International Journal of Engineering, Science and Mathematics
Vol. 6 Issue 8, December 2017 (Special Issue)
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com Double-Blind
Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

An Empirical Comparison Of Attribute Reduction, Rough Set Theory And Machine Learning Algorithms

M.Sudha*
A.Kumaravel**

Abstract

An increasing amount of data is becoming available on the internet. Each and every one of us is con producing and releasing data about something. Big Data can develop into a problem when different sou data are match up for commercial use in targeted advertising processes. Rough set theory is a new mathe approach to imperfect knowledge. It was proposed by *Pawlak(1982)*.The benefit of rough set theory analysis is that it can easily find the reducts using approximations without need any preliminary or ad information about data. The proposed approach provides efficient algorithms for deducting unseen patt data. We tested medical data set Thyroid using Rough set based tool inordertoreducethe attributes witho of informationand extractdecisionrules.Theresultsofthisstudyprovidetherelatedmembersanddecisionmakerstoreduceandprev hypothyroid problemof patientsby predicting them.

Keywords:

Rough set theory; Attribute reductuion; Rule induction; Prediction accuracy; Mining classifiers.

Author correspondence:

M.Sudha,
1Research Scholar, Department of Mathematics.
Amet University,Kanathur,Chennai-600112, India

1. Introduction (10pt)

Data mining isfinding of hidden predictive information from large databases. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into an useful information. Data mining is an interdisciplinary field which cover-up the areas statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, etc. The major motivation behind data mining is autonomously extracting useful information or knowledge from large data sets [1]. Attribute selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information [2]. In many real world problems attribute selection is must due to the plenty of deafening, unrelated or confusing features. For instance, by removing these factors, learning from data techniques can benefit.

The most important data mining technique which search through the entire dataset is an association rule generator that finds the rules revealing the nature and frequency of relationships between data entities. Rough

* Doctorate Program, Linguistics Program Studies, Udayana University Denpasar, Bali-Indonesia (9 pt)
** STIMIK STIKOM-Bali, Renon, Depasar, Bali-Indonesia

set can be used as a tool of data mining that is used for rule generation.The rough set approach[3] to data analysis has many important advantages that provides efficient algorithms for finding hidden patterns in data, finds minimal sets of data, evaluates significance of data, generates sets of decision rules from data etc., ROSE2 [4] is a tool based on rough set theory.

2. Methods and Materials

2.1 Rough set theory:

RST can be defined by means of approximations called lower and upper. The set of instances is called universe U and we assume an equivalence relation R to represents the knowledge about instances in U. To characterize the set X with respect to R, we need the concept of rough set theory that is lower and upper approximations [5].

2.2 LowerandUpperapproximations

Rough set theoryanalysisisbasedon two approximationssuchas upper and lower approximations[6].

Lowerapproximation is the union of elements possibly belonging to a concept (set) with respect to R. They definitely belong to the set.

$$R_*(x) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\} \qquad (1)$$

Upperapproximation is the union of elements possibly and partially belonging toaconcept(set) with respect to R. i.e., they roughly are in the set.

$$R^*(x) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \varnothing\} \qquad (2)$$

Boundary set is the set of all objects that can be neither classified as X and nor X compliment with respect to R. that is the boundary region of the set is the difference between the lower and upper approximations [7].

$$RN_R(x) = R^*(x) - R_*(x) \qquad (3)$$

2.3 Quality and accuracy of approximations.

Using lower and upper approximations one can calculate the quality and accuracy of approximation [8]. 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 approximation is defined as

$$Accuracyrelatedtodecisionclass = \frac{No.ofobjectsbelongstolowerapproximationofdecision\ class}{No.ofobjectsbelongstoupperapproximationofdecisionclass} \qquad (4)$$

The quality of classification is defined as

$$Qualityrelatedtodecision\ class = \frac{No.of\ objects\ correctly\ classified\ as\ both\ classes\ by\ the\ attributes}{No.ofobjectsintheuniversalset} \qquad (5)$$

The accuracy of approximation decides whether the set is a rough set or a crisp set with respect to set of attributes. If Accuracy is equal to 1, then the set is rough otherwise it is a crisp set.

2.4 AnApplicationof Rough set theory

Rough set theory has reached widely in the last decade. The applications of rough set theory covers the area of information sciences, decision analysis, medical diagnosing[9], software safety analysis, economic and financial prediction, environmental production, signal and image processing, social sciences, molecular

biology and chemistry(pharmacy) etc. [10] .In this paper we deal with decision analysis using RST which become a main application of RST now a days. To experiment this we have taken Thyroid problem affected patients records as data.

Toprepare the data asaninformationsysteminRST,itisneededtoidentify theparameters called condition and decision.Afterdefiningofconditionanddecisionattributes,informationsystemis created asshown in Table 2.Thistableiscalledasdecisionsystem where eachrowrepresents some condition attributes and decision attributes [11].

3 Data Experiment

In order to compare Rough set theory with Naivebayes, Knn and J48 classifiers we tested them on Thyroid data set selected from the UCI machine learning repository [12]. Since this data set was already used and tested by many machine learning algorithms and methods which says that thyroid data is perfect for classification and the sequential floating forward selection method achieved good classification results. The best accuracy obtained by the Knn algorithm during feature selection was 97.99%, using 6 of the 21 available features. 100 bootstrap tests on this feature set yielded a mean bootstrap accuracy of 98.06%. This accuracy is similar to those obtained by the various methods reported by Weiss and Kapouleas [13]. Table 1 summarizes the datasets used for experiments.

Table 1 Data sets

Data set	Number of instances	Number of attributes	Number of classes
Thyroid	7200 (3772 training instances, 3428 testing instances)	22 (21 numeric ,1 nominal decision)	3 (1,2,3)

Many tools are using rough set theory tools in which we use ROSE 2 tool [4]. This tool only has the fundamentals of rough set theory than others [14]. The set of objects/instances which can be certainly classified as objects of class1or 2 or 3, employing the attributes of models and the set of objects which can be possibly classified as elements of classes, 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 using equations (4) and (5) [15]. 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 approximation is defined as

Table 2 Data set-Thyroid

	Lower approx.	Upper approx.	Accuracy of approx.
Class 1	165	167	0.9880
Class 2	354	372	0.9516
Class 3	6661	6681	0.9970
Quality of classification		0.9972	

The quality of classification 0.9972 says about consistency and the numbers of equivalence classes are 1767 which should not be changed while removing the attributes. 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.

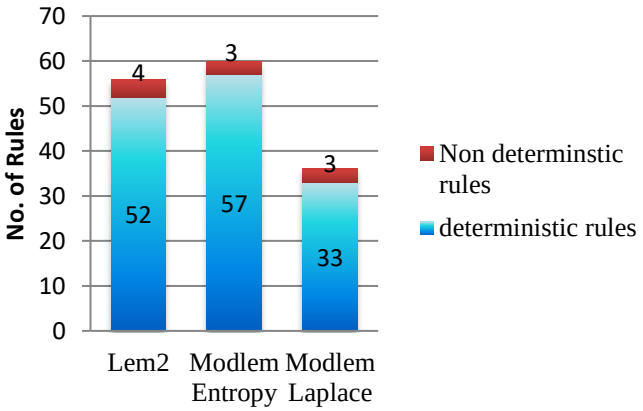
3.1 Concept of attribute reduction

The next step of the rough set analysis is to construct the minimal subset of attributes called Reduct that conforming the same quality of classification as the condition attributes of original set [16]. That means the number of equivalence classes of the reduct set of attributes must be equal to the number of equivalence class of the original attribute set and our experiment results on reduction shows that 12 attributes are relevant to the data which conformed by both lattice search and discernibility matrix. According to the reduct, remaining attributes are removed and rough set theory was applied using approximations and classification performances are observed. For rule induction, the LERS system based algorithms Lem2, Modlem-Entropy and Modlem-Laplace are used and the number of certain and approximate rules were recorded in the table

Thyroid data	LEM2	MODLEM-Entropy	MODLEM-Laplace
Number of Rules	56	60	34
Prediction accuracy	99.49%	99.39%	99.53%

Since the number of rules are minimum in Modlem-Laplace and the consistency of rules were also high compared with other rule induction algorithms. These rule were validated using minimal covering and tested by 10 cross validation to have percentage of correctly classified instances.

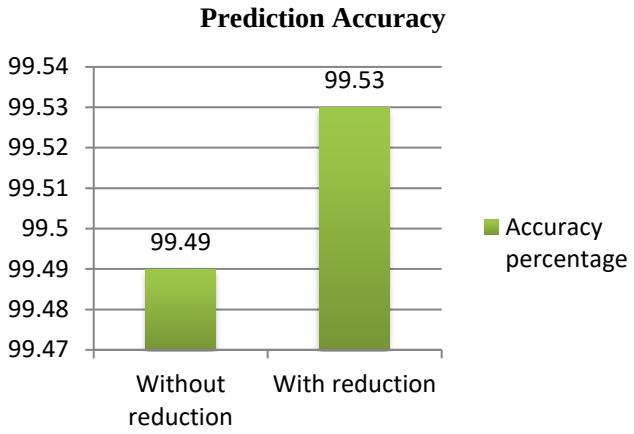
Chart 1 Rule induction on reduct



A reduct generated the unique and approximation rules which are given in chart 1. The unique rules are deterministic to define the decision rules. Generally data analyst wants to know what generated rules are worthy, that is how fine they can classify objects.

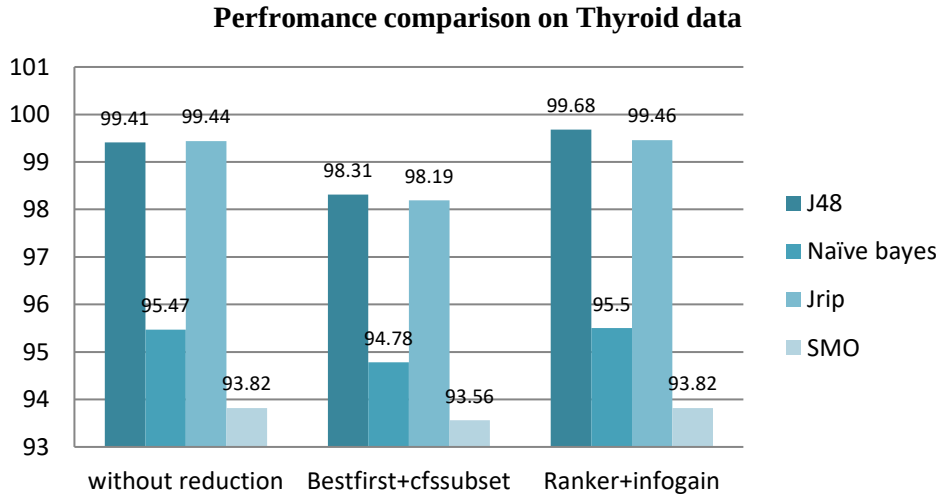
4. Results and Discussion

Chart 2 performance comparison on feature reduction.



The prediction accuracy was improved from 99.49% to 99.53% using rough set theory reduction concept which is shown in chart 2. Overall the performance of Rough set theory is appreciable on Thyroid data compared to other data mining techniques. The performance of those techniques was analyzed with Genetic algorithm in [17]. For the thyroid data, the GA–Knn was more effective than all, but required only three of the attributes to make the classification. The GA–Knn attained classification accuracy within 1% of the best technique reported. On the another research [13], the same data was analyzed with neural nets. The neural nets did perform well, and they were the only statistical classifiers to do well on the thyroid problem. However, overall they were not the best classifiers; they consumed enormous amounts of time; so it may be probably that performance can be improved. Now the same data is measured by rough set theory and with other mining algorithms. For purposes of comparison of the mathods, the prediction accuracy attained by them were filled in with line chart 3

Chart 3 Prediction accuracy of feature selection methods



Overall the performance of J48 is better than other algorithms but it requires 13 attributes which was extracted from 21 attributes by Infogain attribute evaluator + ranker method. Also we observed that the algorithms achieved poor accuracy in the reduct of cfssubset + Bestfirst/greedystepwise than the data without reduction. But the method extracted only 5 attributes from 21 available features. When comparing these results with Rough set theory, the reduction of RST was effective and efficient in classification and prediction was proved empirically using Thyroid data.

5. Conclusion

The aim of this paper was to find out the aspects of rough set theory for analysis of text classification, in particular reduction of attributes. In this paper, we investigated the Thyroid data for pattern recognition. The same data was investigated with KNN and genetic algorithm reported in [17]. Here in our experiment we used Rough set approach which reduced the size of the attributes without loss of information. RST reduction concept shows better prediction accuracy than other pattern recognition methods. Also RST attains 99.53% while Knn and GA feature extraction reported 95.3% and 98.4% respectively. With the minimum number of attributes we could derive the decision rules which provide us the useful information about the patient whether they referred to the clinic for hypothyroid and help to diagnosis the new instances in future. The decision rules generated by the précised data can give significant information forthe researchers and decision makers. Still the classification techniques remain to be investigated with RST in the domain like text mining. We hope this study will help

the researches in pattern extraction and believe that the next step in such analysis should further investigate the decision rules.

ACKNOWLEDGEMENT

The first author would like to thank of AMET University for the support and the encouragement for this research work.

3. Results and Analysis (10pt)

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [2], [5]. The discussion can be made in several sub-chapters.

3.1. Sub section 1

Sub section should be written without bold type. The result and analysis are presented by present form. Please avoid too many paragraph in this section.

3.2. Sub section2

Sub section should be written without bold type. The result and analysis are presented by present form. Please avoid too many paragraph in this section.

4. Conclusion (10pt)

Provide a statement that what is expected, as stated in the "Introduction" chapter can ultimately result in "Results and Discussion" chapter, so there is compatibility. Moreover, it can also be added the prospect of the development of research results and application prospects of further studies into the next (based on result and discussion).

References(10pt)

[1]. Huan Liu, Hiroshi Motoda., 2008 "Computational Methods of Feature Selection" by Taylor & Francis Group, LLC., pp 23-26

[2]. IPeter scully, Dr. Richard Jensen, 2011 "Investigating rough set feature selection for gene expression analysis" pds7@aber.ac.uk

[3]. J. W. Grzymala-Busse, 1992 LERS – A system for learning from examples based on rough sets. In Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory, 3–18

[4]. W.Ziarko. IDSS, 1999, 'ROSE User Manual: A Tool for Building', www.rosecompiler.org, Source-to-Source Translators

[5]. Z. Pawlak, «Rough Sets and Intelligent Data Analysis,» Information Sciences, Vol147, no. 1, pp. 1-12, 2002

[6]. Z. Pawlak, Rough sets, International JournalofComputer andInformationSciences, Vol 11, pp. 341-356, 1982.

[7]. M. Sudha , A. Kumaravel, 'Performance Comparison based on Attribute Selection Tools for Data Mining', *Indian Journal of Science and Technology, Vol 7(S7), 61–65, November 2014.*

[8]. Maciocha, A and Kisielnicki, J. "Intellectual Capital and Corporate Performance" The Electronic Journal of Knowledge Management Volume 9 Issue 3 (271-283)

[9]. G. Ilczuk ve A.Wakulicz-Deja, 'Rough SetsApproach to Medical DiagnosisSystem' AWIC, pp. 204-210, 2005

[10]. Caner Erden, Fatih Tüysüz, 'An Application Of Rough Sets Theory On Traffic Accidents'An InternationalConference on Engineering and Applied Sciences Optimization, 2014.

[11]. Z. Pawlak, «Rough Sets,» Int. Journal ofComputerand Information Sciences,Vol 11, pp. 341-356, 1982.

[12]. http://archive.ics.uci.edu/ml/datasets/thyroid+disease

[13]. S. M. Weiss and I. Kapouleas, "An empirical comparison of pattern recognition, neural nets, and machine learning classification methods," in Proc. 11th Int. Joint Conf. Artif. Intell., N. S. Sridharan, Ed. Detroit, MI: Morgan Kaufmann, 1989, pp. 781–787.

[14]. Zain Abbas, Aqil Burney, 'A Survey of Software Packages Used for Rough Set Analysis',Journal of Computer and Communications, 2016, 4, 10-18.

[15]. Pawlak Z., Slowinski R., "Decision Analysis Using Rough Sets", International Transaction Operational Research Vol. 1, No. 1, 1994.
[16]. Ahmad F., Hamdan A.R., Bakar A.A., "Determining Success Indicators of E-Commerce Companies Using Rough Set Approach", The Journal of American Academy of Business , Cambridge, September 2004.
[17]. Dimensionality Reduction Using Genetic Algorithms Michael L. Raymer, William F. Punch, Erik D. Goodman, Leslie A. Kuhn, and Anil K. Jain, IEEE transactions on evolutionary computation, vol. 4, no. 2, july 2000.