International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Construction Of Opinion Models For E-Learning Courses By Rough Set Theory And Text Mining

C.S.Sasikumar,A.Kumaravel

Abstract: ***Extracting knowledge through the machine learning techniques in general lacks in its predictions the level of perfection with minimal error or accuracy. Recently, researchers have been enjoying the fruits of Rough Set Theory (RST) to uncover the hidden patterns with its simplicity and expressive power. In RST mainly the issue of attribute reduction is tackled through the notion of 'reducts' using lower and upper approximations of rough sets based on a given information table with conditional and decision attributes. Hence, while researchers go for dimension reduction they propose many methods among which RST approach shown to be simple and efficient for text mining tasks. The area of text mining has focused on patterns based on text files or corpus, initially preprocessed to identify and remove irrelevant and replicated words without inducing any information loss for the classifying models later generated and tested. In this current work, this hypothesis are taken as core and tested on feedbacks for e-learning courses using RST's attribution reduction and generating distinct models of n-grams and finally the results are presented for selecting final efficient model.***

Keywords:Text Mining, n-grams; Rough Set Theory; attribute reduction; prediction accuracy; correlation.

I. INTRODUCTION

Data mining is a process of finding useful information and used to find the patternsand correlationsbetween huge datasets to predict outcomes which are hard to extract. These reviews are varying from user to user and they are full of features which help to analyze the courses and their difficulties. The data mining needs an information system where these features as condition attributes and the reviews labeled with decision class by experts as decision attribute are structured in a matrix form. Identifying the features of the feedbacksis being done by machine learning tools like WEKA[20]. The authors [22] try to use learning models with randomized and synthetic data sets.Identifying a subset of features which are important and contributing to the final output is computationally intensive and has exponential complexity. This will become more difficult when features from imprecise and incomplete text of opinions.There are many algorithms to reduce the dimension of the search space practically. Among which the elegance of Rough Set Theory makes the efforts put forth by the researchers more effective. Moreover, their applications are aligned in machine learning domain [1].

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

C.S.Sasikumar*, Research Scholar, Department of Computer Science and Engineering Bharath Institute of Higher Education and Research, Chennai, India

A.Kumaravel, Professor, Dean, School of Computing, Bharath Institute of Higher Education and Research, Chennai, India

Uncertainty by the presence of superfluous features, RST helps to find the important attributes which leads to attribute reduction [2]. Moreover RST is capable of optimizing through soft computing approaches [3]. In this study,experimental results indicate the model based on diagrams and uni-grams are effective in text mining classification.

II. METHODS AND MATERIALS

2.1 Data Source

E-learning happens to be a vital and familiar learning environment and it has been developed without human resources, typically through online. For the exercises in this study, we download the feedbacks for E learning courses available in EC council university which is an online learning source [19]. The courses offered by this forum are based on the cyber security professions for example Network defender, ethical hacker, Threat intelligence analyst, Security Analyst, Penetration tester, Forensic investigator, etc.The feedbacks of the learners nearly 700 are given in the website of EC council. Among them, we distinguished 340 are highly rated, 325 are medium rated and remaining are neutral. In order to achieve an impartial data distribution for our binary classification, we have measured only 320 high rating and 320 medium rating feedback documents. These documents are carried out through preprocessing by the filter stop words, stemming and tokenization which results in an amount of words. These words are becoming attributes, among them 184 are unigrams, 62 are bigrams and 17 are trigrams. The N-gram model is the most important tools in speech and language processing[5].In order to learn the weight of words in text classification by rough set theory, we used the three models that were developed using the n-grams combinations of words which are mentioned in Table 1.The Model I is with only unigrams, Model II is with unigrams and bigrams and Model III is with unigrams, bigrams, and trigrams. The models are created based on the frequency of words in each document and therefore the processed data models are obtained with discrete values [4].

Table 1: Models with their types of attributes

	Total no. of instances	High Rating instances	Medium Rating instances	Attributes type	No. of attributes
MODEL TYPE I	640	320	320	Unigrams = 184	184
MODEL TYPE II	640	320	320	Unigrams + bigrams = 184+62	246
MODEL TYPE III	640	320	320	Unigrams + bigrams + trigrams = 184 +62+17	253

2.2 Rough Set theory

Theory based on Rough Sets is a new mathematical approach to an imperfect knowledge. If the knowledge is not perfect, then it is imperfect knowledge[6]. The real world is unpredictable. Sensors and actuators may not be perfect. So in this dynamic environment, something may change without our control and knowledge. It may invalidate our knowledge sometimes. This can lead to incorrect perceptions and uncertainty which is a state of having limited knowledge where it is impossible to describe the future outcomes. To get rid of these problems, finally, a polish computer scientist Zadeh proposed the theory called fuzzy set theory[7]. After that, the new theory was proposed by Pawlak in 1981 is the Rough set theory which is expressed by a boundary region of a set and defined in terms of topological operations called approximations. It offers mathematical tools to discover pattern hidden in data. Over 2300 paper has been published on rough sets and their applications so far.

2.3 Information system and Approximation of sets

An information table can be seen as a decision table which has condition attributes(C) and decision attributes(D). The decision table is deterministic if and only if C implies D, otherwise non-deterministic. In our experiment, our models are acting as three different information systems and the approximation of each decision tables are found using Rough set theory.Two kinds of approximations are formed the rough set. The lower approximation consists of all objects which certainly belong to the set and the upper approximation contains all objects which probably belong to the set. The difference between the upper and the lower approximation forms the boundary region of the rough set[8].Many tools are using rough set theory in which we used ROSE 2tool[21].This tool only has the fundamentals of Rough set theory than others [9, 10].The set of objects/instances which can be certainly classified as objects of positive/negative, employing the attributes of models and the set of objects which can be possibly classified as elements of positive/negative, using the attributes of described models are given in Table 2. Using lower and upper approximations, one can calculate the quality and accuracy of approximation[11]. The values will be the numbers between [0,1] and this will describe the instances using the information prescribed in the original data.

The accuracy of the approximation is defined as

The accuracy of the approximation is defined as

$$Accuracyrelatedtoneg/posclass = \frac{No.ofobjectsbelongstolowerapproximationofneg/posclass}{No.ofobjectsbelongstoupperapproximationofneg/posclass}$$

The quality of approximation is defined as

$$Qualityrelatedtodecisionclass = \frac{No.ofobjectscorrectlyclassifiedasbothclassesbytheattributes}{No.ofobjectsintheuniversalset}$$

Table 2: Accuracy and Quality of classification

	MODEL I (184)			MODEL II (246)			MODEL III (253)		
	Lower approx.	Upper approx.	Accuracy of approx.	Lower approx.	Upper approx.	Accuracy of approx.	Lower approx.	Upper approx.	Accuracy of approx.
Negative class	243	396	0.6136	259	344	0.7529	260	343	0.7580
Positive class	249	383	0.6501	266	361	0.7368	258	357	0.7227
Qual. of classification	0.7688			0.8203			0.8094		

If the value of quality of approximation equals 1 says that the classification is acceptable otherwise the elements of the sets have been vaguely classified to the positive region using the set of attributes. Our results show that0.7688, 0.8203 and 0.8094sizes of objects are correctly classified as positive and negative using the attributes of Model I, Model II and Model III respectively.

2.4 Concept of attribute reduction

The next step of the Rough set analysis is to construct the minimal subset of attributes called Reduct that confirming the same quality of classification as the condition attributes of the original set[12].That means the number of equivalence classes of the reduct set of attributesmust be equal to the number of equivalence class of the original attribute set and our experimental results on reduction is given in table 3.

Table 3: Details of reducts

	# of reducts	Min length	Max length	# of core attributes	# of describing attributes
MODEL TYPE I	45	97	102	93	184
MODEL TYPE II	62	150	155	148	246
MODEL TYPE III	86	166	171	164	253

In Model I, the no. of reducts are 45, from that we have deducted the set which satisfiesthe properties of reduction in Rough set theory. We know that the quality of classification for the reducts should be the same as the original set. To do that we need to know the core attributes. The core attributes are the main attributes of the system and it can be found at the intersection of all reducts. We should not eliminate any of the attributes from the core otherwise the quality of approximation will be disturbed. In the Model I, the no. of core attributes are 93 and lengths of the reducts vary as97,98,99,100 and 101. Here we could see 11 attributes are more in the reducts apart from the core attributes. Since the core attribute does not attain the same quality as the original set, we have to select the important attributes from the 11 attributes to reach the quality of approximation. The important of attributes is calculated using the frequency percentage of the attributes in the reduct set.We could see that from column 5 and 6 in table 3, many attributes are redundant.If the no. of occurrences are high in the reduct may improve the quality and accuracy of the classification[13].The highest frequency of attributes is 100% refer the attributes of a core. Since the core attributes failed to ensure the same quality of the original set, at least 50% frequency of attributes in the reducts are characterized and form a minimal reductthat satisfied the properties of reduction in Rough set theory. Since this reduct attains the original quality without the attribute "low", we removed that as redundant.Thisreduct has the same approximation of decision classes 0.7688 as the original set Model I. But the core was only 0.7189 quality of approximation which may lose some information from the original set.In this case, all the reducts and the core should be presented for consideration in the tables in view of getting an opinion about what reduct should be used to create decision rules from the reduced decision table.For each model, we can view a significant reduction in terms of the number of attributes positioned in the reduct. In a similar way, we found the other two perfect reducts with Model II and Model III.

There we got the accurate approximation for the attributes which occurs more than 15 times in the 62reducts of Model II. In case of Model III, we adjusted the frequency level upto 15% and we found 27 more attributes are indispensable as core attributes.Finally, in each model, we got a reduct that satisfies the property of reduction that means each minimal set of attributes of Model reach the quality of classification of the original set. This ensures that the decision rules derived from these reducts preserve the exact information as the universal set.

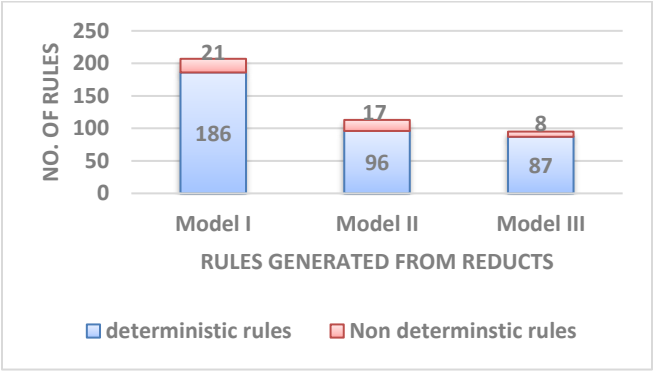

Figure 1: Rules generated by Modlem-entropy

2.5 Prediction accuracy of rules

One of the most important elements of data analysis is rule generation. We used MOLEM2minimal covering rules for rule induction[10]. MODLEM is the modified learning from examples method. It uses rough set theory to manage inconsistent objects and calculates a single local covering for each approximation of the concept[14, 15].

Each Model's reduct generated the unique and approximation rules which are given in chart 2. The unique rules are deterministic to define the decision rules, whereas the approximation rules are possible to define. Generally,the data analyst wants to know which generated rules are worthy that is how fine they can classify objects. The prediction accuracy is evaluated based on the number of correctly classified objects. We have taken the cross-validation type of test to examine the accuracy based on a minimal covering algorithm that says the minimal number of possibly shortest rules covering all the objects. This validation test results of rules of experimented models are charted below figure2.

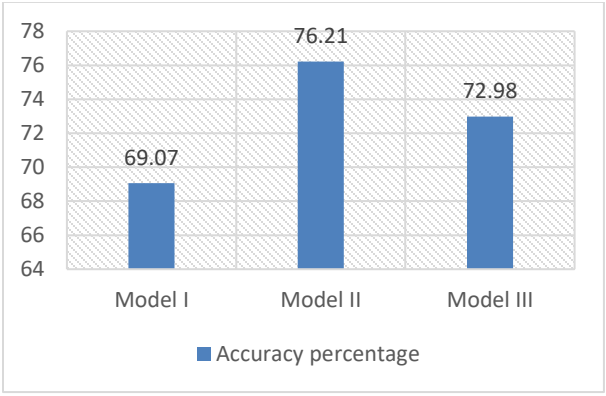

Figure 2: Accuracy of the reducts of each Model

III. RESULTS AND DISCUSSIONS

We got reducts for each Model that constructed by using n-grams and their performances on the classification of objects were evaluated using a cross-validation test.

The test results help us to get the knowledge about the generated rules that how they are worth to make a decision and how far they are good to classify the objects as correctly and incorrectly classified. Accordingly, the classification accuracy of objects was obtainedfor each Model using their rules, which were derived from their reducts using ModLem2 algorithm, which were shown in the chart 3.

The prediction accuracy of each model is shown in chart 4 where Model I is increased from 68.06% to 69.07, Model II is increased from 72.22% to 72.98% and Model III is increased from 71.94% to 76.21%. Among them, the accuracy of Model II has a better prediction accuracy, which tells us the combination unigrams, and bigrams can boost the precision of objects than other combinations. Also, we could find that the trigrams are not enhancing the classification much as others. For the reason that the reduct set of Model III contains only one trigram however the original set has 9 trigrams and the core set of attributes also has one which causes that the importance of trigrams in text classification using rough set theory is less.The rough set theory is a powerful tool in classification problems [16].Few results have shown that merging individual classifiers is an efficient method for progressing classification accuracy [17, 18]. But through this paper, we are showing that the combination of Rough set theory approach with any type of classifier will improve the classification accuracy in opinion mining.

IV. RESULTS AND DISCUSSIONS

We got reducts for each Model that constructed by using n-grams and their performances on the classification of objects were evaluated using a cross-validation test. The test results help us to get the knowledge about the generated rules that how they are worth to make a decision and how far they are good to classify the objects as correctly and incorrectly classified. Accordingly, the classification accuracy of objects was obtainedfor each Model using their rules, which were derived from their reducts using ModLem2 algorithm, which were shown in the chart 3. The prediction accuracy of each model is shown in chart 4 where Model I is increased from 68.06% to 69.07, Model II is increased from 72.22% to 72.98% and Model III is increased from 71.94% to 76.21%. Among them, the accuracy of Model II has a better prediction accuracy, which tells us the combination unigrams, and bigrams can boost the precision of objects than other combinations. Also, we could find that the trigrams are not enhancing the classification much as others. For the reason that the reduct set of Model III contains only one trigram however the original set has 9 trigrams and the core set of attributes also has one which causes that the importance of trigrams in text classification using rough set theory is less.The rough set theory is a powerful tool in classification problems [16].Few results have shown that merging individual classifiers is an efficient method for progressing classification accuracy [17, 18]. But through this paper, we are showing that the combination of Rough set theory approach with any type of classifier will improve the classification accuracy in opinion mining.

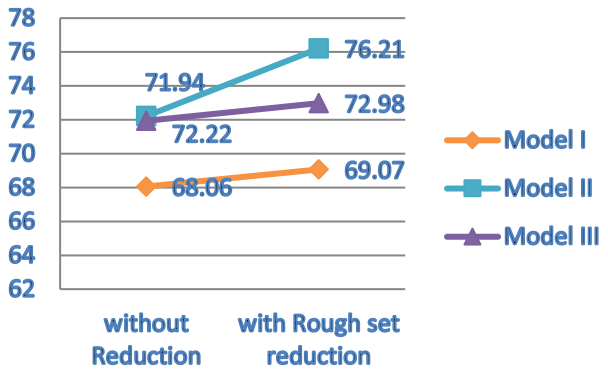

Figure 3: RST prediction accuracy of Models

The performance difference of Rough set theory rule-based classifier and other individual classifiers were found to be significant for all three models as shown in Figure 4.. Also, we could see thatthe degree of correlation is highly positive for the Model II. Since this paper has the aim to find the performance level of Rough set theory, we focused on the degree of correlation of other classifiers with RST only and we found Naïve bayes classifier is highly correlated, which results in case of analyzing the performance of Rough set theory, we have to use classifiers like Naïve bayes in order to improve the accuracy. Since the performance of classifiers is evaluated for product reviews, this needs to be done with other application domains.

V. CONCLUSIONS

The aim of this paper was to find out the aspects of Rough set theory in the analysis of text mining tasks. A massive amount of new information and data are generated every day through economic, academic and social activities. Techniques such as text mining and analytics are required to exploit this potential. So this paper may help the researchers those involved in research of text mining to get a better way of classifying texts using rough set theory. Especially whenmining product reviews, a prediction accuracy of decision rules is improved for perfect preprocessing. Here we found, grouping of unigrams and bigrams words helps to get better accuracy and will lead to acquiring a better knowledge when applying Rough set theory for mining. Consequently, the Rough set theory proves once again a better approach to furnish reducts of independent measure having the same capability of approximating the decision as the whole set and can be used to stipulate a solution.Weexpect this research will contribute to the further research of the Rough set theory in opinion mining and believe the next step in such analysis should investigateadvancethiswith rule-basedclassifiers for the development.

REFERENCES

1. Jensen, Richard, and QiangShen. "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches." Knowledge and Data Engineering, IEEE Transactions on 16.12 (2004): 1457-1471.
2. Polkowski, L. (2013) Rough Sets: Mathematical Foundations. Springer Science & Business Media, Berlin.
3. Qinghua Zhang, Qin Xie, Guoyin Wang, A survey on rough set theory and its applications, CAAI Transactions on Intelligence Technology 1 (2016) 323-333.
4. G. Vinodhini, R M Chandrasekaran, 'Opinion mining using principal component analysis based ensemble model for e-commerce application' , CSIT (November 2014) 2(3):169–179.
5. Daniel Jurafsky& James H. Martin , "N-Grams" Chapter 4 of 'Speech and Language Processing'., Draft chapter in progress , January 2017.
6. Pawlak, Z., Rough sets, J. Comput. Information Sciences, vol.11, pp.341-345, 1982.
7. Pawlak, Z., Rough sets and fuzzy sets, Fuzzy Sests and Systems, vol.17, pp.99-102, 1985.
8. Z. Pawlak, Rough Sets – Theoretical Aspects of Reasoning about Data. Boston, London, Dordrecht: Kluwer, 1991.
9. Zain Abbas, Aqil Burney, 'A Survey of Software Packages Used for Rough Set Analysis',Journal of Computer and Communications, 2016, 4, 10-18.
10. M. Sudha , A. Kumaravel, 'Performance Comparison based on Attribute Selection Tools for Data Mining',
11. Indian Journal of Science and Technology, Vol 7(S7), 61–65, November 2014.
12. Pawlak Z., Slowinski R., "Decision Analysis Using Rough Sets", International Transaction Operational
13. Research Vol. 1, No. 1, 1994.
14. Ahmad F., Hamdan A.R., Bakar A.A., "Determining Success Indicators of E-Commerce Companies Using
15. Rough Set Approach", The Journal of American Academy of Business , Cambridge, September 2004.
16. Dimitras A, Slowinski R., Susmaga R., Zopounidis C., "Business failure prediction using rough sets",
17. European Journal of Operational Research 114, 1999.
18. Grzymala-Busse, J.W. &Stefanowski, J., (2001), Three Discretization Methods for Rule Induction.
19. International Journal of Intelligent Systems, 16, 29-38.
20. MertBal, 'Rough Sets Theory as Symbolic Data Mining Method: An Application on Complete Decision
21. Table', Inf. Sci. Lett. 2, 1, 35-47 (2013).
22. Xiaohan Li, 'Attribute Selection Methods in Rough Set Theory' project of San José State University,may
23. 2014.
24. Xia Rui, ZongChengqing, Li Shoushan, 'Ensemble of feature sets and classification algorithms for opinion
25. classification', InfSci 181:1138–1152(2011).
26. Whitehead M, Yaeger L (2010) ''Sentiment mining using ensemble classification models.'' In: Innovations
27. and advances in computer sciences and engineering, Springer, Netherlands, 509–514.
28. Data source: https://blog.eccouncil.org/members-feedback/
29. Available from: http://weka.sourceforge.net/doc.dev/weka/attributeSelection
30. ROSE Software: http://idss.cs.put.poznan.pl/site/rose.html
C.S.Sasikumar, A.Kumaravel, Attribute Selection on Student Academic and Social Attributes Based on Randomized And Synthetic Dataset, International Journal of Engineering & Technology, 7 (4.39) (2018), 1069-107