Comparative Study among Data Reduction Techniques over Classification Accuracy

Ibrahim M. El-Hasnony
Faculty of Computer Science & Information Systems,
Mansoura University,
Mansoura, Egypt

Hazem M. El Bakry
Faculty of Computer Science & Information Systems,
Mansoura University,
Mansoura, Egypt

Ahmed A. Saleh
Faculty of Computer Science & Information Systems,
Mansoura University,
Mansoura, Egypt

ABSTRACT
Nowadays, Healthcare is one of the most critical issues that need efficient and effective analysis. Data mining provides many techniques and tools that help in getting a good analysis for healthcare data. Data classification is a form of data analysis for deducting models. Mining on a reduced version of data or a lower number of attributes increases the efficiency of system providing almost the same results. In this paper, a comparative study between different data reduction techniques is introduced. Such comparison is tested against classification algorithms accuracy. The results showed that fuzzy rough feature selection outperforms rough set attribute selection, gain ratio, correlation feature selection and principal components analysis.

General Terms
Data mining, bioinformatics

Keywords
Fuzzy rough feature selection, rough set attribute reduction, principal component analysis, correlation feature selection, gain ratio

1. INTRODUCTION
The revolution in medical data volumes is considered a problem not just for the enormous size, but also for the incremental speed of the data creation and complexity [1]. There are many sources of medical data such as mobile applications, capturing devices, and sensors that all results from new technologies development. Such huge medical data will be troublesome for processing or examination utilizing basic database management tools. Clearly, catching, putting away, seeking, and breaking down medical huge data to discover valuable results of knowledge will enhance the results of the social insurance frameworks. Also through intelligent decisions and effective explanatory algorithms health awareness cost will be lowered. Because of the huge amount of data that reaches to several gigabytes or more, it is possible for medical databases to be exposed to many problems such as noise, missing and data inconsistency [3]. Data pre-processing enhances the quality of data, along these lines serving to improve the precision and proficiency of the consequent mining procedure.

Knowledge discovery process depends mainly on data pre-processing. The decision's efficiency depends mainly on the quality of data, hence data pre-processing is considered an imperative stride in the learning disclosure process. The process of detecting data abnormalities, redressing them early, and data reduction prompts tremendous adjustments for decision making. If the data used in the analysis process is large, the data mining process will be slow. Data reduction acquires a decreased representation for the data sets that have volume smaller than the original, yet delivers almost the same results or analytical output.

Dimension reduction or attribute reduction[2] of substantial data sets has dependably been a search area, particularly for the data sets included in the healthcare field. The attributes of these data sets are not all applicable for the purpose of classification. From the perspective of classification, it is essential to hold just those attributes that maximize the classification effectiveness. Data reduction handles not only reducing the number of attributes but also reducing the instances as well. The major of data reduction depends on attributes reduction. Hence when the pre-processing is done for reducing attributes, the most important aspect is producing the reduct with the same effectiveness as the original data set. The reduct is the lowest number of attributes that the original data depends on.

The proposed model evaluates data reduction techniques along with classification algorithms with metric accuracy. Such model composes of pre-processing phase and classification phase. The pre-processing phase handles noisy data and makes comparative study among different features reduction techniques such as gain ratio, rough set attribute reduction (RSAR), fuzzy rough feature selection (FRFS), principal components analysis (PCA) and correlation feature selection (CFS).

The model is tested against classification algorithms accuracies such as C4.5, fuzzy rough nearest neighbor, Multi-layer perceptron (MLP), Nearest-neighbor-like algorithm using non-nested generalized exemplars(NNGE), Fuzzy nearest neighbor, sequential minimum optimization(SMO), classification via clustering , NB-tree and naïve Bayes algorithms.

The results showed that fuzzy rough feature selection technique for data reduction is reasonable more than other algorithms with medical data. Also comparison showed that classification algorithms depending on FRFS achieve higher accuracies than those depending on other data reduction algorithms. Moreover, CFS in many cases has achieved good results.

The rest of this paper is organized as follows. Section 2 highlights the most recent researches in medical data pre-processing and classification. Section 3 presents materials and methodologies which the proposed model depends on. Section 4 introduces the proposed model framework. Experimental results and conclusions are showed in sections 5 and 6 respectively.

2. RELATED WORK

Because of the healthcare critical issues, huge amount of medical data and the speed in which data generated, many research papers tried to overcome the problems result from such huge data.

Mangai et al. [4] proposed a new strategy for attributes reduction utilizing a measure of ward's minimum variance. Ward's minimum variance measure is initially used for recognizing redundant features clusters in a web page. This strategy held only the best illustrative features and removing the other features. Classification is affected by the resource utilization that is free of redundant features. By comparing with PCA, the model has achieved better results for reduction and classification accuracy.

Patil and Sane [2] performed a comparative study for effective classification after data reduction. This study discussed in brief the techniques of reduction with performing a comparison of accuracy after dimension reduction. They used the weka mining tool for applying built in filters and fuzzy rough approach. The results showed that fuzzy rough feature selection improves the accuracy of artificial neural network classifiers.

Sakthivel et al. [5] presented a comparison between traditional and nonlinear techniques for dimension reduction. In their paper, the decision tree classifier applied on the reduced data is compared with Bayes net, KNN, and naive Bayes classifiers. Moreover, the results displayed that nonlinear dimension reduction techniques cannot outperforms the traditional linear ones such as PCA.

Karegowda et al. [6] introduced a comparison between gain ratio and correlation feature selection. The study addressed the effect of those feature reduction methods for Pima Indian diabetic data set classification. The classification algorithms used for testing are back propagation neural network and radial basis function network. The results proven superiority of CFS compared to gain ratio.

Dai and Xu [7] proposed dimension reduction method with respect to fuzzy gain ratio based on fuzzy rough set theory. Three data sets for real world tumer in gene expression were used. The paper proved the efficiency of their model with classification accuracy.

Porkodi [8] performed a comparative study among ReliefF (RF), Information Gain (IG), Gain Ratio, Gini Index (GI), and Random Forest (RF) on lung cancer data sets. The results showed that random forest outperforms the other techniques of reduction for its efficiency.

This paper depends on more than one feature reduction technique from more than one previous study and holding a comparison among 5 data reduction techniques to provide general view about the effectiveness of each technique in enhancing the classification efficiency. This study relied on two standard data sets from UCI machine learning repository.

3. MATERIALS AND METHODS

3.1 Methodology

In this paper, a comparative study among different data reduction algorithms is introduced. The study addressed the differences among data reduction techniques together with the contribution of each one and its effectiveness against classification efficiency. The proposed dimension reduction algorithms are gain ratio, rough set, correlation feature selection, principal components analysis and fuzzy rough feature selection. The comparison tested with respect to classification accuracy for C4.5, fuzzy rough nearest neighbor, Multi-layer perceptron(MLP), Nearest-neighbor-like algorithm using non-nested generalized exemplars(NNGE), Fuzzy nearest neighbor, sequential minimum optimization(SMO), classification via clustering , NB-tree and naïve Bayes algorithms.

3.2 Data Reduction Techniques

- Correlation Feature Selection (CFS).

Correlation based feature selection (CFS) [10] surveys the estimation of attribute subsets by considering the individual prescient capacity of each component nearby the level of repetition between them. From previous studies, CFS displays good results in dealing with features that are noisy, immaterial or repetitive. CFS works well for reducing medical data set features. Classification accuracy that depends on feature reduction algorithms in many cases is better than working with the whole features. CSF depends on a heuristic to determine the features for elimination. The eliminated features surly do not contribute to the class determination. Formalization of heuristic can be shown in equation (1):

$$Merit_s = \frac{t\overline{r_{cf}}}{\sqrt{t + t(k-1)\overline{r_{ff}}}} \quad (1)$$

Where $Merit_s$ is the heuristic 'merit' (worst) of a feature Subset S containing t features, $\overline{r_{ff}}$ is the average feature-feature inter-correlation, and $\overline{r_{cf}}$ is the mean feature class correlation ($f \in S$)

- Gain Ratio

Gain ratio [6] is extension to information gain. Information gain attempts to choose which attribute used for testing in each node of the tree. The splitting attributes for decision tree determined by higher gain ration attributes. Choosing attributes with extensive number of values is a measure of information gain.

Let S be set composed of s samples of data with m distinct classes. The expected information needed for classification of a specific sample is given by:

$$I(s) = -\sum_{i=1}^{m} p_i \log_2(pi) \quad (2)$$

Where p_i is the probability that an arbitrary sample belongs to class C_i

The encoding information that would be gained by branching on A is:

$$gain(A) = I(S) - E(A) \quad (3)$$

Where E (A) is the entropy, or expected information based on the partitioning into subsets by A, is given by:

$$E(A) = -\sum_{i=1}^{m} I(S)\frac{S_{1i} + S_{2i} + \ldots S_{mi}}{S} \quad (4)$$

Where attribute A has v distinct values, s_{ij} be number of samples of class C_i in a subset S_j. S_j contains those samples in S that have value a_j of A.

- Principal Components Analysis (PCA).

PCA [9] is a traditional method for data analysis. It transforms data to linear while maintaining the maximal variance amount. PCA probabilistic formulation gives a decent establishment for taking care of missing values. Utilizing PCA helps for compressing data or vectors of high dimension to vectors of lower dimensional. PCA applied in many fields such as medical data.

- Rough Set Attribute Reduction (RSAR).

Rough set theory [11] provides approximate description of data analysis objects through techniques that characterized by being mathematical and rigorous. A rough set considers an estimate of an ambiguous by upper approximations and lower approximations concepts. The lower approximation is a depiction of the space objects which are known with sureness to have a place with the subset of interest, while the upper approximation is a depiction of the objects which potentially have a place with the subset. Rough set works on data sets with discrete values for removing repetitive and restrictive or conditional attributes and in the same time don not lose their information content.

- Fuzzy Rough Feature Selection (FRFS).

Fuzzy-rough feature selection [12] [13] [15] solves many problems that have faced rough set attribute reduction. Rough set handles only data sets of discrete values. Moreover rough set cannot find the chance to work with noisy data. Fuzzy rough feature selection overcomes this problems and with no additional supplied data from user. Fuzzy rough feature selection has numerous favorable circumstances to handle noisy and continuous or discrete terms of data. Fuzzy rough feature selection makes use of ambiguity and indiscernibility from fuzzy sets and rough sets respectively. Fuzzy rough feature selection uses FRQuickReduct for implementation as shown in figure (1).C indicates the conditional attributes and D decision attributes set. The algorithm depends on dependency equation (5) for fuzzy attributes and equivalence classes' dependency degree calculation.

$$\gamma'_P(Q) = \frac{\sum_{x \in U} \mu_{POS_{R_P}(Q)}(x)}{|U|} \quad (5)$$

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{x \in U/Q} \mu_{R_P x}(x) \quad (6)$$

Where

$R \leftarrow \{\}, \gamma'_{optimal} = 0, \gamma'_{old} = 0$
do
$T \leftarrow R$
$\gamma'_{old} \leftarrow \gamma'_{optimal}$
$\forall x \in (C - R)$
$if\ \gamma'_{R \cup \{x\}}(D) > \gamma'_T(D)$
$T \leftarrow R \cup \{x\}$
$\gamma'_{optimal} \leftarrow \gamma'_T(D)$
$R \leftarrow T$
$until\ \gamma'_{optimal} \leftarrow \gamma'_{old}$
$return\ R$

Fig1: FRQuickReduct Algorithm

3.3 Classification Algorithms

Classification [3] is a type of data analysis for models deduction that depicts imperative data classes. The process of classification composes of learning and classification steps. Learning step addresses developing the classification model whereas predicting new instances class labels is performed in the classification step. Data classification is considered a form of supervised training where the class label for each training record is available. Classification accuracy is determined by the percentage of correctly classified tuples. There are many classification algorithms but in this paper a set of well-known classifiers are utilized such as C4.5, fuzzy rough nearest neighbor, Multi-layer perceptron(MLP), Nearest-neighbor-like algorithm using non-nested generalized exemplars(NNGE), Fuzzy nearest neighbor, sequential minimum optimization(SMO), classification via clustering , NB-tree and naïve Bayes algorithms.

4. PROPOSED FRAMEWORK

The medical data is gathered from many resources like mobile applications, capturing devices, and sensors which may suffer from many problems such as noise, missing values and redundant features. These problems need to be carefully handled in order to improve the healthcare decision making process. This paper concentrates on dealing with one of the most challenging problems which is data reduction. The study aims to choose the most suitable data reduction technique that improves the decision making process in terms of accuracy complexity. The comparison of data reduction techniques is applied through classification algorithms accuracy. The model starts by preprocessing data by handling missing and noisy data then selecting the features by applying different data reduction algorithms. The classifiers applied on each reduct to evaluate the contribution of each data reduction technique and its effectiveness in classification accuracy. Figure 2 shows the framework for the proposed system.

The framework is composed of

a. Data pre-processing
    1. Dealing with missing values
    2. Reduction process
b. Classification process
    1. Building training model
    2. Testing the classification model
c. Evaluating the performance

Missing values have many problems especially in statistical reasoning. Methods, that handle missing values, belong either to sequential methods that are performed during pre-processing or parallel methods where missing attribute values are taken into account during extracting knowledge. There are many methods for missing value handling such as deletion or replacement according to the nature of processing data and size of data missed.

The goal of feature reduction discovers the least number of attributes provided that the extracted probability distribution of the data classes is as close as possible to the original distribution got utilizing all attributes. Data mining on a small number of attributes has many advantages. It limits the quantity of attributes showing up in the discovered patterns, making the patterns simpler to get and understandable. Further, it enhances the precision of classification and its learning runtime.

The classification process builds the classification model through the training data instances. Different techniques use several methodologies in building the classification model. For example, C4.5 algorithms use decisions trees to express model while MLP uses a neural network structure. The classification model helps in the decision making process by classifying new instances. This is accomplished during the testing process. Upon the test results, the overall system performance is evaluated in terms of accuracy metric.

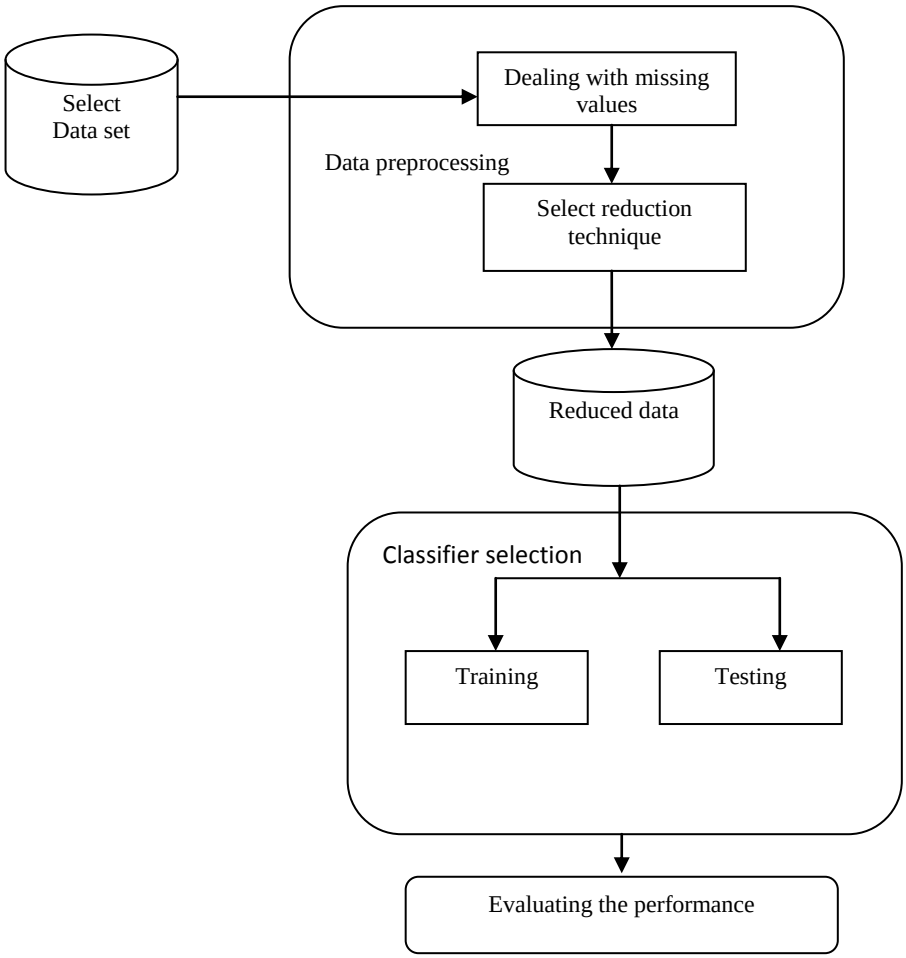

Fig 2: The components of the proposed system

The proposed model addresses the comparison among CFS, RSAR, FRFS, PCA, and gain ratio. The study tends to explain what the effective method is or the technique that achieves more reasonable reduct. The effectiveness of the techniques is measured in terms of increasing the performance of classification accuracy. To test the effectiveness of feature reduction techniques , C4.5, fuzzy rough nearest neighbor, Multi-layer perceptron(MLP), Nearest-neighbor-like algorithm using non-nested generalized exemplars(NNGE), Fuzzy nearest neighbor, sequential minimum optimization(SMO), classification via clustering , NB-tree and naïve Bayes algorithms are used. The comparison of feature reduction algorithms described the difference between RSAR and a hybrid of using rough set with fuzzy sets to produce effective techniques for feature reduction that is FRFS.

5. EXPERIMENTAL RESULTS

5.1 Data Set

There are two data sets used from UCI machine learning repository [14] for examining the proposed model are breast cancer and Thoracic Surgery. Table 1 and table 2 shows a description for Breast cancer and Thoracic Surgery data sets.

Table 1: Breast cancer data set description

Dataset characteristics	Multivariate	**Attributes**	10
Attribute characteristics	Integer	**Instances**	699
Missing values	Yes	**Class**	2

Table 2: Thoracic Surgery data set description

Dataset characteristics	Multivariate	**Attributes**	17
Attribute characteristics	Integer , real	**Instances**	470
Missing values	No	**Class**	2

Breast cancer data set contains 699 cases for patients who had experienced surgery for breast cancer while thoracic surgery data set was gathered reflectively at Wroclaw thoracic surgery center for patient who underwent major lung resections for essential lung cancer.

5.2 Classification Results on Original and Reduced Data

The proposed model is composed of two main phases. The first phase is the pre-processing which handles the data problems like noise, missing data and the most important is redundant features. Data reduction process aims to eliminate redundant features keeping only those affecting the decision making process.

Removing redundant features helps a lot in decreasing space and time consumption specially in challenging fields like medical and healthcare. The second phase is the classification phase where the system learns the classification model using the training data then predicts the classes of the new instances during the testing process.

In the first step data sets are preprocessed for noisy and missing values. The idea behind this paper is to evaluate some data reduction techniques to study if a reduction of the original features to lower number have any effect on the performance of classification accuracy or not. From the results shown below in tables 3 and 4, which graphically displayed in figures 3 and 4, there are changes in classification accuracies either by an increase or decrease in the performance.

It is noticeable that the FRFS has superiority on the average accuracy shown in figures 5 and 6. The average accuracy obtained after applying FRFS is 96.6% and 76% for breast cancer and thoracic surgery data sets respectively. Moreover CFS has the ability to improve performance, achieving the second level after FRFS.

Table 3: The classification accuracy with data reduction techniques using breast cancer dataset

Classifier	ORIGINAL DATA	AFTER CFS	AFTER GAIN RATIO	AFTER PCA	AFTER ROUGH SET	AFTER FRFS
c4.5	**94.6**	**95.2**	**95**	**96.3**	**94.1**	**96**
Fuzzy rough nearest neighbor	**95.7**	**96.9**	**96.8**	**95.3**	**96.8**	**97**
MLP	**95.9**	**96.57**	**95.42**	**95**	**95.2**	**95**
Fuzzy nearest neighbor	**62.1**	**96.8**	**62.1**	**96.4**	**96.9**	**97**
NNGE	**96.4**	**96.2**	**96.6**	**96.6**	**96.2**	**96.9**
SMO	**96.7**	**96.8**	**96**	**96.3**	**96.3**	**96**
classification via clustering	**95.7**	**95.4**	**95.7**	**95.4**	**95.6**	**96.1**
NBTREE	**96.5**	**96.8**	**96.5**	**97**	**96.4**	**97.6**
naïve Bayes	**96.2**	**96.2**	**96.2**	**95**	**96.2**	**96.4**

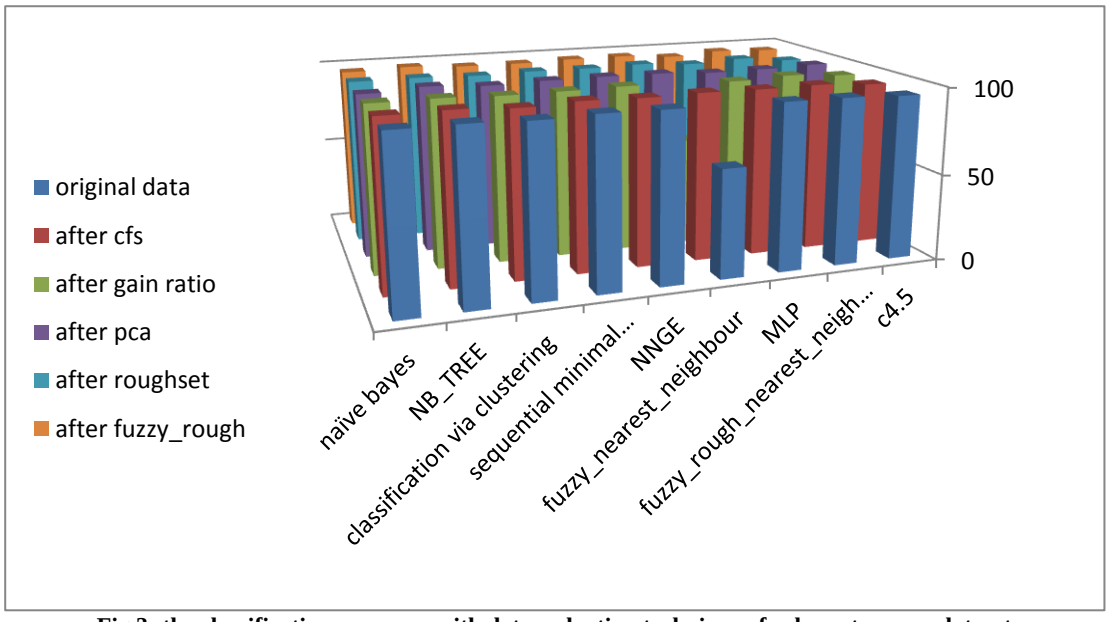

Fig 3: the classification accuracy with data reduction techniques for breast cancer dataset

Table 4: The classification accuracy with data reduction techniques using thoracic surgery dataset

Classifier	ORIGINAL DATA	AFTER CFS	AFTER GAIN RATIO	AFTER PCA	AFTER ROUGH SET	AFTER FRFS
c4.5	75.5	74	72.3	75	72.3	75.5
Fuzzy rough nearest neighbor	75.4	77	74.8	74.5	74.1	74.5
MLP	77.6	77.2	75	75	76	78
Fuzzy nearest neighbor	72.4	73.9	73.8	74.3	72	74.4
NNGE	75.6	75.9	75.7	75.7	75.7	75
SMO	72.4	72.4	72.4	72.4	72.4	72.4
classification via clustering	72	75.4	74.3	77.4	72.8	78
NBTREE	73.2	74	74.3	72.2	76.3	77
naïve Bayes	76.2	75	74.7	77.7	75	79

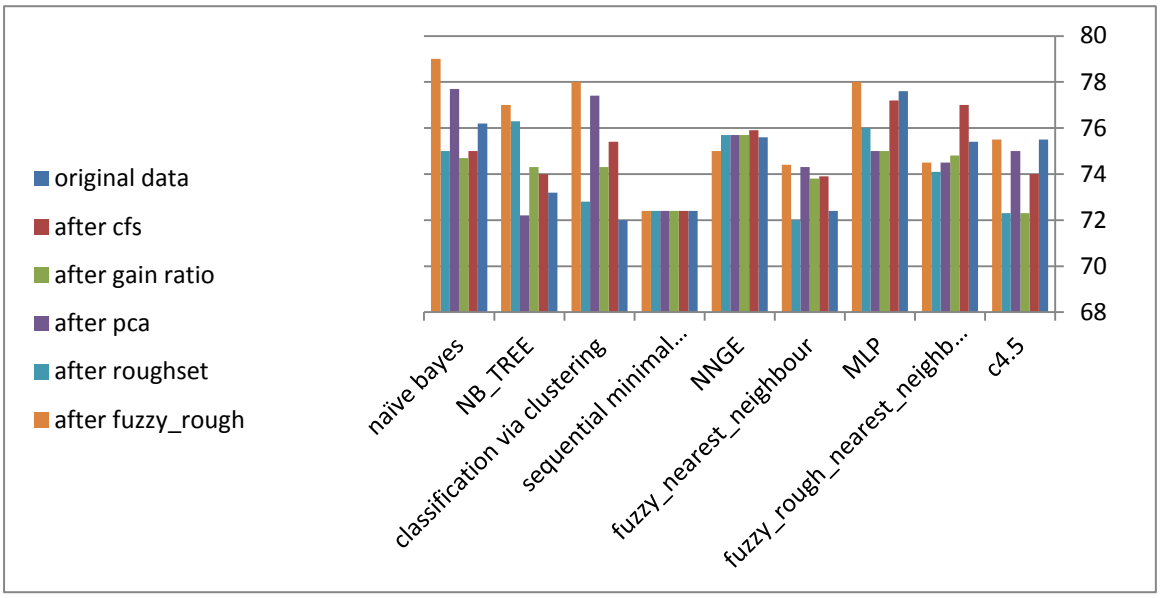

Fig 4: the classification accuracy with data reduction techniques for thoracic surgery dataset

Figure 5 and figure 6 shows the averages of classification algorithms accuracy with comparison to data reduction algorithms. These figures displays that FRFS in both data sets has average accuracy greater than the other techniques together with CFS.

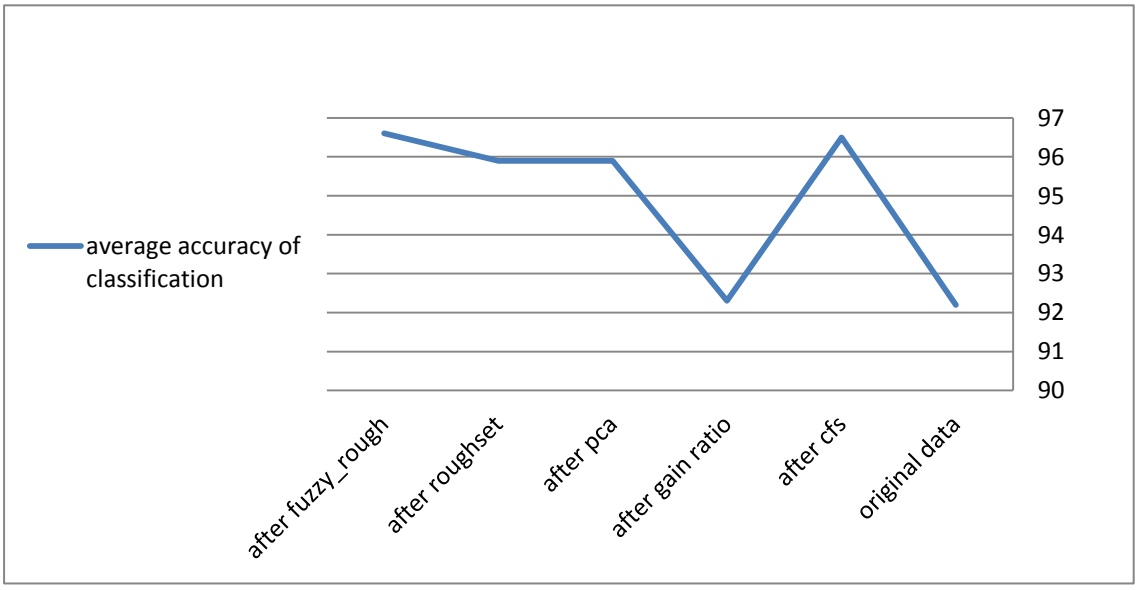

Fig 5: Average of classification algorithms accuracy with each reduction technique (breast cancer)

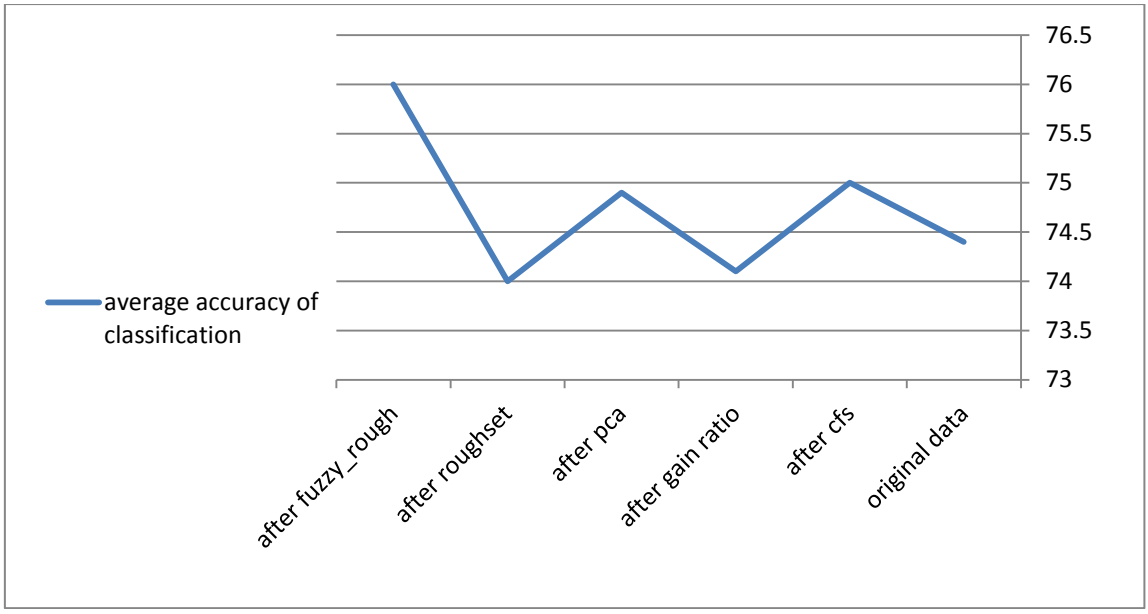

Fig 6: Average of classification algorithms accuracy with each reduction technique (thoracic surgery)

6. CONCLUSION

Feature selection techniques have many benefits for data analysis such as constructing straightforward and more understandable models, useful for understanding concentrated data, and enhance data mining efficiency. In this paper, a comparative study and analysis of different data reduction techniques were introduced. Such comparison included FRFS, RSAR, PCA, CFS, and gain ratio data reduction techniques. These techniques were tested over classification algorithms accuracy with breast cancer and thoracic surgery data sets. Classification algorithms like C4.5, fuzzy rough nearest neighbor, Multi-layer perceptron(MLP), Nearest-neighbor-like algorithm using non-nested generalized exemplars(NNGE), Fuzzy nearest neighbor, sequential minimum optimization(SMO), classification via clustering , NB-tree and naïve Bayes were introduced in the study. The results showed that FRFS outperformed the other techniques in reducing medical data. CFS gave good results compared to RSAR, PCA or gain ratio.

7. REFERENCES

[1] Ahmed E. Youssef," A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments", International Journal of Ambient Systems and Applications (IJASA), Vol.2, No.2, (2014), pp.: 1-11.

[2] Mohini D Patil, Dr. Shirish S. Sane," Effective Classification after Dimension Reduction: A Comparative Study", International Journal of Scientific and Research Publications, Vol.4, No7, (2014), pp.: 1-4.

[3] Han, J., M. Kamber, and J. Pei., "Data Mining, third Edition: Concepts and Techniques ",The Morgan Kaufmann Series in Data Management Systems. ISBN-13: 978-0-12-381479-1 (2012).

[4] J. Alamelu Mangai, V. Santhosh Kumar, S. Appavu alias Balamurugan, "A Novel Feature Selection Framework for Automatic Web Page Classification", International Journal of Automation and Computing,Vol.9,No.4, (2012), pp.:442-448.

[5] N.R. Sakthivel a, Binoy B. Nairb, M. Elangovana, V. Sugumaranc and S. Saravanmurugan , "Comparison of dimensionality reduction techniques for the fault diagnosis of mono block centrifugal pump using vibration signals", Engineering Science and Technology, an International Journal,Vol.17 ,(2014),pp.:30-38.

[6] Asha Gowda Karegowda, A. S. Manjunath& M.A.Jayaram introduced ," comparative study of attribute selection using gain ratio and correlation based feature selection", International Journal of Information Technology and Knowledge Management, Vol.2, No. 2,(2010), pp.: 271-277.

[7] Jianhua Dai, Qing Xu," Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification", Applied Soft Computing,Vol.13 ,(2013),pp.:211–221.

[8] Porkodi, R.,"comparison of filter based feature selection algorithms: An overview", international journal of innovative research in technology& science, VOl.2, No.2, (2014), pp.:108-113.

[9] Alexander Ilin, Tapani Raiko," Practical Approaches to Principal Component Analysis in the Presence of Missing Values", Journal of Machine Learning Research,Vol.11,(2010),pp.:1957-2000.

[10] Hall, Mark A., and Lloyd A. Smith., "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper." In FLAIRS conference, (1999), pp.: 235-239.

[11] Jensen, Richard, and Qiang Shen.," Fuzzy–rough attribute reduction with application to web categorization ", fuzzy sets and systems Vol.141, no.3,(2004),pp.:469-485.

[12] Jensen, Richard, and Qiang Shen. "New approaches to fuzzy-rough feature selection.", Fuzzy Systems, IEEE Transactions ,Vol.17, No.4,(2009),pp.: 824-838.

[13] Kumar, R. Kavitha, and R. M. Chadrasekaran. "Attribute correction–data cleaning using association rule and

clustering methods." International Journal of Data Mining & Knowledge Management Process (IJDKP) .Vol.1, no.2, (2011), pp.: 22-32.

[14] Ics.uci.edu, (2015). Donald Bren School of Information and Computer Sciences @ University of California, Irvine. [online] Available at: http://www.ics.uci.edu [Accessed 29 May 2015].

[15] Mona Gamal,Ahmed Abou El-Fetouh, Shereef Barakat ."A Fuzzy Rough Rule Based System Enhanced By Fuzzy Cellular Automata". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, no.5, (2013), pp.:1-11.