IX. CONCLUSION

Dropout of student from educational institution is not a good sign for educational institutional institution and society. In this research paper dropout problem is discussed on five factors i.e. family and personal problem (FP), financial problem (FiP), Institutional Problem (IP), health problem (HP) and course problem (CP). These are not limit for dropping out from educational institution. Above mentioned five factors are studied here using interestingness measure i.e. support, confidence, lift, leverage, coverage, conviction and cosine. Data is analyzed

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

in R programing language, which offer package for interestingness measure analysis. The study reveals that health problem is not significant reason for dropping out from educational institution. Second inference is that institutional problem is most significant reason from drop out analysis. Third inference is that course problem and institutional problem is most associated, it means that when a student face problem related with course then s/he face problem with institution also. Thus, institution must address course and institution problem faced by the student to reduce the dropout from institution.

REFERENCES

[1.] Brin, Sergey, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur (1997). Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA

[2.] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In: *Knowledge Discovery in Databases*, pages 229–248

[3.] Agrawal, R., H Mannila, R Srikant, H Toivonen, AI Verkamo (1996). Fast Discovery of Association Rules.*Advances in Knowledge Discovery and Data Mining* 12 (1), 307–328.

[4.] Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313.

[5.] Campbell, J., &Oblinger, D. (2007). Academic analytics. Washington, DC: Educause.

[6.] Chyung, Y. (2001). Systemic and systematic approaches to reducing attrition rates in online higher education. The American Journal of Distance Education, 15(3), 36–49.

[7.] Chyung, Y., Winiecki, D. J., &Fenner, J. A. (1998). A case study: Increase enrollment by reducing dropout rates in adult distance education (ERIC Document Reproduction Service No. ED 422 848).

[8.] SanjeevRao, Priyanka Gupta, ―Implementing Improved Algorithm over APRIORI Data Mining Association Rule Algorithm‖ , ISSN: 0976-8491 (Online) | ISSN: 2229-4333 (Print) IJCST Vol. 3, Issue 1, Jan. - March 2012.

[9.] Doo, M., & Kim, Y. (2000). The effect of relevance-enhanced messages on learning in Web-based training. Korean Association for Educational Information and Broadcasting, 6(2), 73–90.

[10.] Jun, J. (2005). Understanding dropout of adult learners in e-learning. Unpublished doctoral dissertation, University of Georgia, Athens, Georgia, USA.

[11.] Mahmood A. Rashid, MdTamjidulHoque, Abdul Sattar, "Association Rules Mining Based Clinical Observations ",Griffith University Nathan, QLD, Australia, {m.rashid, t.hoque, a.sattar}@griffith.edu.au, 2010

[12.] Levy, Y. (2007). Comparing dropouts and persistence in e-learning courses. Computers & Education, 48, 185–204.

[13.] Menager-Beeley, R. (2004). Web-based distance learning in a community college: The influence of task values on task choice, retention and commitment. (Doctoral dissertation, University of Southern California). Dissertation Abstracts International, 64(9– A), 3191.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[14.] Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs. Internet and Higher Education, 6, 1–16.

[15.] Park, J. (2007). Factors related to learner dropout in online learning. In Nafukho, F. M., Chermack, T. H., & Graham, C. M. (Eds.) Proceedings of the 2007 Academy of Human Resource Development Annual Conference (pp. 25-1–25-8). Indianapolis, IN: AHRD.

[16.] Sheenu Toms, Deepa John," Rules Extraction in XML Using Correlation", International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 1 ISSN 2250-3153

[17.] Park, J., & Choi, H. (2007). Differences in personal characteristics, family and organizational supports, and learner satisfaction between dropouts and persistent learners of online programs. In G. Richards (Ed.), Proceedings of World Conference on ELearning in Corporate, Government, Healthcare, and Higher Education 2007 (pp. 6444–6450). Chesapeake, VA: AACE.

[18.] Monika Gandhi; Shailendra Narayan Singh, "Predictions in heart disease using techniques of data mining", International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015, IEEE Conference Publications, pp. 520 - 525, DOI: 10.1109/ABLAZE.2015.7154917

[19.] Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition (2nd ed.). Chicago, IL: University of Chicago Press.

[20.] Pandey, U.K., and Pal S. (2011), (2011), "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814.

[21.] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Very Large Data Bases, VLDB 1994, pp. 487–499 (1994)

[22.] Hilderman, R., Hamilton, H.: Knowledge discovery and measures of interest. Kluwer Academic Publishers (2001)

[23.] Huynh, X.-H., Guillet, F., Blanchard, J., Kuntz, P., Briand, H., Gras, R.: A Graph-Based Clustering Approach to Evaluate Interestingness Measures: A Tool and a Comparative Study. In: Guillet, F.J., Hamilton, H.J. (eds.) Quality Measures in Data Mining. SCI, vol. 43, pp. 25–50. Springer, Heidelberg (2007)

[24.] Lee, Y.K., Kim, W.Y., Cai, Y., Han, J.: CoMine: efficient mining of correlated patterns. In: IEEE International Conference on Data Mining, pp. 581–584 (2003)

[25.] Omiecinski, E.: Alternative interest measures for mining associations in databases. IEEE Transaction on Knowledge and Data Engineering 15(1), 57–69 (2003)

[26.] Piatetsky-Shapiro, G., Steingold, S.: Measuring lift quality in database marketing. SIGKDD Explorations 2(2), 76–80 (2000)

[27.] Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4), 293–313 (2004)

[28.] Park J & Choi H J, "Factors Influencing Adult Learners' Decision to Drop Out or persist in Inline Learning", International Forum of Educational Technology & Society 12(4). 207-217 ISSN 1436-4522