www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC
Gestão, Inovação e Tecnologias

Features Selection for Supervised Learning Using Centrality Measures

Saba Saleem[1]; Mehmood Ahmed[2]; Luqman Shah[3]; Ali Imran Jehangiri[4]; Muhammad Naeem[5]; Yousaf Saeed[6]; Muhammad Junaid[7]; Fahad Ali Khan[8]

[1]Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan.
[1]sabasaleem91@yahoo.com
[2]Department of IT, The University of Haripur, Pakistan.
[3]Department of IT, The University of Haripur, Pakistan.
[4]Department of IT, Hazara University, Mansehra, Pakistan.
[5]Department if IT, Abbottabad University of Sciences and Technology, Abbottabad, Pakistan.
[6]Department of IT, The University of Haripur, Pakistan.
[7]Department of IT, The University of Haripur, Pakistan.
[7]mjunaid@uoh.edu.pk
[8]Department of IT, The University of Haripur, Pakistan.

Abstract

The data mining methods have been extensively used in the process of decision making. The popularity of data mining methods is due to availability of high speed algorithms, processing and storage power of computers. The effective use of data mining methods help in mining datasets and taking better decisions. The data need to be preprocessed before applying data mining methods. Some datasets require little preparation like dealing with missing and redundant instances while some high-dimensional datasets require strong processing like dimensionality reduction. One of the techniques used for dimensionality reduction is feature selection. This study uses graph based centrality measure for feature selection. Graph based centrality measures are used for ranking features which is used for removing irrelevant attributes. After comparison of results with other approaches, it has been found that the proposed approach results in reduction of feature space without compromising accuracy. The results also shows that proposed approach performs better than some other feature selection approaches not only in terms of accuracy but also on the basis of larger reduction in feature space.

Key-words: Online Sources, Emails, Social Networks, Search Queries.

1. Introduction

Today, the concept of big data and its analysis is widespread in modern science and business. In general data is generated by online sources, emails, logs, social networks, search queries, health data, mobile phone and different apps. Increase in the use of digital technology has resulted in production of large volume of data. Some facts about data explosion are mentioned below.

- About 90% of the data of world produced today was created in the past 2 years [1].
- Total amount of data stored in world in year 2000 was 800,000 PB. But it is expected that this number reaches to 40 zettabytes (ZB) by year 2020 [1].

The low cost of data storage and ease of data capturing power such as barcode reader, credit card swipe make storage of data easy. This huge amount of data is useless till the mining and extraction of knowledge. Although, efficient online transaction processing (OLTP) systems are in place, for capturing and storing application specific data, but there exists vast value oriented, data mining opportunities on this data. There exists an inverse relation between data processing and data creation i.e. as the amount of generated data increases, the amount of data that is analyzed and processed decreases [2]. This increase in amount of data requires strong processing tools for knowledge extraction that makes knowledge discovery process easy. Analysis of data leads to good prediction which can be used to take better decisions.

As more information is collected from different sources, the likelihood of working with high dimensional data increases. High dimensional data containing attributes in hundreds or thousands are normal now a days. For example genomic libraries contain the attributes in thousands but the mining of such a high dimensional dataset causes many problems.

- Increase in the number of features causes increase in the computational cost of both supervised and unsupervised learning algorithms exponentially.
- All the attributes are not relevant for particular problem. Some of them are not suitable for classification task, so need to be removed from feature set.

Feature selection (FS) helps to overcome these issues by selecting relevant features and removing redundant and irrelevant features. It helps to decrease the dimensionality of feature space and computational cost as well. Filter, wrapper and hybrid are the main approaches used for selecting features. Filter approach selects the relevant features with the help of statistical analysis. It normally select the features without depending on any learning algorithm. Wrapper approaches are dependent on learning algorithms for removing irrelevant features. Learning algorithms help in evaluating the features and choose the optimal subset of feature. Whereas hybrid approach is combination of both

Received: 08.06.2021 – Accepted: 09.07.2021

filter and wrapper approach. It ranks features according to filter part and use of algorithm comes in the category of wrapper part where irrelevant features are removed incrementally.

Reduction in feature space results in improving accuracy, efficiency and scalability of algorithm. In recent years, different social networking techniques gain interest in selection of relevant features by removing redundant and irrelevant features from high dimensional datasets. In our approach, we used network based techniques i.e. centrality measures for ranking features. Different centrality measures are used to find important nodes in a network. This ranking is done on the basis of properties of network. The feature having high value is most influential in a network and represents important information. These ranked features are step by step given to learning algorithms for selecting optimal subset of features. Different evaluation measures like accuracy and error rate are used for evaluating each subset of feature.

The rest of the paper is organized as follows: Section II reviews the literature of feature selection methods. Section III presents the methodology used in this paper. The step by step method of selecting features is discussed in this section. Section IV evaluates the methodology. The results of four datasets are compared in Section V. Section VI conclude the work.

2. Existing Work

The process of feature selection comprises of four main steps i.e. subset generation, evaluation, stopping criteria and results validation [3]. The first step of feature selection involves searching of the feature space for the selection of subset that is most likely to predict the class of data provided. If there are N features then the possible subsets for N features are 2^N. This exhaustive search of the feature space is suitable for small number of features but is impractical for large search space as there exist 2^N possible subsets for N attributes. A more feasible approach to use for large number of N is heuristic approach as compared to exhaustive search but it doesn’t guarantee the finding of optimal subset [4]. Generated subset is evaluated on the basis of different evaluation measure. The different methods of evaluation are used in different approaches. Some criteria like classification accuracy and error rate are set to halt the algorithms. After selection of features’ subset, results are compared with other feature selection approaches for results validation.

The process of feature selection can be categorized into four methods namely filter, wrapper and hybrid. This classification is based on the usage of machine learning algorithm in the feature selection process.

A. Filter Approach

Filter based methods do not depend on any classification algorithm for feature selection and evaluate the relevance of features on the basis of properties of data. A ranking criteria is set to score the attributes and remove the irrelevant attributes on the basis of selected threshold. After removal of irrelevant attributes, the selected subset is given as input to classification algorithms. Different algorithms use different measures like distance, correlation, information gain and consistency to give weight to features and rank them for selection. In [5], relief algorithm was proposed, that used distance based metrics. In this algorithm, weight is given to features on the basis of statistical relevancy of feature with class. It did not handle the issue of redundant features and multiclass. To handle the problem of multiclass, Kononenko in [6] proposed a modified version of relief algorithm called Relief F. It was also capable of dealing with noisy as well as incomplete dataset.

In [7], researchers proposed correlation based feature selection heuristic (CFS) for selecting relevant features. In this method subset of features are evaluated by selecting the subset having features that are not correlated with each other but are highly correlated with classes. Each feature is evaluated independently on the basis of its predictive ability and degree of redundancy. The comparison of results shows that the proposed method is many times faster than wrapper method.

MRMR stands for minimum redundancy and maximum relevancy and was proposed by [8]. As its name implies, it tries to maximize the relevancy of features with target class and minimize the redundancy in each class. Mutual information is used to find relevance between features and target.

Filter based features selection methods are fast, scalable and are independent of learning algorithms. These properties result in one time selection of features and then evaluated with different classifiers [9]. The filter based methods generate general results and have lower classification accuracy due to lack of interaction with classifiers [9].

B. Wrapper Approach

Wrapper based approach depends on supervised learning algorithm. Subsets of features are generated using any searching technique and then these subsets are evaluated on the basis of evaluation measures like classification error or accuracy by any supervised learning algorithm. The disadvantage is that it has overfitting risk as compared to filter approach and is computationally expensive [10].

George et al. [11] were the first in using wrapper as a general framework for feature selection in machine learning. They presented definition of feature relevance and claimed that wrapper can discover relevant features. In [4], the proposed method generates subsets using search engine and used classification algorithms for evaluating subsets. Accuracy was used for measuring performance with the help of decision tree and naïve bays algorithms.

A new approach based on combination of Support Vector Machine (SVM) with Kernel function was proposed in [12]. The generated subsets are evaluated on the basis of classification error to identify best subset. Wrapper methods are considered computationally expensive but many experimental results show that performance of wrapper methods is better [13].

C. Hybrid Approach

Filter and wrapper based methods combine to give hybrid method. In this approach features are ranked using different measures as part of filter approach and then wrapper method is applied to select the desired number of features [14].

In [15], statistical measures were used for ranking the attributes, based on relevancy. Top order attributes are then given to wrapper method, so that the required number of evaluations remains linear and results in reducing complexity of medical data classification. Similarly, [16] developed an algorithm based on hybrid approach called gene selection algorithm for selecting significant genes. Statistical approach is used to filter features and then these filtered features are fed to wrapper approach that results in recognition of significant genes causing cancer.

In [17], artificial neural network input gain measurement approximation (ANNIGMA) was proposed. It ranks features with the mutual information (MI) and select the required subset using artificial neural network (ANN). A hybrid model for text classification was proposed in [18] that used information gain (IG), Mutual information (MI), Chi-Square and document frequency for ranking features and SVM and Genetic algorithm (GA) for selecting features.

It is based on reducing computational time required for classification of different subsets which is done in wrapper methods. Hybrid approach takes the advantage of both the filter and wrapper approaches. Like filter approach they are less computationally intensive than wrapper approach and like wrapper approach they include interaction with classification model [19].

D. Graph Based Feature Selection Approach

Graph mining has been used in different fields such as web, social sciences and bioinformatics. Researchers proposed different algorithms for graph clustering, graph classification, graph querying, finding motifs and dense pattern in a graph [20]. Out of these different fields, graph classification has been focused greatly by researchers. The biggest challenge faced by researchers in graph classification is extraction of relevant features. Researchers used different graph based techniques along with other statistical measures for feature selection.

In [21], social network based method for feature selection was proposed. The proposed method used combination of community detection algorithm and newly proposed centrality measures for feature selection. In [22], researchers proposed a graph based method for selecting features for improving medical diagnosis. They also used community detection algorithm for clustering features and selected the best representative features from each cluster using fisher score.

In [23], Multidimensional interaction information (MII) is used for selecting subset of relevant features. Nodes in a graph represent features, and the relation between nodes is determined on the basis of similarity between features. Then they performed dominant set clustering for clustering the features. On each dominant set, apply MII criteria and select the top K features based on the value of incremental gain.

In [24], consensus based clustering for feature selection was proposed. They proposed their own algorithm known as Best of-K (BOK) consensus clustering algorithm for finding the best partition.

Page rank is a social networking technique used for selecting features [25]. Most of the feature selection techniques had been applied on supervised data as compared to un-supervised data. With the advancement in technology, amount of unsupervised data available on the web is large. It results in increase in interest in features selection for unsupervised data.

Centrality measures are the most important measures used in social networking technique for finding influential node in a network. These centrality measures are used to rank features on the basis of their influence in a whole network. Different centrality measures like degree, betweenness, Eigen-vector, and closeness centrality are used for finding influential node in network. But till now to the best of our knowledge, out of these centrality measures only Eigen vector centrality [26] is used in feature selection. In our work, we will consider other standard centrality measures for feature selection.

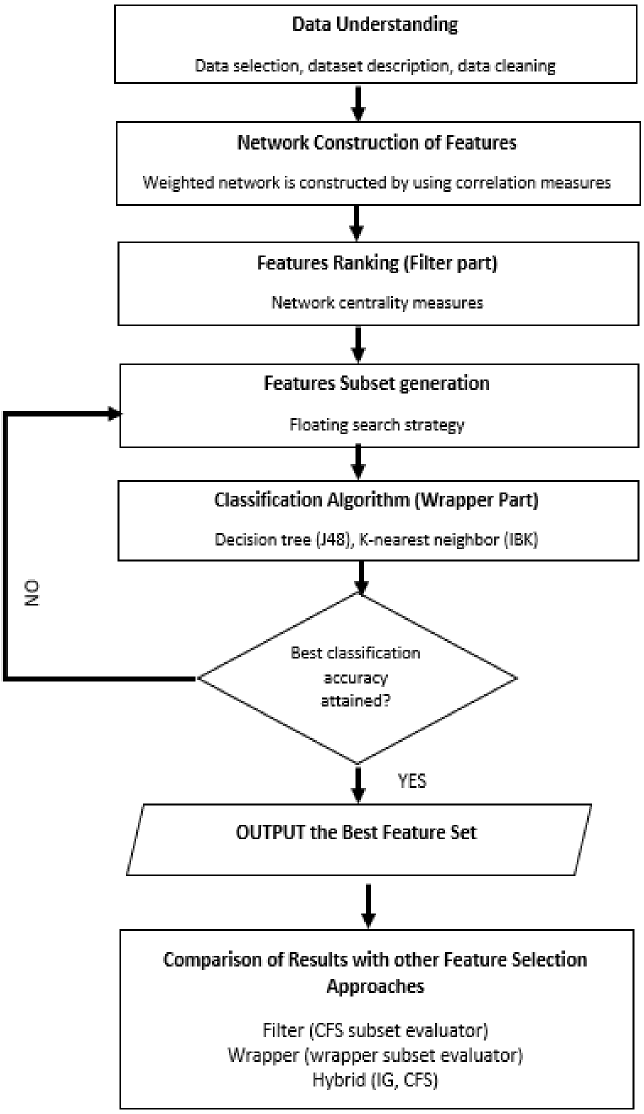
3. Methodology

Figure 1 shows the framework of our proposed approach. The process of selecting relevant features is divided into three main phases. In first step, all the features are represented in the form of network. After network construction, the second phase involves the application of social networks method for finding influential nodes in the network. For ranking nodes in a network, different centrality measures are utilized in this phase. This centrality measures ranks the features according to their importance in the network. Last phase is the involvement of the machine learning algorithms for the selection of relevant features from the features' set on the basis of evaluation measures. The final step of our approach is comparison of results of our proposed approach with other feature selection approaches.

A. Network Construction

After data selection, next step in our approach is the representation of features in the form of graph for feature selection. Feature selection is a part of data preprocessing stage. The increase in the size of data requires representation of data in efficient way for reducing complexity and making analysis easy.

Fig. 1 - Framework of Proposed Approach


Data Understanding
Data selection, dataset description, data cleaning
Network Construction of Features
Weighted network is constructed by using correlation measures
Features Ranking (Filter part)
Network centrality measures
Features Subset generation
Floating search strategy
Classification Algorithm (Wrapper Part)
Decision tree (J48), K-nearest neighbor (IBK)
NO
Best classification accuracy attained?
YES
OUTPUT the Best Feature Set
Comparison of Results with other Feature Selection Approaches
Filter (CFS subset evaluator)
Wrapper (wrapper subset evaluator)
Hybrid (IG, CFS)

Received: 08.06.2021 – Accepted: 09.07.2021

In graph of features, a feature is represented by a node. Two nodes are connected with each other if and only if they are similar. Different similarity measuring techniques are used to measure similarity between nodes. The network of features can be represented as G=< F, E> where F= {F1, F2, F3… Fn} are set of features and E represents set of edges between features. The value of Pearson correlation is used by different researchers [24] as an edge weight between two features and link is established between the vertices on the basis of similarity. The edge represents level of similarity between two features.

Correlation is one of the common statistic measures for establishing the degree of relation between variables. The relationship between two variables varies from weak to strong or none. The strong relation between two variables F1 and F2 means that value of variable F1 can be used to predict the value of F2. If the relation between two variables F1 and F2 is weak then the known value of variable F1 doesn't help to predict the value of variable F2. The four common types of correlation measures are [27]:

- Pearson's correlation: It is widely used correlation measure and is used to measure the strength of association between two continuous variables.
- Spearman's correlation: If the variables are ordinal then the spearman correlation is used to find association between them.
- Point-bi-serial correlation: It measures the strength of relationship between two variables from which one is continuous and other is dichotomous. The variables having two values are termed as dichotomous variables. Gender is a dichotomous variable.
- Phi (φ) correlation: If both variables are dichotomous then phi-correlation is used to find the strength of association between two variables [27].

The datasets we used in our study have continuous values.

So for this purpose we used Pearson's correlation measure as given in

$$r_{F1F2} = \frac{N\sum F1F2-(\sum F1)(F2)}{\sqrt{[N\sum F1^2-(\sum F1)^2][N\sum F2^2-(\sum F2)^2]}} \qquad (1)$$

In the Equation 1, F1 and F2 represents two variables (attributes). Where

N: Total number of values in data

ΣF1F2: Sum of product of values of all pairs of variable

ΣF1: Sum of all values of F1 variable

ΣF2: Sum of all values of F2 variable

$\Sigma F1^2$: Sum of square of all values of F1 variable

$\Sigma F2^2$: Sum of square of all values of F2 variable

The value of r is used to determine strength of association. It lies between -1 and 1. Negative value of r indicates that the relation between variables is negatively correlated i.e. increase in one value results in decrease in other value. Positively correlated value shows that increase in value of variable results in increase in value of other variable and vice versa. On the basis of values of r, the strength of association is categorized into different groups [28].

- 0.00-0.19 “very weak”
- 0.20-0.39 “weak”
- 0.40-0.59 “moderate”
- 0.60-0.79 “strong”
- 0.80-1.0 “very strong”

B. Features’ Ranking (Filter Part)

After visualization of features in the form of graphs, the most informative features from set of feature list are selected by applying centrality metrics. The selection of a central node in different networks has different usage in different fields. In the process of feature selection, ranking of features come under the category of filter approach. We proposed hybrid approach which is combination of filter and wrapper approach. In the filter approach, for ranking features we are using graph based centrality measures. These centrality measures are helpful in finding influential node in network.

Degree centrality (C_d) is simplest centrality measure. It measures the number of edges going in or out from particular node. The degree centrality is simplest centrality measures. This advantage is due to the reason that it only considers the local structure of a node. Its limitation is that it doesn’t consider the global structure of a node. For example, though a node is connected to many other nodes, it might not be in a position to quickly reach to other nodes of a network to access resources, knowledge and information [29].

Closeness Centrality (C_C) measures how close a node is to other nodes with respect to its position. It represents the closeness of a node to all other nodes of network or sum of the distance to all other nodes in the network in case of weighted graph. In unweighted network, smallest number of ties required to reach all other nodes are counted. A main disadvantage of this approach is that this measure cannot be applied to a network having disconnected component because distance between two disconnected nodes become infinite.

The betweenness centrality (C_B) of a node i is defined as number of shortest paths to all nodes that passes through node i divided by all possible shortest paths between starting and ending node. It is used to measure how often a node lies between the nodes. It overcomes the problem faced in degree centrality by taking into account the global structure of a network. Global structure helps the node to reach towards other non-adjacent nodes of a network for accessing resources and knowledge while local structure only consider the nodes directly connected to particular node. C_B could also be applied to a network having disconnected components which is the problem of closeness centrality. Although it overcome the issues of degree and closeness centrality but it is not free from limitations. Majority of the nodes in a network doesn't come in any shortest path and results in same betweenness score of 0.

We used C_B for ranking the features as it considers global structure of a network and could also be applied to a network having disconnected components. C_B controls the flow of resources within the network. So, the node having high C_B will be more representative of many other nodes as it is included in shortest paths to majority of other nodes and deletion of a low ranked node will have less or no effect on the classification accuracy. This will results in most representative features at the end. According to [30], the problem we faced with C_B is that majority of nodes doesn't lie in shortest path to many node, so result in same C$_B$ of 0. To rank the features having same C_B, we used the weighted degree centrality measure as both centrality measures are highly correlated [31]. This will help in ranking those features having same C_B score.

C. Feature Searching

Next step after features' ranking is the selection of best subset of features that is representative of full feature set. Different researchers use different approaches for selecting features. Sequential forward (SFS) and Sequential backward (SBS) searches are simplest searching approaches. In SFS, search may start with empty set and in each iteration, a new feature is added in the subset. After adding feature in the subset, subset is evaluated on the basis of chosen evaluation criteria. It is also called as successive addition process. SBS is reverse of SFS approach. In this approach, search is started with full set of features and features are removed step by step on the basis of evaluation criteria. This searching approach is also called as successive elimination mechanism. In another approach called Bidirectional approach, searching starts from both ends where features are added to one side and removed from other side simultaneously. Although the sequential search strategy is simple and it has fast implementation but it doesn't allow re- selection of removed features later i.e. once the feature is selected, it can't be deleted from the subset as in SFS and if the feature is deleted then it can't be re-added like in SBS. This problem is called nesting effect. To overcome this problem, floating search strategy was proposed. This strategy allows re-addition of deleted feature and removal of already selected features. The floating search strategy is computationally efficient and its performance has been found to be good as compared to other approaches [32]. For this reason, we use floating search strategy as it is good in comparison of other approaches in term of performance. We used classification accuracy as an evaluation measure. Accuracy is defined as the percentage of instances correctly classified by the classifier. Addition and deletion of a feature is determined on the basis of classification accuracy. If the addition of feature results in increase in accuracy then the feature is permanently included in subset but if it degrades the performance then it will be removed. Mathematically, it can be expressed as:

$$\begin{cases} \mathrm{Acc}(\mathrm{S}-\{\mathrm{A}\}) >= \mathrm{Acc}\,(\mathrm{S}), & \mathrm{S}-\{\mathrm{A}\} \\ \mathrm{Acc}(\mathrm{S}-\{\mathrm{A}\} < \mathrm{Acc}\,(\mathrm{S}), & \mathrm{S} \end{cases} \qquad (2)$$

In equation 2, A = a_1, a_2,….a_n, S represents full set of features, arranged according to ranked criteria. "*A*" represents individual element of set of feature. In different iterations, features are removed from the set on the basis of Function 2 and accuracy is determined with the help of classifier. If the deletion of a feature results in decreasing accuracy of a particular classifier, then the deleted feature is restored. But if the deletion of a feature results in increasing accuracy of classifier or the accuracy remains the same then the feature is considered irrelevant and removed from the feature set. After removal of a feature in current iteration, the updated subset S has 1 less feature as compared to previous iteration.

D. Classification Algorithms (Wrapper Part)

After ranking of features using centrality measures, the final features are selected using classification algorithm on the basis of some evaluation measure. The use of classification algorithm for feature selection comes in wrapper approach as it is dependent on classification algorithm for feature selection. Hybrid approach is combination of filter and wrapper approach, so it is also dependent on classification algorithm for selecting features. For wrapper part of our approach, we used two different benchmark classifiers namely decision tree, and K-nearest neighbor. These algorithms are chosen from list of top 10 most commonly used algorithms list [33].

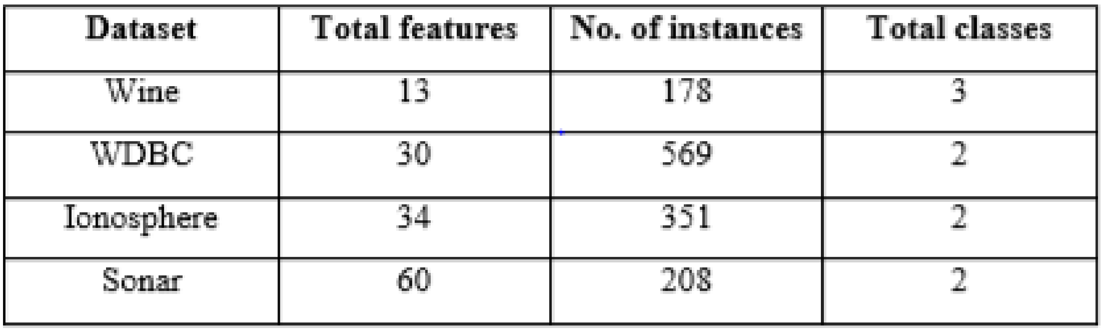
Decision Tree is flow-chart like structure, where the top most node of a tree is called root node, internal nodes represent features and leaf nodes represent the classes. On each internal node, condition is tested and results of condition are represented by branches that emerge from the node. The construction of decision tree is based on divide and conquers approach. A person having no technical knowledge of the process can even trace the flow of tree. Iterative Dichotomiser 3 (ID3), C4.5, Classification and Regression Tree (CART) etc. are some popular decision tree based algorithms. We used WEKA's J48 classifier that is implementation of C4.5 algorithm for our study.

Decision tree algorithm is known as eager learner because when training data is given to classifier it builds classification model before giving test tuple for classification. They are ready and eager to classify unseen data. While k-nearest neighbors are known as lazy learner because they don't build classification model when training data is given and wait for the test tuple. When lazy learner sees test tuple, it starts building classification model and classifies tuple. Lazy learner do less work when the training tuple is given and more work when test tuple is given so are also called instance based learner. The tuples in data having n attributes are represented in n-dimensional space. Every tuple of a dataset is represented as a point in n-dimensional space. When test tuple is supplied, it compares the test tuple with K- training tuples and selects the k-closest tuples from the dataset. These k closest tuples are called k-nearest neighbor. Closeness is defined in terms of distance. Different formulas are used for calculating distance between tuples.

E. Evaluation Performance

After building classification model, one is interested in determining how good the classifier is in prediction of class labels of test instances. Different evaluation measures are used by different researchers according to the requirements of the problem. The most common measure, used for estimating the performance of new instance is classification accuracy [34]. Accuracy and error rate are the main evaluation criteria used in this approach for selecting relevant features. Error rate defines the percentage of instances that are misclassified by the classifier. Feature selection is considered successful if the reduction in features does not deteriorate the accuracy or in simple words accuracy remains same or increase after deleting irrelevant features [34].

Table I - Datasets Description



Dataset	Total features	No. of instances	Total classes
Wine	13	178	3
WDBC	30	569	2
Ionosphere	34	351	2
Sonar	60	208	2

Finally, results are compared with other FS approaches like Filter CFS (correlation based feature selection), Wrapper Subset evaluator, Hybrid IG (Information Gain) and Hybrid Correlation using WEKA.

4. Evaluation

A. Data Selection

We used 4 different datasets in our study. The selection of dataset depends on the number of attributes. Benchmark datasets have been collected from UCI machine learning repository. Wine (< 15 features), WDBC and ionosphere (> 15 and <50 features) and sonar (>50 and <100 features) datasets are used in this study. The detail of chosen datasets are mentioned in Table 1.

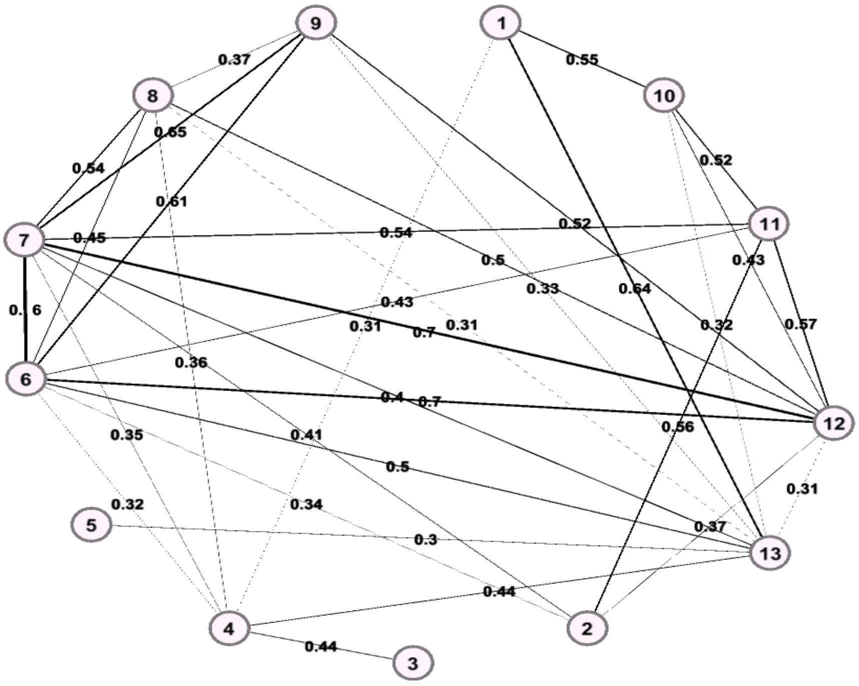
Wine Dataset comprises of results of chemical analysis of wine cultivated in the same area in Italy. The dataset was collected from three different cultivars. It is a classification problem and the task is to classify the instance in either of 3 types depending upon the quantities of 13 constituents in each of three types of wine. WDBC Dataset is complete with 569 patterns and 30 different attributes. All attributes are continuous. These 30 attributes are computed from digitized image of fine needle aspirate (FNA) of a breast mass. They described characteristics of the cell nuclei present in the image. There are two classes in this dataset i.e. malignant and benign. The task is to classify the cell into either of the two categories depending on the characteristic of nucleus.

Ionosphere dataset comprises of 351 instances and 34 attribute without any missing value. All attributes are continuous. Ionosphere is a layer of earth’s atmosphere which contains high concentration of free electrons and ions. They have ability to reflect radio waves. The radar data was collected by a system in Goose Bay, Labrador. It target free electrons present in this layer. The signals are then categorized into good and bad. Good radar shows some type of structure in the ionosphere while bad doesn’t return any structure and their signals pass through the ionosphere. The task is to classify the instances as either good or bad depending on the signals received. Sonar dataset contains 60 attributes and 208 instances. The attributes are real. Task in this dataset is to discriminate sonar signals. The signals that bounced off a metal cylinder (Mine) are categorized as M and signals that bounced off a rock are categorized into R.

B. Feature Selection

This section describes the step by step method of feature selection in detail on wine dataset. In the graph of Figure 2, 13 nodes represent 13 different features. The strength of edge between two features show level of similarity between them. The features having high correlation are highly connected and represented by edge having more weight. In this graph, high similarity exist between features 6 and 7 and the edge connecting these two features has high weight.

Fig. 2 - Graph of *Features* of Wine *Dataset*



Hybrid approach is a combination of filter and wrapper approach. Betweenness centrality (C_B) measure is used for ranking the features which is part of filter based approach. The most influential node in a network will have high betweenness centrality and will be ranked first. As the network is weighted, so we used weighted betweenness centrality value.

Table 2 shows the betweenness centrality value of all features arranged in descending order. Node 13 has highest weighted betweenness centrality in the graph and is influential node of a network. The order of ranking is (13, 7, 4, 12, 10, 11, 1), where feature 13 has highest value and feature 1 has lowest value. Features (2, 3, 5, 6, 8, 9) have 0 centrality. It means, these nodes don't lie in between shortest path of any node and results in C_B of 0. So, it becomes difficult to rank such type of features.

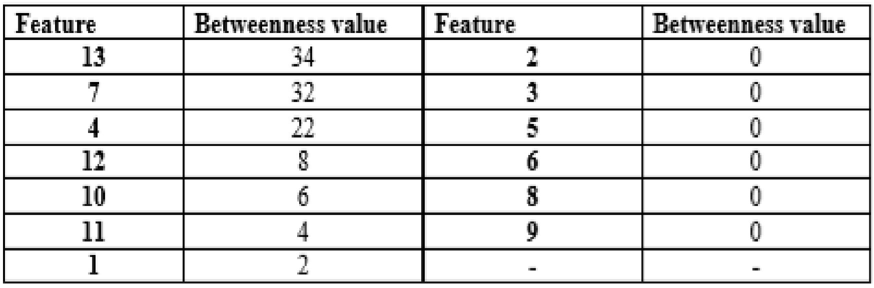
To overcome this issue, we chose degree centrality (C_d) for ranking such types of features as degree and betweenness centralities are highly correlated. It means that both of these centrality

Received: 08.06.2021 – Accepted: 09.07.2021

measures vary together. All the features having 0 betweenness value are ranked again with respect to degree centrality and final ranking results are shown in Table 3. Now, the features, which were ranked 0 by C_B are ranked again according to C_d and they are ranked as (6, 8, 9, 2, 3, and 5) where feature 6 has given high rank and feature 5 is ranked lowest among features. After features' ranking, irrelevant features are removed step by step by applying wrapper approach.

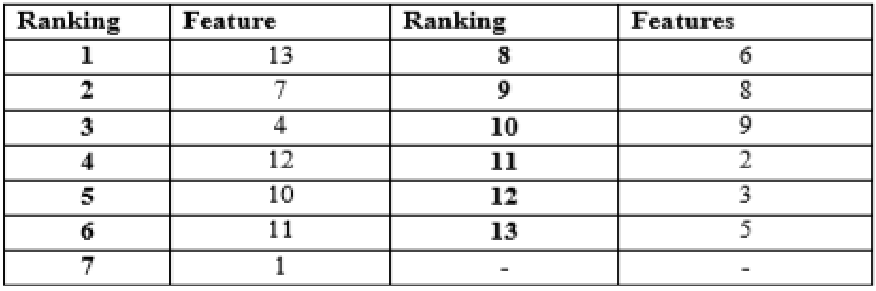
Wrapper approach is dependent on classifier for selecting features. In our approach, we chose classification accuracy as feature selection criteria. The results of IBK and J48 classifiers for wine dataset are listed in Table 4. We started with full feature set, ranked according to centrality value and calculate accuracy. In first iteration, top ranked feature is removed and accuracy is calculated with the classifier and check the relevance of this feature according to Equation 2. In Table 4, for J48 classifier, removal of feature 13 results in decreasing accuracy, so the feature will be re-added in the subset. The removal of feature "7" in second iteration also results in decreasing classification accuracy, so this feature is also restored but in the third iteration, removal of feature 4 results in increasing accuracy and is removed from subset. Letter "R" is mentioned for removed features and "Re-add" for features included in final subset under the column "Status of feature". This process continues till the last features. Final subset of feature only contains relevant features. In case of IBK, accuracy of 97.753% is achieved with 5 features (13,12,11,1,5) and in case of J48 classifier, accuracy of 96.067% is achieved with 3 features (13,7,10).

Table II - Features Ranking According to C_B Score of Wine Dataset



Feature	Betweenness value	Feature	Betweenness value
13	34	2	0
7	32	3	0
4	22	5	0
12	8	6	0
10	6	8	0
11	4	9	0
1	2	-	-

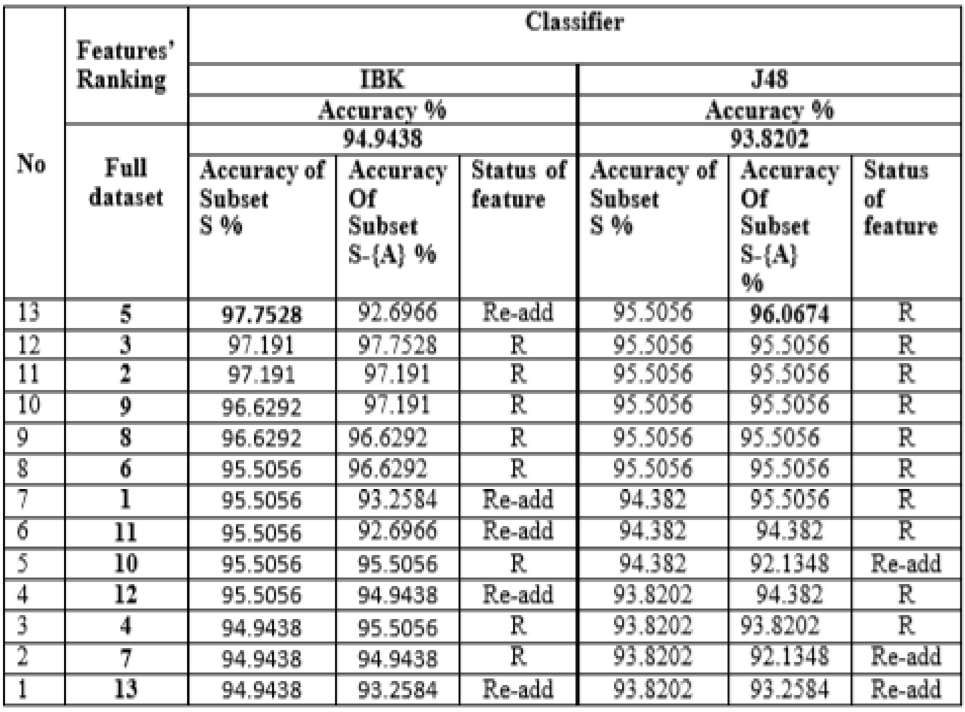
Table III - Ranking Features of Wine Dataset



Ranking	Feature	Ranking	Features
1	13	8	6
2	7	9	8
3	4	10	9
4	12	11	2
5	10	12	3
6	11	13	5
7	1	-	-

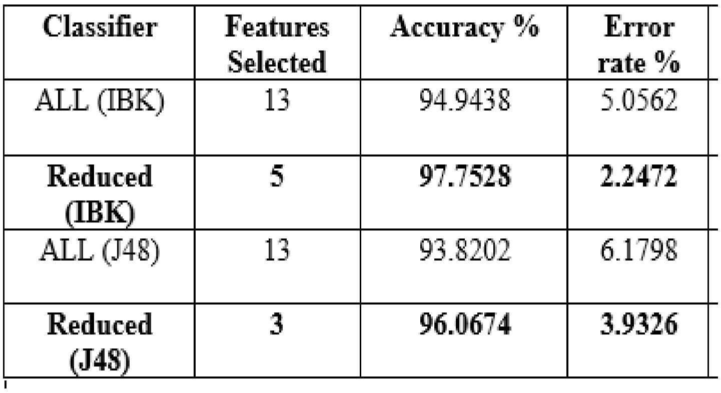
Full dataset of features and reduced dataset of features are tested by using WEKA's classifiers namely decision tree and k-nearest neighbor. In WEKA, the classifiers used for k-nearest neighbor and decision tree are IBK and J48 respectively. All the parameters' settings are default. With both classifiers, reduction of features results in increase in accuracy. In IBK, features are reduced from 13 to 5 and achieve accuracy of 97.75%. Similarly, with J48 classifier feature size is reduced to 3 and accuracy achieved is 96.07% which is greater than accuracy of full dataset. Error rate also decrease after removing irrelevant features. Results of these measures are shown in Table 5.

Table IV- Selection of Relevant Features of Wine Dataset



No	Features' Ranking	Classifier					
		IBK			J48		
		Accuracy %			Accuracy %		
	Full dataset	94.9438			93.8202		
		Accuracy of Subset S %	Accuracy Of Subset S-{A} %	Status of feature	Accuracy of Subset S %	Accuracy Of Subset S-{A} %	Status of feature
13	**5**	**97.7528**	92.6966	Re-add	95.5056	**96.0674**	R
12	**3**	97.191	97.7528	R	95.5056	95.5056	R
11	**2**	97.191	97.191	R	95.5056	95.5056	R
10	**9**	96.6292	97.191	R	95.5056	95.5056	R
9	**8**	96.6292	96.6292	R	95.5056	95.5056	R
8	**6**	95.5056	96.6292	R	95.5056	95.5056	R
7	**1**	95.5056	93.2584	Re-add	94.382	95.5056	R
6	**11**	95.5056	92.6966	Re-add	94.382	94.382	R
5	**10**	95.5056	95.5056	R	94.382	92.1348	Re-add
4	**12**	95.5056	94.9438	Re-add	93.8202	94.382	R
3	**4**	94.9438	95.5056	R	93.8202	93.8202	R
2	**7**	94.9438	94.9438	R	93.8202	92.1348	Re-add
1	**13**	94.9438	93.2584	Re-add	93.8202	93.2584	Re-add

Table V - Performance of Proposed Approach on Wine Dataset



Classifier	Features Selected	Accuracy %	Error rate %
ALL (IBK)	13	94.9438	5.0562
Reduced (IBK)	**5**	**97.7528**	**2.2472**
ALL (J48)	13	93.8202	6.1798
Reduced (J48)	**3**	**96.0674**	**3.9326**

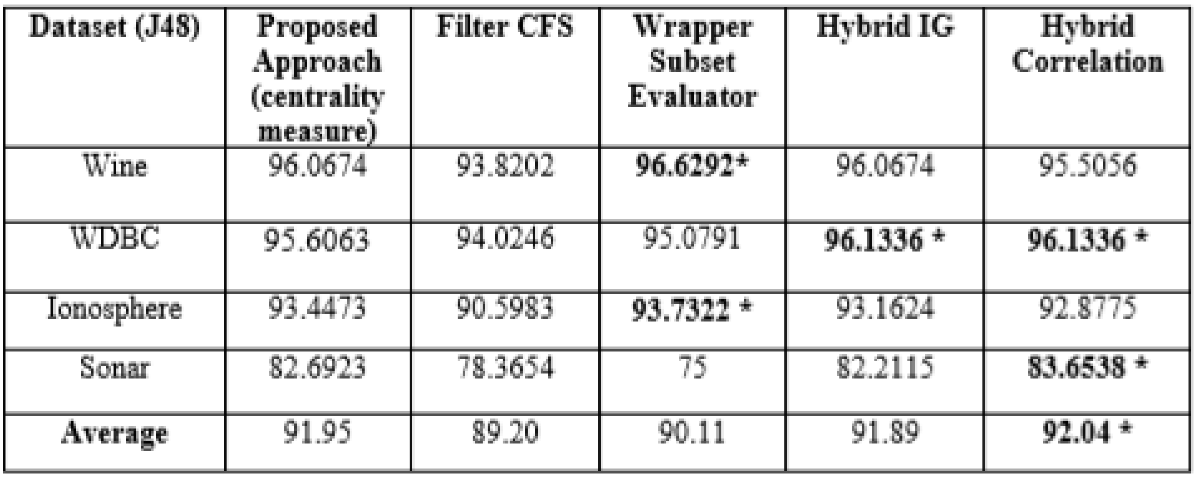
5. Results Comparison

A. Classification Accuracy

The results of proposed approach and other filter, wrapper and hybrid approaches are summarized in Table 6 and 7 using J48 and IBK algorithms respectively. Filter CFS is correlation based feature selection technique. This approach check for the features that are highly correlated with the class but have minimum correlation in between features. The algorithm used in WEKA for wrapper approach is Wrapper subset evaluator. The hybrid approaches used for comparison are hybrid CFS and hybrid information gain. In these approaches, features are ranked according to correlation and information gain respectively. After ranking features, machine learning algorithms are applied to remove irrelevant features. The asterisk sign in this table highlights the approaches having good accuracy against each dataset.

The results shows that accuracy of wrapper approach is greater for some datasets because it is based on generating all possible subsets. Generation of all subsets results in high computational cost. The accuracy of filter approach is less than other approaches due to the reason that it doesn’t depend on classifier, so results in less classification accuracy. In case of J48, among hybrid approaches, for wine and ionosphere dataset, our proposed approach shows better results and for other two datasets results of hybrid correlation are better. In case of IBK classifier, our proposed approach gives better results for all datasets. The results depend on the choice and order of feature chosen, so it is difficult to tell the exact reason for the difference in accuracy.

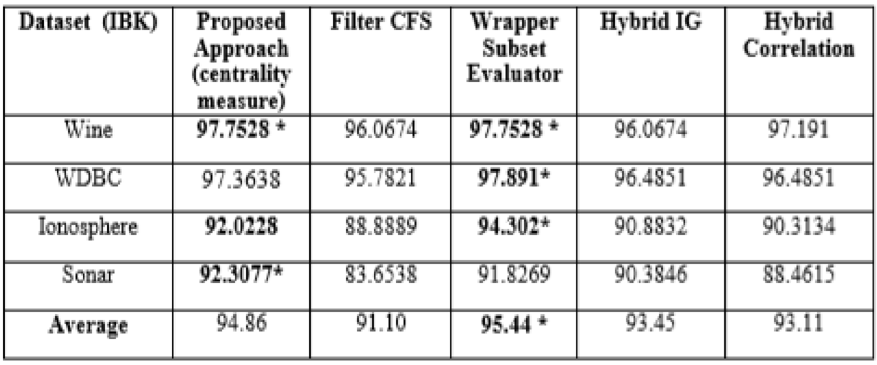
Table VI - Comparison of Accuracy of FS Approaches using J48



Dataset (J48)	Proposed Approach (centrality measure)	Filter CFS	Wrapper Subset Evaluator	Hybrid IG	Hybrid Correlation
Wine	96.0674	93.8202	**96.6292***	96.0674	95.5056
WDBC	95.6063	94.0246	95.0791	**96.1336 ***	**96.1336 ***
Ionosphere	93.4473	90.5983	**93.7322 ***	93.1624	92.8775
Sonar	82.6923	78.3654	75	82.2115	**83.6538 ***
Average	91.95	89.20	90.11	91.89	**92.04 ***

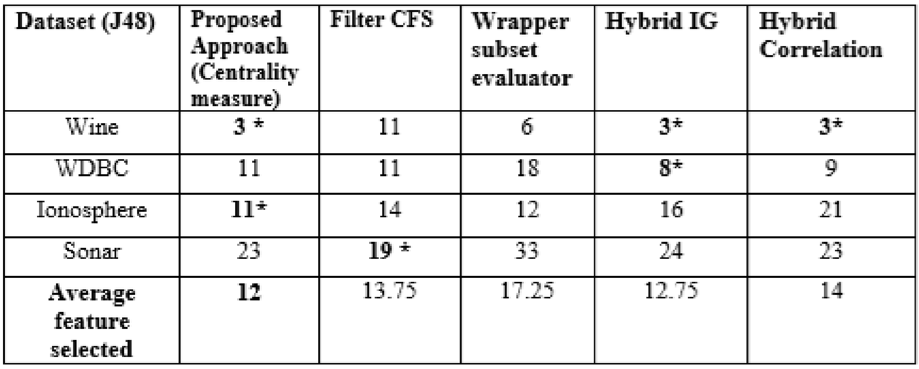
ISSN: 2237-0722
Vol. 11 No. 4 (2021)
Received: 08.06.2021 – Accepted: 09.07.2021

Table VII - Comparison of Accuray of FS apporaches using IBK



Dataset (IBK)	Proposed Approach (centrality measure)	Filter CFS	Wrapper Subset Evaluator	Hybrid IG	Hybrid Correlation
Wine	**97.7528 ***	96.0674	**97.7528 ***	96.0674	97.191
WDBC	97.3638	95.7821	**97.891***	96.4851	96.4851
Ionosphere	**92.0228**	88.8889	**94.302***	90.8832	90.3134
Sonar	**92.3077***	83.6538	91.8269	90.3846	88.4615
Average	94.86	91.10	**95.44 ***	93.45	93.11

Table VIII - Comparison of Number of features selected using J48



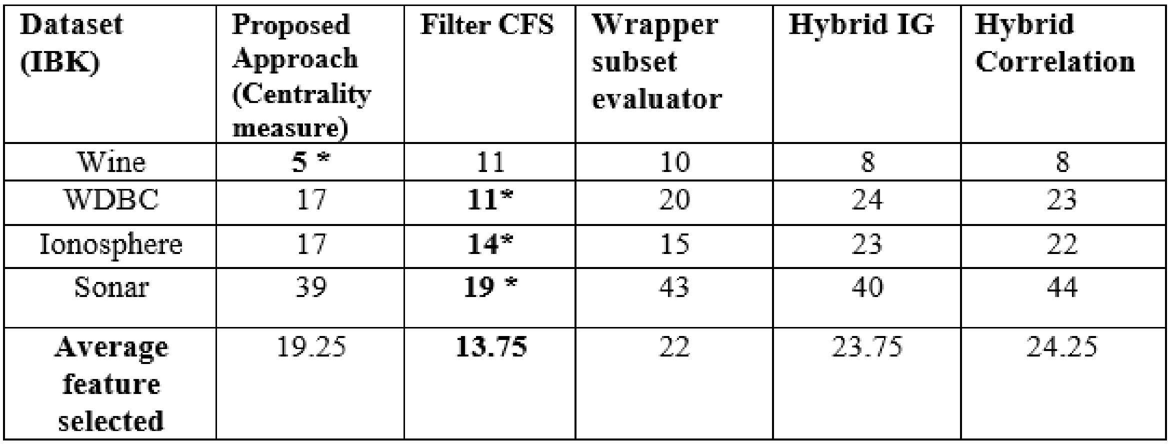
Dataset (J48)	Proposed Approach (Centrality measure)	Filter CFS	Wrapper subset evaluator	Hybrid IG	Hybrid Correlation
Wine	**3 ***	11	6	**3***	**3***
WDBC	11	11	18	**8***	9
Ionosphere	**11***	14	12	16	21
Sonar	23	**19 ***	33	24	23
Average feature selected	**12**	13.75	17.25	12.75	14

B. Number of Features Selected

The increase in size of data demands the powerful processing tool for extracting information from it. Each of the FS approach selects different number of features. The results in Table 8 represent the number of features selected by each approach using J48 classifiers. The asterisk sign is used to highlight the approach which selects less number of features. Although accuracy of wrapper approach is more than other approaches but the number of features removed by this approach are less as compared to other FS approaches discussed in this paper. CFS is independent of learning algorithm, the number of features selected in case of sonar dataset are less but the accuracy is also low i.e. 78.3654% as compared to other approaches. Among hybrid approaches, our approach selects less number of features for wine, ionosphere and sonar dataset but for WDBC, hybrid IG approach selects less features i.e. 8. On an average, for all the four datasets, our proposed approach selects less number of features that is 12.

Table 9 shows the number of features selected by each FS approach using 4 different datasets. The classifier used is IBK for selecting relevant features. The accuracy of wrapper subset evaluator is greater than other FS approaches but according to Table 9, number of features removed by this approach is less. For WDBC, ionosphere, and sonar dataset, the CFS approach removes more features but the accuracy achieved for this approach is low. Among hybrid approaches, the number of features removed by our approach is more than other two hybrid approaches. Our proposed hybrid approach selects less number of features among wrapper and other two hybrid approaches. While among all approaches, CFS selects least number of features by compromising the accuracy value. We can conclude from our results that ranking obtained from graph based centrality measures can perform better than the ranking obtained from statistical measures like information gain, correlation etc. as our graph based approach performs better than other two hybrid approaches in most of the cases.

Table IX - Comparison of Number of features selected using IBK



Dataset (IBK)	**Proposed Approach (Centrality measure)**	**Filter CFS**	**Wrapper subset evaluator**	**Hybrid IG**	**Hybrid Correlation**
Wine	**5 ***	11	10	8	8
WDBC	17	**11***	20	24	23
Ionosphere	17	**14***	15	23	22
Sonar	39	**19 ***	43	40	44
Average feature selected	19.25	**13.75**	22	23.75	24.25

6. Conclusion

Feature selection is an important pre-processing step for removing irrelevant data. Removal of irrelevant data helps to reduce computational complexity problem which data scientists has to face due to data explosion. It also helps the analyst to focus on only relevant features by ignoring the irrelevant features. This work proposed a graph based method for selecting relevant features. The method comprises of three main phases including representation of features in the form of graphs, calculation of centrality measures for ranking features and the last phase includes the use of learning algorithm for the removal of irrelevant features. Decision tree and nearest neighbour algorithms were used for calculating relevancy of features on the basis of accuracy. Finally after features were selected, the results of proposed approach were compared against four other feature selection approaches. The results show that in most cases, the filter base feature selection approach results in removal of more features but with the compromise on accuracy. Its accuracy was found less than all other approaches discussed in this study. Use of learning algorithm increased the accuracy of wrapper approach and the number of selected features was greater in this case. It has been found that graph based feature selection approach results in deletion of large number of irrelevant features than other two hybrid approaches and also the classification accuracy achieved was better than other hybrid approaches and filter approach studied in this work. So, it can be concluded that graph based measures can also be used in the process of feature selection without compromising the accuracy and can perform better than other approaches.

References

B. Marr, "Big data: The eye-opening facts everyone should know," 25 September 2014. https://www.linkedin.com/pulse/20140925030713-64875646-big-data-the-eye-opening-facts-everyone-should-know

P. Zikopoulos, C. Eaton, D. De Roos, G. Lapis and T. Deutsch, Understanding big data: Analytics for enterprise class hadoop and streaming data, McGraw-Hill Osborne Media, 2011.

H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering,* 491-502, 2005.

R. Kohavi and G.H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* 273-324, 1997.

K. Kira and L.A. Rendell, "Kira, K & Rendell, LA 1992. A practical approach to feature selection," In *Proceedings of the ninth international workshop on Machine learning,* Aberdeen, Scotland, UK, 1992.

I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF," *In Proceeding of European Conference on Machine Learning,* Catania, 1994.

M.A. Hall and L.A. Smith, "Practical feature subset selection for machine learning," In Hall, M.A. & Smith, L. A. 98 *Proceedings of the 21st Australasian Computer Science Conference ACSC,* Berlin, 1998.

H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1226-1238, 2005.

Y. Saeys, I. Inza and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* 2507-2517, 2007.

M.W. Mwadulo, "A review on feature selection methods for classification tasks," *International Journal of Computer Applications Technology and Research,* 395 – 402, 2016.

R.K. a. P.P.G.H. John, "Irrelevant features and the subset selection problem.," *In Machine Learning: Proceedings of the Eleventh International Conference,* 1994.

S. Maldonado and R. Weber, "A Wrapper method for feature selection using support vector machines," *Information Science,* 2208-2217, 2009.

L. Zhuo, J. Zheng, F. Wang, X. Li, B. Ai and J. Qian, "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine," *Geographical Research,* 493-501, 2008.

M.S. Pervez and D.M. Farid, "Literature review of feature selection for mining tasks," *International Journal of Computer Applications,* pp. 30-33, 2005.

P. Bermejo, J.A. G´amez and J.M. Puerta, "On incremental wrapper-based attribute selection: experimental analysis of the relevance criteria," *in Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems,* Melaga, 2008.

R. Ruiz, J.C. Riquelme and J.S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition,* 2383--2392, 2006.

S. Huda, J. Yearwood and A. Stranieri, "Hybrid wrapper-filter approaches for input feature selection using maximum relevance-minimum redundancy and artificial neural network input gain measurement approximation (ANNIGMA)," *in Proceedings of the Thirty-Fourth Australasian Computer Science Conference,* Darlinghurst, Australia, Australia, 2011.

S. Gunal, "Hybrid feature selection for text classification," *Turkish Journal of Electrical Engineering and Computer Sciences,* 1296-1311, 2012.

J. Tang, S. Alelyani and H. Liu, "Feature selection for classification: A review," *Data Classification: Algorithms and Applications,* 2014.

Y. Keneshloo and S. Yazdani, "A relative feature selection algorithm for graph classification," Advances in Databases and Information Systems, pp. 137-148, 2013.

S. Azadifar and S.A. Monadjemi, "Feature selection using social network technique," *in 7th Conference on Information and Knowledge Technology (IKT),* Urmia, 2015.

A.R. Noruzi and H.R. Saheb, "A graph-based feature selection method for improving medical diagnosis," *Advances in Computer Science: an International Journal (ACSIJ),* 36-40, 2015.

Z. Zhang and E.R. Hancock, "A graph-based approach to feature selection," *in Graph-Based Representations in Pattern Recognition,* 2011.

D.S. Rani, T.S. Rani and S.D. Bhavani, "Feature subset selection using consensus clustering," *in Eighth International Conference on Advances in Pattern Recognition (ICAPR),* Kolkata, 2015.

D. Ienco, R. Meo and M. Botta, "Using Page Rank in feature selection," *in Proceedings of the Sixteenth Italian Symposium on Advanced Database Systems,* Mondello, PA, 2008.

G. Roffo and S. Melzi, "Ranking to learn: Feature ranking and selection via eigenvector centrality," New Frontiers in Mining Complex Patterns, *Fifth International workshop, nfMCP2016. Lecture Notes in Computer Science,* 19-35, 2017.

F.L. Coolidge and F.L. *Coolidge, Statistics: A gentle introduction,* SAGE, 2012.

J.D. Evans*, Straight forward statistics for the behavioural sciences,* Pacific Grove: Brooks/Cole Pub. Co., 1996.

S.P. Borgatti, "Centrality and network flow," in Social Networks, 2005.

T. Opsahl, F. Agneessens and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks,* 245-251, 2010.

B. Cheng, "Using social network analysis to investigate potential bias in editorial peer review in core *journals of comparative/international education (Doctoral dissertation),*" 2006.

M.M. Islam, M.M. Kabir and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, 3273-3283, 2010.

X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, 1–37, 2008.

M.A. Hall (2), "Correlation-based feature selection for machine learning (Doctoral dissertation)," April 1999. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Received: 08.06.2021 – Accepted: 09.07.2021