Journal of Engineering Science and Technology Review 10 (6) (2017) 154 - 162

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Datasets

Sai Prasad Potharaju* and M.Sreedevi

Dept of CSE , K L University, Guntur, Andhra Pradesh, India

Received 25 October 2017; Accepted 7 December 2017

Abstract

In recent days, due to the advancements in technology, a massive amount of data is generating in every area of study, including the medical field. This massive amount of data contains a large number of attributes and instances in it. It is not an easy task for classification and prediction from this high dimensional data. Because, all the attributes in the dataset can't give an impressive result in classification and prediction. Now, it is unavoidable to reduce the high dimensional data for better classification result, which is possible by feature selection and reduction techniques .In this research paper, a novel M-Cluster feature selection (Mcfs) based on Symmetrical Uncertainty (SU) Attribute Evaluator is proposed for improving the classification accuracy of medical datasets. The proposed approach divides the total feature space into 'M' clusters, each cluster has a finite set of attributes in it without any duplication. Feature subset formed by proposed technique is tested using Dermatology and Breast Cancer medical datasets, and compared with an existing filter-based feature selection techniques(Information Gain (IG), Chi- Squared (Chi), Gain Ratio Attribute Evaluator (GR), ReliefF (Rel)).Experimental results displayed an improved performance with some of the clusters formed by proposed method than existing methods. For experimenting proposed technique, KNN-Lazy learner, Naive Bayes (NB) Classifier, J48-Rule based learner, JRip – Tree based learners are used.

Keywords: Data mining; Feature Reduction; Classification; SMOTE; Symmetrical Uncertainty

1. Introduction

Data mining (DM) is an assured technique for finding the interesting results from the available space of data. Classification is one of the DM techniques to predict unknown interesting patterns from the available data by generating a classification model. Before generating classification model, data has to be pre-processed, which is the second stage of DM. Data pre-processing is an obligatory stage in DM for producing quality result[1]. Dimensionality reduction is one of the techniques in pre-processing apart from missing values, class imbalance, noisy data, and missing labels.

Since a decade, data mining is becoming very popular in every field of study (business, education, finance, marketing, healthcare, etc.)[2]. Data mining is a useful approach in the medical field for several reasons. In mining, classification is one of the important and very useful techniques for predicting a record or instances, which class it belongs to. Many authors have proposed various intelligent classification methods to reduce the manual intervention for classifying the data points. In the medical study, classification techniques can be applied for finding the more insights of the patient disease. Data mining approaches used to predict the cancer type by the various researchers [3]. Few medical researchers applied mining methods to predict the class of Fetal Heart Rate (FHR) of the cardiotocographic dataset using ensemble approaches [4]. Some of the medical practitioners applied the mining techniques to know the chances of heart stroke [5].

For classification, initially, classifiers have to be trained over the initial data set collected for creating a learning model. Then, learning model has to be tested to record its performance. For creating a learning model, all the attributes of dataset is not essential. Because dataset may have weak attributes and they may not be useful to strengthen the learning model. Instead, they may create confusion and dilute the model also [6]. Few attributes in the initial dataset may be duplicated and few may be irrelevant (noisy attributes). It is always suggested to select the best features and remove the duplicated and noisy features.

This can be achieved using FS techniques. There are three types of FS methods; those are Filter, Wrapper, and Hybrid. Using feature selection, irrelevant or redundant features can be discarded, the memory required to analysis is decreased, the speed of the model generation can be increased, and the accuracy of classification model can be increased [7].

Filter method uses the concept of information theory and assigns the rank to each feature in the data space based on the information worth of the feature. In the case of wrapper technique, a learning algorithm is used as a subroutine for evaluating the attributes importance in prediction over validation dataset. Embedded technique combines the both filter and wrapper. These three approaches have its own merits and demerits in terms of computation speed and the probability of over fitting. In terms of speed of computation, filter methods are comparatively inexpensive.

In this research paper, filter methods are considered for comparing the feature subsets formed by proposed technique.

*E-mail address: psaiprasadcse@gmail.com

doi:10.25103/jestr.106.20

Below Table .1 shows the list of filter based feature selection methods considered with their functional view.

Table 1 . List of FS methods considered.

Name	Functional View
Information Gain (IG)	Ranker
Chi- Square (Chi)	Ranker
Gain Ratio Attribute Evaluator (GR)	Ranker
ReliefF (Rel)	Ranker
Symmetrical Uncertainty (SU) Attribute Evaluator	Ranker

In pre-processing, class imbalance also one of the serious issue, which may deviate or biased towards majority class if the classifier is trained over the imbalanced dataset. It can be addressed using the oversampling technique called SMOTE [8]. Few researchers used the concept of SMOTE in the medical field to boost the classification performance by ensemble methods [9,10]. In the current research also we have considered the concept of SMOTE over the dermatology and Breast Cancer datasets.

FS applied by many researchers in different studies of the medical field. To overcome the high dimensionality problem in breast cancer data set, FS is applied and reduced the feature set. Ensemble techniques are applied over reduced dataset; thereby performance was boosted [11]. To determine diabetes, authors considered various classifiers and compared the performance [12]. Bi-level dimensionality reduction method is used for prediction of diabetes (normal or Pima diabetes) [13]. For this, authors considered PCA-Principal Component Analysis, IG, Cfs subset evaluation. Results show that PCA has given better performance. FCBF-Fast Correlation-Based Filter method is applied in the study for the prediction of Type-II diabetes [14]. To get the greater classification accuracy and minimum response time cfs and FCBF methods are applied over dermatology dataset [15]. Using those methods, a minimum subset of features are collected and classification model is generated with those features. FS approaches are applied for analysis of microarray gene expression datasets, as it contains few thousands of features it is not an easy job to analyze that much huge dataset with those many features. In such cases, FS is a necessary approach for better result and shorter the response time [16,17].

For the proposed framework Symmetrical Uncertainty (SU) is the key criteria. Entropy is the key foundation of SU, IG, and GR ranking methods, which is the concept of information theory measure [18]. All the ranking techniques given in Table 1 gives the rank to each attribute of the data set.An attribute which contains more weight will have top rank and less weighted attribute has the least rank. Depending on the need and type of application top 'N' attributes will be considered for analysis, and remaining attributes will be discarded.

In literature filter based ranking techniques used by many researchers for different purposes. Authors of [19] applied IG, CHI, GR, SU, oneR, Rel for generating ranks of each attribute of Austria and German credit data. Based on the property and information measure, each technique produced different ranks to each attribute. Few attributes have common rank by few methods. FCFilter is proposed by authors of [20] for text mining. FCFilter discards the number of clusters to input by combining the words in an availability of the sufficiently large number of clusters. To optimize the groups, Genetic algorithm (GA) is applied, which will produce the best feature set. Dimensionality reduction is used for analyzing the medical datasets also. GA based FS is proposed by authors of [21] to increase the performance of classification of a medical dataset. Their proposed method removes unwanted features, thereby dimensions of the dataset are reduced. In the article [22], researchers proposed hybrid feature selection (HFS) technique. This technique is based on MKL-multiple kernel learning.HFS is used to measure the accuracy on expression datasets.

The key criteria what we considered to form the 'M' clusters of features is SU, which can be defined as

SU=2*IG/(H(Y)+H(X))
H(X) is Entropy of X
H(Y) is Entropy of Y

SU takes the value in the range [0,1].SU value 1 indicates one attribute can predict completely others, 0 indicates two attributes are uncorrelated.

2. Dataset Description

To test the proposed framework, Dermatology and Breast Cancer medical datasets are collected from UCI machine learning repository. The initial dermatology dataset has 366 records, 34 features, and Class label. The initial breast cancer dataset has 569 records, 30 features, and class label. Both the datasets description is given in Table 2.

Table 2. Datasets description [23, 24]

Dataset	Dermatology	Breast cancer
Total # Records	366	569
Total # Features	34	30
Total # Classes	6	2

Dermatology class has six diseases codes in it. Those are 1 (Psoriasis), 2 (Seboreic dermatitis), 3 (Lichen planus), 4 (Pityriasis rosea), 5 (Cronic dermatitis), 6 (Pityriasis rubra pilaris). Class distribution of the initial dermatology dataset is given in Table 3. Breast cancer class has two diagnosis values (M = malignant, B = benign). Class distribution of the initial breast cancer dataset is given in Table 4.

Table 3. Class Distribution of initial dermatology dataset

Class code	# Instances	%
Psoriasis	112	30.60
Seboreic dermatitis	61	16.66
Lichen planus	72	19.67
Pityriasis rosea	49	13.38
Cronic dermatitis	52	14.20
Pityriasis rubra pilaris	20	5.46

Table 4. Class Distribution of initial breast cancer dataset

Class code	# Instances	%
M	212	37.25
B	357	62.75

With the Table 3 and Table 4 statistics, it is clear that, both the datasets having class imbalance problem. Class 1 has more records than all other classes in case of dermatology dataset. Whereas, class B has more records than class M in the case of breast cancer dataset. Minority class records need to be increased to meet the majority class instances for the better result. To balance this dataset, SMOTE has applied on initial datasets. SMOTE, runs based on the KNN algorithm and create the synthetic records; it requires the percentage of the synthetic instance to be added and K value. For this, experiment K (nearest neighbours to be considered) =5. Table 5 gives the balanced class distribution of datasets after applying SMOTE.

Table 5. Balanced dataset class distribution.

Dermatology 🟨			Breast Cancer 🟧		
Class code	% of instances Increased	Total Instances ormulated	Class code	% of instances Increased	Total Instances formulated
1	0	112	M	60	339
2	100	122	B	0	357
3	55	111			
4	120	107			
5	120	114			
6	500	120			

Now, the modified datasets are almost balanced. After balancing the datasets by applying the SMOTE, dermatology dataset has total 686 instances and breast cancer dataset has 696 records in it.

3. Proposed Methodology

The intention of suggested framework is to minimize the data region. If whole data set consists of 'R' features, from 'R' features if there is a requirement to select most popular 'S' features without any duplication, in such scenario total C(R, S) number of groups (subsets) can be generated. Analyzing those many groups in case of the high dimensional dataset is not an easy task. But alternatively, filter based ranking techniques can be utilized to give the rank to each feature and then most popular 'S' features can be considered for analysis. Other than the features selected by existing techniques, we proposed a novel framework for generating a subset of features. Proposed method is as per the flowchart given below.

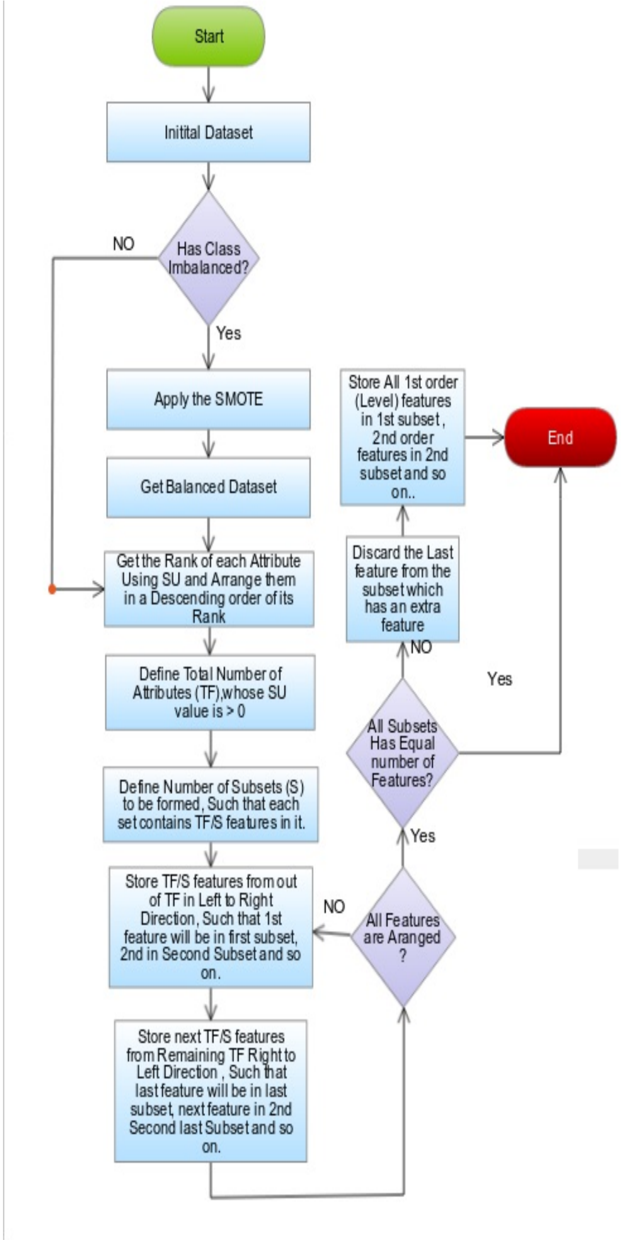
As the base for our framework is SU, Table 6. Describes the SU value of each feature including the obtained rank of each feature by IG, Chi, Rel, GR of the initial dermatology dataset (Imbalanced data set).

Table 6. SU value of each feature and Rank of each feature by IG, Chi, Rel, GR of dermatology dataset (Imbalanced data set).

Rank	SU Value	Feature No By SU	Feature No by IG	Feature No by GR	Feature No by Chi	Feature No by Rel
1	.4778	21	21	12	33	21
2	.4672	22	20	29	29	33
3	.4489	20	22	33	27	22
4	.4328	33	33	15	12	20
5	.4291	29	29	27	15	28
6	.427	27	27	31	31	27
7	.426	12	12	6	25	29
8	.4188	25	25	25	6	6
9	.4147	6	6	8	22	12
10	.3944	8	8	22	20	16
11	.3739	15	9	21	8	25
12	.3288	9	16	30	21	8
13	.3197	28	15	20	30	15
14	.2979	16	28	7	16	9
15	.2904	10	10	24	9	4
16	.28	24	24	10	7	14
17	.2505	14	14	28	10	10
18	.2244	5	5	34	34	5
19	.2159	31	26	9	28	24
20	.2094	26	3	14	24	3
21	.1868	7	31	16	26	26
22	.1825	30	19	5	14	19
23	.1726	23	23	23	3	7
24	.1692	3	7	26	5	11
25	.1447	34	30	11	19	2
26	.1441	19	2	4	23	31
27	.1341	4	4	3	2	18
28	.1301	2	34	19	4	23
29	.1066	11	11	2	11	30
30	.0641	1	1	13	1	17
31	.0597	13	13	1	13	34
32	.0495	17	18	17	18	1
33	.0483	18	17	18	17	13
34	0	32	32	32	32	32

Total # Features: 34

Flowchart 1.Proposed framework flowchart


Start
Initial Dataset
NO
Has Class Imbalanced?
Yes
Apply the SMOTE
Get Balanced Dataset
Get the Rank of each Attribute Using SU and Arrange them in a Descending order of its Rank
Define Total Number of Attributes (TF),whose SU value is > 0
Define Number of Subsets (S) to be formed, Such that each set contains TF/S features in it.
Store TF/S features from out of TF in Left to Right Direction, Such that 1st feature will be in first subset, 2nd in Second Subset and so on.
Store next TF/S features from Remaining TF Right to Left Direction , Such that last feature will be in last subset, next feature in 2nd Second last Subset and so on.
NO
All Features are Arranged ?
Yes
All Subsets Has Equal number of Features?
NO
Discard the Last feature from the subset which has an extra feature
Store All 1st order (Level) features in 1st subset, 2nd order features in 2nd subset and so on..
Yes
End

# Total Features whose SU > 0 (TF): 33

Note: Feature ID 32 (Rank 34) has SU value zero, It has to be discarded.

Assume # groups to be formed (S) is 4, then each subset has 33/4 =8 features in it,

According to the proposed methodology features in each subset will be formed as below Table 7.

Table 7. Formation of features in each group by proposed framework over imbalanced dataset

	1st Level Features	2nd Level Features	3rd Level Features	4th Level Features	Direction
→	21	22	20	33	Left to Right
←	25	12	27	29	Right to Left
→	6	8	15	9	Left to Right
←	24	10	16	28	Right to Left
→	14	5	31	26	Left to Right
←	3	23	30	7	Right to Left
→	34	19	4	2	Left to Right
←	17	13	1	11	Right to Left
→	18				Left to Right
Group ID (Subset)	IS41	IS42	IS43	IS44	

From the above Table 7, IS41 group has an additional attribute i.e feature id 18, which has to be discarded. After this process store, all 1st order attributes in group IS41, 2nd order features in group IS42, 3rd order features in subset IS43, 4th order features in subset IS44. Below Table. 8 show the features in each subset after grouping them.

Table 8. Subsets of features, Where # Subsets are 4 over dermatology imbalanced dataset

Subset ID	Features in it
IS41	21, 25, 6, 24, 14, 3, 34, 17
IS42	22, 12, 8, 10, 5, 23, 19, 13
IS43	20,27, 15, 16, 31, 30, 4, 1
IS44	33, 29, 9, 28, 26, 7, 2, 11
IG*	21, 20,22,33,29,27,12,25
GR*	12,29,33,15,27,31,6,25
Chi*	33,29,27,12,15,31,25,6
Rel*	21,33,22,20,28,27,29,6

* Top 8 features derived by existing methods (Refer Table 6)

4. Experiment

For testing and analysing the strength of proposed framework, we considered S=3,4,5 . For subset of features formed when # subsets are 3 refer Table. 9 , for # subsets are 4 refer Table. 8 and # subsets are 5 refer Table. 10

Table 9. Subsets of features, Where # Subsets are 3 over dermatology imbalanced dataset

Subset ID	**Features in it**
IS31	21, 27, 12, 9, 28, 5, 31, 3, 34, 1, 13
IS32	22, 29, 25, 15, 16, 14, 26, 23, 19, 11, 17
IS33	20, 33, 6, 8, 10, 24, 7, 30, 4, 2, 18
IG@	21, 20,22,33,29,27,12,25,6,8,9
GR@	12,29,33,15,27,31,6,25.8.22.21
Chi@	33,29,27,12,15,31,25,6,22,20,21
Rel@	21,33,22,20,28,27,29,6,12,16,25

@ Top 11 features derived by existing methods (Refer Table 6)

Table 10. Subsets of features, Where # Subsets 5 over dermatology imbalanced dataset

Subset ID	Features in it
IS51	21, 8, 15, 26, 7, 1
IS52	22, 6, 9, 31, 30, 11
IS53	20, 25, 28, 5, 23, 2
IS54	33, 12, 16, 14, 3, 4
IS55	29, 27, 10, 24, 34, 19
IG^	21, 20,22,33,29,27
GR^	12,29,33,15,27,31
Chi^	33,29,27,12,15,31
Rel^	21,33,22,20,28,27

^ Top 6 features derived by existing methods (Refer Table 6)

For examining the strongness of each subset of attributes, an equal number of top attributes formed by the existing techniques (IG, Chi, Rel, GR) are considered. IS31, IS32, IS33 subsets have 11 attributes in it. So, top 11 attributes formed by the existing techniques are chosen to know the performance of those subsets. In a similar way, remaining subsets are measured by analyzing with JRip, J48, Naive Bayes, KNN classifiers.

The same framework is tested with dermatology balanced dataset(BD), breast cancer imbalanced dataset(IBC), and breast cancer balanced dataset(BBC) also. Symmetrical uncertainty (SU) is applied on those datasets to find out the rank of each attribute. The order of features after applying SU is given in Table 11.

Table 11. Order of features after applying SU on each dataset

Dataset	**Order of Features**
Balanced dermatology (BD)	21, 31, 7, 15, 33, 29, 27, 9, 12, 25, 6, 30, 8, 20, 22, 5, 28, 34, 10, 14, 16, 26, 11, 24, 4, 3, 2, 23, 19, 1, 18, 17, 13, 32 .
Imbalanced Breast Cancer(IBC)	23, 21, 24, 28, 8, 3, 7, 4, 1, 27, 14, 11, 13, 6, 26, 17, 2, 18, 22, 25, 29, 16, 5, 30, 9, 19, 20, **10, 12, 15**
Balanced Breast Cancer(BBC)	28, 23, 21, 8, 24, 27, 7, 4, 14, 3, 1, 11, 13, 6, 26, 17, 2, 18, 22, 16, 5, 25, 29, 9, 30, 20, 12, 19, 15, **10**

All the attributes of BD dataset have SU value greater than Zero, so all are considered to form the subsets.

SU value of feature **10, 12, 15** over IBC dataset and feature **10** over BBC dataset is zero. So, these features need to be discarded to form the clusters. 3, 4, 5 subsets of features are formed over all these datasets as per the proposed framework. For 3 subsets of features refer Table.12, 4 subsets of features refer Table.13 and 5 subsets of features refer Table.14.

Table 12. Subsets of features, Where # Subsets are 3

Dataset	Subset ID	Features in it
Balanced dermatology (BD)	BD31	21,29,27,30,8,34,10,24,4,1,18
	BD32	31,33,9,6,20,28,14,11,3,19,17
	BD33	7,15,12,25,22,5,16,26,2,23,13
	BD3 IG	21,9,7,31,20,28,25,15,33,29,27
	BD3 CHI	21,7,25,31,33,29,27,12,9,15,6
	BD3 GR	31,12,29,33,27,15,6,8,22,30,25
	BD3 REL	33,21,28,29,27,31,6,7,15,9,12
Imbalanced breast cancer(IBC)	IBC31	23, 3, 7, 11, 13, 18, 22, 30, 9
	IBC32	21, 8, 4, 14, 6, 2, 25, 5, 9
	IBC33	24, 28, 1, 27, 26, 17, 29, 16, 20
	IBC IG	23, 24, 21, 28, 8, 3, 4, 1, 7
	IBC3 CHI	23, 21, 24, 28, 8, 3, 4, 1, 7
	IBC3 GR	23, 21, 24, 28, 8, 7, 27, 3, 4
	IBC3 REL	21, 28, 23, 22, 1, 3, 8, 24, 4
Balanced breast cancer(BBC)	BBC31	28, 27, 7, 11, 13, 18, 22, 9, 30
	BBC32	23,24,4, 1, 6, 2, 16, 29, 20
	BBC33	21, 8, 14, 3, 26, 17, 5, 25, 12
	BBC3 IG	23, 24, 28, 21, 8, 3, 7, 1, 4
	BBC3 CHI	23, 28, 24, 21, 8, 7, 3, 4, 1
	BBC3 GR	28, 23, 21, 8, 27, 24, 7, 4, 14
	BBC3 REL	21, 23, 28, 3, 1, 8, 24, 22, 4

Table 13. Subsets of features, Where # Subsets are 4

Dataset	Subset ID	Features in it
Balanced dermatology (BD)	BD41	21,9,12,5,28,24,4,17
	BD42	31,27,25,22,34,11,3,18
	BD43	7,29,6,20,10,26,2,1
	BD44	15,33,30,8,14,16,23,19
	BD4 IG	21,9,7,31,20,28,25,15
	BD4 CHI	21,7,25,31,33,29,27,12
	BD4 GR	31,12,29,33,27,15,6,8
	BD4 REL	33,21,28,29,27,31,6,7
Imbalanced breast cancer(IBC)	IBC 41	23, 4, 1, 17, 2, 30
	IBC 42	21, 7, 27, 24, 18, 5
	IBC 43	24, 3, 14, 6, 22, 16
	IBC 44	28, 8, 11, 13, 25, 29
	IBC4 IG	23, 24, 21, 28, 8, 3
	IBC4 CHI	23, 21, 24, 28, 8, 3
	IBC4 GR	23, 21, 24, 28, 8, 7
	IBC4 REL	21, 28, 23, 22, 1, 3
Balanced breast cancer(BBC)	BBC41	28, 4, 14, 17, 2, 9, 30
	BBC42	23, 7, 3, 26, 18, 29, 20
	BBC43	21, 27,1, 6, 22, 25, 12

	BBC44	8, 24, 11, 13, 16, 5, 19
	BBC4 IG	23, 24, 28, 21, 8, 3, 7
	BBC4 CHI	23, 28, 24, 21, 8, 7, 3
	BBC4 GR	28, 23, 21, 8, 27, 24, 7
	BBC4 REL	21, 23, 28, 3, 1, 8, 24

Table 14. Subsets of features, Where # Subsets are 5

Dataset	Subset ID	Features in it
Balanced dermatology (BD)	BD51	21,25,6,14,16,1
	BD52	31,12,30,10,26,19
	BD53	7,9,8,34,11,23
	BD54	15,27,20,28,24,2
	BD55	33,29,22,5,4,3
	BD5 IG	21,9,7,31,20,28
	BD5 CHI	21,7,25,31,33,29
	BD5 GR	31,12,29,33,27,15
	BD5 REL	33,21,28,29,27,31
Imbalanced breast cancer(IBC)	IBC 51	23, 27, 14, 25, 29
	IBC 52	21,1, 11, 22, 16
	IBC 53	24, 4, 13, 18, 5
	IBC 54	28,7, 6, 2, 30
	IBC 55	8, 3, 26, 17, 1
	IBC5 IG	23, 24, 21, 28, 8
	IBC5 CHI	23, 21, 24, 28, 8
	IBC5 GR	23, 21, 24, 28, 8
	IBC5 REL	21, 28, 23, 22, 1
Balanced breast cancer(BBC)	BBC51	28, 3, 1, 16, 5
	BBC52	23, 14, 11, 22, 25
	BBC53	21, 4, 13, 18, 29
	BBC54	8, 7, 6, 2, 9
	BBC55	24, 27, 26, 17, 30
	BBC5 IG	23, 24, 28, 21, 8
	BBC5 CHI	23, 28, 24, 21, 8
	BBC5 GR	28, 23, 21, 8, 27
	BBC5 REL	21, 23, 28, 3, 1

5. Results and Discussion

The performance of each classifier (KNN, JRip, NB, J48) against the each subset of features and their ranks are given in this section with discussion. Rank of each subset by the selected classifier is denoted with **/ (slash).**

From the Table 15, it is cleared that, IS31 subset of features got boosted performance with the JRip. It is observed that IS32 subset of features also registered greater performance with KNN, NB, J48 when compared with existing feature selection methods. With this three subsets approach, only 33 % of features can be trained for model generation and time for training, and memory consumption can also be reduced. BD33 subset of features displayed the highest accuracy than all existing methods over the balanced dataset with all classifiers.

From the Table 16, it has been observed that almost all subsets of features registered greater accuracy with the almost all classifiers when compared with existing feature selection techniques. Especially IS43 subset placed in 1st position than all. Remaining all subsets occupied in top 4 positions. Over the balanced dataset, KNN, Jrip, J48 recorded the best performance with BD41 subset of features. With this 4 subsets approach, only 25 % of features can be trained for model generation.

From the Table 17, subset IS53 performed better than all existing techniques with KNN, J48 and Jrip. IS51 recorded better performed when analyzed with NB and J48. Over the balanced dataset, BD54 boosted the performance of all classifiers. Remaining subsets performance can also be interpreted in the same way. With this 5 subsets approach, only 20 % of features can be trained for model generation.

From the Table 18, it is cleared that, IBCREL (ReliefF) subset of features got boosted performance with the KNN. IBC32 performing better than other existing methods except ReliefF. IBC32 recorded improved performance with Jrip and J48. All the existing methods performed well over the imbalanced dataset. ReliefF performed better than all methods with the KNN and J48, but BBC31 and BBC33 recorded excess accuracy other than ReliefF over the balanced dataset.

From the Table 19, it has been observed that ReliefF method performed better than all methods. IBC42 subset of features recorded improved performance than IG, GR, and CHI with KNN over the imbalanced dataset. JRip performed well with BBC41 subset of features. KNN, J48, NB recorded better accuracy with the BBC43 over the balanced dataset.

From the Table 20, it is cleared that, IBC51 subset of feature performed better than existing IG, GR, CHI with all Classifiers over the imbalanced dataset. BD52 recorded the highest performance than all existing methods with KNN, JRip, J48 over the Balanced dataset. To justify and prove the worth of proposed framework, it is also tested with 5 more real time benchmark datasets. Those result analysis can be found here (Open the URL).

Table 15 . Performance analysis with 3 Subsets over Dermatology data set

Imbalanced					**Balanced**				
ID	**KNN**	**JRip**	**NB**	**J48**	**ID**	**KNN**	**JRip**	**NB**	**J48**
IS31	84.42/4	**86.06/1**	90.43/2	87.43/2	BD31	88.92/3	85.56/3	81.91/6	87.90/3
IS32	**87.43/1**	85.24/2	**90.71/1**	**88.52/1**	BD32	92.27/2	91.10/2	86.29/2	92.27/2
IS33	85.71/3	82.24/5	84.42/4	83.06/4	**BD33**	**97.66/1**	**97.08/1**	**97.81/1**	**97.52/1**
CHI	85.79/2	83.6/3	85.51/3	83.33/3	BD CHI	84.83/6	84.40/6	83.66/5	83.38/6
GR	83.06/5	82.78/4	83.87/5	81.69/5	BD GR	82.94/7	82.50/7	81.04/7	82.50/7
IG	80.05/7	64.48/7	79.23/6	78.68/6	BD IG	87.60/4	85.42/4	86.15/3	85.56/4
REL	80.32/6	72.4/6	79.23/7	74.86/7	BD REL	87.60/5	85.13/5	85.56/4	85.56/5

Note: The existing method performance is given in bottom 4 rows of every table(Table 15 to Table 20)

Table 16 . Performance analysis with 4 Subsets over Dermatology data set

Imbalanced					**Balanced**				
ID	**KNN**	**JRip**	**NB**	**J48**	**ID**	**KNN**	**JRip**	**NB**	**J48**
IS41	82.51/3	84.15/2	86.61/2	86.06/2	**BD41**	**91.25/1**	**91.39/1**	86.44/2	**91.54/1**
IS42	80.6/4	68.57/5	80.32/4	80.87/4	BD42	73.76/6	73.17/6	72.01/6	73.17/6
IS43	**88.25/1**	**87.97/1**	**91.25/1**	**91.53/1**	BD43	87.02/4	88.48/3	81.19/4	88.62/2
IS44	84.15/2	82.51/3	86.33/3	84.15/3	BD44	88.48/2	88.75/2	**87.12/1**	88.48/3
CHI	69.12/7	68.03/6	69.12/7	68.57/7	BD CHI	73.46/7	72.15/7	68.95/7	72.59/7
GR	69.12/7	68.03/6	69.12/7	68.57/7	BD GR	65.59/8	67.49/8	65.59/8	67.20/8
IG	75.95/6	59.83/7	74.86/6	75.95/6	BD IG	87.17/3	85.27/4	85.86/3	85.86/4
REL	78.14/5	75.13/4	78.41/5	76.22/5	BD REL	77.84/5	78.13/5	73.32/5	74.48/5

Table 17 . Performance analysis with 5 Subsets over Dermatology data set

Imbalanced					**Balanced**				
ID	**KNN**	**JRip**	**NB**	**J48**	**ID**	**KNN**	**JRip**	**NB**	**J48**
IS51	85.51/2	85.51/2	**87.97/1**	**87.7/1**	BD51	78.71/2	80.17/2	71.28/6	80.75/2
IS52	69.67/6	54.64/7	70.21/6	70.76/5	BD52	71.57/7	71.20/6	72.30/4	72.15/7
IS53	**86.61/1**	**86.06/1**	87.43/2	**87.7/1**	BD53	69.82/8	66.47/8	70.11/7	67.93/8
IS54	76.77/4	70.49/4	76.5/4	74.31/4	BD54	**81.63/1**	**81.77/1**	**77.69/1**	**81.04/1**
IS55	64.2/8	53.82/8	65.3/8	65.4/7	BD55	76.38/4	72.59/4	73.90/3	74.05/5
CHI	69.12/7	69.12/5	69.12/7	68.85/6	BD CHI	73.61/6	72.44/5	69.97/8	72.59/6
GR	69.12/7	69.12/5	69.12/7	68.85/6	BD GR	67.20/9	67.34/7	65.59/9	67.20/9
IG	76.22/5	59.28/6	74.86/5	76.22/3	BD IG	75.51/5	72.44/5	75.05/2	74.34/4
REL	77.59/3	71.85/3	78.41/3	76.77/2	BD REL	76.53/3	76.53/3	72.15/5	74.92/3

Table 18 . Performance analysis with 3 Subsets over Breast cancer data set

Imbalanced					**Balanced**				
ID	**KNN**	**JRip**	**J48**	**NB**	**ID**	**KNN**	**JRip**	**J48**	**NB**
IBC31	93.14/6	93.32/4	93.67/3	91.91/4	BBC31	*95.68/2*	92.38/6	91.81/6	91.23/4
IBC32	*95.25/2*	**95.43/1**	**96.3/1**	92.97/3	BBC32	93.67/5	93.1/5	*93.24/3*	*93.24/3*
IBC33	94.37/3	93.49/3	93.49/4	93.67/2	BBC33	*95.68/2*	*93.67/3*	*93.39/2*	*93.24/3*
IBC IG	94.20/4	93.49/3	92.26/6	**94.02/1**	BBC IG	95.25/3	93.53/4	93.10/4	93.53/2
IBC CHI	94.20/4	93.49/3	92.26/6	**94.02/1**	BBC CHI	95.25/3	93.53/4	93.10/4	93.53/2
IBC GR	93.32/5	92.09/5	92.79/5	**94.02/1**	BBC GR	94.54/4	93.95/2	93.1/5	**93.82/1**
IBC REL	**95.78/1**	94.55/2	94.90/2	**94.02/1**	BBC REL	**97.27/1**	95.11/1	**94.97/1**	*93.24/3*

Table 19 . Performance analysis with 4 Subsets over Breast cancer data set

Imbalanced					**Balanced**				
ID	**KNN**	**JRip**	**J48**	**NB**	**ID**	**KNN**	**JRip**	**J48**	**NB**
IBC41	93.49/4	94.20/4	92.61/5	93.32/5	BBC41	95.25/2	**95.54/1**	94.97/2	93.96/3
IBC42	*94.55/2*	*94.72/3*	92.97/4	*94.37/3*	BBC42	92.24/5	92.67/5	93.96/3	92.95/5
IBC43	92.79/5	92.61/6	*93.67/2*	92.61/6	BBC43	**96.26/1**	94.82/2	**95.11/1**	**95.11/1**
IBC44	90.15/6	92.79/5	91.21/6	92.44/7	BBC44	92.09/6	92.09/7	92.09/7	90.22/6
IBC IG	94.20/3	95.25/2	93.14/3	93.67/4	BBC IG	94.68/3	94.39/3	93.1/5	93.96/3
IBC CHI	94.20/3	95.25/2	93.14/3	93.67/4	BBC CHI	94.68/3	94.39/3	93.1/5	93.96/3
IBC GR	93.49/4	94.20/4	92.61/5	94.55/2	BBC GR	94.54/4	93.24/4	93.39/4	94.39/2
IBC REL	**95.43/1**	**95.43/1**	**94.20/1**	**94.72/1**	BBCREL	94.54/4	92.95/6	92.81/6	93.39/4

Table 20 . Performance analysis with 5 Subsets over Breast cancer data set

Imbalanced					**Balanced**				
ID	**KNN**	**JRip**	**J48**	**NB**	**ID**	**KNN**	**JRip**	**J48**	**NB**
IBC51	*93.32/2*	*93.84/2*	*94.02/2*	*94.20/2*	BD51	93.39/3	93.82/5	93.96/2	*92.95/3*
IBC52	90.33/7	92.61/5	93.84/3	92.44/4	BD52	**96.69/1**	**96.12/1**	**95.54/1**	89.22/7
IBC53	91.73/5	92.97/4	91.91/5	89.10/7	BD53	93.24/5	92.81/7	93.24/4	90.08/6
IBC54	92.44/3	91.91/6	91.21/6	89.45/6	BD54	91.81/7	92.38/8	90.22/7	90.94/5
IBC55	90.86/6	91.03/7	91.03/7	91.91/5	BD55	91.54/8	93.67/6	93.24/4	92.24
IBC IG	92.26/4	93.67/3	92.61/4	94.20/3	BD IG	92.67/6	95.53/2	92.52/6	94.25/2
IBC CHI	92.26/4	93.67/3	92.61/4	94.20/3	BD CHI	92.67/6	95.53/2	92.52/6	94.25/2
IBC GR	92.26/4	93.67/3	92.61/4	94.20/3	BD GR	93.53/4	95.11/3	93.67/3	**94.97/1**
IBCREL	**95.95/1**	**94.55/1**	**94.20/1**	**95.07/1**	BD REL	94.39/2	94.25/4	93.10/5	92.24/4

6. Conclusion

In this study, a novel M-cluster of dimensionality reduction and ranking framework is proposed. The proposed method is analyzed using real time Dermatology and Breast Cancer dataset. It has been observed that initial dataset is having class imbalance problem. To overcome this problem, the oversampling technique called SMOTE is applied and balanced the dataset. With this framework ‘S’ number of subsets of attributes are formed, each subset has a minimum number of attributes without any duplication. All the subsets of attributes are analyzed using JRip, J48, NB, KNN classifiers, and corresponding results are compared with the existing filter-based feature selection techniques over balanced and imbalanced datasets. Then, ranking for each subset is assigned as per the accuracy. Displayed results show that one of the subsets, in some cases more than one subset giving boosted results than existing methods. With this, we conclude that instead of selecting features using already existing methods, depending on the requirement, the proposed technique can be used to form the subset of features for greater prediction accuracy. The proposed method performance may vary depending on the data set considered. This framework can be implemented with the MapRedece approach in case analysis of large amount of data set using Hadoop framework.

References

1. Saleem, A., Asif, K.H., Ali, A., Awan, S.M. and Alghamdi, M.A., 2014, December. Pre-processing methods of data mining. In *Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on* (pp. 451-456). IEEE.
2. Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016.*Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
3. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 13, 8–17. doi:10.1016/j.csbj.2014.11.005
4. Silwattananusarn, T., Kanarkard, W. and Tuamsuk, K., 2016. Enhanced Classification Accuracy for Cardiotocogram Data with Ensemble Feature Selection and Classifier Ensemble.*Journal of Computer and Communications*, *4*(04), p.20.
5. Masethe, H.D. and Masethe, M.A., 2014, October. Prediction of heart disease using classification algorithms. In *Proceedings of the world congress on engineering and computer science* (Vol. 2, pp. 22-24).
6. Angelis, V., Felici, G. and Mancinelli, G., 2006. Feature selection for data mining.*Data Mining and Knowledge Discovery approaches based on rule induction techniques*, pp.227-252.
7. Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering 40, 16–28. doi:10.1016/j.compeleceng.2013.11.024
8. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*,*16*, pp.321-357.
9. Potharaju, S.P. and Sreedevi, M., 2016. An Improved Prediction of Kidney Disease using SMOTE.*Indian Journal of Science and Technology*, *9*(31).
10. Potharaju, S.P., Sreedevi, M., 2016. Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data. Journal of Engineering Science and Technology Review 9, 201–207.
11. Sarkar, C., Cooley, S. and Srivastava, J., 2014. Robust feature selection technique using rank aggregation.*Applied Artificial Intelligence*, *28*(3), pp.243-257.
12. Shivakumar, B.L. and Alby, S., 2014, March. A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes. In *Intelligent Computing Applications (ICICA), 2014 International Conference on* (pp. 167-173). IEEE.
13. Sheikhpour, R. and Sarram, M.A., 2014. Diagnosis of Diabetes Using an Intelligent Approach Based on Bi-Level Dimensionality Reduction and Classification Algorithms.*Iranian Journal of Diabetesband Onesity*, *6*(2), pp.74-84.
14. Balakrishnan, S. and Narayanaswamy, R., 2009. Feature selection using FCBF in type ii diabetes databases.*International Journal of the Computer, the Internet and the Management*, *17*(1), pp.50-8.
15. Abdul-Rahman, S., Norhan, A.K., Yusoff, M., Mohamed, A. and Mutalib, S., 2012, December. Dermatology diagnosis with feature selection methods and artificial neural network. In *Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on* (pp. 371-376). IEEE.
16. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F., 2014. A review of microarray datasets and applied feature selection methods. Information Sciences 282, 111–135. doi:10.1016/j.ins.2014.05.042
17. Piao, Y., Piao, M., Park, K., Ryu, K.H., 2012. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics28,3306–3315. doi:10.1093/bioinformatics/bts602
18. Abe, N. and Kudo, M., 2005. Entropy criterion for classifier-independent feature selection. In *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 905-905). Springer Berlin/Heidelberg.
19. Novaković, J., 2016. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*,*21*(1).
20. Ferreira, C.H., de Medeiros, D.M. and Santana, F., 2016, July. FCFilter: Feature selection based on clustering and genetic algorithms. In *Evolutionary Computation (CEC), 2016 IEEE Congress on* (pp. 2106-2113). IEEE.
21. Singh, D.A.A.G., Leavline, E.J., Priyanka, R. and Priya, P.P., 2016. Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. I*nternational Journal of Intelligent Systems and Applications*, *8*(1), p.67.
22. Du, W., Cao, Z., Song, T., Li, Y. and Liang, Y., 2017. A feature selection method based on multiple kernel learning with expression profiles of different types. *BioData Mining*, *10*(1), p.4.
23. https://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/
24. https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/