Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 7 (2017), pp. 3249-3260
© Research India Publications
http://www.ripublication.com

Comparative Analysis between Rough set theory and Data mining algorithms on their prediction

M. Sudha
Assisteant Professor, Department of Mathematics.
Amet University,Kanathur,Chennai-600112, India.

A. Kumaravel
Dean, School of Computing,
Bharath University, Selaiyur, Chennai-600073, India.

Abstract

During the last few years there is a remarkable increase in development of Data mining techniques. Nowadays, various organizations store their information as a kind of databases. So huge amount of data and their informations are available in repositories. To interpret these data, we need effective mining techniques for better performance of classifications. This helps us to take best decision and prediction. This paper is prepared to analyze the performance of classification algorithms with help of data mining tool TANAGRA and Rough set theory. At the same time, we can reveal the best tool among them based on their performance level. To experiment this, huge size data has taken which tells that performance of classification tools are affected by the kind of dataset and significant results are discussed over comparative analysis.

Keywords: Tanagra, ROSE2.2 , Naïve Bayes, C4.5, K-NN, CRT, Predicted accuracy

1 INTRODUCTION

Data mining is the process of *extraction of hidden predictive information from large databases.* Generally, data mining is the process of analyzing data from different viewpoint and briefs it into useful information. Data mining is an interdisciplinary field which cover-up the area's 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 stores or sets [5]. Feature selection has been an active research area in pattern detection, and data mining areas. The main idea of feature selection is to choose a subset of input variables by removing attributes with little or no predictive information [6]. In many real-world problems, Feature selection is a must due to plenty of deafening, unrelated or confusing features. For instance, by removing these, learning techniques from data can be an advantage.

The most important data mining technique which searches through the entire data set is an association rule generator who finds the rules revealing the nature and frequency of relationships between data entities. Rough set can be used as a tool to generate rules form decision table in data mining. The rough set approach [7] to data analysis has many important advantages that provide 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 [13] is a tool based on rough set theory.

Tanagra is one of the few academic data mining tools to be able to produce reports that can easily be displayed in office automation software. For example, the tables can be copied into Excel spreadsheets for more computations.

In this paper, we are going to experiment a comparative study on a data set between Rough set theory [6] and data mining tools Tanagra according to their accuracy level. The process is figured below [fig1]

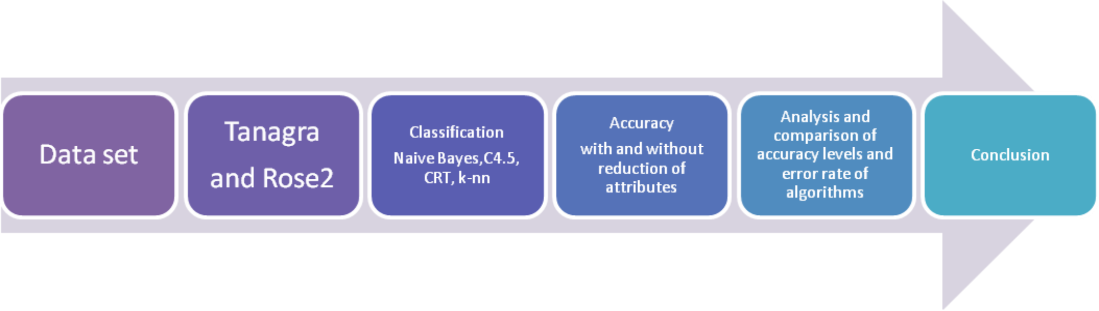

Data set
Tanagra and Rose2
Classification
Naive Bayes,C4.5, CRT, k-nn
Accuracy with and without reduction of attributes
Analysis and comparison of accuracy levels and error rate of algorithms
Conclusion

Figure 1 Process of Comparative analysis

2 METHODS AND MATERIALS

In this section, we describe the data set with various classification methods and compare them by the accuracy an error rate obtained with whole data set with and without selected attributes. Tanagra tool and Rose2 are used to reduce dimensionality with various attribute select techniques and classification models.

2.1. *Data set description*

The data has multivariate characteristics based on classification tasks. Number of instances in this data base is 420 and the number of attributes is 280 plus the class. This data available in (https://archive.ics.uci.edu/ml/datasets/Arrhythmia).

The aim of the data is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal'; classes 02 to 15 refer to different classes of arrhythmia and class 16 refers to the rest of unclassified ones. However, there are differences between the cardiology's and the program's classification. Taking the cardiologist as a gold standard, we aim to minimize this difference by means of machine learning tools.

2.2 *Attribute description*

The attributes of cardiac arrhythmia are enumerated below with their description:
1 Age: Age in years 2 Sex: Sex (0 = male; 1 = female) , nominal 3 Height: Height in centimeters , linear 4 Weight: Weight in kilograms , linear 5 QRS duration: Average of QRS duration in msec., linear 6 P-R interval: Average duration between onset of P and Q waves in msec., 7 Q-T interval: Average duration between onset of Q and offset of T waves in msec., 8 T interval: Average duration of T wave in msec., 9 P interval: Average duration of P wave in msec., 10-15 Vector angles in degrees on front plane of waves. 15 Heart rate: Number of heart beats per minute. Average width, Number of intrinsic deflections, Existence of ragged and Existence of diphasic derivation of waves on Channel DI,DII,DIII,AVR,AVL,AVF,V1,V2,V3,V4,V5 and V6 are the attributes in the range 16-27, 28-39,40-51,52-63,64-75,76-87,88-99,100-111,112-123,124-135,136-147 and 148-159 respectively. Then amplitude,QRSA and QRSTA of the above same channels are attributed in the following ranges 170-179,180-189, 190-199,200-209,210-219,220-229, 230-239,240-249,250-259,260-269 and 270-279 respectively. The decision class distributed by a discrete set as 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15 and 16 which distribution is discribed as follows.
01 Normal
02 Ischemic changes (Coronary Artery Disease)
03 Old Anterior Myocardial Infarction
04 Old Inferior Myocardial Infarction
05 Sinus tachycardy
06 Sinus bradycardy
07 Ventricular Premature Contraction (PVC)
08 Supraventricular Premature Contraction
09 Left bundle branch block
10 Right bundle branch block
11 degree AtrioVentricular block
12 degree AV block
13 degree AV block

14 Left ventricule hypertrophy
15 Atrial Fibrillation or Flutter
16 Others

3 METHODOLOGY

The information collected from the survey will be analyzed and evaluated in several stages in each tool.

3.1 Tanagra tool

TANAGRA is a machine learning [1,11] software proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area [12]. We want to predict a class attribute from discrete and continuous descriptors with or without Feature selection. We select cross validation in order to compare the error rate.

3.2 ***Algorithms used* : *Naïve Bayes***

Naïve Bayes [4] is fast to scan and classify and not sensitive to irrelevant features. It handles real and discrete data store efficiently. To simplify the task, naïve Bayesian classifiers assume attributes have independent distributions and thereby estimate.

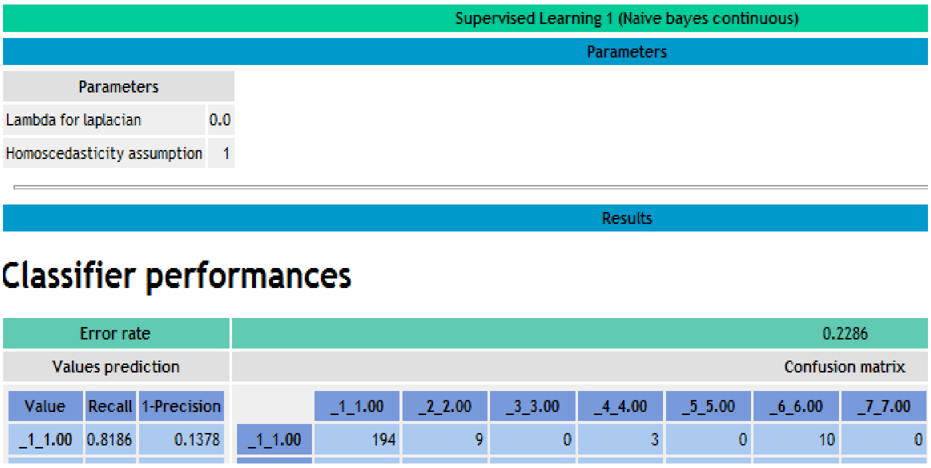
Bayesian classifiers use Bayes theorem, which says $p\left(\frac{c_j}{d}\right) = \frac{p\left(\frac{c_j}{d}\right)p(c_j)}{p(d)}$

$p\left(\frac{c_j}{d}\right) =$ probability of instance d being in class cj p(d | cj) = probability of generating instance d given class cj

$p(c_j)$ = probability of occurrence of class cj

$p(d)$ = probability of instance d occurring.

Since our data includes both discrete and continuous attributes, experiment with the method discrete naives bayes , convert all the continuous attributes into discrete using MDLPC from "Feature construction" tab [10]. Now we have all the attributes are of discrete type. Now apply Naïve Bayes " from Supervised learning tab. The figure 2 shows that supervised naïve bayes for continuous predictors gives the error rate 22.86 % and after reduction it increases to 24.29% which shows the naive ayes does not match with the filter.



Supervised Learning 1 (Naive bayes continuous)

Parameters

Parameters	
Lambda for laplacian	0.0
Homoscedasticity assumption	1

Results

Classifier performances

Error rate	0.2286
Values prediction	Confusion matrix

Value	Recall	1-Precision		_1_1.00	_2_2.00	_3_3.00	_4_4.00	_5_5.00	_6_6.00	_7_7.00
_1_1.00	0.8186	0.1378	_1_1.00	194	9	0	3	0	10	0

Fig 2: Results of Naïve bayes classifier performance and error rate

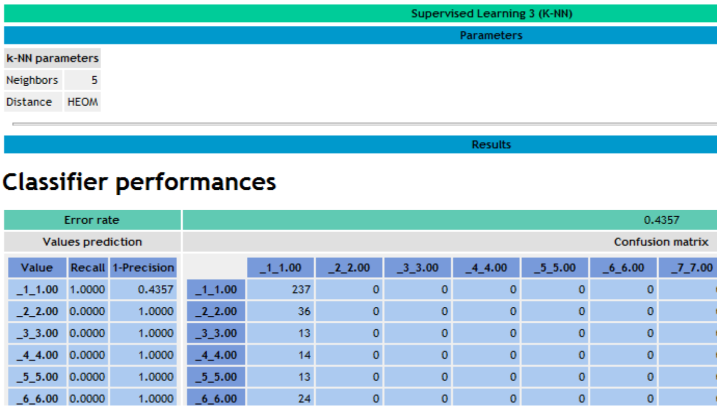
3.3 C4.5

Decision trees are controlling categorization algorithms. Accepted decision tree algorithms consist of C4.5, CRT, and CS-CRT. At the equivalent time as the name implies, this performance recursively separate inspection in branches to build tree for the purpose of improving the calculation accuracy [3].

Our attributes included both discrete and continuous. This algorithm accepted the data to classify; it gives 15.71% error rate and 84.28% accuracy. For feature selection, we apply supervised learning algorithm based approach RELEIF which approach does not take into consideration the redundancy of the input attributes, this filter reduced the 100 attributes to 279 and gave the error rate 15.48% after reduction that shows little improvement in the accuracy.

3.4 K-NN

The nearest neighbor algorithm. The K-nearest neighbor's algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms [13] and classifies cases based on a similarity measure. Now, convert all the discrete attributes in to continuous using 'disc to cont' from Feature construction tab. Now we have all the attributes are of continuous type.



Supervised Learning 3 (K-NN)

Parameters

k-NN parameters	
Neighbors	5
Distance	HEOM

Results

Classifier performances

Error rate	0.4357

Values prediction

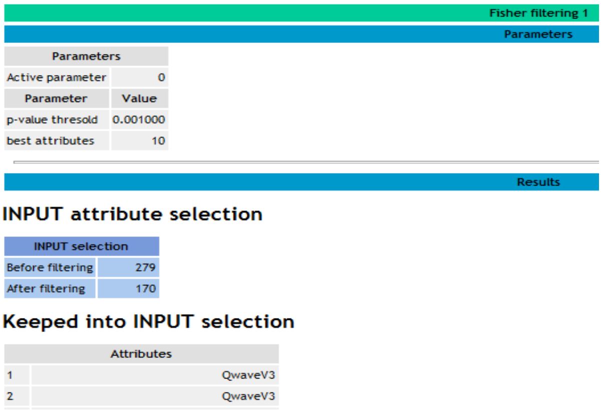
Value	Recall	1-Precision
_1_1.00	1.0000	0.4357
_2_2.00	0.0000	1.0000
_3_3.00	0.0000	1.0000
_4_4.00	0.0000	1.0000
_5_5.00	0.0000	1.0000
_6_6.00	0.0000	1.0000

Confusion matrix

	_1_1.00	_2_2.00	_3_3.00	_4_4.00	_5_5.00	_6_6.00	_7_7.00
_1_1.00	237	0	0	0	0	0	
_2_2.00	36	0	0	0	0	0	
_3_3.00	13	0	0	0	0	0	
_4_4.00	14	0	0	0	0	0	
_5_5.00	13	0	0	0	0	0	
_6_6.00	24	0	0	0	0	0	

Fig 3: Screen shot of K-NN classifier performance

In this experiment K-NN spv learning without feature selection gives error rate 43.57% [fig3]. Now Apply Fisher filtering" from feature selection tab. This filtering technique selects 170 attributes out of 279 as shown in fig.4.



Fisher filtering 1

Parameters

Parameters	
Active parameter	0
Parameter	Value
p-value thresold	0.001000
best attributes	10

Results

INPUT attribute selection

INPUT selection	
Before filtering	279
After filtering	170

Keeped into INPUT selection

Attributes	
1	QwaveV3
2	QwaveV3

Fig 4: Screen shot for K-NN classifier performance with feature selection

The dropped attributes are not relevant for the analysis of Cardiac Arrhythmia. Therefore, these attributes have removed from the data set. Now apply K-NN supervise learning on selected attributes. It is observed that accuracy of K-NN method on cardiology data set has decreased the error rate to 28.10%.

3.5 CRT

C-RT is a very accepted classification algorithm tree learning algorithm. Exactly, it includes all the ingredient of a good learning control: the post-pruning method enables to make the substitution between the bias and the variance; the cost intricacy system allows "smoothing" the looking at of the space of solutions, controlling the first

choice for ease with the standard error rule (SE-rule) etc. The Breiman's algorithm is provided under different designations in the free data mining tools.
C-RT consists of two sets one is growing set and another one is Pruning set. Growing set decreases as the number of leaves increases cannot use this information to select the right model [fig5]. Use Pruning to choose the best model. The tree minimizes the error rate on the Pruning set.

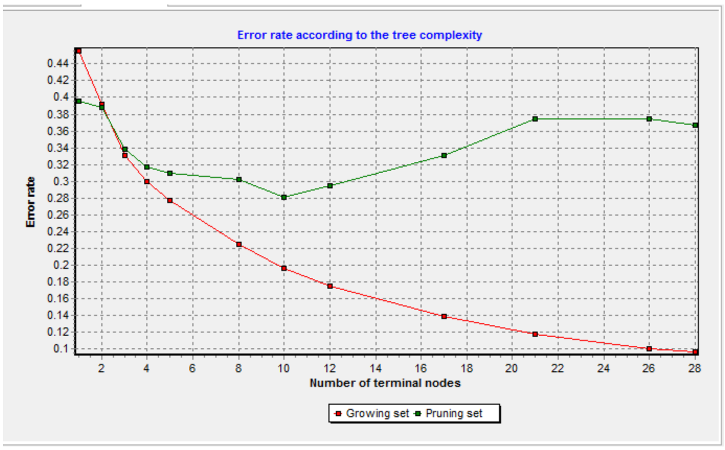


Fig 5: Screen shot for C-RT chart of pruning set

4 ROUGH SET THEORY BASED TOOL ROSE 2

Rose is microcomputer software [16] designed to analyze data by means of rough set theory. It consists of integrated environment and external executable modules. Pre-processing, reducts, rules, classification and similarity relation are the analysis methods in ROSE [8].

4.1 Discretization and Approximation

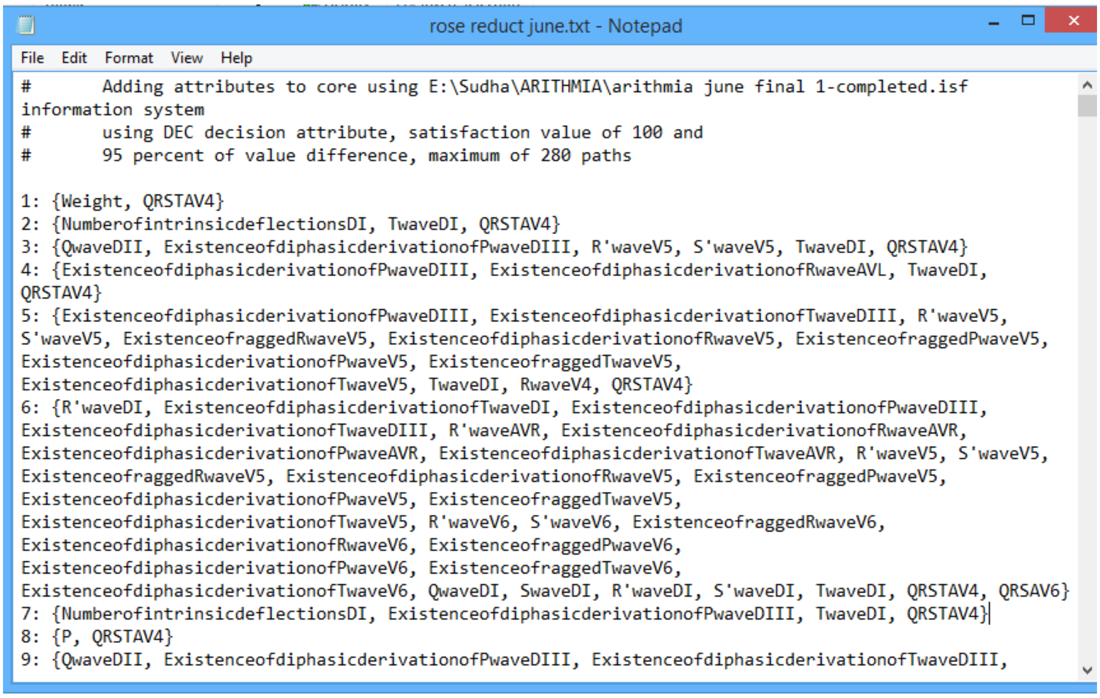
The data transformed into '.isf' format and added to the tool. Now pre-processing data which is used to discretize attributes with continuous domains into the ones with discrete domains. All such attributes in the source data file are processed. It is an entropy based method.

Since the data has some missing values, we processed through missing values where ROSE tool replace the missing value by the most frequent value for the given attribute then we generated rough set approximations by running approximation method and the quality of classification is 1 which tells the set is not rough.

4.2 Reduct and Core

One of the most important contributions of rough set theory to the data analysis field is that it can remove superfluous information. Removal of duplicate data from the information system is one of the central concepts in rough set theory. The concepts which perform this task are reduct and core [15]. Reduct is a set of attributes that preserves partition [6].

The core is the set of all indispensable attributes of the set. The core contains the most significant attributes of the set. We cannot remove any of the elements from the core without losing information from the set. The following is an important property that connects the notion of the core and reducts [5]. Core = ∩ Red (A) where Red (A) is the set of all reducts of set.



```
#       Adding attributes to core using E:\Sudha\ARITHMIA\arithmia june final 1-completed.isf
information system
#       using DEC decision attribute, satisfaction value of 100 and
#       95 percent of value difference, maximum of 280 paths

1: {Weight, QRSTAV4}
2: {NumberofintrinsicdeflectionsDI, TwaveDI, QRSTAV4}
3: {QwaveDII, ExistenceofdiphasicderivationofPwaveDIII, R'waveV5, S'waveV5, TwaveDI, QRSTAV4}
4: {ExistenceofdiphasicderivationofPwaveDIII, ExistenceofdiphasicderivationofRwaveAVL, TwaveDI,
QRSTAV4}
5: {ExistenceofdiphasicderivationofPwaveDIII, ExistenceofdiphasicderivationofTwaveDIII, R'waveV5,
S'waveV5, ExistenceofraggedRwaveV5, ExistenceofdiphasicderivationofRwaveV5, ExistenceofraggedPwaveV5,
ExistenceofdiphasicderivationofPwaveV5, ExistenceofraggedTwaveV5,
ExistenceofdiphasicderivationofTwaveV5, TwaveDI, RwaveV4, QRSTAV4}
6: {R'waveDI, ExistenceofdiphasicderivationofTwaveDI, ExistenceofdiphasicderivationofPwaveDIII,
ExistenceofdiphasicderivationofTwaveDIII, R'waveAVR, ExistenceofdiphasicderivationofRwaveAVR,
ExistenceofdiphasicderivationofPwaveAVR, ExistenceofdiphasicderivationofTwaveAVR, R'waveV5, S'waveV5,
ExistenceofraggedRwaveV5, ExistenceofdiphasicderivationofRwaveV5, ExistenceofraggedPwaveV5,
ExistenceofdiphasicderivationofPwaveV5, ExistenceofraggedTwaveV5,
ExistenceofdiphasicderivationofTwaveV5, R'waveV6, S'waveV6, ExistenceofraggedRwaveV6,
ExistenceofdiphasicderivationofRwaveV6, ExistenceofraggedPwaveV6,
ExistenceofdiphasicderivationofPwaveV6, ExistenceofraggedTwaveV6,
ExistenceofdiphasicderivationofTwaveV6, QwaveDI, SwaveDI, R'waveDI, S'waveDI, TwaveDI, QRSTAV4, QRSAV6}
7: {NumberofintrinsicdeflectionsDI, ExistenceofdiphasicderivationofPwaveDIII, TwaveDI, QRSTAV4}
8: {P, QRSTAV4}
9: {QwaveDII, ExistenceofdiphasicderivationofPwaveDIII, ExistenceofdiphasicderivationofTwaveDIII,
```

Fig 6: Screen shot of reduction sets generated by rough set

Since all reduction of attributes has no intersection, the number of core attributes is zero. The quality of classification remains same for all number of reduced set of attributes [fig6]. Merely 6100 sets were reducted while lattice search reduction. Among them some take into consideration for classification where we could see the error rates while validation are not significant. The approximation of the decision can be defined by constructing the decision rules. From the reduct computation, decision rules can be generated for clustering of the objects. Decision rules is created by combining rule reduct attributes. Each row of reduct viewer verifies a decision rule, which specifies the decision that must be taken when conditions are indicated and the condition attributes are fulfilled.

To know what are the generate rules worth and how good they can classify objects we use LEM2. In ROSE system we can able to do the classification test using any of the available methods for rule generation [9]. But we use Extended minimal covering rules (ModifiedLEM2) which generate less rules than LEM2 where entrophy and laplace measures are used for evaluating conditions, gives 26 an 53 rules respectively.

The data is classified now by similarity rule strength where k=fold cross validation has applied and various performances are acquired using the classifications parameters which shown below in chart 1 are significant. Among them we could find that parameter used 20 most similar partially matched rules gives better accuracy perhaps the parameter can be selected individually according to the kind of the data.

The accuracy of the cardiac arrhythmia data as without reduction and with reduction were calculated using basic minimal covering with stratified division where the distribution of the decision classes in each fold is the same as in the whole learning input and with the rule support strength into similarity which boosting accuracy from 60.93% to 70.25%. Even though the considered reducts were not shown the significant results. The performance of reduced data (attributes) gives little worse accuracy than the original data. The range and no. of rules were huge in basic minimal covering which affects the rule based classification. By setting some opt threshold according to the data we can reveal better results.

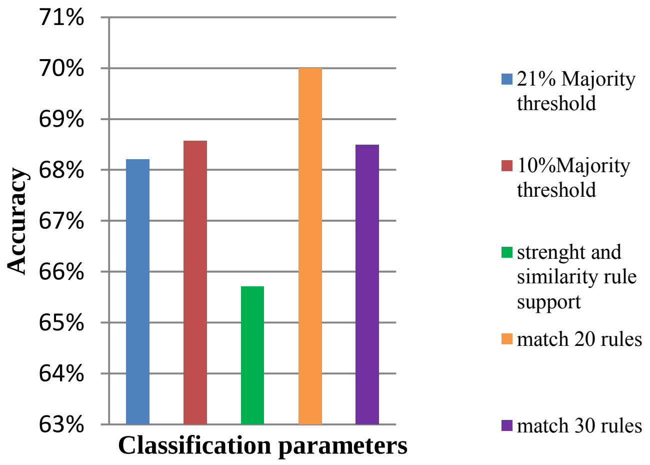
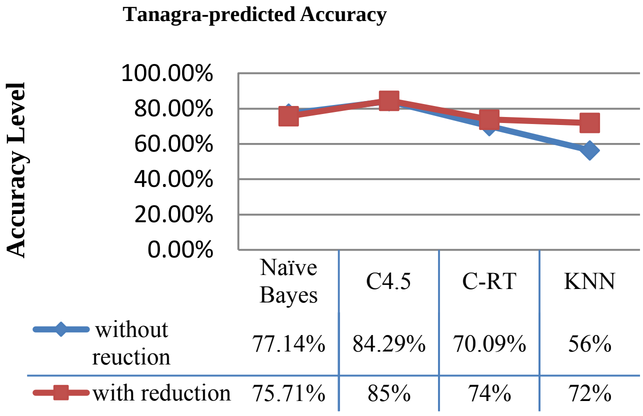


Chart 1: Accuracy of 10-fold cross validation test using minimal covering algorithm.

5. RESULTS AND DISCUSSIONS

Classification is one of the data mining techniques, which gives the decision for diagnosing process. There are different algorithms which are practiced in two different data mining tools such as ROSE2 and Tanagra. The performance parameters such as accuracy and error rate are calculated using both the data mining tools. Among them, the classification algorithms are well executed in ROSE2 compare to Tanagra with the quality of classification 1. In Tanagra, C4.5 and K-NN supervised learning algorithms [2] reach more than 70% accuracy, and the time taken to build the algorithms is almost low[chart 2]. In ROSE2, the rule based validation shows below and up to 70% accuracy and takes less time to execute an algorithm. Performance of the algorithms are compared and shown in the following chart 3.



	Naïve Bayes	C4.5	C-RT	KNN
without reuction	77.14%	84.29%	70.09%	56%
with reduction	75.71%	85%	74%	72%

Chart 2: Accuracy level between algorithms of Data mining

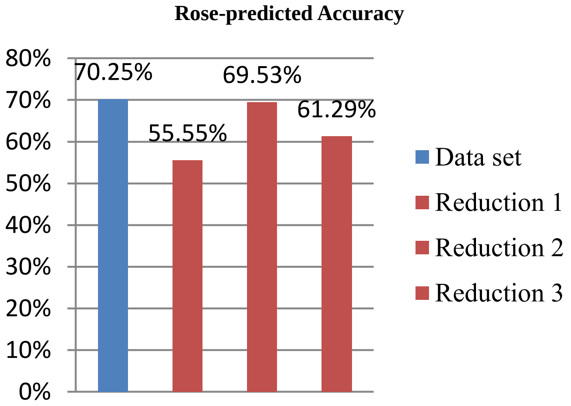


Chart 3: Accuracy level between some Reduction sets of Rough set theory

6. CONCLUSION

In this paper four classification techniques of Tanagra tool of data mining were compared with Rough set theory base classification method. We used instances involved in arrhythmia disease to predict the disease in patients. They names are : C4.5, K-NN, CRT, and Naïve Bayes. These techniques are compared by their classification accuracy and error rate (True Positive, True Negative, False Positive and False Negative). In ROSE, we classified some reduction sets made by lattice search gives comparable accuracy levels. The following chart 4 shows the accuracy level between the mining tools ROSE and Tanagra with and without feature selection.

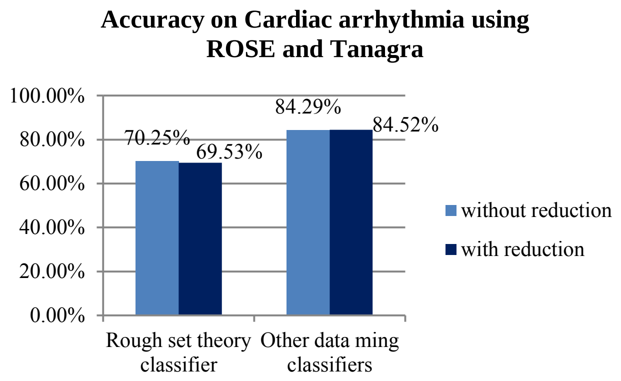


Chart 4: Accuracy between RST and Mining algorithms

But comparatively, C4.5 supervised learning algorithm of Tanagra shows improved prediction accuracy than ROSE tool algorithms. Also we may find that, if the number of attributes is more than 50% of number of objects, then the reducts will not give better accuracy than the whole set. And we noticed that by increasing the strength of rules we may produce better results. Since it's a rule base classification, the decisions rules are generated by the rules are huge in size. So, by pruning the decision tree formed by the rules, we can reduce number of similarity rules that would get more accurate diagnosis result in future and to improve the performance of these classification methods.

REFERENCES

[1] Alpaydin Ethem, 2004. "Introduction to Machine Learning". Cambridge, Massachusetts, London, England: MIT Pr.

[2] Chuanyi ji , sheng Ma, 1999, "Performance and efficiency: recent Advances in Supervised learning",. Proceedings of the ieee, vol. 87, no. 9, 1519-1535

[3] Divya Jain, 2014‘A Comparison of Data Mining Tools using the implementation of C4.5 Algorithm’, , (IJSR), ISSN (Online): 2319-7064

[4] Eamonn Keogh 2006. ‘Naïve Bayes Classifier’,UCR, and Christopher Bishop “Pattern Recognition Machine Learning”, Springer-Verlag.

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

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

[7] 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

[8] L. Polkowski and A. Skowron 1998, ‘ROSE - Software Implementation of the Rough Set Theory’, (Eds.): RSCTC'98, LNAI 1424, pp. 605-608,. c Springer-Verlag Berlin Heidelberg

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

[10] Ritu Ganda, Vijay Chahar ,2013 “A Comparative Study on Feature Selection Using Data Mining Tools” International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, 9:26-33

[11] T.Mitchell, 1997, ‘Machine Learning’: MIT Press.

[12] Vishal jain, gagandeep singh narula & mayank singh 2013, ‘Implementation Of Data Mining In Online Shopping System Using Tanagra Tool’, , (IJCSE), ISSN 2278-9960 Vol. 2, Issue 1, 47-58

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

[14] Z.Pawlak: 1982, “Rough sets, International Journal of Computer and Information Sciences”, 11, 341-356

[15] Z.Pawlak, 2002 “Rough sets and intelligent data analysis”, Information Sciences 147, 1–12, Elsevier Science Inc.

[16] Z.Pawlak , 1997, Rough Sets and Data Mining, Proceedings of the Australiasia-Pacific Forum on Intelligent Processing and Manufacturing of Materials, pp. 663-667