Research Article GU J Sci 31(3): 775-787 (2018)

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Correlation Coefficient Based Feature Selection Framework Using Graph Construction

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

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

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

Article Info

Received: 13/09/2017
Accepted: 16/04/2018

Keywords

Classification
Correlation Coefficient
Feature Selection
Machine Learning
Symmetrical Uncertainty

Abstract

In machine learning, selecting the best features for classification is a critical issue. It is a necessary task to reduce the number of attributes/features existed in the initial feature space for achieving the outstanding classification accuracy, to minimize the computing power, and to reduce the memory size. In this present research, a novel methodology is proposed based on the concept of Symmetrical Uncertainty (SU) and Correlation Coefficient (CCE) by constructing the graph to select the reduced feature set. The recommended features by the proposed methodology are clubbed into finite number of groups (clusters) by measuring their CCE and considering the highest SU score of the feature. From each group, a feature which has maximum SU value is picked up and rest of the features in the same group are ignored. The proposed structure was inspected with ten (10) real world data sets available in the public domain. Experimental outcomes guarantees that the proposed method is recorded the better performance than most of the traditional filter based feature selection methods. The proposed method performed better than traditional methods such as Information Gain and Chi-Square on 70 % of the data sets. It is also produced better result than traditional Gain ration method on 80 % of the data sets and competing with traditional ReliefF approach on 50 % of the data sets. This methodology is assessed using Lazy, Tree Based, Naive Bayes, and Rule Based learners.

1. INTRODUCTION

Data Mining (DM) is a promising field of study in all sectors including marketing, education, health, protection, consultancy, investment, etc. DM is holding many intelligent methods such as association rule mining, classification, regression, clustering for heterogeneous reasons. DM is a valuable analytic knowledge-based method to procure the more intuition of data for productive decision making. Basically, DM episode includes data composing from various sources, pre-processing the collected data, applying various data mining techniques, review the results, and finally visualization for better understanding. There are distinct resources for composing the data set. Composed data set required to be varnished for enhanced results, as it contains noisy (clamorous), imbalanced labels, missing labels, missing values, and high dimensional (more number) features in pre-processing phase. After the pre-processing step, based on the data set gathered and type of the problem statement, analytical methods will be applied to build the model, then results will be reviewed and visualized in various forms to find more interesting patterns.

This present research focused on the classification technique of data mining and high dimensional issue of pre-processing. A data set which contains more number of features or attributes is called as high

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

dimensional data set. This type of data set may cause the different complications to the learning model such as creating confusion, improper prediction, bias towards one class, etc. Generally, all the attributes in the primary feature space may not be productive. Some attributes may be noisy and replicated. These features do not deliver any additional insights, instead there is a possibility of producing the uncertainty to the model created and also it heads to degrade the efficacy of model. High dimensional data set need an extra processing power and memory. This type of issues can be resolved by feature selection (FS) or feature reduction techniques.

The fundamental thought of FS is to suggest the strongestfeatures [1]. Noisy and redundant features are useless and they need to be truncated in FS process. If a classification performance can't be increased with the addition of a feature, we can say, the feature is useless. But, key point is how to know those redundant and noisy features? In existing study, few methods are available to give the solutions to this question.

Filter and Wrapper are frequently used FS methods. Filter method assigns the score (weight) to each feature based on the information worth it holds. Based on the score assigned to the feature, a rank/position will be awarded, there by top 'K' positioned features can be taken for model generation [2]. Information Gain (IG), ReliefF (Rel), Gain Ratio (GR) Chi-Square (Chi) are some of the filter-based algorithms [3]. In different approach, wrapper method is generally time consuming process, as it need to consider some searching (best first, greedy stepwise, etc.) criteria and learning algorithms for nominating the best candidate feature set [4]. In this process, features which are producing an insignificant accuracy by learning algorithms will be discarded from the primary data set. In this current study, we tried to produce the best candidate feature set that can mount the classification performance using CCE and SU by grouping the features into multiple clusters.

Euclidean distance is frequently applied measurement as a similarity metric in the clustering analysis. But, to know the relationship between two random variables, CCE can be used. In this study, instead of Euclidean distance, we selected CCE to measure the relationship between two features. According to mutual information concept, if two features are mutually dependent, we can select only one of them for classification as they share common properties and give the almost equal result.

The procedure to construct the cluster of features and nominating the best feature from each cluster is described in the methodology section. The proposed methodology is tested with Tree Based, Rule Based, Lazy, and Bayes learners over ten popular data sets available in public domain. In second portion, brief existing literature review and related work is discussed. In third section, proposed framework with an example is discussed. In section four, experimental procedure is illustrated. Result analysis with discussion is given in fifth section.

2. LITERATURE

To get the reduced feature set for the classification problem, various feature selection methods (Filter and Wrapper) applied by many researchers in order to improve the classifier's accuracy. Particle swarm optimization feature selection algorithm is proposed for text clustering [5]. Wrapper based method for selecting the best genes from the microarray data set is proposed with markov blanket approach [6]. The authors achieved effective results with their approach. Embedded feature selection is another approach in addition to filter and wrapper which is also widely applied for classification problems. SVM-RFE, LASSO,Random-Forest are some of the popular embedded FS methods. SVM-RFE is applied for cancer classification [7]. LASSO is proposed by the researchers to draw the minimum features for effective results using stability arguments [8]. Random-Forest method is applied for land coverage classification [9]. Authors of [10] applied filter and wrapper techniques for solving the protein disordered region prediction issue. The considered data set initially has 440 attributes in it. Initially, IG and F-Score is employed over the dataset, later wrapper method is applied to find the better classification performance.
In our present study, we measured the proposed framework with some of the popular traditional feature selection techniques (IG, Chi, GR, Rel). The idea of IG is on the basis of information theory. It examines the association between features and classes for removing the redundant attributes, and the extremely

independent attributes with the class label. IG based feature selection method is applied over kidney disease and voting dataset by the authors [11]. The authors of [12] proposed a FS method on the basis of mutual information. MaxDep concept was proposed by the researcher of [13]. It computes the subset statistical dependency with the target class label. This method aims to select 'n' features that jointly have the maximum dependency with the target class.

For integrated high dimensional protein data, Maximum Correlation Information-Recursive Feature Elimination (MCI-RFE) method is proposed by the researchers [14]. In MCI-RFE method, the significance of every attribute is calculated by maximizing the correlation information (MCI). Then, MCI is merged with recursive feature elimination (RFE) to generate the strong subset of feature. MCI-RFE is highly competitive with SVM-RFE, ReliefF-RFE and Random Forest.

FS has become an alternative task for many researchers to address the critical issues linked with more number of attributes in the field of pattern recognition to secure the better results [15]. Correlation-Based Selection (CFS) approach is applied by the researchers for various purposes. CFS is applied to predict the electricity demand in Australia [16]. They applied neural networks, tree based algorithms over two years of time series load data. Authors proposed FAST algorithm based on the SU and CCE to get the optimal subset. FAST is a clustering based algorithm which works in two steps. In initial step, graph theory clustering method is applied to generate the attributes into clusters. In the second step, prims algorithm is applied to choose the most optimal attributes [17].

This present study also mainly on the basis of two statistical techniques used in the FAST algorithm. Those are: CCE and SU. CCE is considered to know the relationship (Weight) between two variables and also to form the cluster of features. SU is considered to fix the minimum (threshold) value of weight and also to know the best feature in each cluster. The procedure to measure the weight of feature is discussed in next section. Out of 'n' observations, CCE of two random variables X and Y can be derived as below equation

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (1)$$

If'r' value is close to 1, we can say X and Y are strongly depending. If 'r' $\frac{n}{2}$ value is 0, then we can say there is no relationship between X and Y. In our research, we considered positive 'r' value to measure the weight of the feature.
Symmetrical Uncertainty (SU) can defined as below

$$SU = 2*IG/(H(X)+H(Y)) \quad (2)$$

Where IG is Information Gain, H(X) is the entropy of a discrete random variable X and H(Y) is the entropy of a discrete random variable Y. If the prior probability of each element of X is p(x), then H(X) can be calculated by

$$H(X) = -\int p(x)\log (p(x))\, \partial x \quad (3)$$

If the prior probability of each element of Y is p(y), then H(Y) can be calculated by

$$H(Y) = -\int p(y)\log (p(y))\, \partial y \quad (4)$$

IG can be defined as information that is derived by drawing the score of the feature, which is the subtraction of the entropy distribution before the split and distribution after the split. A value 1 of SU(X, Y) indicate that knowledge of the object value strongly represent the values of other and the SU(X, Y) value 0 indicate the independence of X and Y. In this paper, we also deal with continuous features by

normalized in proper discrete form. In the next section, proposed methodology is discussed with an example

3. PROPOSED METHODOLOGY

The target of the proposed framework is to derive the best candidate subset which can maximize the classification accuracy. Proposed framework is as per the below algorithmic steps.

Algorithm:

1.Derive the SU score of every feature and place it in it's descending order of SU score
2. Select the mid feature's SU score as Threshold (T).
3. Construct the Correlation Coefficient Symmetrical matrix (CCE(X_i,Y_i)) of primary data set .
4. Convert the CCE(X_i,Y_i) matrix to weighted binary matrix (WB) with the following steps.
for(i=*1 to n*)
for(j=*1 to n*)
if(CCE(X_i,Y_i) >T)
WB(X_i,Y_i)=1
else
WB(X_i,Y_i)=0
End
End
5. Construct the graph (G) such that, Edge between two nodes will be existed iff CCE(X,Y)==1. Then, define the degree or weight of a node (feature) W(F_i).
for(i=*1 to n*)
for(j=*1 to n*)
W(F_i)=$\sum$ WB(X_i,Y_i)
End
End
6. Group the node which are having same degree (W(F))
Cluster$_i$={F_{i1}, F_{i2},...F_{ik}} /* i is the cluster id, increment i by 1 until all features are formed */
7. Choose the best node (feature which has maximum SU value) from each cluster and form the final candidate subset
for(i=*1* to last cluster)
F_i= MAX SU(cluster$_i$)
Candidate Feature set (CFS)<- F_i
End

Definition:

Degree of a node: With how many number of features it is correlated.
As per the presented algorithmic steps, an example to construct the best candidate feature subset is given below.

Example:

Consider there are ten features (a, b, c, d, e, f, g, h ,i, j) in primary data set.
1. SU score of every feature is given in Table.1 (As per step 1)

Table 1 . *SU score of all features in primary data set*

SU	Rank	Fid
.19	1	J
.19	2	H
.19	3	G
.18	4	I
.15	5	B
.09	6	A
.07	7	D
.06	8	C
.06	9	E
.02	10	F

2. Threshold (T) =.15 , as 'b' is the middle feature . (As per step 2)
3. CCE(X_i,Y_i) matrix of the primary data set is given in below Table 2. (As per setp3)

Table 2. *CCE(X_i,Y_i) matrix*

Feature Id	A	B	C	D	E	F	G	H	I	J
A	1	0.21	0.25	0.22	-0.23	0.01	-0.23	0.06	0.07	0.09
B	0.21	1	-0.15	-0.9	0.03	0.04	-0.09	0.05	0.04	0.07
C	0.25	-0.15	1	0.08	0.05	0.24	-0.09	0.26	0.01	0.13
D	0.22	-0.9	0.08	1	0.27	0.06	0.2	-0.21	0.02	0.03
E	-0.23	0.03	0.05	0.27	1	0.03	0.06	-0.1	0.06	0.17
F	0.01	0.04	0.24	0.06	0.03	1	0.04	0.19	0.03	-0.08
G	-0.23	-0.09	-0.09	0.2	0.06	0.04	1	0.03	0.19	0.02
H	0.06	0.05	0.26	-0.21	-0.1	0.19	0.03	1	0.04	0.02
I	0.07	0.04	0.01	0.02	0.06	0.03	0.19	0.04	1	-0.06
J	0.09	0.07	0.13	0.03	0.17	-0.08	0.02	0.02	-0.06	1

4. Convert the CCE(X_i,Y_i) matrix to weighted binary matrix (WB), refer Table 3 (As per step 4)

Table 3. *Weighted binary matrix*

Feature Id	A	B	C	D	E	F	G	H	I	J
A	1	1	1	1	0	0	0	0	0	0
B	1	1	0	0	0	0	0	0	0	0
C	1	0	1	0	0	1	0	1	0	0
D	1	0	0	1	1	0	1	0	0	0
E	0	0	0	1	1	0	0	0	0	1
F	0	0	1	0	0	1	0	1	0	0
G	0	0	0	1	0	0	1	0	1	0
H	0	0	1	0	0	1	0	1	0	0
I	0	0	0	0	0	0	1	0	1	0
J	0	0	0	0	1	0	0	0	0	1

5. Construct the graph and define the degree of each node. Figure 1 shows the graph constructed for the Table 3 matrix. Here self-node count also considered which is optional. The same can be demonstrated without constructing the graph also. i.e. calculate the sum of every column or row and define it as its feature weight.

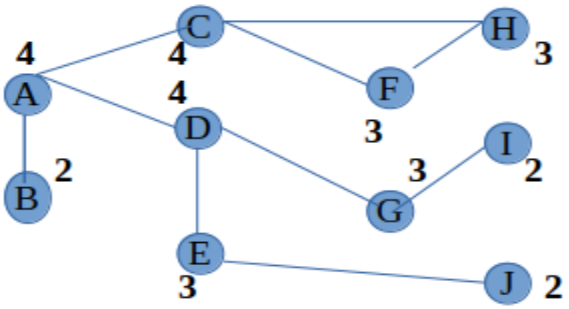

Figure 1. Graph with it's Degree

6. Form the clusters and select the best feature in each cluster. Sorted list of feature and weight is given in below Table 4. (As per step 6)

Table 4. *Sorted list of feature and weight*

Cluster Id	Weight	FID/ Node	Selected Features From each Cluster
1	2	I	J
	2	J	
	2	B	
2	3	E	H
	3	F	

	3	G	
	3	H	
3	4	A	A
	4	C	
	4	D	

6. Form the final candidate feature set (CFS)
CFS= {J, H, A}

4. EXPERIMENT

To examine the proposed framework, ten (10) real-time benchmark data sets are taken into consideration. The list of data sets and their brief description is given in Table 5.

Table 5. *Data sets description*

Data set ID	Name of the Data Set	# Instances	# Features	# Class
1	Ionosphere	351	34	2
2	Dermatology	366	34	6
3	Biodegradation	1055	41	2
4	Cardiotocography	2126	22	3
5	Lung Cancer	33	56	3
6	Libras Movement	360	90	15
7	Connectionist Bench(Sonar)	208	60	2
8	Spambase	4601	57	2
9	Breast Cancer(WDBC)	569	30	2
10	Musk (V 2)	476	166	2

The proposed framework is evaluated using popular open source machine learning tool WEKA. We applied 10-fold cross validation for all the data sets. After employing this framework reduced number of features obtained. The number of features formed as a result of proposed framework is given in Table 6.

Table 6*. Number of features formed by proposed method*

Data set ID	Name of the Data Set	# Features in Original Data set	#Features formed by Proposed Method (S)
1	Ionosphere	34	13
2	Dermatology	34	13
3	Biodegradation	41	23
4	Cardiotocography	22	12
5	Lung Cancer	56	15

6	Libras Movement	90	21
7	Connectionist Bench(Sonar)	60	28
8	Spambase	57	16
9	Breast Cancer(WDBC)	30	15
10	Musk (V 2)	166	54

To compute the quality of the proposed framework, Top 'S' features extracted by traditional approaches are selected. For calculating the CCE value between the features R statistical programming is used. SU and performance of classifiers with the derived features set is evaluated using WEKA.

5. RESULTS AND DISCUSSION

In this section, classification performance of proposed and traditional methods using Jrip, Ridor, J48, Simple Cart, IBK, Naive Bayes classifiers with different data sets are presented with brief discussion. Result analysis over Ionosphere dataset is given in Table 7a.

Table 7a. *Result Analysis Over Ionosphere dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	89.74	91.45	87.46	91.73	86.89	84.9	88.7
Chi	90.88	90.31	88.88	91.16	88.03	88.03	89.55
GR	89.45	90.02	88.6	90.59	90.31	86.03	89.17
Rel	91.16	91.16	91.45	94.01	89.17	90.02	91.16
Proposed	90.31	88.6	90.02	92.02	88.88	89.17	89.83

Initially Ionosphere data set has 34 features. After applying the proposed method 13 best features are derived. Over Ionosphere data set the proposed method outruns than traditional methods such as IG and GR with Ridor classifier. With J48, Simple cart and NB classifiers the proposed method has produced the best performance than traditional methods like IG, Chi, GR but recorded less accuracy than traditional ReliefF method. With Instance based learner (IBK), the proposed method performed better than IG and Chi. Overall average performance of the proposed method has demonstrated good accuracy than traditional IG, Chi, GR, but not than Relief. Result analysis over Dermatology dataset is given in Table 7b.

Table 7b. *Result Analysis Over Dermatology dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	82.24	79.23	80.32	80.87	82.51	83.33	81.42
Chi	83.33	83.6	83.33	83.06	85.24	84.15	83.79
GR	83.33	83.6	83.33	83.06	85.24	84.15	83.79
Rel	79.23	76.77	77.32	77.86	81.14	81.42	78.96
Proposed	91.25	89.07	89.89	91.8	92.34	94.53	91.48

Initially Dermatology dataset has 34 features. After applying the proposed method 13 best features are derived. With the 13 best features the proposed method has produced the highest accuracy than all

traditional methods with the all classifiers. Authors of [18] proposed the clustering based feature selection using SU. They classified with Jrip and secured 86% of accuracy With IBK classifier 87% accuracy is achieved. Result analysis over Biodegradation dataset is given in Table 7c.

Table 7c. *Result Analysis Over Biodegradation dataset*

	Ridor	Jrip	SC	J48	NB	IBK	Avg
IG	81.51	82.27	83.79	83.79	73.64	82.08	81.18
Chi	81.32	82.18	83.31	83.69	73.45	82.27	81.04
GR	81.51	81.61	82.18	83.69	74.02	83.31	81.05
Rel	81.42	82.18	83.22	84.26	75.16	83.12	81.56
Proposed	80	81.99	83.5	83.79	74.5	82.84	81.1

Initially Biodegradation dataset has 41 features. After applying the proposed method 23 best features are derived. With these 23 best features the proposed method has produced the highest accuracy than traditional GR methods with the Jrip classifiers. Proposed method recorded better accuracy than traditional Chi, GR, Rel methods with Simple cart. Also, with NB classifier, proposed approach outperforms than traditional IG, Chi, GR methods. Result analysis over Cardiotocography dataset is given in Table 7d.

Table 7d. *Result Analysis Over Cardiotocography dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	98.4	98.82	98.63	98.58	97.83	88.33	96.77
Chi	98.11	98.91	98.54	98.82	97.78	89.46	96.94
GR	98.11	98.44	98.49	98.82	97.69	90.21	96.96
Rel	98.44	98.73	98.63	98.63	96.94	90.54	96.99
Proposed	98.49	98.73	98.54	98.63	97.22	89.93	96.92

Initially Cardiotocography dataset has 22 features. After applying the proposed method 12 best features are derived. The average performance of the proposed method is little improved than traditional IG method. Also, it is competing with all other existed methods. Result analysis over Lung Cancer dataset is given in Table 7e.

Table 7e. *Result Analysis Over Lung Cancer dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	53.12	59.37	62.5	59.37	56.25	65.62	59.37
Chi	59.37	53.12	62.5	62.5	62.5	62.5	60.42
GR	59.37	53.12	62.5	62.5	62.5	62.5	60.42
Rel	68.75	53.12	56.25	56.25	68.75	71.87	62.5
Proposed	56.25	56.25	68.75	59.37	50	71.87	60.42

Initially Lung Cancer dataset has 56 features. After applying the proposed method 15 best features are derived. The average performance of the proposed method is competing with the all other traditional methods except Rel . Especially, with NB and SC classifier proposed approach is recorded best accuracy than existing methods. Result analysis over Libras Movement dataset is given in Table 7f.

Table 7f. *Result Analysis Over Libras Movement dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	46.38	44.16	55.55	56.66	73.05	44.16	53.33
Chi	48.88	45.55	54.16	55.83	72.22	44.16	53.47
GR	47.77	43.61	52.5	57.5	72.22	41.38	52.5
Rel	49.72	48.61	58.88	60	76.38	44.16	56.29
Proposed	57.22	51.94	60.55	65.83	84.44	60.83	63.47

Initially Libras Movement dataset has 90 features. After applying the proposed method 21 best features are derived. With the 21 best features, the proposed method has produced the highest accuracy than all traditional methods with the all classifiers. Result analysis over Connectionist Bench(Sonar) dataset is given in Table 7g.

Table 7g. *Result Analysis Over Connectionist Bench(Sonar) dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	72.11	77.4	72.59	74.51	87.98	70.19	75.8
Chi	72.11	77.4	72.59	74.51	87.98	70.19	75.8
GR	72.11	77.4	72.59	74.51	87.98	70.19	75.8
Rel	76.44	75.96	73.55	74.03	87.5	69.23	76.12
Proposed	74.03	81.25	78.36	76.92	85.09	72.59	78.04

Initially Sonar dataset has 60 features. After applying the proposed method 28 best features are derived. With the 28 best features, the proposed method has produced the highest accuracy than all traditional methods with the all classifiers except with the IBK. Researchers applied clustering based feature selection on the same dataset and classified with various classifiers[19]. The proposed method is competing with their approach. Result analysis over Spambase dataset is given in Table 7h.

Table 7h. *Result Analysis Over Spambase dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	91.15	91.98	91.87	92.91	89.48	88.24	90.94
Chi	91	91.54	91.76	93.02	89.89	86.06	90.55
GR	89	90.06	90.28	90.61	88.39	70.68	86.5
Rel	85.41	86.3	87.58	87.37	85.98	68.31	83.49
Proposed	90.15	90.52	91.45	91.48	88.93	75.94	88.08

Initially Spambase dataset has 57 features. After applying the proposed method 16 best features are derived. The average performance of the proposed method is little improved than traditional GR and Rel. Result analysis over Breast Cancer(WDBC) dataset is given in Table 7i.

Table 7i. *Result Analysis Over Breast Cancer(WDBC) dataset*

	Ridor	Jrip	SC	J48	IBK	NB	Avg
IG	92.26	91.91	92.26	92.79	92.97	92.44	92.44
Chi	92.26	92	92.26	92.79	92.97	92.44	92.45

GR	92.26	92	92.26	92.79	92.97	92.44	92.45
Rel	94.55	93.84	92.97	93.67	96.13	94.55	94.29
Proposed	94.9	93.49	93.84	94.37	95.07	93.32	94.17

Initially Breast Cancer dataset has 30 features. After applying the proposed method 15 best features are derived. The average performance of the proposed method is better than traditional IG, GR and Chi. WithRidor, Jrip, SC and J48 classifiers, the proposed method recorded little improved classification accuracy than features derived by all traditional methods. In the research article [18], authors applied feature selection on the same data set and applied various classifiers. Result analysis over Musk (V 2) dataset is given in Table 7j.

Table 7j. *Result Analysis Over Musk (V 2) dataset*

	Ridor	Jrip	SC	J48	NB	IBK	Avg
IG	72.89	78.99	79.41	83.61	75.84	84.87	79.27
Chi	73.31	74.36	80.25	82.35	76.05	85.71	78.67
GR	75.42	74.78	77.94	80.67	67.43	80.61	76.14
Rel	74.78	76.26	81.09	82.35	72.05	82.77	78.22
Proposed	73.52	75.84	81.72	81.09	74.78	85.5	78.74

Initially Musk (V 2) dataset has 166 features. After applying the proposed method 54 best features are derived. The average performance of the proposed method is little better than traditional Chi, GR and Rel. The average competences of our method with the existing methods is given in below Table 8 With Win(W), Draw (D), and Loss(L).

Table 8. *Average competences of proposed method with existing methods*

	Data set ids			Out of 10 Data sets		
Method	Win	Draw	Loss	Win%	Draw%	Loss %
IG	1,2,4,5,6,7,9	nil	3,8,10	70	nil	30
Chi	1,2,3,6,7,9,10	5	4,8	70	10	20
Gr	1,2,3,6,7,8,9,10	5	4	80	10	10
Rel	2,6,7,8,10	nil	1,3,4,5,9	50	nil	50

From the Table 8 statistical analysis, the proposed method performed better than traditional IG, Chi methods on 70 % of the data sets. ReliefF performed better on 50 % of the data sets than proposed method. GR performed better on 80 % of the data sets than proposed method.

6. CONCLUSION

In this research, we have presented a feature selection framework to minimize the data set dimensionality by nominating the optimal features for enhancing the classification performance. For this proposed framework, two statistical approaches namely correlation coefficient and Symmetrical Uncertainty is considered to select the optimal features. The proposed framework was compared with four traditional filter based methods namely, Chi- Square (Chi), Information gain (IG). Grain Ratio (GR), and ReliefF (Rel). The proposed method was applied on ten different real time data sets which are available in public domain. Features derived by the proposed method is evaluated with six different classifiers namely, Jrip,

Ridor, J48, Simple cart, Naive Bayes, IBK. After careful investigation of classifiers accuracy on all the datasets, it is proved that the presented framework performed better than traditional IG and Chi on 7 data sets. It is also performed stronger than traditional GR on 8 data sets. It is also competing with ReliefF method on 5 data sets. The same technique can be implemented using Hadoop framework, which is our future work.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Liao, S.H., Chu, P.H., Hsiao, P.Y.,"Data mining techniques and applications–A decade review from 2000 to 2011", Expert systems with applications, 39(12):11303-11311,(2012).

[2] Jović, A., Brkić, K., Bogunović, N., "A review of feature selection methods with applications", 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),1(1): 1200-1205,(2015).

[3] Sharma, A., Dey, S., "A comparative study of feature selection and machine learning techniques for sentiment analysis", In Proceedings of the 2012 ACM research in applied computation symposium, 1(1):1-7,(2012).

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

[5] Abualigah, L. M., Khader, A. T., "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering", The Journal of Supercomputing, 73(11):4773-4795,(2017).

[6] Wang, A., An, N., Yang, J., Chen, G., Li, L.,Alterovitz, G.," Wrapper-based gene selection with Markov blanket", Computers in biology and medicine, 81(1):11-23,(2017).

[7] Duan, K. B., Rajapakse, J. C., Wang, H., Azuaje, F. ," Multiple SVM-RFE for gene selection in cancer classification with expression data", IEEE transactions on nanobioscience, 4(3):228-234,(2005).

[8] Thakurta, A. G., Smith, A.,"Differentially private feature selection via stability arguments, and the robustness of the lasso", In Conference on Learning Theory,1(1): 819-850,(2013).

[9] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M.,Rigol-Sanchez, J. P.,"An assessment of the effectiveness of a random forest classifier for land-cover classification", ISPRS Journal of Photogrammetry and Remote Sensing, 67(1): 93-104,(2012).

[10] Hsu, H. H., Hsieh, C. W., Lu, M. D. , "Hybrid feature selection by combining filters and wrappers", Expert Systems with Applications, 38(7): 8144-8150,(2011).

[11] Potharaju, S. P., Sreedevi, M.," A Novel Cluster of Feature Selection Method Based on Information Gain" , IJCTA, 10(14): 9-16,(2017).

[12] Maji, P., Garai, P., "On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance", Applied Soft Computing, 13(9): 3968-3980,(2013).

[13] Ding, C., Peng, H.," Minimum redundancy feature selection from microarray gene expression data", Journal of bioinformatics and computational biology, 3(2):185-205,(2005).

[14] Yuan, M., Yang, Z., Huang, G., Ji, G.," Feature selection by maximizing correlation information for integrated high-dimensional protein data", Pattern Recognition Letters, 92(1), 17-24,(2017).

[15] Partila, P., Voznak, M., Tovarek, J.," Pattern recognition methods and features selection for speech emotion recognition system", The Scientific World Journal, 15(1):1-7,(2015).

[16] Koprinska, I., Rana, M., Agelidis, V. G.," Correlation and instance based feature selection for electricity load forecasting", Knowledge-Based Systems, 82(1): 29-40,(2015).

[17] Mudaliar, P. U., Patil, T. A., Thete, S. S., Moholkar, K. P.,"A Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data", International journal of emerging trend in engineering and basic science, 2(1):494-499,(2015).

[18] Potharaju, S. P.,Sreedevi, M.,"A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Datasets", Journal of Engineering Science & Technology Review, 10(6):154-162,(2017).

[19] Potharaju, S. P., Sreedevi, M.,"A Novel Subset Feature Selection Framework for Increasing the Classification Performance of SONAR Targets", Procedia Computer Science, 125(1): 902-909,(2018).