GU J Sci 31(2): 456-470 (2018)

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

A Novel Cluster of Quarter Feature Selection Based on Symmetrical Uncertainty

Sai Prasad POTHARAJU[1,*], Marriboyina SREEDEVI[2]

[1]*Computer Science and Engineering Department, K L University, 522502 Guntur (AP), India*

[2]*Computer Science and Engineering Department, K L University, 522502 Guntur (AP), India*

Article Info

Received: 02/05/2017
Accepted: 19/12/2017

Keywords

Data Mining
Feature Selection
Filter
Pre-Processing
Symmetric Uncertainty

Abstract

Due to the diversity of sources, a large amount of data is being produced. The captured data associated with several problems including mislabeled data, missing values, imbalanced class labels, noise and high dimensionality. In this research article, we proposed a novel framework to address the high dimensionality issue with feature selection to increase the classification performance of various lazy learners, rule-based induction, Bayes, and tree-based models. In this research, we proposed robust Quarter Feature Selection (QFS) framework based on Symmetrical Uncertainty Attribute Evaluator. Our proposed technique analyzed with Six real world datasets. The proposed framework, divides the whole data space into 4 sets (Quarters) of features without duplication. Each such quarter has less than or equals 25 % features of whole data space. Practical results recorded that, one of the quarter, sometimes more than one quarter recorded improved accuracy than the traditional feature selection methods in the literature. In this research, we used filter-based feature selection methods such as Gain Ratio (GRAE), Information Gain (IG), Chi Squared (CHI 2), Relief to compare the quarter of features created by proposed technique.

1. INTRODUCTION

Data mining (DM) is an effective process for getting hidden knowledge (interesting patterns) from large datasets. It is currently gaining a massive deal of focus. DM also became a salient analysis tool [1]. In recent days, DM techniques are applied in diverse areas such as human resource management (HRM) , telecommunications, stock market analysis, supermarkets, banking, health care management (HCM), traffic management, education institutes and others. In DM, prediction and classification are most oftenly used for forecasting and analyzing the present progression. Data mining is a broad concept, it has several stages, one of the stage is data preprocessing, in this stage, noise will be minimized, missing values are normalized, missing labels are corrected. After this stage, mining methods such as clustering, association rule mining, classification, and others are employed on processed data set. Results from such mining methods are evaluated and interpreted for better decision making. Due to the multiple intermixed platforms, a large amount of data is being generated, it associated following difficulties with it.

1. Mislabeled data

As data increases, the chance of mislabeled data points increases as well. When considering such large data points, it is not simple to cross check whether all of such training data points are labeled or not, and training models on such incorrect data points will leads to weak accuracy.

*Corresponding author, e-mail: psaiprasadcse@gmail.com

2. Missing values
Similar to mislabeled data points, missing values also leads to weak accuracy model generation when clustering algorithms are applied. This problem can be generally minimized either by removing the data points permanently or using imputation techniques.

3. Noise
Noisy data suffer from over fitting. Clustering methods can help to check the noisy data.

4. Imbalanced data points
In classification problems, the imbalanced problem happens during the training phase, if majority data points belong to single class. This problem heads to less accurate learners. This can be addressed using SMOTE (Synthetic Minority over Sampling Technique).

5. High dimensionality
This problem can emerge when the features are more, or data points are very large. To manage this issue Feature selection (FS) techniques and Principal Component Analysis (PCA) can be employed.
In this research, we contributed a novel feature selection framework to address high dimensionality issue of preprocessing. FS methods permit to conceive the proper predictive model. It helps the decision makers by preferring features that can launch the preferable performance with limited features. Generally, less number of features are encouraged to generate the model, because it reduces the complexity of the computation, and also simple model is easy to understand. FS has below listed advantages, such as:

- Accelerating the performance of the algorithm.
- Simple to understanding.
- Easy to getting hidden knowledge.
- Easy to visualize the process.
- Minimizing the processing and storage costs.
- Maximize the speed of generation of the model .

FS is an essential experimental technique applied on high dimensional data. Basically, FS is used to cutting down meaningless (not important) features and redundant features [2]. As the number of features raising, competency level of data analysis method reduced. Hence, FS is applied in preprocessing step of analysis to find the best subset of features. Since the accuracy of the classifier is influenced by immaterial and redundant feature, it is obligatory to select the best subset of features by eliminating those unwanted features. Meaningless features do not give additional strength to learning model, and also there is a scope of leading to distraction at the time the classification. In existed study of literature, three modes of FS techniques are presented by various researchers. Those include Filter, Wrapper, and Embedded.

Filter:
This mode uses some evaluation criteria in order to remove noise and immaterial features that describe the dataset strongly. Filter methods accomplish the feature selection procedure as a preprocessing step with no induction algorithm. The common properties like the distance between classes or statistical dependencies of the training data are used to select features. This model is quicker than the wrapper technique. As it acts independently of the induction algorithm, it produces better generalization results. However, this method selects the high number of subsets of features, a threshold value is recommended to select the subset. These methods, designates the rank to the each variables based on usefulness. Examples of such methods includes Information Gain (IG), Chi-Square Attribute Evaluator (Chi), Relief, and Gain Ratio Attribute Evaluator (GRAE). In this current study, our proposed technique is compared with these existing filter methods.

Wrapper:
This method uses some searching techniques in order to choose the candidate subset. The idea behind the wrapper approach is, the induction algorithm is considered as a black box. The induction algorithm is executed on the data set, usually original feature space is divided into different sets of features. Induction

algorithm will run on all set of features. The feature subset with the highest evaluation is selected as the final set on which algorithm is applied. Typically used evaluation techniques by wrapper approach are best first search, Hill climbing, Genetic search.

Embedded:
In this method, selection of attribute is based on the ensembling of any classifiers like nearest neighbor or support vector machine etc. On the basis of competency of classifier used, selection of best feature will takes place.

Interact, Fast correlation based filter (FCBF), Correlation-based feature selection (CBFS), Fast clustering based feature subset selection, are few algorithms belong to filter category. Interact algorithm uses Symmetric Uncertainty (SU) as selection criteria. It increases the accuracy as the number of dimensions increases. FCBF proposed by Yu [3], it uses the correlation between features as the best measure based on SU. It offers the huge reduction of features, but it fails to address redundant features. According to Cui [4], CBFS also uses SU as criteria for measuring the correlation between feature to class and feature to feature. It performs good on low dimensional datasets but fails to address numerical class problems. It has an advantage of addressing both meaningless and redundant features. Relief is extensible to datasets with increasing dimensionality. It has the disadvantage of removing redundant attributes. FS based on wrapper uses sequential floating forward Selection (SFFS) as searching criteria for feature selection, and SVM for evaluation. It gives good accuracy and also offers rapid computation [5].

Recently ensembling approaches gaining popularity as they combine the advantages of multiple classifiers or multiple techniques. Bagging and Boosting are two basic ensembling techniques. However, in this present research, we used SU as key criteria to form the features, which will be discussed in next section.

The remaining portion of this article is structured as follows. Section 2 contains the existed literature and background subject. Section 3 describes the proposed methodology for the intended research, which includes the QFS approach. Section 4 includes the experimental analysis on real-world datasets. Section 5 presents the results and discussion on datasets by various classifiers. Conclusions and recommendations are presented in the final section.

2. LITERATURE

Roffo [6] proposed Infinite feature selection (Inf-Fs) by considering weight of feature relevance and redundancy. Inf-Fs technique recorded best performance against filter, wrapper and hybrid methods. Genetic algorithm based feature selection is applied for Stock market analysis. According to this study, GA achieved best performance than PCA. In current research, SU is used as primary criteria to measure the strongness of features of whole data space. SU calculates strongness between feature and the target class. The feature which has maximum value of SU gets high priority for selection. It can be defined as

$$SU(A,B) = 2 * MI(A,B) / H(A) + H(B)$$
$$H(A) = -\partial\ p(a) \log (p(a))\ \partial x$$
$$MI(A,B) = H(A,B) - H(A|B) - H(B|A)$$

where H(A) is the entropy of a discrete random variable A. MI is mutual information , used to measure how two attributes are correlated.

Authors proposed Ant Colony Optimization (ACO) technique [7]. In their technique, subset of features are selected using ACO and selected features are evaluated on the basis of SU .This method is scrutinized on 17 real-world datasets. Out of 17 datasets, their proposed technique displayed better on 15 datasets when compared with various implementations using ACO, Particle Swarm Optimization, Genetic Algorithm. A technique based on the memetic framework for feature selection is proposed [8] in literature .In the memetic framework, for local search ranking method is considered. This method is compared with Genetic Algorithm and few other existed feature selection methods. Their proposed framework is

performed better than those cited. In article [9], authors worked for DDoS detection with various feature selection methods. They considered 4 training sets and 4 different subsets of features. For the experiment, authors considered Decision Tree with SU, and Chi-Square techniques. As per their experiment, only 7 features are recording 95 % accuracy to detect DDoS attack patterns.

In research article [10], the authors presented a study on airline database and applied a different filter, wrapper-based feature selection techniques available in weka software tool. As per their study, PCA Transformer would perform better than other attribute evaluators on airline data. Authors of the article [11], worked to predict the customer reordering demand in direct marketing. For their investigation, three techniques namely, CFS- Correlation based feature selection, SU, and SC- Subset Consistency are employed. According to obtained outcome, SU is performing better than other two techniques used .In the article [12], authors used two real-world datasets and six ranking based methods based on entropy and statistical measurement. To build a model, four algorithms namely, C4.5, RBF network, IB1 and Naive Bayes. As per experimental result, different ranking feature selection methods given different results with different learning algorithms.

Intrusion detection systems (IDSs) deal with a huge amount of data, It is an important task to select the best features which represent the whole data space without redundant. Authors of [13] proposed cuttlefish optimization algorithm (CFA) to select best features for the same. To judge the features selected by CFA, decision tree (DT) classifier is used. The degree of redundancy and independence among features using correlation information entropy is used to construct correlation matrix. Then, eigenvalues are calculated. Therefore, the sorting technique of features and an adaptive feature subset selection technique combining with the parameters are proposed [14]. Information theory is a popular consideration in FS due to its scalability, classifier independence, and computational efficacy. Based on information theory, authors of [15] proposed JMIM-Joint Mutual Information Maximization and NJMIM-Normalized Joint Mutual Information Maximization and tested on 11 real datasets compared with existing techniques. The results show that the JMIM technique performs better than the other methods on most real datasets. A dynamic mixed strategy based on filter based approach is proposed in the research article [16], this approach joins the mutation operators. Mutation operators are formed with standard deviation (SD) and cardinality of candidate subset of features.

Authors of the article [17] proposed feature selection technique based on dependency margin. This technique is performing better than existed traditional methods. Feature Selection using mutual information is proposed in the article [18], to select the compact subset of features. In this method, mutual information between features is calculated with respect to the target class. Based on relevance measure, a multi-objective evolutionary algorithm with class-dependent redundancy for feature selection (MECY-FS) is proposed. Pareto optimality is employed in the MECY-FS algorithm to evaluate candidate feature subsets to select best feature subsets with both the maximal relevance and the minimal redundancy. Wrapper FS based on GA is proposed by authors of [19], in their research, a parallel GA is used to examine and evaluate a huge number of candidate subset features simultaneously. This method is experimented on heterogeneous biomedical data sets and produced a remarkable reduction of the number of features without minimizing performance. FS based on a two-tier approach using information gain to discover a list of attacks in intrusion detection is proposed in the article [20]. Ranking of each feature in decreasing order is calculated using high information gain entropy in the first tier. The next tier stretch out additional features with a finer discriminative ability than the priorily ranked features. Researchers of [21] proposed SYMON method, it uses SU and harmony search. SU is used to measure the weight of features with respect to their dependency to class labels. Harmony search is used as an optimization problem to select the best possible combination of features.

Motivation:
The objective of proposed methodology is, to reduce the searching of feature space by 25%. The proposed method is inspired from ensembling approaches like Bagging and Boosting. These approaches combine the two or more classifiers and improve the overall performance. In the similar fashion, instead of considering only strong features, if weak or average strength.

3. PROPOSED METHODOLOGY

The proposed methodology is based on the requirement that, if there is a requirement to select maximum 1/4th features from whole space, which features has to be selected? In literature, to select the top subset of features there are few existing techniques are available. Those includes: Information Gain (IG), Chi 2 Attribute Evaluator (Chi), Relief , and Gain Ratio Attribute Evaluator (GRAE). Other Than these available techniques, in this current research we proposed a new framework for feature selection based on Symmetric Uncertainty.

Our proposed framework methodology is as follows.
1. Generate the weight and Rank of each feature using SU
2 Remove the features, whose weight is Zero (0) as it can't influence the learners.
Follow the below steps to form the subset of features in 4 Quarters.
Step 1: Arrange the first 4 features in descending order of Ranks from left to right in Level 1
Step 2: Arrange the next 4 features in descending order of Ranks from right to left in Level 2.
Step 3: Repeat the Step 1 then step 2 for next Levels until the all features are arranged.
Step 4: Group, all vertically first order features of all levels in First Quarter, Second order features of all levels in Second Quarter, and so on.
Step 5: Balance the number of features of each quarter by removing last feature from the quarter which has an extra feature, if not balanced.

Generalized cpp pseudo code for the above algorithm is given below

```
        int n, t, nc, list[50], c[10][10] ;
    cout<<"Enter No. of Terms : " ;
    cin>>n ;
    cout<<"Enter The Terms : " ;
    int i = 0, j = 0, k = 0, fb = 0, wj[10] ;
    for ( i = 0 ; i < 10 ; i++ )
        wj[i] = 0 ;
    for ( i = 0 ; i < n ; i++ )
    {
        cin>>list[i] ;
    }
    cout<<"Enter Number of Cluster : " ;
    cin>>nc ;
    for ( i = 0 ; i < n ; i++ )
    {
        c[j][k] = list[i] ;
        wj[j]++ ;
        if ( fb == 0 )
        {
            j = j + 1 ;
            if ( j == nc )
            {
                k = k + 1 ;
                fb = 1 ;
                j = j - 1 ;
            }
        }
        else
        {
            j = j - 1 ;
            if ( j == -1 )
            {
                k = k + 1 ;
```

```
                    fb = 0 ;
                    j = j + 1 ;
                }
            }
        }
        cout<<"\n\n\t Display \n" ;
        for ( i = 0 ; i < nc ; i++ )
        {
            cout<<"Cluster "<<i+1<<"\t" ;
            for ( j = 0 ; j < wj[i] ; j++ )
            {
                cout<<c[i][j]<<"\t" ;
            }
            cout<<endl ;
        }
        cout<<"\n\n Balanced Cluster \n" ;
        int temp = n / nc ;
        for ( i = 0 ; i < nc ; i++ )
        {
            cout<<"Cluster "<<i+1<<"\t" ;
            for ( j = 0 ; j < temp ; j++ )
            {
                cout<<c[i][j]<<"\t" ;
            }
            cout<<endl ;
        }
```

Example:
Consider the following example to form 4 sets (Quarters) of subset of features.
Assume total number of features (N) is 20. Table 1 describes the weight and rank of each features generated using SU.

Table 1. *Sample Weight and Rank of each feature*

Weight	Rank	Attribute Name	Weight	Rank	Attribute Name	Weight	Rank	Attribute Name
0.92	1	A	0.52	8	H	0.1	15	O
0.91	2	B	0.5	9	I	0	16	P
0.88	3	C	0.49	10	J	0	17	Q
0.86	4	D	0.48	11	K	0	18	R
0.85	5	E	0.4	12	L	0	19	S
0.6	6	F	0.3	13	M	0	20	T
0.59	7	G	0.2	14	N			

As per the proposed method, features having weight zero need to be removed from feature space. In the given example, features P,Q,R,S,T, has weight zero. So, discard them from the feature space and apply the algorithm to form the quarters. Below table 2 describes the formation of features in each quarter**.**

Table 2. *Example of subset of features in each Quarter*

Level	Q1	Q2	Q3	Q4	Direction
1	A	B	C	D	Left to Right
2	H	G	F	E	Right to Left
3	I	J	K	L	Left to Right
4		O	N	M	Right to Left
#	3	4	4	4	

Total Number of features in each Quarter

According to step 5 of framework, all quarters should contain equal number of features. Q2, Q3, Q4 has 4 features and Q1 has 3 features. To balance the all quarters, remove the last feature from Q2, Q3, Q4 i.e. O from Q2, N from Q3, M from Q4 has to be removed. After this procedure Q1 has (A,H,I),Q2 has (B,G,J),Q3 has (C, F, K), Q4 has (D, E, L) set of features.

4. EXPERIMENT

To experiment our proposed methodology, 6 benchmark data sets available at UCI machine learning repository are used. Data sets considered to test the proposed methodology are described in table 3. For analyzing the proposed framework popular machine learning tool WEKA is used with all default settings.

Table 3. *Description of datasets used*

Dataset	#Instances	# Attributes	# Class
Spambase	4601	57	2
Musk (V2)	6598	168	2
Dermatology	366	34	6
Bio Degeneration	1054	41	2
Libras Movement	360	91	15
Connectionist Bench	208	60	2

Features of each Quarter are analyzed using Tree, Lazy, Bayes, Rule based classifiers and compared with existing FS methods like IG, GRAE, CHI2, Relief. Table 4 describes the classifiers used to analyze the datasets for proposed and existing techniques.

Table 4. *List of Classifiers*

Type	Classifiers
Rule	Jrip, OneR, Ridor
Tree	J48, Simplecart
Bayes	Naive Bayes
Lazy	IBK

To test the strength of proposed feature selection framework against existing techniques, equal number of top features of traditional techniques are considered .i.e. if proposed framework derives 'N ' features, then

top 'N' features of existing techniques are considered. Description of features selected by proposed technique for each dataset is given in below table 5.

Table 5. *Number of features selected for each dataset*

Dataset	# Features selected
Spambase	14
Musk (Version 2)	36
Dermatology	8
Bio Degeneration	10
Libras Movement	18
Connectionist Bench	5

Sample features formed by proposed and traditional techniques for the Spambase data set is given in table 6.

Table 6. *Sample subset of features obtained by proposed and existing techniques*

Dataset	T	S	Q1	Q2	Q3	Q4	IG	CHI	GRAE	Relief
Spambase	57	14	2	5	4	1	5	3	7	2
			6	13	9	3	7	5	11	9
			7	15	10	11	16	7	16	11
			8	17	12	16	19	16	20	12
			14	21	20	18	21	19	21	21
			25	26	22	19	23	21	23	23
			27	28	23	32	24	23	24	25
			29	30	24	35	25	24	25	26
			31	33	37	39	27	25	26	27
			34	43	41	40	52	52	27	28
			36	46	42	50	53	53	29	30
			38	48	44	51	55	55	50	32
			45	49	56	52	56	56	52	34
			54	53	57	55	57	57	53	40

T : Total number of features.
S: Features formed by proposed technique
Q1: Features of first Quarter
Q2: Features of second Quarter
Q3: Features of third Quarter
Q4: Features of fourth Quarter
IG: Top 'S' features formed by Information Gain

CHI: Top 'S' features formed by Chi Square Attribute Evaluator
GRAE: Top 'S' features formed by Gain Ratio Attribute Evaluator
Relief: Top 'S' features formed by Relief

5. RESULTS AND DISCUSSION

In this section, results obtained by proposed and existing techniques are given.

Spambase

Table 7. *Accuracy of classifiers for the dataset Spambase*

	Jrip	OneR	Ridor	J48	SC	NB	IBK
Q1	85.22	76.26	84.37	86.06	86.37	68.28	84.43
Q2	86.56	79.00	84.52	87.00	86.43	64.92	85.59
Q3	83.78	75.17	82.28	85.80	85.93	79.09	83.35
Q4	87.02	78.30	85.50	88.08	87.50	52.22	84.48
IG	91.58	78.30	91.65	92.67	91.78	85.91	90.06
CHI	91.82	78.30	90.91	92.63	91.11	82.61	89.39
GRAE	90.26	78.30	89.00	90.80	90.26	71.48	88.58
Relief	85.04	75.48	84.02	85.50	85.76	68.11	84.11

Hierarchical accuracy of each classifier against each subset of features formed by proposed and existing techniques for the dataset Spambase is given in table 8.

Table 8. *Hierarchical accuracy of each classifier for the dataset Spambase*

Classifier	Hierarchical accuracy
Jrip	CHI >IG >GRAE >**Q4 >Q2 >Q1**>Relief >Q3
OneR	**Q2>Q4** >IG >CHI > GRAE>**Q1** >Relief >Q3
Ridor	IG>CHI>GRAE>**Q4>Q2> Q1**>Relief>Q3
J48	IG >CHI>GRAE>**Q4>Q2>Q1>Q3**>Relief
Simple cart	IG >CHI>GRAE>**Q4>Q2>Q1>Q3**>Relief
Naive Bayes	IG>CHI>**Q3>**GRAE>**Q1**>Relief>Q2>Q4
IBK	IG>CHI>GRAE>**Q2>Q4>Q1**>Relief>Q3

Q4, Q2, Q1 subset of features recorded increasing performance than existing technique Relief with Jrip. Q2, Q4 performed better than all existing techniques, and Q1 recorded improved accuracy than Relief with OneR. Q4, Q2, Q1 subset of features displayed increasing performance than existing technique Relief with Ridor. All the subset of features performed better than existing Relief with tree based J48 and Simplecart. With Naive Bayes, Q3 recorded better than GRAE and Relief, Q1 displayed better than Relief. Q2, Q4, Q1 subset of features displayed increasing performance than existing technique Relief with IBK. In the similar fashion performance with other datasets can be interpreted.

Bio Degeneration

Table 9. *Accuracy of classifiers for the dataset Bio Degeneration*

	Jrip	OneR	Ridor	J48	SC	NB	IBK
Q1	79.81	77.15	80.28	80.00	80.66	70.14	77.44
Q2	81.80	72.03	77.63	80.66	81.80	77.53	80.47
Q3	82.18	71.65	81.61	84.36	83.69	73.08	81.89
Q4	79.90	69.57	78.67	83.60	82.46	70.04	81.61
IG	78.95	77.15	78.10	79.81	80.56	72.79	79.24
CHI	80.94	77.15	79.71	81.42	80.56	72.79	78.48
GRAE	80.47	77.15	78.95	78.86	81.13	61.13	77.34
Relief	83.98	77.15	82.27	85.02	84.07	68.15	82.74

Hierarchical accuracy for the dataset Bio degeneration is given in table 10.

Table 10. *Hierarchical accuracy of each classifier for the dataset Bio degeneration*

Classifier	Hierarchical accuracy
Jrip	**Q3 > Q2>** CHI> GRAE > **Q4 > Q1 >**I G> Relief
OneR	**Q1** > IG> CHI > GRAE >Relief>Q2>Q3>Q4
Ridor	Relief>**Q3>Q1**>CHI>GRAE>**Q4**>IG>Q2
J48	Relief>**Q3>Q4**>CHI>**Q2>Q1**>IG>GRAE
Simple cart	Relief>**Q3>Q4>Q2**>GRAE>**Q1**>IG>CHI
Naive Bayes	**Q2>Q3**>IG>CHI>**Q1>Q4**>Relief>GRAE
IBK	Relief>**Q3>Q4>Q2**>IG>CHI>**Q1**>GRAE

On Bio degenaration Q3, Q2 displayed boosted performance than all traditional methods with Jrip, NB. Q3, Q4 recorded enhanced accuracy than all existing methods except Relief with Ridor, J48, SC.

Dermatology

Table 11. *Accuracy of classifiers for the dataset Dermatology*

	Jrip	OneR	Ridor	J48	SC	NB	IBK
Q1	84.15	50.27	80.05	86.06	85.24	86.61	82.51
Q2	68.57	49.72	78.68	80.87	80.60	80.32	80.60
Q3	87.97	47.54	88.79	91.53	90.98	91.25	88.25
Q4	82.51	49.72	80.32	84.15	85.24	86.33	86.06
IG	59.83	49.72	75.13	75.95	74.86	74.86	75.95
CHI	68.03	48.90	68.57	68.57	68.57	69.12	69.12
GRAE	68.03	48.90	68.57	68.57	68.57	69.12	69.12

Relief	75.13	50.27	76.22	76.22	77.59	78.41	78.14

Hierarchical accuracy for the dataset Dermatology is given in table 12.

Table 12. *Hierarchical accuracy of each classifier for the dataset Dermatology*

Classifier	Hierarchical accuracy
Jrip	**Q3>Q1>Q4**>Relief>**Q2**>CHI>GRAE>IG
OneR	**Q1**>Relief>**Q2>Q4**>IG>CHI>GRAE>Q3
Ridor	**Q3>Q4>Q1>Q2**>Relief>IG>CHI>GRAE
J48	**Q3>Q1>Q4>Q2**>Relief>IG>CHI>GRAE
Simplecart	**Q3>Q1>Q4>Q2**>Relief>IG>CHI>GRAE
Naive Bayes	**Q3>Q1>Q4>Q2**>Relief>IG>CHI>GRAE
IBK	**Q3>Q4>Q1>Q2**>Relief>IG>CHI>GRAE

On Dermatology data set almost all clusters formed by proposed framework outperforms than all traditional feature selection methods.

Musk (Version 2)

Table 13. *Accuracy of classifiers for the dataset Musk (Version 2)*

	Jrip	OneR	Ridor	J48	SC	NB	IBK
Q1	76.26	60.08	73.52	80.88	80.04	72.68	82.35
Q2	77.31	65.54	73.94	79.83	79.20	66.38	83.82
Q3	71.42	58.61	72.05	82.35	74.15	73.94	84.66
Q4	73.73	59.66	73.94	83.19	75.00	69.11	81.51
IG	75.42	62.18	75.84	80.04	79.20	75.63	84.03
CHI	72.26	60.71	75.42	81.72	80.25	75.84	83.40
GRAE	73.52	65.75	71.21	77.73	74.36	58.61	82.98
Relief	70.79	61.97	72.05	73.73	74.57	71.63	77.94

Hierarchical accuracy for the dataset Musk (Version 2) is given in table 14.

Table 14. *Hierarchical accuracy of each classifier for the dataset Musk (Version 2)*

Classifier	Hierarchical accuracy
Jrip	**Q2>Q1**>IG>**Q4**>GRAE>CHI>**Q3**>Relief
OneR	GRAE>**Q2**>IG>Relief>CHI>Q1>Q4>Q3
Ridor	IG>CHI>**Q2>Q4>Q1>Q3**>Relief>GRAE
J48	**Q4>Q3**>CHI>**Q1**>IG>**Q2**>GRAE>Relief
Simplecart	CHI>**Q1>Q2**>IG>**Q4**>Relief>GRAE>Q3
Naive Bayes	CHI>IG>**Q3>Q1**>Relief>**Q4>Q2**>GRAE

IBK	**Q3**>IG>**Q2**>CHI>GRAE>**Q1>Q4**>Relief

Libras Movement

Table 15. *Accuracy of classifiers for the dataset Libras Movement*

	Jrip	OneR	Ridor	J48	SC	NB	IBK
Q1	53.61	21.11	59.72	63.05	59.72	60.83	81.66
Q2	54.72	19.16	57.5	66.11	60.00	55.55	85.27
Q3	51.38	21.38	56.38	68.61	65.27	58.88	82.5
Q4	55.27	23.33	60.00	66.38	68.05	62.5	85.83
IG	44.44	21.11	48.88	55	53.61	40	71.11
CHI	44.44	21.11	48.88	55	53.61	40	71.11
GRAE	35	22.22	42.77	47.77	46.94	35	60.27
Relief	45.55	21.11	49.16	60.55	54.72	37.22	69.72

Hierarchical accuracy for the dataset Libras Movement is given in table 16.

Table 16. *Hierarchical accuracy of each classifier for the dataset Libras Movement*

Classifier	Hierarchical accuracy
Jrip	**Q4>Q2>Q1>Q3**>Relief>IG>CHI>GRAE
OneR	**Q4**>GRAE>**Q3**>Q1>IG>CHI>Relief>Q2
Ridor	**Q4>Q1>Q2>Q3**>Relief>IG>CHI>GRAE
J48	**Q3>Q4>Q2>Q1**>Relief>IG>CHI>GRAE
Simplecart	**Q4>Q3>Q2>Q1**>Relief>IG>CHI>GRAE
Naive Bayes	**Q4>Q1>Q3>Q2**>IG>CHI>Relief>GRAE
IBK	**Q4>Q2>Q3>Q1**>IG>CHI>Relief>GRAE

On Libras movement data set almost all clusters formed by proposed framework outperforms than all traditional feature selection methods.

Connectionist Bench

Table 17. *Accuracy of classifiers for the dataset Connectionist Bench*

	Jrip	OneR	Ridor	J48	SC	NB	IBK
Q1	73.07	62.98	69.23	71.15	74.51	62.01	72.11
Q2	72.11	66.82	69.71	71.63	69.71	71.63	69.23
Q3	70.67	59.61	72.59	73.55	71.63	62.98	64.90
Q4	69.71	62.98	68.75	69.23	69.71	71.15	74.03
IG	74.03	61.53	69.71	70.19	72.59	70.19	67.30

CHI	74.03	61.53	69.71	70.19	72.59	70.19	67.30
GRAE	72.11	61.53	68.75	72.11	72.59	68.26	70.19
Relief	74.03	61.53	68.26	70.19	71.15	71.15	76.44

Hierarchical accuracy for the dataset Connectionist Bench is given in table 18.

Table 18. *Hierarchical accuracy of each classifier for the dataset Connectionist Bench*

Classifier	Hierarchical accuracy
Jrip	IG>CHI>Relief>**Q1>Q2**>GRAE>Q3>Q4
OneR	**Q2>Q1>Q4**>IG>CHI>GRAE>Relief>Q3
Ridor	**Q3>Q2**>IG>CHI>**Q1>Q4**>GRAE>Relief
J48	**Q3**>GRAE>**Q2>Q1**>IG>CHI>Relief>Q4
Simplecart	**Q1>**IG>CHI>GRAE>**Q3**>Relief>Q2>Q4
Naive Bayes	**Q2>Q4**>Relief>IG>CHI>GRAE>Q3>Q1
IBK	Relief>**Q4>Q1**>GRAE>**Q2**>IG>CHI>Q3

The strength of the proposed method is, it minimizes the searching space (Combinations). It can give new set of features which can give better performance than features derived by traditional methods. It can provide few more set of options to decide with group can be considered for better classification. This framework is tested with lower dimensions i.e with 3 groups and with higher dimensions i.e with 5 clusters. Those results can be found here Click here.

6. CONCLUSION

In this paper, a novel framework of subset of feature selection using Symmetric Uncertainty has been proposed. Proposed technique forms 4 sets (Quarters) of features without any repetitions. All the four quarters are analyzed using Tree, Rule, Bayes, lazy classifiers. Each quarter has 'N' number of features. To compare each quarter ,top 'N' features derived by existing filter based methods like IG, GRAE, CHI, Relief are considered and analyzed using same classifiers. To examine the proposed framework, 6 real-world datasets were considered. Experimental results show that, at least one quarter of features, sometimes more than one quarter of features performing better than existing techniques. With this we conclude that, instead of considering existing feature selection methods, Quarter Feature selection (QFS) can also be considered depending on classifier and dataset used. It is also suggested that, instead of considering Symmetric Uncertainty as a key criteria, IG or CHI or GRAE or Relief can also be used for forming subset of feature, which is our future work.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Rahman, H., " Data mining applications for empowering knowledge societies", IGI Global,1(1):25-32, (2008).

[2] Chandrashekar, G., Sahin, F., "A survey on feature selection methods", Computers & Electrical Engineering, 40(1): 16-28, (2014).

[3] Yu, L., Liu, H., " Feature selection for high-dimensional data: A fast correlation-based filter solution", In Proceedings of the 20th international conference on machine learning, 20(1): 856-863, (2003).

[4] Cui, Y., Jin, J. S., Zhang, S., Luo, S., & Tian, Q., " Correlation-based feature selection and regression", In Pacific-Rim Conference on Multimedia,1(1): 25-35, (2010).

[5] Sun, J., Liu, J., & Wei, X., "Feature selection algorithm based on SVM" , In 35th Chinese Control Conference (CCC),35(1): 4113-4116, (2016).

[6] Roffo, G., Melzi, S., & Cristani, M., "Infinite feature selection", In Proceedings of the IEEE International Conference on Computer Vision,1(1): 4202-4210, (2015).

[7] Ali, S . I., Shahzad, W., "A feature subset selection method based on symmetric uncertainty and ant colony optimization", In Emerging Technologies (ICET),26(1): 1-6,(2012).

[8] Kannan, S. S., Ramaraj, N., "A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm", Knowledge-Based Systems, 23(6): 580–585,(2010).

[9] Balkanli, E., Zincir-Heywood, A. N., & Heywood, M. I., "Feature selection for robust backscatter DDoS detection", Local Computer Networks Conference Workshops (LCN Workshops),10(1): 611–618, (2015).

[10] Christopher, A. A., & alias Balamurugan, S. A., "Feature selection techniques for prediction of warning level in aircraft accidents", In International Conference on Advanced Computing and Communication Systems (ICACCS),1(1): 1–6,(2013).

[11] Tan, D. W., Yeoh, W., Boo, Y. L., & Liew, S. Y., "The Impact of Feature Selection: A Data-Mining Application In Direct Marketing: The Impact Of Feature Selection", Intelligent Systems in Accounting, Finance and Management, 20(1): 23–38,(2013).

[12] Novaković, J., "Toward optimal feature selection using ranking methods and classification algorithms" ,Yugoslav Journal of Operations Research, 21(1): 119–135, (2011).

[13] Eesa, A. S., Orman, Z., & Brifcani, A. M. A. , "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems", Expert Systems with Applications, 42(5): 2670–2679, (2015).

[14] Hongbin, D., Xuyang, T., & Xue, Y., "Feature Selection Based on the Measurement of Correlation Information Entropy" ,Journal of Computer Research and Development, 53(8): 1684–1695, (2016).

[15] Bennasar, M., Hicks, Y., & Setchi, R., "Feature selection using Joint Mutual Information Maximisation", Expert Systems with Applications, 42(22): 8520–8532, (2015).

[16] Dong, H., Teng, X., Zhou, Y., & He, J.,"Feature subset selection using Dynamic Mixed Strategy", IEEE Congress on Evolutionary Computation in CES,7(1):672-679,(2015).

[17] Liu, Y., Tang, F., & Zeng, Z. , "Feature Selection Based on Dependency Margin", IEEE Transactions on Cybernetics, 45(6): 1209–1221, (2015).

[18] Wang, Z., Li, M., & Li, J., "A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure", Information Sciences, 307(1): 73–88,(2015).

[19] Soufan, O., Kleftogiannis, D., Kalnis, P., & Bajic, V. B. , "DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm," Plos One, 10(2): 79-88,(2015).

[20] Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F.j, "Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation", Plos One, 11(11): 66-77,(2016).

[21] Moayedikia, A., Ong, K. L., Boo, Y. L., Yeoh, W. G., & Jensen, R., "Feature selection for high dimensional imbalanced class data using harmony search", Engineering Applications of Artificial Intelligence, 57(1): 38-49,(2017).