DOI: 10.21917/ijsc.2023.0404

SEF-USIEF FEATURE SELECTOR: AN APPROACH TO SELECT EFFECTIVE FEATURES AND UNSELECT INEFFECTIVE FEATURES

Subitha Sivakumar[1], Sivakumar Venkataraman[2] and Asherl Bwatiramba[3]

[1]Quality Assurance and Program Review, New Era College, Botswana

[2,3]Faculty of Health and Education, Botho University, Botswana

Abstract

Feature selection is a method in Data mining to reduce the features from the original dataset by removing the noisy features from the dataset to improve the performance of the classifiers in terms of improving the prediction accuracy. Classification tasks often include, among the large number of features to be processed in the datasets, many irrelevant and redundant ones, which can even decrease the efficiency of classifiers. Feature Selection (FS) is the most common pre-processing technique utilized to overcome the drawbacks of the high dimensionality of datasets. The proposed SEF-USIEF Feature Selector: An approach to Select Effective Features and Unselect Ineffective Feature methods increases the performance of the classification methods by eliminating the irrelevant features from the dataset. This proposed SEF-USIEF method is implemented for the numerical datasets. This method is derived from the ward Minimum Variance cluster method and in this experiment Minimum Variance is used as the feature selector method. The numerical datasets are obtained from the UCI repository and WebKB repository. The results obtained by the proposed SEF-USIEF method are compared with the existing feature selection methods to analyze whether the features are reduced by the SEF-USIEF method or not. Then, features are given as input to the classifiers to check whether the classifier performance is increased or not. Based on the compared analysis of the results, the SEF-USIEF method proved that the performance of the classifiers is increased, and also the selected features are reduced when compared to the existing feature selection methods.

Keywords:

Feature Selection, Data Mining, Classifiers, Minimum Variance Cluster Method, Minimum Variance Feature Selector Method

1. INTRODUCTION

The classification problem is a major challenge in Data mining due to the increase of features in the datasets. Due to the diversity of the features, the classification performances are affected, and the accuracy of the prediction is affected. Many researchers have proved that the reduction in features improves classification performance. In Data mining, feature selection is a method that helps to reduce the features by selecting the appropriate features from the dataset.

Data mining is a technique used to extract related features from a large volume of datasets [17]. The data mining tool is used to convert uncooked information to valuable information for analysis. Data mining is used to find the irregularities, patterns, and relations among the features from the huge dataset in predicting the results. Data mining can be used in different data like statistical data, text data, media data, website data, and audio and video data.

Feature Selection is an important method in data mining that can be used to reduce the original features from the datasets by eliminating the unwanted features [7]. Selecting the optimal features plays an important role in increasing the prediction accuracy and the performance of the classifiers. It is important to remove the irrelevant and redundant features from the original dataset because it affects the performance of the classifiers [15] and [14].

Classifiers are used to evaluate the accuracy of the Supervised Learning Algorithms (SLA) in terms of performance. The SLA has two groups feature subset-based method and feature ranking methods like filter, wrapper, embedded, and hybrid methods [5]. The filter method analyses the relations among the features and the class label and does not include the learning algorithms for the evaluations. The wrapper method analyses the performance of the learning algorithm in terms of classification during the evaluation. The main goal is to have a smaller number of features as subset features and to improve the accuracy of the classifiers. An embedded method is implemented with the SLA as a feature selection method. The embedded method is cheaper than the wrapper method and the filter method is cheaper than the wrapper method but embedded in the high generality. The hybrid method combines two methods which can reduce the original features and can improve the classification accuracy. The hybrid method can combine the feature ranking method with the feature selection method to have a selected feature subset.

This study proposed a SEF-USIEF Feature Selector: An approach to Select Effective Features and Unselect Ineffective Feature Methods to select the optimal features from the dataset and to improve the classifiers accuracy. The SEF-USIEF method is a statistical method that can be used for numerical datasets. This method creates a cluster of features as two groups with predictive features. The first groups cluster with the predictive features without the class label and the second cluster group with predictive feature along with the class label. The features in the first group without the class label are considered irrelevant features and are not used to evaluate the performance of the classifiers. The features in the second cluster group with the class label are considered the relevant features and these features are used to evaluate the classifiers.

2. LITERATURE REVIEW

Researchers have developed several feature selectors which help to reduce the size of the original features by selecting the relevant features in different fields like Healthcare, Engineering and Science, E-commerce, and more.

A new feature selection method with the Genetic algorithm and with a neural network classifier to obtain the relevant features called as Alzheimer's disease (AD) recognition method implemented by [19]. This method is used to obtain the leading features from the dataset and the obtained features are used as the input for the network. This method proved and concluded to be used for primary detections of Alzheimer's disease.

A novel feature selection method to reduce the features and deals with high-dimensional health dataset suggested by [22]. The authors developed a new feature section method for text classification based on independent space search. First, the RDTFD method is used separate the features into two subset-independent features. Second, the PSO method is used to select the best features to improve the text classifier performance. The proposed method proved that the subset features are selected with optimal features which showed better classification performances.

A medical dataset classifier by implementing the decision tree concept to select the features and to improve the classification accuracy when compared to the existing algorithms [3]. This study includes decision trees like ID3, C4.5, and C5.0. This method exposed good accuracy and effective result. In comparison to other classifiers, the C5.0 classifier produces efficient classification in less time.

The PCA method [8] as an adaptive classification approach to reduce the size of the dataset by using the data analysis method. This method uses the Eigen matrix and Eigenvectors in the process of reducing the original features. This method proves that the original size of the dataset is reduced by selecting only the right features. Further investigation reveals that feature extraction by PCA is advantageous, particularly for data with several balanced classes.

A novel feature selector technique designed by [2] for the one-way ANOVA F Test on mail-spam clarification in terms of improving the Support Vector Machine (SVM) limitations, reducing the computational complexity, and improving the classification accuracy of the classifiers. The author conducted this experiment by using the Spam-base English Database. The research outcomes proved the improved Support Vector Machine (SVM) named Feature Selector Support Vector Machine (FESVM) suggestively outperforms the Support Vector Machine (SVM) results.

A novel feature selector method that combines the filter methods for classification issues [18]. By using this method, a subset of features is selected by using several filter methods, and then the exhaustive search method is implemented to obtain the best features. The suggested method is evaluated using widely used datasets from the UCI repository as well as two datasets from the industrial setting. The proposed novel hybrid method is tested with various datasets and the results proved that this method selects the best features for classification accuracies.

A new feature selector technique founded using the neural network and machine learning to select the relevant features was proposed by [12]. This approach uses the new weightage methods for the selected features which are given as input in the neural network to obtain the relevant features from the databases. These new feature selector methods show the best features for classification accuracy are selected as subset features.

A new feature selection approach by using H-Ratio where to find the relevant features from the database and to improve the performance of the classifiers [9]. This H-Ratio method is developed to improve the nominal datasets for the classifier founded by using formal concept analyses. The results comparison shows that the new feature selector method result is better than the results of the two earlier research works results on nominal classifiers based on Formals Concepts Analysis.

An ensemble feature selection method suggested by [16] to improve the classification performance Chronic Obstructive Pulmonary Disease (COPD) images dataset. An ensemble feature selection-based classification model was built in this research, employing image features collected from COPD patients' CT scan pictures, to categorize illness into "Severity level" and "Normal level" categories, signifying their likelihood of developing COPD. Five classification methods and three state-of-the-art ensemble classifiers are applied to the COPD dataset and the classification performance is evaluated. The evaluation result proved that the proposed ensemble feature selection method has a high accuracy value for Chronic Obstructive Pulmonary Disease data classifications.

A novel hybrid feature selector method developed by [20], which is an improved feature selector method by using the Filter and Wrapper method. In this method, firstly the Filter method is used to select the features by using the ranking criteria, secondly, the wrapper method is used to obtain the relevant features as subset features. Comparison results of the novel hybrid feature selector method show a better output for the classification performance by selecting the best features.

A novel hybrid feature selection method as "signature" founded by using SVM and t-statistic in selecting the best features to improve the performance of the classification algorithms for the microarray dataset got from Gene Expression Omnibus [13]. Authors discovered that "signature" properly predicts the signature of four of the six microarray data sets, but the other techniques predict the signature of fewer data sets. As a result, "signature" outperforms similar algorithms in detecting differentially significant features in microarray data sets.

A novel, simple and effective feature selection method used to select the relevant features based on the class-wise data in each class. For the selected features from Stanford, the Twitter dataset are evaluated [6]. On the Stanford Twitter dataset, the proposed feature selection approach surpasses existing feature selection methods in terms of classification accuracy. Similarly, in most of the feature subsets on the Ravikiran Janardhana dataset, the suggested strategy outperforms existing feature selection methods in terms of classification accuracy. The novel feature selection method proved that the classification performance of the classifiers is increased.

A novel feature selector method based developed by [11] on a Genetic Algorithm (GA) which has three stages. In the first stage, GA-based community exposure is done, and similarities are found between the features. In the second stage, the features are grouped as a cluster based on the relation. In the third stage, the related features are picked by using the GA. The authors evaluated the performance of this method by checking the classifiers accuracy.

A new, adaptive, and hybrid feature selection method to produce a more generic answer by combining and utilizing many separate approaches. Several state-of-the-art feature selection approaches are extensively provided with examples of their applications, and an extensive evaluation is carried out to quantify and compare their performance with the suggested methodology [23]. While the separate feature selection approaches may perform well in a wide range of test circumstances, the combined algorithm consistently gives a significantly superior result.

A new hybrid feature selection method [10] that enhances the performance of a collection of classifiers by making use of wrapper subset evaluation at a cheaper cost. This new method uses the filter methods combined with the feature subset evolution method to obtain the finest features. The experimentation is done with different types of UCI repository datasets and observed that the results are better in the classification performance.

A hybrid feature selection method called the IWSS method and Shuffled Frog Leaping Algorithm (SFLA) [4]. This method is used to select the optimal features and this method uses the two phases like filter method and the wrapper method. The hybrid method is experimented with by using the gene expression dataset. The experimental findings show that, in comparison to similar approaches, the suggested strategy achieves a more compact collection of characteristics while maintaining excellent accuracy.

A hybrid feature selector method [21] by using machine learning methods and knowledge graphical technologies. This hybrid uses the supervised, unsupervised, and knowledge graph to model from the point of view of data and text features. The text-based feature weights were created using knowledge graph technology, and the weight sets were merged to obtain a feature matrix with high explanatory qualities that fulfill both the data and text features. The results of the experiments show that the strategy described in this research has good effects and is easy to understand.

The review of the literature proves that the enhanced or novel feature selector selects the optimal features to increase the performance of the classifiers. In this study a new SEF-USIEF Feature Selector: An approach to Select Effective Features and Unselect Ineffective Feature methods is implemented to obtain the optimal features from the dataset and to evaluate the performance of the classifiers.

3. METHODOLOGY

For this experiment, the mixed datatype of datasets is used for evaluating the performance of the classifiers for the selected features. The datasets like Iris, Vehicle, Glass, Shuttle Landing, Wisconsin breast cancer, and Mammographic masses are obtained from the UCI repository [1], and Webb, WebKb2, and WebKB4 are obtained from the WebKB Dataset repository. To find the relevant features and to check the performance of the classifiers, the WEKA tool is used. To obtain the relevant features for the SEF-USIEF Feature Selector: An approach to Select Effective Feature and Unselect Ineffective Feature Method JAVA code is written and the selected features are used in WEKA to check the performance of the clarifiers.

SEF-USIEF Feature Selector: An approach to Select Effective Features and Unselect Ineffective Feature method process.

Variance is part of statistical theory; the variance is a statistical concept related to the spread or dispersion of a set of data. Variance describes how much a random variable differs from its expected value. Variance analysis is usually associated with explaining the difference (or variance) between actual and expected output. It is tremendously important to visualize and understand the data being considered.

In our approach, the concept of variance was used as part of the feature selection process. In the proposed method the variance value of the attributes pair is computed. The proposed approach measures the variance in each feature pair. It claims that predictive features with minimal variance without class labels may not have any pattern in them. Therefore, we can discard those features regarding their lowest variance and predictive features with the minimum variance that are associated with the class label may have the pattern in it, therefore those features are considered relevant features. The proposed method significantly improved the performance of various classifiers.

3.1 EXPERIMENTAL PROCEDURE

- Datasets Vehicle, Glass, Shuttle Landing, Wisconsin breast cancer, Mammographic masses, Webb, WebKB2, and WebKB4 are used for this experiment.
- To select the relevant features for this dataset the existing methods and the proposed SEF-USIEF method are used.
- Information Gain (IG), Gain Ratio (GR), ReliefF (RF), Symmetric Uncertainty (SU), OneR (OR), Correlation Attribute (CA), and Principal Components (PCA) are existing methods
- Naïve Bayes (NB), Multilayer Perceptron Classifier (MLP), K Nearest Neighbor (KNN), Decision Tree (DT), SMO, J48, KStar, JRip, and OneR classifiers are used to check the performance for the features selected by existing methods and the proposed method.
- To check the performance the results are compared.

4. RESULTS AND FINDINGS

The datasets like Vehicle, Glass, Shuttle Landing, Wisconsin breast cancer, Mammographic masses, Webb, WebKB2, and WebKB4 are loaded in the WEKA tool. The number of original features and the instances for each dataset are noted in Table.1.

Table.1. Datasets: Features and Instances

Dataset	Features	Instances
Iris	5	150
Vehicle	19	846
Glass	10	214
webkb	6	56
Webkb2	7	92
Webkb4	10	298
Shuttle Landing Control	7	6
Wisconsin breast cancer	10	699
mammographic masses	6	961
Total	80	3322

By using the WEKA tool, the Information Gain (IG), Gain Ratio (GR), ReliefF (RF), Symmetric Uncertainty (SU), OneR (OR), Correlation Attribute (CA), Principal Components (PCA) feature selection method is used and the selected features are noted in Table.2 and a threshold value 80% is used on the ranking method to select the features as shown in Table.3.

Features used for the experiments are given in Table.4. The Table.5 and the features not used for the experiment are given in Table.6, Table.7. After selecting the features by using the existing method and proposed method, the selected features are evaluated by using the different classifiers Naïve Bayes (NB), Multilayer Perceptron Classifier (MLP), K Nearest Neighbor (KNN), Decision Tree (DT), SMO, J48, KStar, JRip and OneR classifiers. The results are produced from Table.7 to Table.15.

Table.2. Number of features selected

DS/FS	ORF	PCA	IG	GR	RF	SU	OR	CA	SEF-USIEF
Iris	5	3	4	4	4	4	4	4	1
Vehicle	19	9	18	18	18	18	18	18	15
Glass	10	10	9	9	9	9	9	9	8
webkb	6	5	5	5	5	5	5	5	4
Webkb2	7	6	6	6	6	6	6	6	5
Webkb4	10	9	9	9	9	9	9	9	8
Shuttle Landing Control	7	4	6	6	6	6	-	6	3
Wisconsin breast cancer	10	7	9	9	9	9	9	9	6
mammographic masses	6	5	5	5	5	5	5	5	2
AVG	8.9	6.4	7.9	7.9	7.9	7.9	8.1	7.9	5.8

Table.3 Threshold 80% for ranking methods

DS/FS	ORF	PCA	IG	GR	RF	SU	OR	CA	SEF-SIEF
Iris	5	3	3.2	3.2	3.2	3.2	3.2	3.2	1
Vehicle	19	9	14.4	14.4	14.4	14.4	14.4	14.4	15
Glass	10	10	7.2	7.2	7.2	7.2	7.2	7.2	8
webkb	6	5	4	4	4	4	4	4	4
Webkb2	7	6	4.8	4.8	4.8	4.8	4.8	4.8	5
Webkb4	10	9	7.2	7.2	7.2	7.2	7.2	7.2	8
Shuttle Landing Control	7	4	4.8	4.8	4.8	4.8	-	4.8	3
Wisconsin breast cancer	10	7	7.2	7.2	7.2	7.2	7.2	7.2	6
mammographic masses	6	5	4	4	4	4	4	4	2
AVG	36	6.4	6.3	6.3	6.3	6.3	6.5	6.3	5.8

Table.4. Features used for the experiment

DS/FS	ORF	PCA	IG	GR
Iris	5	1,2,3	3,4,1	4,3,1
Vehicle	19	1,2,3,4,5,6,7,8,9	12,7,8,11,9,3,6,2,1,4,13,10,14,17	11,9,12,7,4,8,6,3,5,18,13,14,1,2
Glass	10	1,2,3,4,5,6,7,8,9,10	4,3,6,7,8,2,1	8,3,2,4,6,7,1
webkb	6	1,2,3,4,5	5,2,3,4	5,2,3,4
Webkb2	7	1,2,3,4,5,6	5,1,4,2,6	5,1,4,2,6
Webkb4	10	1,2,3,4,5,6,7,8,9	4,2,9,7,1,8,6	4,7,8,2,9,6,1
Shuttle Landing Control	7	1,2,3,4	5,4,2,6,3	5,4,2,6,3
Wisconsin breast cancer	10	1,2,3,4,5,6,7	2,3,6,7,5,8,1	8,5,2,6,3,7,9
mammographic masses	6	1,2,3,4,5	1,4,3,2	1,3,4,2

Table.5. Features used for the experiment

DS/FS	RF	SU	OR	CA	SEF-USIEF
Iris	4,3,1	4,3,1	4,3,1	3,4,1	4
Vehicle	8,18,7,12,9,3,	12,7,8,11,	12,9,7,8,	3,8,7,	1,2,3,4,

	10,12,17,1,13,4,6	9,6,3,4,1,13,2,14,10,17	11,3,1,2,17,6,14,4,10,18	12,11,9,4,14,1,18,16,13,2,6	5,6,7,8,9,10,11,12,13,14,18
Glass	3,4,8,7,2,1,5	3,4,8,6,7,2,1	4,7,1,6,8,2,3	3,4,8,2,9,7,1	1,2,3,4,5,7,8,9
webkb	5,3,1,4	5,2,3,4	5,2,3,4	5,2,3,4	1,2,4,5
Webkb2	5,1,4,2,3	5,1,4,2,6	5,6,3,4,1	5,1,4,3,6	1,2,4,5,6
Webkb4	9,2,8,6,1,7,5	4,7,2,9,8,6,1	4,7,9,1,2,8,3	4,7,2,9,1,8,6	1,2,3,5,6,7,8,9
Shuttle Landing Control	5,4,2,6,3	5,4,2,6,3	-	6,5,2,3,4	1,2,6
Wisconsin breast cancer	6,1,3,2,8,7,4	2,6,5,8,3,7,4	2,3,7,6,8,5,4	3,2,6,7,1,8,4	1,2,3,4,5,9
mammographic masses	2,5,4,3	1,3,4,2	1,4,3,2	4,3,2,1	1,5

Table.6. Features not used for the experiment

DS/FS	ORF	PCA	IG	GR
Iris	5	4	2	2
Vehicle	19	10,11,12,13,14,15,16,17,18	18,5,16,15	16,10,15,17
Glass	10	-	9,5	9,5
webkb	6	-	1	1
Webkb2	7	-	3	3
Webkb4	10	-	5,3	5,3
Shuttle Landing Control	7	5,6	1	1
Wisconsin breast cancer	10	10	8,9	4,9
mammographic masses	6	-	5	5

Table.7. Features not used for the experiment

DS/FS	RF	SU	OR	CA	SEF-USIEF
Iris	2	2	2	2	1,2,3
Vehicle	14,15,5,16	18,5,16,15	5,13,16,15	15,10,5,17	15,16,17
Glass	6,9	9,5	5,9	6,5	6
webkb	2	1	1	1	3
Webkb2	6	3	2	2	3
Webkb4	4,3	5,3	6,5	5,3	6
Shuttle Landing Control	1	1	-	1	4,3,5
Wisconsin breast cancer	4,1	5,9	1,9	1,9	5,9
mammographic masses	1	5	5	5	3,2,4

Table.8. Classification Accuracy - PCA Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	88	94	88.66	90.66	93.33	94	92.66	92.66	93.33
Vehicle	46.45	75.17	68.32	71.74	71.51	65.95	65.95	50.94	70.68
Glass	48.59	67.75	56.07	70.56	75.23	68.22	68.69	57.94	66.82
webkb	100	100	100	100	96.42	100	100	100	100
Webkb2	93.47	92.39	96.73	93.4	94.56	96.73	96.73	96.73	96.73
Webkb4	96.64	93.28	96.64	93.62	94.29	94.29	94.29	87.24	95.63
Shuttle Landing Control	83.33	66.66	83.33	50	83.33	83.33	83.33	83.33	83.33
Wisconsin breast cancer	96.13	95.42	96.7	95.85	95.42	95.27	95.85	92.7	94.84
mammographic masses	78.35	80.95	79.29	75.23	81.26	82.31	82.51	81.89	82.1
AVG	81.22	85.07	85.08	82.34	87.26	86.68	86.67	82.60	87.05

Table.9. Classification Accuracy - IG Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	94.66	92.66	92.66	92	96
Vehicle	41.84	77.77	71.74	69.85	70.56	64.3	67.61	51.89	72.1
Glass	72.89	64.95	52.8	77.57	78.03	66.82	68.69	57.94	69.62
webkb	100	100	100	100	98.21	100	100	100	100
Webkb2	90.21	94.56	96.73	95.65	94.56	96.73	96.73	96.73	96.73
Webkb4	94.29	94.29	95.3	95.3	95.3	94.29	94.63	87.24	95.63
Shuttle Landing Control	83.33	33.33	66.66	83.33	83.33	83.33	83.33	66.66	83.33
Wisconsin breast cancer	96.13	94.56	96.42	95.7	94.7	94.99	95.13	91.55	94.56
mammographic masses	78.35	81.37	79.6	77.31	83.03	82.2	82.51	81.89	81.68
AVG	83.67	82.02	83.92	87.78	88.04	86.15	86.81	80.66	87.74

Table.10. Classification Accuracy - GR Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	94.66	92.66	92.66	92	96
Vehicle	45.03	78.36	71.51	70.09	70.33	65.6	67.61	51.89	73.99
Glass	72.89	64.95	52.8	77.57	78.03	66.82	68.69	57.94	69.62
webkb	100	100	100	100	98.21	100	100	100	100
Webkb2	90.21	94.56	96.73	95.65	94.56	96.73	96.73	96.73	96.73
Webkb4	94.29	94.29	95.3	95.3	95.3	94.29	94.63	87.24	95.63
Shuttle Landing Control	83.33	33.33	66.66	83.33	83.33	83.33	83.33	66.66	83.33
Wisconsin breast cancer	96.42	95.85	96.85	95.42	96.28	95.27	96.13	92.7	94.27
mammographic masses	78.35	81.37	79.6	77.31	83.03	82.2	82.51	81.89	81.68
AVG	84.06	82.23	83.94	87.78	88.19	86.32	86.92	80.78	87.92

Table.11. Classification Accuracy - RF Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	94.66	92.66	92.66	92	96
Vehicle	41.84	78.6	70.68	72.81	70.92	65.72	68.08	51.89	71.04
Glass	46.72	70.56	51.4	75.23	76.63	66.35	63.55	57.94	68.22
webkb	100	100	100	100	96.42	100	100	100	100
Webkb2	92.39	96.73	96.73	94.56	96.73	96.73	96.73	96.73	96.73
Webkb4	94.29	93.95	95.63	94.29	94.29	93.95	93.28	83.22	93.95
Shuttle Landing Control	83.33	33.33	66.66	83.33	83.33	83.33	83.33	66.66	83.33
Wisconsin breast cancer	96.13	95.42	96.28	94.56	95.85	95.27	94.56	92.7	94.56
mammographic masses	77.1	80.95	78.87	74.81	80.12	78.98	80.12	76.06	80.85
AVG	80.87	82.99	83.58	87.21	87.66	85.89	85.81	79.69	87.19

Table.12. Classification Accuracy - SU Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	94.66	92.66	92.66	92	96
Vehicle	41.84	77.77	71.74	69.85	70.56	64.3	67.61	51.89	72.1
Glass	72.89	64.95	52.8	77.57	78.03	66.82	68.69	57.94	69.62
webkb	100	100	100	100	98.21	100	100	100	100
Webkb2	90.21	94.56	96.73	95.65	94.56	96.73	96.73	96.73	96.73

Webkb4	94.29	94.29	95.3	95.3	95.3	94.29	94.63	87.24	95.63
Shuttle Landing Control	83.33	33.33	66.66	83.33	83.33	83.33	83.33	66.66	83.33
Wisconsin breast cancer	96.7	95.27	96.85	95.13	95.56	95.42	95.7	92.7	95.27
mammographic masses	78.35	81.37	79.6	77.31	83.03	82.2	82.51	81.89	81.68
AVG	83.73	82.10	83.96	87.72	88.14	86.19	86.87	80.78	87.82

Table.13. Classification Accuracy - OR Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	94.66	92.66	92.66	92	96
Vehicle	42.43	81.44	73.52	71.86	71.27	65.6	67.49	51.89	73.87
Glass	41.84	77.77	71.74	69.85	70.56	64.3	67.61	51.89	72.1
webkb	72.89	64.95	52.8	77.57	78.03	66.82	68.69	57.94	69.62
Webkb2	93.47	94.56	96.73	94.56	96.73	96.73	96.73	96.73	96.73
Webkb4	94.96	94.63	94.96	93.62	95.3	94.29	94.29	87.24	95.63
Shuttle Landing Control	83.33	33.33	66.66	83.33	83.33	83.33	83.33	66.66	83.33
Wisconsin breast cancer	96.13	94.84	95.7	94.27	94.99	94.99	94.84	91.55	93.84
mammographic masses	78.35	81.37	79.6	77.31	83.03	82.2	82.51	81.89	81.68
AVG	77.71	80.02	80.86	84.19	85.32	82.32	83.13	75.31	84.76

Table.14. Classification Accuracy - CA Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	94.66	92.66	92.66	92	96
Vehicle	41.48	79.07	69.5	69.03	70.68	65.24	65.72	49.64	72.57
Glass	45.32	65.88	57	69.15	78.03	66.82	60.74	54.2	65.42
webkb	100	100	100	100	98.21	100	100	100	100
Webkb2	90.21	94.56	96.73	95.65	94.56	96.73	96.73	96.73	96.73
Webkb4	94.29	94.29	95.3	95.3	95.3	94.29	94.63	87.24	95.63
Shuttle Landing Control	83.33	33.33	66.66	83.33	83.33	83.33	83.33	66.66	83.33
Wisconsin breast cancer	96.13	94.84	95.7	94.27	94.99	94.99	94.84	91.55	93.84
mammographic masses	78.35	81.37	79.6	77.31	83.03	82.2	82.51	81.89	81.68
AVG	80.57	82.30	84.05	86.60	88.09	86.25	85.68	79.99	87.24

Table.15. Classification Accuracy - SEF-USIEF Method

DS/FS	NB	MLP	SMO	KNN	KSTAR	DT	JRIP	ONER	J48
Iris	96	97.33	96	95.33	95.33	92.66	92.66	92	96
Vehicle	41.84	77.77	71.74	69.85	70.56	64.3	67.61	51.89	72.1
Glass	49.06	64.95	52.8	77.57	78.03	66.82	68.69	77.74	69.62
webkb	100	100	100	100	98.21	100	100	100	100
Webkb2	93.47	94.56	96.73	94.56	96.73	96.73	96.73	96.73	96.73
Webkb4	94.29	94.29	95.3	95.3	95.3	94.29	94.63	87.24	95.63
Shuttle Landing Control	83.33	83.33	66.66	83.33	83.33	83.33	83.33	83.33	83.33
Wisconsin breast cancer	96.13	94.56	96.42	95.7	94.7	94.99	95.13	91.55	94.56
mammographic masses	78.35	81.37	79.6	77.31	83.03	82.2	82.51	81.89	81.68
AVG	81.39	87.57	83.92	87.66	88.36	86.15	86.81	84.71	87.74

Table.15. Average performance of the classifiers

CA/FS	PCA	IG	GR	RF	SU	OR	CA	SEF-USIEF
NB	81.22	83.67	84.06	80.87	83.73	77.71	80.57	81.39
MLP	85.07	82.02	82.23	82.99	82.10	80.02	82.30	87.57
SMO	85.08	83.92	83.94	83.58	83.96	80.86	84.05	83.92
KNN	82.34	87.78	87.78	87.21	87.72	84.19	86.60	87.66
KSTAR	87.26	88.04	88.19	87.66	88.14	85.32	88.09	88.36
DT	86.68	86.15	86.32	85.89	86.19	82.32	86.25	86.15
JRIP	86.67	86.81	86.92	85.81	86.87	83.13	85.68	86.81
ONER	82.60	80.66	80.78	79.69	80.78	75.31	79.99	84.71
J48	87.05	87.74	87.92	80.85	87.82	84.76	87.24	87.74
AVG	84.89	85.20	85.35	83.84	85.26	81.51	84.53	86.03

The Table.15 shows the average performance of the different classifiers against the selected features. From this table, it shows that the proposed Feature selector MVM has the better performance than the other feature selection methods.

5. DISCUSSION

The 9 datasets provided in Table.1 are used for the experimentation investigation. The datasets are subjected to feature selection techniques as Information Gain (IG), Gain Ratio (GR), ReliefF (RF), Symmetric Uncertainty (SU), OneR (OR), Correlation Attribute (CA), and Principal Components Analysis (PCA). The datasets are used to conduct the SEF-USIEF technique as well. The number of features chosen by the SEF-USIEF approach, and the feature selection methods are displayed in Table.2. Comparing the SEF-USIEF average value to those of the other classifiers reveals that it is lower (5.8) than those of the other classifiers, as shown in Table.2. This demonstrated that the SEF-USIEF technique is superior to other classifiers.

An 80% of threshold value is utilized on the chosen characteristics, as indicated in Tables 1.3, 1.4, and 1.5, to assess the efficacy of the SEF-USIEF approach. The performance of the classifiers is evaluated using the same datasets and several classification techniques such as Naive Bayes (NB), Multilayer Perceptron Classifier (MLP), Sequential Minimal Optimization (SMO), K Nearest Neighbor (KNN), KStar, Decision Tree (DT), JRip, OneR, and J48. The Table.7 displays the classification accuracy of the PCA method, Table.8 displays the IG method, Table.9 displays the GR method, Table.10 displays the RF method, Table.11 displays the SU method, Table.12 displays the OR method, Table.3 displays the CA method, and Table.14 displays the SEF-USIEF method. Table.15 displays the classifiers' combined average performance, and SEF-average USIEF's score (87.66) is higher than those of the other classifiers. This shown that SEF-USIEF outperformed the other classifiers in terms of classification performance.

6. CONCLUSION

This study proposed a SEF-USIEF Feature Selector: An approach to Select Effective Features and Unselect Ineffective Feature selector method to select the minimum features that are consider as relevant features from the dataset and these selected relevant features are evaluated to check the performance in the different classifiers. The results are compared to find the accuracy with the existing feature selection methods and the existing classification methods. From the comparison, it has proved that the SEF-USIEF Feature Selector: An approach to Select Effective Features and Unselect Ineffective Feature method shows a better performance when compared to the existing methods.

REFERENCES

[1] D. Dua and C. Graff, “UCI Machine Learning. Repository”, Available at http://archive.ics.uci.edu/ml, Accessed at 2019.

[2] N. Elssied and A. Osman, “A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification”, *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 12, pp. 625-638, 2014.

[3] A. Galathiya and C. Bhensdadia, “Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning”, *International Journal of Computer Science*, Vol. 3, No. 2, pp. 3427-3431, 2012.

[4] P. Jamshid and O. Mohammad Hossein, “An Efficient Hybrid Filter-Wrapper Metaheuristic-Based Gene Selection Method for High Dimensional Datasets”, *Scientific Reports*, Vol. 9, pp. 1-15, 2019.

[5] A. Jovic and N. Bogunovic, “A Review of Feature Selection Methods with Applications”, *Proceedings of International Convention on Information and Communication Technology, Electronics and Microelectronics*, pp. 1-7, 2015.

[6] K.H. Keerthi and B. Harish, “A New Feature Selection Method for Sentiment Analysis in Short Text”, *Journal of Intelligent Systems*, Vol. 12, pp. 1122-1134, 2020.

[7] T. Khawla and Z. Azeddine, “Feature Selection Methods and Genomic Big Data: A Systematic Review”, *Journal of Big Data*, Vol. 12, No. 2, pp. 1-14, 2019.

[8] K.I. Ludmila and F.J. William, “PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data”, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 25, No. 1, pp. 69-80, 2014.

[9] T. Marwa, "A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis", *Procedia Computer Science*, Vol. 78, pp. 186-194, 2017.

[10] N. Mehdi, B. Amir-Masoud and V. Touraj, "A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms", *International Journal of Computer Applications*, Vol. 69, No. 17, pp. 1-13, 2013.

[11] R. Mehrdad and F. Saman, "A Novel Community Detection Based Genetic Algorithm for Feature Selection". *Machine Learning*, Vol. 13, No. 1, pp. 1-16, 2020.

[12] C. Nicole and B. Pierre, "New Feature Selection Method based on Neural Network and Machine Learning", *IEEE International Multidisciplinary Conference on Engineering Technology*, pp. 81-85, 2016.

[13] D. Pijush and R. Susanta, "sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data using Support Vector Machine and T Statistic", *Journal Frontiers in Genetics*, Vol. 14, No. 1, pp. 1-14, 2020.

[14] E.D. Preetha, G. Deepti Raj and T. Rajendran, "Feature Subset Selection for Irrelevant Data Removal using Decision Tree Algorithm", *Proceedings of International Conference on Advanced Computing*, pp. 268-274, 2013.

[15] R. Radha and S. Muralidhara, "Removal of Redundant and Irrelevant Data from Training Datasets using Speedy Feature Selection method", *International Journal of Computer Science and Mobile Computing*, Vol. 15, pp. 359-364, 2016.

[16] B. Raja and T. Babu, "A Novel Feature Selection Based Ensemble Decision Tree Classification Model for Predicting Severity Level of COPD Disease", *Biomedical and Pharmacology Journal*, Vol. 23, No. 1, pp.1-18, 2019.

[17] P. Selwyn, "Evaluating Feature Selection Methods for Learning in Data Mining Applications", *Proceedings of International Conference on Computing, Artificial Intelligence and Information Technology*, pp. 483-494, 2006.

[18] C. Silvia and V. Marco, "A Hybrid Feature Selection Method for Classification Purposes", *Proceedings of International Conference on European Modelling*, pp. 1-13, 2016.

[19] C. Sunyoung and K. Hyuntaek, K., "Automatic Recognition of Alzheimer's Disease using Genetic Algorithms and Neural Network", *Lecture Notes in Computer Science*, pp. 695-702, 2003.

[20] M.S. Suresh and N. Athi, "Improving Classification Accuracy using Combined Filter+Wrapper Feature Selection Technique", *Proceedings of International Conference on Electrical, Computer and Communication Technologies*, pp. 1-6, 2019.

[21] P. Xiaoqing and C. Yaokai, "Hybrid Feature Selection Model based on Machine Learning and Knowledge Graph", *Journal of Physics: Conference Series*, Vol. 23, No. 1, pp. 1-14, 2021.

[22] P. Yonghong and J. Jianmin, "A Novel Feature Selection Approach for Biomedical Data Classification", *Journal of Biomedical Information*, Vol. 23, No. 1, pp. 15-23, 2010.

[23] Zsolt János, V., Krisztián Balázs, K,, Ádám, F., and Máté István, B, "Adaptive, Hybrid Feature Selection (AHFS)", *Pattern Recognition*. Volume 116, 2021.