Comparative study on classification power of the attributes and reducts

Nilesh Mahajan [1] ,Jyoti Namdeo[2]

[1] Professor IMED, Bharati Vidyapeeth Deemed University Pune, India,
[2] Research Scholar IMED, Bharati Vidyapeeth Deemed University Pune, India
Nilesh.mahajan@bharatividyapeeth.edu , jyotianamdeo@gmail.com
www.ijcaonline.org

Abstract

In rough set theory, reduct are those attributes which are important and are able to represent the whole range of attributes. They show up the important features of database. In the present study of educational data mining, we have used the reduct of the rough set theory to find out the important features of a particular course work. The features of course work are the subjects taught in that course work. We conclude in this paper that if all the subjects are taken, the classificatory power remains same as in case where only the reduct is taken.

Keywords— Rough set theory, reduct, educational data mining and classification

I. INTRODUCTION

In the educational data mining, every institute tries hard to make its name by recognizing it customers as its great assets. It provides all basic amenities which are needed for making its customers satisfied by giving them training through well qualified faculty staff, well equipped lab, sports facility etc. But the name of any institute goes with the result of student and hence good result is the main target area of institute.

In the educational institute exponential growth of student data has been observed in the past two decades due to digitalization of institute. Educational data mining offers to mine the data for getting the valuable knowledge from this huge data.

www.educationaldatamining.org defines:

"Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in."

Baker (2008) defines EDM as a field of scientific research that focused on development of methods for investigating a particular type of data obtained from educational settings, and to use those methods to improve students' learning and the context, in which they learn.

EDM studies mostly concentrate on detecting patterns in educational data, but there are studies that investigate the ways of using these patterns in student modeling (Amarshi and Conati, 2009).

II. RELATED WORK

Romero and Ventura (2007) categorized EDM studies as the follows:

1. Statistics and visualization
2. Web mining:
    i) Clustering, classification, and outlier detection
    ii) Association rule mining & sequential pattern mining
    iii) Text mining

Another viewpoint on educational data mining is given by Baker (2008), which classifies work in educational data mining as follows

1. Prediction:
    i) Classification
    ii) Regression
    iii) Density estimation
2. Clustering
3. Relationship mining
    i) Association rule mining
    ii) Correlation mining
    iii) Sequential pattern mining

iv) Causal data mining

4. Distillation of data for human judgment

5. Discovery with models

EDM studies have focused mostly on tutoring systems where structured problem solving (Baker et al., 2008) or drill and practice activities (Beck, 2005) are supported. Recommendation system was developed based on data mining techniques to help students to take decisions on their academic itineraries, to choose course, based on experience of previous students with similar academic achievements Vialardi et. al. (2009). Ayesha et. al. (2010) studied K-means clustering algorithms to discover knowledge from education data mining. They recommended that all correlated information of class quiz, mid & final exam should be conveyed so that dropout ratio can be reduced and student performance can be improved.

Rough set theory has also been used as one technique for the data mining. Rough set theory was introduced by Z. Pawlak as a mathematical tool for data analysis. Rough sets have many applications in the field of Knowledge Discovery: feature selection, discrimination process, data imputations and create decision Rules. Rough set have been introduced as a tool to deal with, uncertain Knowledge in Artificial Intelligence Application.

Ramasubramanian et. al. (2009) proposed a concept map for each student and staff. This map finds the result of the subjects and also recommends a sequence of remedial teaching. Similarly, Narli (2010) gave detailed analysis of quantitative and categorical data using rough set theory and concluded that rough set approach can be applied to analyze educational research data for the investigation of attitudes, behaviors or beliefs could reveal more comprehensive information about the data. Venkatasubbareddy et. al. (2010) conducted a research on student result oriented learning process evaluation system based on distributed data mining and decision tree algorithm. They concluded that a rule-discovery approach was suitable for the student learning result evaluation and should be employed into practice. In the present work we have used rough set theory for finding the important attributes of a course undertaken by the students of MCA department.

III. METHOD

The data of student result was taken from the MCA department of one of the institute. Student result was obtained in the form of numbers as secondary data. Selected theory subjects were chosen as conditional attributes and given the code as 101, 102 … 601, 602 etc. Marks were converted into grade according to credit system employed by the institute. This data was than given to Rosetta tool kit, data mining software based on the rough set theory.

Rough set theory is a new mathematical approach to imperfect knowledge. Rough set theory (RST) has been used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information (Pawlak, 1982, 1991; Polkowski, 2002). Given a dataset with discretized attribute values using RST, it is possible to find a subset called as reduct of the original attributes that are the most informative and all other attributes can be removed from the dataset with minimal information loss.

Description of Attributes

Attribute Code	Description
101	Elementary Algorithm
103	Procedure Oriented Programming
201	Data Structure
202	Operating System
203	Data Base Management System
301	Software Engineering
302	Computer Networking
303	Object Oriented Programming
401	UML
402	Unix
501	Software Project Management
502	Artificial Intelligence

A. Reduction of Attributes

Data reduction keeps only those attributes that preserve the indiscernibility relation. The rejected attributes are surplus since their removal doesn't affect the classification.

Let S = (U, A) be an information system,

B ⊆ A, and let a∈ B.

Then a is dispensable in B if

INDS (B) = IND (B − {a}) S; otherwise a is indispensable in B.

A set B is called independent if all its attributes are indispensable.

Any subset B' of B is called a reduct of B if B' is independent and IND (B') S = IND (B) S.

Thus reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes. In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe. There are number of such subsets of attributes. But those subsets which are minimal are called reducts.

Out of many algorithms for finding the reduct in rough set, we have applied the Johnson Reducer algorithm (Øhrn, 1999) which is based on discernibility matrix. A discernibility matrix (Skowron and Rauszer, 1992) of a decision table $\boldsymbol{D = (U, C \cup d)}$ is a symmetric $|\boldsymbol{U}| \times |\boldsymbol{U}|$ matrix with entries defined:

$$c_{ij} = \{a \in C \mid a(x_i) \neq a(x_j)\} \quad i, j = 1,...,|U|$$

Where D is decision table, U is universe of objects (students), C is conditional attributes (courses/subjects) and d is decision attribute (result).

Each $\mathbf{c_{ij}} \in \mathbf{U}$ contains those attributes that differ between objects $\mathbf{x_i}, \mathbf{x_j} \in \mathbf{U}$ and $\mathbf{a} \in \mathbf{C}$.

At first Johnson Reducer algorithm makes the reduct set R to be empty set. The number of times the conditional attribute appears in the discernibility matrix is referred as the count. The attribute appears along with other attributes and together this is called as clause. This is taken into account. The attributes that appear more frequently are considered as more significant. Thus attribute with the highest heuristic value (i.e. count of the number of appearances) is added to the reduct candidate. All clauses in the discernibility function containing this attribute are removed. This process iterates and when all clauses are removed from the discernibility matrix, algorithm terminates and returns with reduct R.

B. Johnson Algorithm

Johnson(C, f_D)

C, the set of conditional attributes

f_D, the discernibility function.

$R \leftarrow \emptyset$; best c=0;

W**hile (**f_D not empty)

F**or each** $a \in C$ that appears in f_D

c = heuristic(a)

If (c > best c)

Best c = c; best Attribute ← a

$R \leftarrow$ R $\cup$ a

$f_D \leftarrow$ remove Clauses(f_D, a)

R**eturn** R

After getting the reduct, it was given as input to the classification algorithm J48 using WEKA software which is open access software. Final result after sixth semester of the student was taken as decision attribute and reduct attribute were taken as the conditional attribute for the classification algorithm. The classification accuracy obtained was 56%. That means only on the basis of reduct alone student result can be 56% correctly classified.

IV. RESULT

Johnson Reducer algorithm gave one reduct giving the set of most important attributes as 101, 201 and 303. These subjects were elementary algorithm, data structure, object oriented programming.

Further, we want to test whether the classification accuracy was same when all the attributes have been taken as conditional attributes. So now instead of taking only the reduct attribute, now we have taken all the conditional attributes and applied the same classification algorithm. Again the classification accuracy obtained was 56%.

Below two run information are given. In the first all the 12 attributes along with decision attribute result are taken as UGP101,..., UGP502. UGP stands for University Grade Point. In the second run information only reduct attribute is provided. The classification accuracy is same 56%.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.75 -M 2

```
Relation:     12SUBUGPRESWISEI-V
Instances:    51
Attributes:   13
              UGP101
              UGP103
              UGP201
              UGP202
              UGP203
              UGP301
              UGP302
              UGP303
              UGP401
              UGP402
              UGP501
              UGP502
              RES
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
UGP202 = F
|   UGP201 = C+: SC (1.0)
|   UGP201 = B: PC (0.0)
|   UGP201 = A: PC (0.0)
|   UGP201 = A+: PC (0.0)
|   UGP201 = B+: PC (0.0)
|   UGP201 = O: PC (0.0)
|   UGP201 = C: SC (1.0)
|   UGP201 = D
|   |   UGP303 = F: PC (4.0)
|   |   UGP303 = C: PC (3.0)
|   |   UGP303 = A: SC (1.0)
|   |   UGP303 = A+: PC (0.0)
|   |   UGP303 = B+: PC (0.0)
|   |   UGP303 = B: PC (0.0)
|   |   UGP303 = O: PC (0.0)
|   |   UGP303 = D: PC (1.0)
|   |   UGP303 = C+: PC (0.0)
|   UGP201 = F
|   |   UGP501 = O: FL (0.0)
|   |   UGP501 = A: PC (1.0)
|   |   UGP501 = B+: FL (0.0)
|   |   UGP501 = A+: FL (0.0)
|   |   UGP501 = B: FL (0.0)
|   |   UGP501 = C: FL (2.0)
|   |   UGP501 = D: FL (0.0)
|   |   UGP501 = C+: FL (0.0)
|   |   UGP501 = F: FL (4.0)
```

```
UGP202 = D
|   UGP302 = C+: PC (3.0/1.0)
|   UGP302 = O: SC (1.0)
|   UGP302 = A: FC (1.0)
|   UGP302 = B: HS (2.0)
|   UGP302 = A+: SC (1.0)
|   UGP302 = B+: PC (0.0)
|   UGP302 = F: PC (1.0)
|   UGP302 = C: PC (2.0)
|   UGP302 = D: PC (1.0)
UGP202 = B: FC (3.0/1.0)
UGP202 = C+
|   UGP401 = C+: HS (1.0)
|   UGP401 = B: FC (3.0)
|   UGP401 = C: FC (2.0/1.0)
|   UGP401 = B+: FC (0.0)
|   UGP401 = A: FC (0.0)
|   UGP401 = D: FC (0.0)
|   UGP401 = F: SC (1.0)
UGP202 = B+: FC (4.0/1.0)
UGP202 = C: HS (5.0)
UGP202 = A: FC (1.0)
UGP202 = A+: DC (1.0)
Number of Leaves  :         46
Size of the tree :   52
Time taken to build model: 0.39 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          29               56.8627 %
Incorrectly Classified Instances        22               43.1373 %
Kappa statistic                          0.4489
Mean absolute error                      0.1561
Root mean squared error                  0.3487
Relative absolute error                 59.0711 %
Root relative squared error             96.0179 %
Total Number of Instances               51
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.286    0.182      0.2       0.286     0.235      0.541    SC
                 0.636    0.1        0.636     0.636     0.636      0.795    FC
                 0.5      0.024      0.833     0.5       0.625      0.726    HS
                 0.563    0.229      0.529     0.563     0.545      0.69     PC
                 1        0.022      0.857     1         0.923      0.985    FL
                 0        0          0         0         0          0.49     DC
Weighted Avg.    0.569    0.126      0.595     0.569     0.572      0.73
```

```
=== Confusion Matrix ===
 a b c d e f   <-- classified as
 2 0 0 5 0 0 | a = SC
 2 7 1 1 0 0 | b = FC
 2 2 5 1 0 0 | c = HS
 4 2 0 9 1 0 | d = PC
 0 0 0 0 6 0 | e = FL
 0 0 0 1 0 0 | f = DC
=== Run information ===
Scheme:weka.classifiers.trees.J48 -C 0.75 -M 2
Relation:     train2007
Instances:    51
Attributes:   4
              UGP101
              UGP201
              UGP303
              RES
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
UGP201 = C+: SC (3.0/2.0)
UGP201 = B
|   UGP303 = F: PC (1.0)
|   UGP303 = C: SC (1.0)
|   UGP303 = A: HS (2.0)
|   UGP303 = A+: HS (0.0)
|   UGP303 = B+: HS (2.0)
|   UGP303 = B: SC (2.0/1.0)
|   UGP303 = O: HS (0.0)
|   UGP303 = D: HS (0.0)
|   UGP303 = C+: FC (1.0)
UGP201 = A: FC (4.0/1.0)
UGP201 = A+: FC (2.0)
UGP201 = B+
|   UGP101 = C: HS (1.0)
|   UGP101 = B: FC (2.0)
|   UGP101 = C+: HS (2.0/1.0)
|   UGP101 = A: HS (0.0)
|   UGP101 = F: SC (1.0)
|   UGP101 = B+: HS (0.0)
|   UGP101 = D: HS (2.0/1.0)
|   UGP101 = A+: HS (0.0)
|   UGP101 = O: HS (0.0)
UGP201 = O: FC (3.0/1.0)
UGP201 = C: SC (3.0/1.0)
UGP201 = D
|   UGP303 = F: PC (4.0)
```

```
|   UGP303 = C: PC (3.0)
|   UGP303 = A: SC (1.0)
|   UGP303 = A+: PC (0.0)
|   UGP303 = B+: PC (0.0)
|   UGP303 = B: PC (0.0)
|   UGP303 = O: PC (0.0)
|   UGP303 = D: PC (2.0)
|   UGP303 = C+: PC (0.0)
UGP201 = F: FL (9.0/3.0)
Number of Leaves  :         33
Size of the tree :  37
Time taken to build model: 0.11 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          29               56.8627 %
Incorrectly Classified Instances        22               43.1373 %
Kappa statistic                          0.4577
Mean absolute error                      0.1634
Root mean squared error                  0.3505
Relative absolute error                 61.8652 %
Root relative squared error             96.5361 %
Total Number of Instances               51
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.286    0.114      0.286     0.286     0.286      0.594    SC
                 0.727    0.125      0.615     0.727     0.667      0.781    FC
                 0.4      0.171      0.364     0.4       0.381      0.674    HS
                 0.563    0.057      0.818     0.563     0.667      0.839    PC
                 1        0.067      0.667     1         0.8        0.941    FL
                 0        0          0         0         0          0.48     DC
Weighted Avg.    0.569    0.102      0.578     0.569     0.561      0.766
=== Confusion Matrix ===
 a b c d e f   <-- classified as
 2 0 3 2 0 0 | a = SC
 1 8 2 0 0 0 | b = FC
 2 4 4 0 0 0 | c = HS
 2 0 2 9 3 0 | d = PC
 0 0 0 0 6 0 | e = FL
 0 1 0 0 0 0 | f = DC
```

V. CONCLUSION

The classification accuracy of reduct as well as the classification accuracy of the whole set of attributes was coming same as 56%. This indicates that reduct obtained was the right reduct and it has the same classificatory power as that of the whole set of conditional attributes. In fact this proves the very essential characteristic of reduct.

VI. FUTURE WORK

Reduct property was checked on the educational data set of student result. This can be checked in by taking data set from different domains of educational data apart from result. Also different algorithms can be taken for finding out the reduct and validating their classification accuracy using other software.

REFERENCES

[1] E. Baker, International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier. 2008, pp-123-37.
[2] C. Romero, S. Ventura, Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications 33, 2007 pp-135–146.
[3] J. Beck. Engagement Tracing: Using Response Times to Model Student Disengagement. In proceedings of the International Conference on Artificial Intelligence in Education. 2005. pp-46-54.
[4] C.Vialardi et. al. Recommendation in Higher Education Using Data Mining Techniques. Educational Data Mining, 2009.pp- 143-175.
[5] S. Ayesha and T. Mustafa. Data Mining Model for Higher Education System. European Journal of Scientific Research 43, 1, 2010 pp-24-29.
[6] Z. Pawlak Rough Sets. International Journal of Information and Computer Sciences, Vol. 11, No. 5, 1982 pp. 341-356.
[7] P Ramasubramanian. et.al, Teaching result analysis using rough sets and data mining. Journal of computing, VOLUME 1, ISSUE 1, .2009. pp-78-83. ISSN: 2151-9617
[8] S. Narli, An alternative evaluation method for Likert type attitude scales: Rough set data analysis. Scientific Research and Essays 5(6), 2010 pp-519-528.
[9] V. Pallamreddy and V. Sreenivasarao, The Result Oriented Process for Students Based On Distributed Data Mining. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1, No. 5, November, 2010. pp-34-41.
[10] Z.Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishing, Dordrecht, 1991. pp-87-93.
[11] L. Polkowski, Rough sets: Mathematical Foundations. Advances in Soft Computing, Physica Verlag, Heidelberg, Germany, 2002.pp-42-47.
[12] A. Øhrn, Discernibility and rough sets in medicine: Tools and Application. Department of Computer Science and Information Science. Trondheim, Norway. Norwegian University of Science and Technology, 1999 pp. 239.
[13] A.Skowron and C. Rauszer, The discernibility matrices and functions in information systems. Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, 1992. pp. 331–362.

AUTHORS PROFILE

Dr. Nilesh Mahajan: Dr. Mahajan is Professor in Institute of Management and Enterpreanurship development (IMED), Bharati Vidyapeeth Deemed University Pune, India

Mrs. Jyoti Namdeo: She is postgraduate in Computer Applications (MCA) and research scholar in Institute of Management and Enterpreanurship development (IMED), Bharati Vidyapeeth Deemed University Pune, India