Relevance Index for Inferred Knowledge in Higher Education Domain Using Data Mining

Preeti Gupta, Deepti Mehrotra and Tarun Kumar Sharma

Abstract Optimizing the real-life scenarios facilitate knowledge building. Developing a knowledge model for optimizing certain output criteria enhances the benefits by many folds. Even a non-profit sector like education needs to define knowledge models that optimize their functioning and eventually help in knowledge building. Quantifying the factors determining the academic well-being of the students in any educational organization is of prime importance. The paper exemplifies the implementation of Data Mining Technique to deduce knowledge through classification rules and further assign relevance index to inferred knowledge.

Keywords Higher education · Knowledge · Data mining · Classification

1 Introduction

It is often said that we are drowning in data but starving for knowledge [1]. Extraction of information from data facilitates knowledge building. Information which can be termed as a subset of data stimulates action in an entity, whereas knowledge defines the action of an entity in a particular setting [2]. A number of researchers have classified knowledge on different basis, sometimes defining the manner of codification and occurrence [3], or on the basis of know-what, know-how, know-why and know-when aspect of knowledge [4]. Some have even mapped knowledge in diverse domains [5].

P. Gupta (✉) · T.K. Sharma
Amity University Rajasthan, Jaipur, India
e-mail: preeti_i@rediffmail.com

T.K. Sharma
e-mail: taruniitr1@gmail.com

D. Mehrotra
ASET, Amity University Uttar Pradesh, Noida, India
e-mail: mehdeepti@gmail.com

M. Pant et al. (eds.), *Soft Computing: Theories and Applications*,
Advances in Intelligent Systems and Computing 584,
https://doi.org/10.1007/978-981-10-5699-4_27

There are varieties of ways for representing knowledge [6]. Using production rules written in form of IF-THEN rules is one of the most popular approach used for knowledge representation [7]. The IF-THEN rules adopt a modular approach, each defining principally independent and a relatively minor piece of knowledge. A rule-based system will include universal rules and actualities about the knowledge domain covered.

Knowledge building in education domain can be achieved by adopting procedures that optimize their functioning.

The research work is undertaken with an objective of deducing relevance index for inferred knowledge. The case of education sector is taken in particular while inferring knowledge related to student's academic performance in a technical subject at level of higher education. It is important to deduce relevance index to inferred knowledge as it is a clear depiction of the existing system and further helps in decision making.

The paper is organized as follows. Section 2 elaborates the methodology adopted for rule induction and further rule evaluation in the higher education set-up. Finally, the conclusions are drawn and presented in Sect. 3.

2 Adopted Methodology in Higher Education Scenario

Educational organizations strive to achieve the higher academic output for the students. Many researchers have strived hard to predict the factors affecting the academic results of students [8–12]. Identification of such critical parameters, which could improve the academic attainment of students, supports an effective academic planning.

In case the individuals of a population can be separated into different classes, generation of a classification rule is a system in which the individuals of the population are each allocated to one or the other class.

In the study, knowledge is represented through classification rules [13], which exist in the form of IF-THEN rules. The work starts by identifying the variables and collecting the data in the context of these variables. The values of the attributes are then encoded on an 8-level scale. Rule induction is initiated through JRip, which implements a propositional rule learner, repeated incremental pruning to produce error reduction (RIPPER). The rules are then evaluated on the basis of the metrics *Net Benefit* which takes into account both classification and misclassification witnessed by the knowledge rule.

2.1 Variable Identification and Data Collection

This dataset has 5000 records and five independent attributes, all of which are categorical. The independent attribute names in the dataset are as follows:

ContinuousEvaluationMarks, SGPA_II, Practical_orient, Attendance, *Base_Sub_Marks.*

The independent attributes affect the dependent attribute of *End_Term_Marks* and are reflected in Table 1.

The attributes were encoded on the 8-level scale, depicted in Table 2.

Table 1 Attributes of the study

Attribute name	Description
ContinuousEvaluationMarks	Performance of the students continuously evaluated by the faculty member with respect to class assignments, marks obtained in class test, performance in viva voce, etc. (maximum marks—30)
SGPA_II	Semester grade point average (SGPA) measures the academic performance of the student in the previous semester on a 10-point scale
Attendance	It reflects the presence of the student in the class of the subject under study
Practical_orient	It reflects the ability of the students to solve the problems related to the subject under the study in a practical manner
Base_Sub_Marks	Performance of the student in physics (base subject) studied in the earlier semester
End_Term_Marks	Performance of the student in the end-term exam of the subject under scrutiny

Table 2 Encoding of the attributes

ContinuousEvaluationMarks (Maximum value 30)		SGPA_II (Maximum value 10)	
Marks range	Encoding	SGPA range	Encoding
0–3	000	0–2	000
4–7	001	2.1–4	001
8–11	010	4.1–5	010
12–15	011	5.1–6	011
16–19	100	6.1–7	100
20–23	101	7.1–8	101
24–27	110	8.1–9	110
28–30	111	9.1–10	111
Base_Sub_Marks (maximum value 100)		**Attendance (maximum value 100%)**	
Marks range	Encoding	Attendance range	Encoding
0–20	000	Below 75%	000
21–40	001	75.1–77%	001
41–50	010	77.1–80%	010

(continued)

Table 2 (continued)

ContinuousEvaluationMarks (Maximum value 30)		SGPA_II (Maximum value 10)	
Marks range	Encoding	SGPA range	Encoding
51–60	011	80.1–83%	011
61–70	100	83.1–85%	100
71–80	101	85.1–90%	101
81–90	110	90.1–95%	110
91–100	111	95.1–100%	111
Practical_orient (maximum value 100)		**End_Term_Marks (maximum value 100)**	
Marks range	Encoding	Marks range	Encoding
0–20	000	0–20	000
21–40	001	21–40	001
41–50	010	41–50	010
51–60	011	51–60	011
61–70	100	61–70	100
71–80	101	71–80	101
81–90	110	81–90	110
91–100	111	91–100	111

2.2 Rule Induction

In the year 1995, Cohen proposed JRip which implemented a propositional rule learner, repeated incremental pruning to produce error reduction (RIPPER) [14].

Error reduction can be witnessed in JRip since the process of incremental pruning examination of the classes is done in the increasing order of their size. The initial ruleset is generated on the basis of incremental reduced error. Initially, JRip (RIPPER) treats all the instances from the training dataset related to a particular judgment as a class and deduces a ruleset that covers all the members of that class. The procedure is repeated for all the classes.

Initialization

Initialize RS = {}, and from each class from the less frequent one to the most frequent one.

Repeat

{

1. ***Building phase: Repeat the phases given below, grow phase and prune phase until there are no positive instances or error rate increases more than 50%.***

   1.1 ***Grow phase: Follow the greedy approach of adding conditions to the rule until the accuracy of the rule reaches 100%.***

1.2 ***Prune phase: Incremental pruning approach should be followed for each rule. The pruning metrics can be measured in terms of $2p/(p + n) - 1$, where p—number of positive instances covered in the ruleset and n—number of negative instances covered in the ruleset.***

2 ***Optimization Phase: On generation of the initial ruleset $\{R_i\}$, two variants of each rule are to be generated and pruned from randomized data using procedures Grow and Prune. The generation of the first variant is done from an empty rule, and the next variant is created by adopting a greedy approach of adding conditions to the original rule. The metrics of Description Length (DL) are computed for each variant. The final representation of the ruleset is done by the rule having the minimal DL. After the examination of all the rules in R_i, Building phase is again used for generating more rules if there are still residual positives.***

3 ***Those rules that increase the DL of the complete ruleset are then deleted from the ruleset, and the final ruleset is added to RS.***

}

In the study, JRip was implemented using Weka 3.8.0 and the following ruleset of 87 rules was generated. A snapshot of the rules and the output achieved is shown in Fig. 1.

2.3 Rule Analysis and Interpretation

For each of the 87 rules acquired by implementing JRip on the dataset, the value of classification (true positive, TP) and misclassification (false positive, FP) was recorded [15].

True positive (TP)— the number of examples satisfying A and C
False positive (FP)—the number of examples satisfying A, but not C

where *A—antecedent of the rule, C—consequent of the rule*

The rules were further evaluated on the basis of *Net Benefit* [16] considering a range of thresholds and calculating the NB across these thresholds. The result was then plotted against Rule Number and Net Benefit. For each threshold P_t, the Net Benefit was calculated as per Eq. 1:

$$\text{Net Benefit (NB)} = \frac{\text{TP}}{\text{N}} - \frac{\text{FP}}{\text{N}}\left(\frac{P_t}{1 - P_t}\right) \tag{1}$$

On evaluating the rules for Net Benefit for different values of P_t, the following observations were met and are depicted through Fig. 2.

On cross-tabulating the rule count for P_t = 0.1–0.6, the NB values for all the 87 rules can be witnessed in Table 3.

=== Run information ===

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation: jrip 1
Instances: 5000
Attributes: 6
ContinuousEvaluationMarks
SGPAII
Practical_orient
Attendance
Base_Sub_Marks
End_Term_Marks
Test mode: evaluate on training data

=== Classifier model (full training set) ===

JRIP rules:
===========

(Base_Sub_Marks = 110) and (Attendance = 110) and (SGPAII = 010) => End_Term_Marks=111 (5.0/1.0)
(SGPAII = 111) and (Base_Sub_Marks = 000) and (Attendance = 000) => End_Term_Marks=111 (4.0/0.0)
(Base_Sub_Marks = 110) and (ContinuousEvaluationMarks = 101) and (Practical_orient = 010) and (SGPAII = 110) => End_Term_Marks=111 (5.0/0.0)
(Base_Sub_Marks = 111) and (SGPAII = 000) and (Attendance = 101) and (ContinuousEvaluationMarks = 011) => End_Term_Marks=111 (5.0/0.0)
(Attendance = 111) and (ContinuousEvaluationMarks = 000) and (Practical_orient = 011) => End_Term_Marks=110 (7.0/2.0)
.
.
.
.
.
Time taken to build model: 4.47 seconds

Fig. 1 Weka implementation

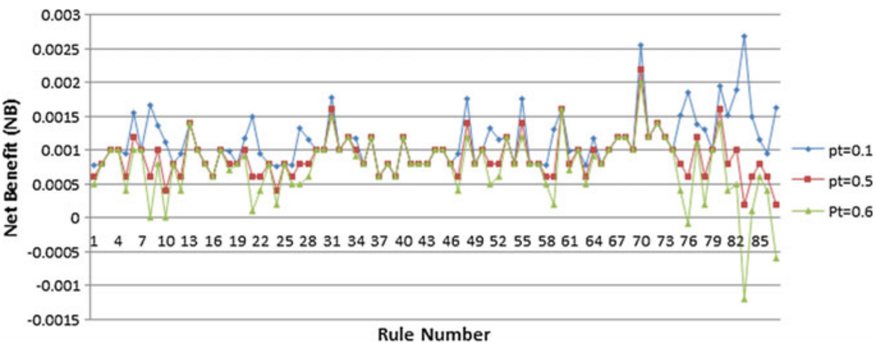

Fig. 2 Consolidated plot across P_t values (0.1, 0.5, 0.6)

Table 3 Analysing NB for the rules

Criteria	$P_t = 0.1$	$P_t = 0.2$	$P_t = 0.3$	$P_t = 0.4$	$P_t = 0.5$	$P_t = 0.6$
Number of rules with NB < 0	0	0	0	0	0	3
Number of rules with NB $\geq$ 0 and NB < 0.0005	0	0	0	0	4	14
Number of rules with NB $\geq$0.0005 and NB < 0.001	32	32	33	41	44	36
Number of rules with NB $\geq$ 0.001 and NB < 0.0015	41	44	46	40	35	31
Number of rules with NB $\geq$ 0.0015 and NB < 0.002	12	9	7	5	3	2
Number of rules with NB $\geq$ 0.002 and NB < 0.0025	0	1	1	1	1	1
Number of rules with NB $\geq$ 0.0025 and NB < 0.003	2	1	0	0	0	0
Total rules	87	87	87	87	87	87

The consolidated plot depicting the NB values for all 87 rules across thresholds ($P_t = 0.1$, 0.5, 0.6), shown in Fig. 2, depict that the Net Benefit of the rule having maximum Net Benefit across all the threshold values of P_t ($P_t = 0.1$–0.6) decreases as we increase the threshold value (P_t) from 0.1 to 0.6. In fact at $P_t = 0.6$, some of the rules exhibit the negative NB.

$P_t = 0.5$ signifies that FP and TP are weighted equally. Hence, maintaining a $P_t = 0.1$ signifies assigning more weightage to the classification, i.e. true positive (TP), rather than to misclassification, i.e. false positive (FP).

The study selects $P_t = 0.1$. Maximum NB and distinct peaks are achieved on selecting a $P_t = 0.1$. It is also observed that NB value decreases as we move from $P_t = 0.1$ to $P_t = 0.6$. Moreover, the NB value also shows a negative growth in case of $P_t = 0.6$. $P_t = 0.6$ signifies the assignment of more weightage to misclassification rather than to classification.

However, for $P_t = 0.1$, the rule that acquires the highest benefit is:

Base_Sub_Marks = 010 and Attendance = 001 and ContinuousEvaluationMarks = 101 => End_Term_Marks = 011

On decoding the rule, it can be stated as:

Base_Sub_Marks is between 41 and 50 and Attendance between 75.1 and 77% and ContinuousEvaluationMarks between 20 and 23 => End_Term_Marks between 51 and 60.

The relevance index assigned to the knowledge rule is on the basis of its Net Benefit (NB), keeping into account the classification and misclassification done by the rule. The Net Benefit (NB) for the above said rule at a threshold value P_t of 0.1 is 0.002689.

The reason for using Net Benefit (NB) to assign relevance index to inferred knowledge is:

1. The prediction model incorporates consequences and hence can be used to infer a decision on the usage of the given model.
2. It can be directly applied to the validation set and does not need any additional information.
3. Even if the model outcome is in binary or continuous form, the method for evaluation is applicable.

3 Conclusion

Rule induction can deduce the relationship existing between the various attributes. The influence of the independent variables on the dependent variable can be observed. Rules with a higher relevance index are much more apt to the system and can be used for appropriate syllabus planning, designing structured lesson plans, structuring criteria for the evaluation of the student's performance and adoption of suitable teaching pedagogy for the improvement in the overall academic performance of the students. The knowledge derived in the form of rules bears relevance in the context of the domain and hence can be added to the knowledge set that can supplement the process of decision making in a knowledge base environment.

References

1. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann Publishers, Canada (2000)
2. Boisot, M.H.: Knowledge Assets Securing Competitive Advantage in the Information Economy. OUP Oxford New Edition (2006)
3. Polanyi, M.: The Tacit Dimension. Routledge and Kegan Paul, London (1966)
4. Nickols, F.W.: The knowledge in knowledge management. In: Cortada, J.W., Woods, J.A. (eds.) The Knowledge Management Yearbook 2000–2001, pp. 12–21. Butterworth-Heinemann, Boston, MA (2000)
5. Gupta, P., Mehrotra, D., Singh, R.: Achieving excellence through knowledge mapping in higher education institution. Int. J. Comput. Appl. 5–10 (2012)
6. Jong, T.D., Ferguson Hesler, M.G.M.: Types and quality of knowledge. Educ. Psychol. **31**,105–113(1996)
7. Rich, E., Knight, K., Nair, S.B.: Artificial Intelligence. TMH, New Delhi (2010)
8. Kabakchieva, D.: Student performance prediction by using data mining classification algorithms. Cybern. Inf. Technol. **13**, 61–72 (2013)

9. Kumar, S.A., Vijayalakshmi, M.N.: Efficiency of decision trees in predicting student's academic performance. Int. J. Comput. Sci. Inf. Technol. **23**, 335–343 (2011)
10. Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., Wani, E.: Prediction of student academic performance by an application of data mining techniques. In: International Conference on Management & Artificial Intelligence, vol. 6, pp. 110–114 (2011)
11. Ramanathan, L., Dhanda, S., Kumar, S.D.: Predicting students' performance using modified ID3 algorithm. Int. J. Eng. Technol. (IJET) **5**(3), 2491–2497 (2013)
12. Gupta, P., Mehrotra, D., Sharma. T.K.: Genetic based weighted aggregation model for optimization of student's performance in higher education. In: Advances in Intelligent Systems and Computing, pp. 877–887. Springer, Singapore (2015)
13. Gupta, P., Mehrotra, D.: Effective curriculum development through rule induction in knowledge centric higher education organization. In: Confluence 2013: The Next Generation Information Technology Summit (4th International Conference), Noida. IET Digital Library, vol. 2013, issue 647 CP, pp. 475–480 (2013)
14. Cohen, W.W.: Fast effective rule induction. In: Twelfth International Conference on Machine Learning, pp. 115–123 (1995)
15. Freitas, A.A.: A survey of evolutionary algorithms for data mining and knowledge discovery. In: Advances in Evolutionary Computing, pp. 819–845, Springer, Heidelberg (2003)
16. Vickers, A.J., Elkin, E.B.: Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Making **26**, 565–574 (2006)