Review of feature selection methods for text classification

Article *in* International Journal of Advanced Computer Research · July 2020

DOI: 10.19101/IJACR.2020.1048037

CITATIONS
11

READS
2,839

4 authors:

Muhammad Iqbal
Bahria University Karachi Campus
32 PUBLICATIONS **256** CITATIONS
SEE PROFILE

Malik Muneeb Abid
University of Wah
55 PUBLICATIONS **217** CITATIONS
SEE PROFILE

Muhammad Noman
Bahria University Karachi Campus
14 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Engr. Dr. Amir Manzoor
Karachi School for Business and Leadership
92 PUBLICATIONS **253** CITATIONS
SEE PROFILE

Research Article

International Journal of Advanced Computer Research, Vol 10(49)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2020.1048037

Review of feature selection methods for text classification

Muhammad Iqbal[1*], Malik Muneeb Abid[2], Muhammad Noman Khalid[3] and Amir Manzoor[4]
School of Information Sciences &Technology, Southwest Jiaotong University, Chengdu, PR China[1]
Department of Civil Engineering, College of Engineering and Technology, University of Sargodha, Pakistan[2]
Department of Computer Science, Bahria University Karachi[3]
Department of Management Sciences, Bahria University, Karachi Campus[4]

Received: 07-April-2020; Revised: 20-July-2020; Accepted: 23-July-2020

Abstract
For the last three decades, the World Wide Web (WWW) has become one of the most widely used podium to generate an immense amount of heterogeneous data in a single day. Presently, many organizations aimed to process their domain data for taking quick decisions to improve their organizational performance. However, high dimensionality in datasets is a biggest obstacle for researchers and domain engineers to achieve their desired performance through their selected machine learning (ML) algorithms. In ML, feature selection is a core concept used for selecting most relevant features of high dimension data and thus improve the performance of the trained learning model. Moreover, the feature selection process also provides an effective way by eliminating in appropriate and redundant features and ultimately shrinks the computational time. Due to the significance and applications of feature selection, it has become a well-researched area of ML. Nowadays, feature selection has a vital role in most of the effective spam detection systems, pattern recognition systems, automated organization, management of documents, and information retrieval systems. In order to do accurate classification, the relevant feature selection is the most important task, and to achieve its objectives, this study starts with an overview of text classification. This overview is then followed by a survey. The survey covered the popular feature selection methods commonly used for text classification. This survey also sheds light on applications of feature selection methods. The focus of this study is three feature selection algorithms, i.e., Principal Component Analysis (PCA), Chi-Square (CS) and Information Gain (IG). This study is helpful for researchers looking for some suitable criterion to decide the suitable technique to be used for better understanding of the performance of the classifier. In order to conduct experiments, web spam uk2007 dataset is considered. Ten, twenty, thirty, and forty features were selected as an optimal subset from web spam uk2007 dataset. Among all three feature selection algorithms, CS and IG had highest F1Score (F-measure =0.911) but at the same time suffered with model building time.

Keywords
Feature selection, Binary classification, Feature selection algorithms.

1. Background
With the immense growth of online information due to Internet, text-categorization has become a very significant technology to classify a large number of documents to make them useful. The concept of text classification is to categorize textual data into one or more predefined categories [1−3]. It is not possible to manage manually these big datasets. As a result, we witness the development of various machine learning and data mining methods. These methods are being used for automatic recognition of the patterns in large data sets for gaining useful information.

We can also define this problem as: assigning a document D to some pre-determined set of categories |C|. Where pre-determined set of categories, C, is defined by equation 1.

$$C = \{c_1, c_2, c_{3,\ldots..} c_{|c|}\} \quad (1)$$

In short, text classification is used for categorizing documents into predefined classes based on their contents. This is an automatic assignment process for text categorization. Text classification is the initial requirement of Text Retrieval Systems (TRS), which digs texts in response to user queries. Nowadays, different ML algorithms are in practice to manage

*Author for correspondence

and organize documents for Information Retrieval Systems (IRS) [4].

The biggest challenge for existing supervised learning methods to classify automatically text is that they require sufficient documents to learn accurately.

In most of the cases, a supervised learning approach is used for training a text classifier, where a learning algorithm provides a set of Labeled training instances to process the data. The labeled training instances for term L is defined in equation 2.

$$L = \{(d_i, c_i): \text{where } i = 1,2..L\} \quad (2)$$

Where d_i representing i^{th} training document and c_i is the matching category label of d_i.

After that the classification algorithm produces a prediction function P that maps documents to intended categories, as presented in equation 3 below:

$$P : D \rightarrow C \quad (3)$$

For improving the classifier's accuracy of generalization (e.g. resisting the curse of dimensionality to improve prediction performance), it is common practice to acquire feature selection algorithms. The feature selection is a process to select an optimum subset from larger pool of feature set. *Figure 1* depicts the generic process of feature selection. The searching process on the feature pool is applied in the procedure of creating subsets to obtain best performance from classifier. Searching algorithms could be heuristic search, complete search and random search.

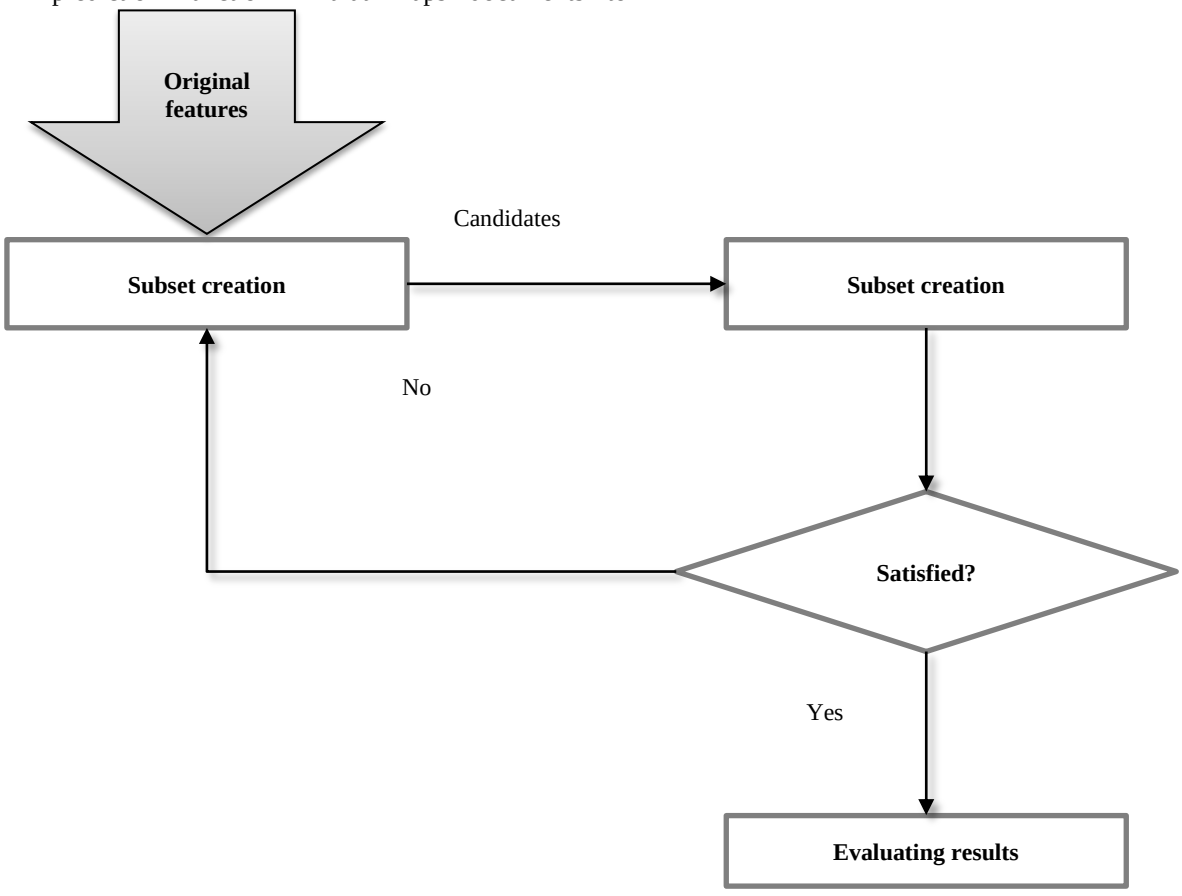

Figure 1 Feature selection process

The term curse of dimensionality refers to a phenomenon. Curse of dimensionality describes the consequences of performance and accuracy due to high-dimensional data in corpus. This problem originates when the used algorithm does not scale well to high-dimensional data. Time and memory are the two exponential variables in processing of data. In most of the cases, the maximum number of

features results in degradation of classifier performance. Studies [5, 6] had pointed out that only solution to this problem can be found by changing the algorithm, or bring data into a lower dimensional space by pre-processing. Feature selection and extraction are the two approaches for dimensionality reduction [7, 8]. The output of the process of feature selection is a subset of available features; while feature extraction refers to a method that digs out subset of new features by inserting some new rules on the existing features (See *Figure 2*).

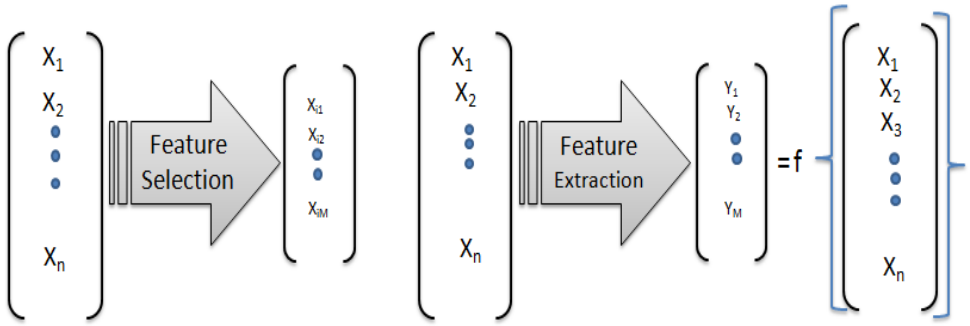

Figure 2 Feature selection and extraction process

The Bayes decision rule deals with the initial space R^N and the reduced space R^M of feature extraction. According to Bayes decision rule, given a feature space xi ∈ R^n find a mapping in:

$$y = f(x): R^N \rightarrow R^M \quad (4)$$

Where M < N.

This transformed feature vector $y_{i\in}R^M$ continues (most of) the information or structure in R^N. An optimal mapping y=f(x) will be one that can result into no addition to the minimum probability of error.

In spite of numerous texts classification applications, spam traffic become a serious challenge for search engines and filtering systems. Better detection of these unwanted messages can only be obtained from improved text classification techniques.

Dimensionality Reduction (DR) is an important task in machine learning to reduce the number of random variables under consideration. DR can be divided into feature selection and feature extraction process (see *Figure 3*).

Figure 3 shows many applications of feature selection and feature extraction techniques. These techniques include text mining, image processing, ECG Signals etc. To obtain better results in a text classification, we have to identify the relevant features from data. Some other approaches, like pruning and clustering, can be used to reduce dimensionality. These techniques can also be helpful in improving performance of text classification.

This study provides some unique contributions to the existing knowledge base. First, this is a comprehensive study conducted to validate the importance of feature selection for achieving prominent results. Second, authors have discussed some typical applications of machine learning; which needs the feature selection process. Third, the discriminative capability of different feature selection process is discussed.

The rest of the paper is structured as follows. In Section 2, we presented related work. In section 3, we have discussed the taxonomy of feature selection approaches. Section 4 covers the applications of feature selection. In section 5, we describe the challenges faced to retrieve appropriate features. In section 6, we present a performance evaluation for comparing three well-known machine-learning approaches. Finally, in section 7 we present concluding remarks.

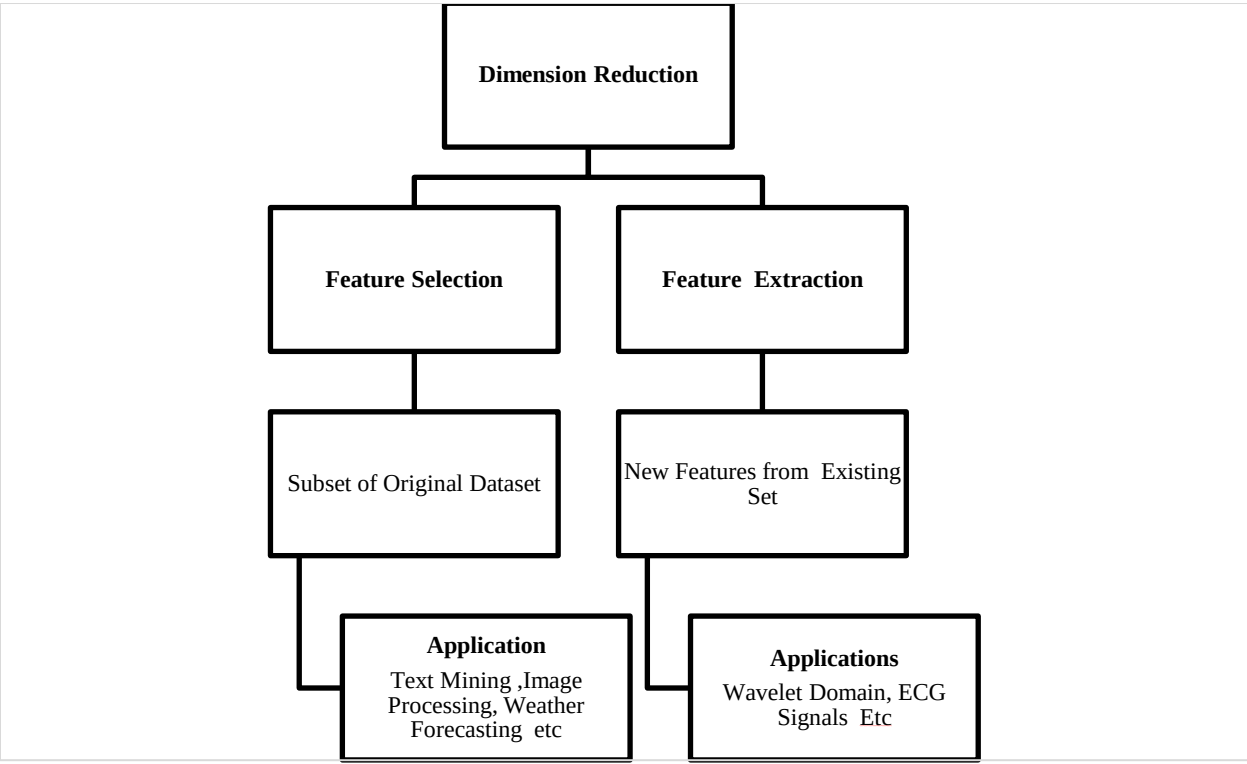

Figure 3 Approaches used for reduction of dimensions

2. Related work

Gilbert et al. [9] coined the idea to use ML algorithms for detection of cybercrime threats. Cybercrimes is an important issue due to its growing global scale and significant impact on people lives and national/global economies. Researchers have also highlighted the importance of unique feature selection for accurate classification of network traffic for malign or benign. The technique of feature selection is one of the active research areas in the domain of data mining. The technique of feature selection has been applied in many domains including image retrieval [10,11],intrusion detection [12], text mining [13],topic detection[14],spam email filtering[15], SMS spam filtering[16], author identification[17], web page classification[18] and sentiment analysis[19].The most well-known application of text classification is web spam filtering. Search engines may also take advantage of text classification techniques to return more accurate and desired results to the end user [20].

Nowadays, two most famous approaches for text classification are supervised and un-supervised text classification. As name implies, a supervise machine learning algorithm needs supervisor. A chunk of training data and chunk of test data is needed, during the training phase. The training phase is supervised as the correct categories or classes have to be assigned first. In un-supervised machine learning methods, the applied algorithm extracts the rules or attributes for classification.

For the accurate categorization of documents, there exist many algorithms of machine learning. Some of such important algorithms are Support Vector Machine (SVM), decision trees, inductive learning, neural networks, k-Nearest Neighbors, and Naïve Bayes classifiers. These algorithms have been used extensively to develop document categorization systems.

A comprehensive survey on text classification was conducted by Sebastiani [21]. Sebastiani discussed three different areas, i.e. document representation, classifier construction, and classifier evaluation. Additionally, the writer reported the new domains of applications, like dealing with noisy data in optical character recognition and the use of automatic text classification in speech recognition.

In their study, Kohavi and John [22] compared different feature selection techniques. They found that chi square statistics and information gain (IG)

were the most effective parameters for optimizing the results of classification. There is a vast literature available on feature selection methods; for instance, see, [23─25] for thorough analysis of feature selection methods and many references.

3. Taxonomy of feature selection

As discussed in background section of the paper that there are two approaches to solve dimension reduction curse from large data sets i.e. feature selection and feature extraction. Feature selection is a well-studied problem and its main objective is to speed up classifier results with accuracy. There are three main types (See *Figure 4*) of feature selection methods; which are mostly discussed in different studies[26, 27].The filter technique uses independent mining algorithms along with general characteristics of training cases [28].The Wrapper technique is based on the analysis of relationship between feature subset selection and relevance. Here, the feature subset selection refers to the optimal feature subset selection. The wrapper technique attempts to find optimal feature subset that can be adapted to some specific mining algorithm [29]. The embedded technique builds on independent criteria. This criterion is used to select optimal feature subsets that have known cardinality [29]. Wrappers and embedded approaches need a frequent classifier interaction during the construction of feature set, while the filter scheme does not interact with classifier during feature construction process. In the feature selection process, more time is required for classifier interaction. The classifier interaction may also be used to adapt feature selection technique to a particular learning model [30].

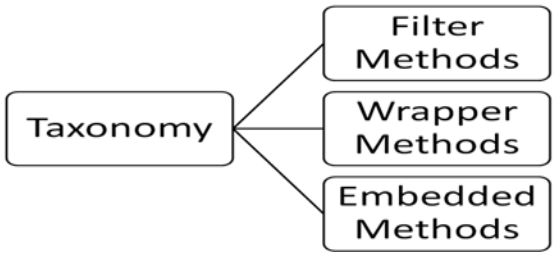

Figure 4 Classification of feature selection

The feature selection process can be thought of as selection of best words from a document that can help to categorize that document. This process consists of four major steps. These steps are generation of subset, evaluation of subset, criteria for stopping, and validating results [31].

Various criteria of feature evaluation are used to select optimal subsets in the feature selection process. The algorithms used in each evaluation criterion may produce different results for a given feature subset. The five important feature evaluation criteria include measures of information, measures of distance, measures of dependency, measures of consistency, and measures of classification error-rate [32].

Feature selection technique is also called attribute or variable selection technique; which is conceived to be a very significant preprocessing step in ML domain. Better selection of features will lead to construction of better model, while irrelevant and redundant features could give poor results to classifier and subsequently will take more computational time [32]. Therefore, the primary objective of most of the researchers in feature selection process is to explore relevant features and remove irrelevant and redundant instances. Through this process, we can speed up the data mining algorithms. This increased speed of algorithms can ultimately improve predictive accuracy of classifier. Many induction algorithms are already incorporating different approaches to eliminate the irrelevant features.

3.1Filter approach

The filter approach is based on the selection of variables regardless of induction model. The feature selection is based on the general characteristics of variables i.e., distance between classes and statistical dependencies i.e., statistical score value is assigned to each feature. This method is computationally efficient because it acts independently to induction method. However, this technique is blind with any interaction among other features. This technique can ignore features that are not important by themselves. These features, when combined with other features, can be important [33]. Filter approach can be further divided into two categories i.e., global and local. This categorization is based on the fact that whether a singular or multi class-based score is assigned to any feature [34]. Chi squared test, information gain, document frequency, improved Gini index and correlation coefficient scores are some good examples of filter method. *Figure 5* depicts the working of the filter model.

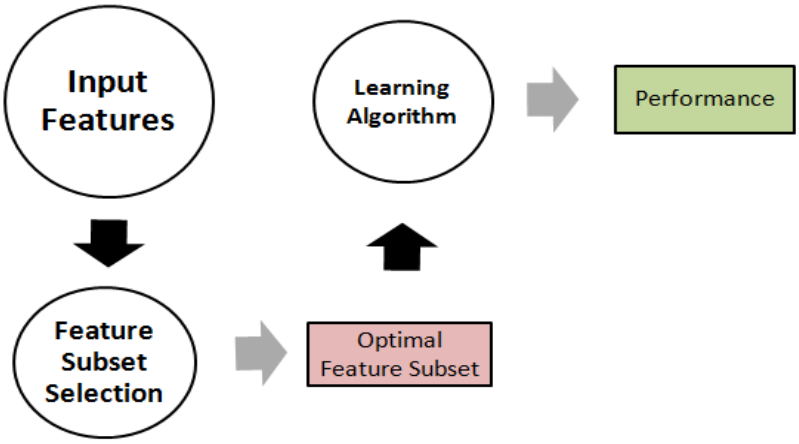

Figure 5 The Process of filter model

Description of a general algorithm of filter approach can be mathematically defined as:
For given data set Ds={F, L} (where F and L represents the feature set and labels respectively)
The algorithm begins with $F^{'}$, such as $F^{'}$= {Ø} or $F^{'} = \{NULL\}$or $F^{'} \subset F$. Mi(Independent Measure) is used to evaluate each generated subset F_g and its comparison with the value of the last optimal subset. The search stops when the stopping criterion Ø is met. The outcome of this process is F_{opt}, the current optimal feature subset.

3.2Wrapper method

Wrapper method is based on Support Vector Machine (SVM) algorithm. It is one of the most widely studied technique in the domain of machine learning. A learning algorithm is used by wrapper method for subset evaluation. *Figure 6* demonstrate the general wrapper approach.

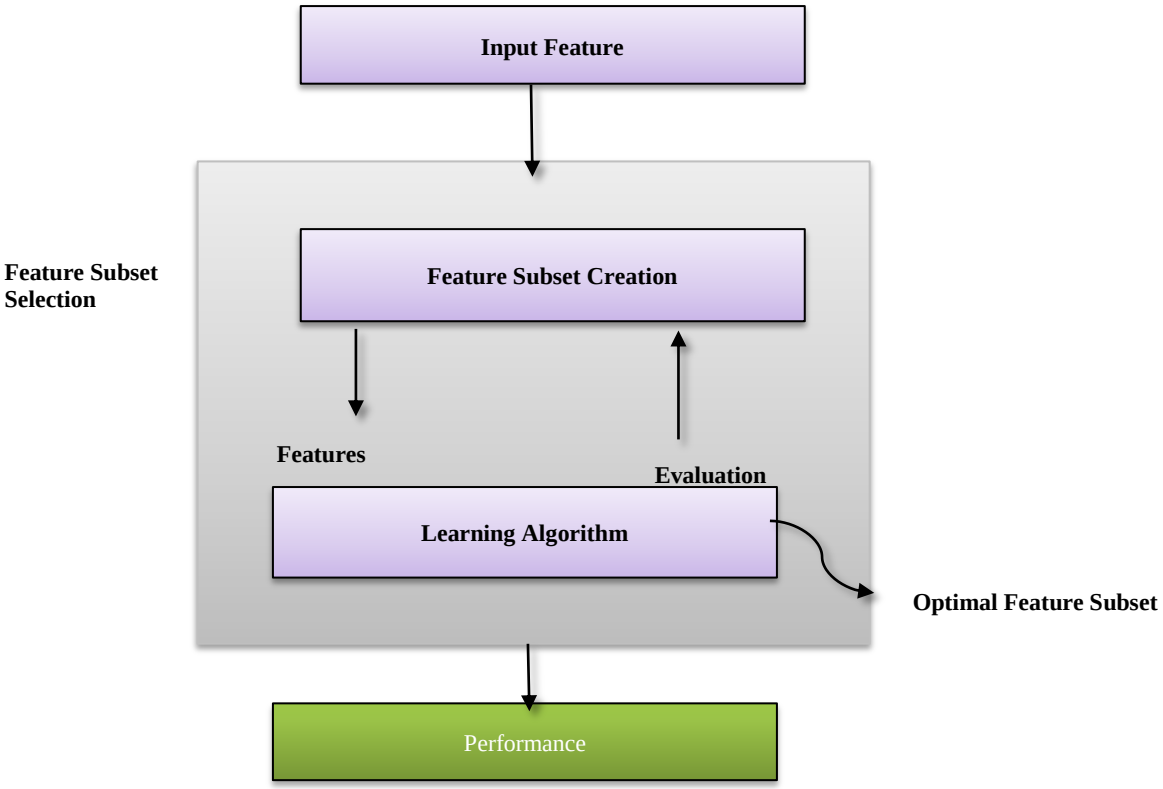

Figure 6 The Process of wrapper model

The wrapper method uses a predefined classifier to search for a subset of features. It then uses this classifier to evaluate the selected subset of features. The selection and evaluation of subsets of features continues until the desired criterion of quality is met.

In wrapper method, the classifiers are predefined and considered as black box. During the search step, the wrapper method produces a set of features. During the evaluation step, the predefined classifier is used to estimate the performance of the subset of features. The results of this step are used to refine the search step in the next iteration of the process. The final feature set is the one whose estimated score is the highest. This feature set is then used to learn the classifier. The resultant classifier is evaluated using an independent feature set. This feature set is the one not used during the training [35]. O(2m) is the size of search space for "m" features in wrapper technique. Therefore, to perform an exhaustive search one needs a small value of "m".

3.3Embedded models

The wrapper technique uses a predefined classifier to evaluate the quality of the subset of features. Multiple time execution of classifier to assess the selected subsets of features makes wrapper model computationally expensive [36]. Embedded method requires less computational cost to interact with learning algorithm as compared to wrapper model. This method is also related to feature dependencies. This technique takes into account the relationship between input/output features. This technique also provides a selection of optimal feature subsets for a known cardinality using an independent criterion [37]. *Figure 7* illustrate the process of feature selection. The embedded model looks into the features that contribute the most to the performance of classification algorithm.

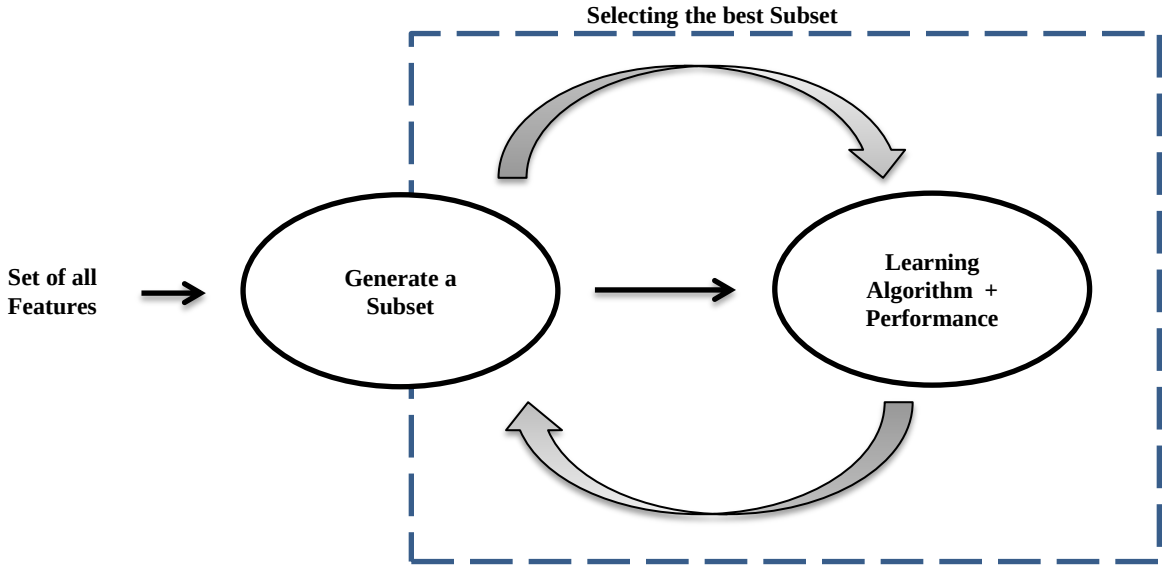

Figure 7 The process of an embedded model

The embedded method technique uses a sequential forward selection method. At the beginning, we have an empty set of features. All possible subsets of cardinality are searched. This search is based on allocation of feature value. This feature value is derived from the remaining feature subsets. Once an optimal subset of cardinality is found, this subset is evaluated using an independent criterion. The evaluation result is compared with the evaluation results of the previous optimal subset of cardinality. The embedded method then uses a learning algorithm to evaluate the performance of the current optimal subset. The results of this performance evaluation are then compared with the performance evaluation results of the optimal subset at cardinality. The output of the embedded method is a final optimal subset of features according to stopping criterion**.**

4. Applications of feature selection

Usually, a number of problems are encountered during data collection. These problems may include high level of feature dependency, a large number of features, and irrelevant features. The process of feature selection deals with these problems using a

ML tools. These tools help to select feature or feature subset that can be used for effective learning of algorithms [38]. Presently, various applications use feature selection process. The consideration of appropriate feature selection in ML is a subject of considerable ongoing research and plays an important role in various applications. We are presenting some successful applications in followed sections.

4.1Text categorization

The size of data on World Wide Web is growing exponentially due to different internet applications like emails, online social networks (OSN) etc. As such, the significance of automatic text categorization and clustering has increased. Kumar and Minz [39] reported that thousands of features exist in mostly every text document. Therefore, in order to use mining algorithms for classifying text documents effectively, feature selection process is considered very important. Studies exist that have used combination of feature selections process and Information Gain Ratio (IGR) for classifying various types of text data such as lyrics [40] and poems [39,41].

4.2 Remote sensing

The remote sensing image classification is another important task in feature selection process. A study [42] has highlighted and explained various issues/challenges in feature selection process and hyper-spectral remote sensing image analysis. Another study [43] proposed pre-processing techniques that can be used for classification of hyper-spectral images (HSI). These techniques emphasize on the use of feature extraction and feature selection. To improve the classification performance of hyper-spectral images, a new method was proposed by [44]. This method is based on selection of a subset of bands from a HSI.

4.3Intrusion detection

The use of internet is exponentially growing all around the world due to esteem benefits of data sharing. However, today network security is the biggest challenge for internet community. Therefore, security of computer networks is an important issue to safeguard these networks against possible intrusion from hacker/attackers [45]. Intrusion detection is one of the ways computer network security can protect computer networks. Feature selection technique is used to classify various system activities as legitimate activities or possible intrusion activities.

4.4 Image retrieval

The exercise of image collection from different sources, like military, medical and civilian equipment has increased with fast pace. These images should be engineered before they can be used for any information-related purposes. This image engineering should be done in such a way that images can be effectively browsed, retrieved, and searched [46]. Silva et al. discussed the importance of image retrieval in medical sciences; especially in diagnosis, surgical planning, therapy process, medical reference, and training [47]. Another study [48] suggested that it is possible to scale the retrieval of content-based images of large size. However, high-dimensionality can create problems. A semantic web search engine (SE) provides an ability to effectively browse, search, and retrieve the web content. A semantic web search engine is heavily dependent on accurate feature selection process for providing better services to its users.

5. Challenges with feature selection

5.1Dealing with large dimensional data

With the exponential growth of internet, the amount of data collected has increased. This data can be in various forms such as text documents, images, and videos. This variety of data indicates high dimensionality of the data denoted by the letter “D”.

A study [49] suggests that if the value of “D” reaches hundreds, the data is considered highly dimensional. The feature selection process has been applied to a very large number of features to extract useful features. Due to high dimensionality problem in large data set, a number of feature selection algorithms suffer with time complexity to maintain scalability; which is the major issue in feature selection process [49]. It is obvious that without the use of some dimensionality reduction methods i.e. feature selection and ranking techniques, it is not possible to produce accurate results. Additionally, over fitting problem can minimize the dimensionality reduction techniques.

One issue with feature selection techniques is the scalability issue of feature selection. The filter technique has less computational complexity as compared with the wrapper technique. As such, filter technique provides better scalability for feature selection process than wrapper technique. As far as the embedded approach is concerned, a study [50] suggested that embedded approach suffers from the same time complexity issue as the filter approach and proposed to use a combination of filter and wrapper

techniques for highly dimensional data environments [51].

5.2Feature selection in sparse data matrix

Sparse data refers to the instances where actual data is not present. When it comes to sparse data, literature [52] suggests that there exist two types of sparsity: controlled and random. In controlled sparsity, data is missing from a range of values of one or more dimensions. In random sparsity, the data is missing randomly in a data variable.

Various business applications such as consumer purchase analysis and email marketing suffer from the problems of sparse data. The sparse data sets include a sparse matrix having large number of attributes. Various sources of content on the Internet, such as emails, news, and customer reviews are the main sources of sparse data. Nowadays, the amount of video streaming data is increasing rapidly. This data is highly dimensional because it comes from a variety of sources such as surveillance cameras and sensors [53]. Many feature selection techniques are not suitable to deal with highly dimensional labeled or unlabeled sparse data sets. As such, the retrieval of feature selection from such data sets is a difficult and time-consuming task. Therefore, it is the need of the hour to develop efficient algorithms for feature selection process than can be used for sparse data sets [53].

6. Experiments and results

The current study focuses on various Feature Selection Techniques (FST). The FST is one of the most important and frequently used steps in data preprocessing for data mining. The main objective of any FST is to remove irrelevant features and to keep those features that hold useful information. In order to judge the performance of different feature selection approaches, we have conducted experimental work on Chi-Square (CS), Information Gain (IG), and Principle Component Analysis (PCA). These feature selection methods are precisely described below.

6.1Chi-square (CS)

The Chi-Square method is applied to determine whether the variables are related or independent. Mathematically we can define Chi-Square as follows:

$$ChiSqr(t,c) = \sum(O-E)^2)/E \qquad (5)$$

Where t is term and c represent category, O is observed frequencies and E is expected frequencies [54].

6.2 Information gain (IG)

The value of IG is a measure of the amount of information achieved by knowing the value of the attribute. A simple mathematical expression of IG is given below in equation 6.

$$IG = (E_{cd} - E_{ca}) \qquad (6)$$

Where E_{cd} represents entropy of distribution before the split, and E_{ca} represents entropy of distribution after it.

Where $E_{cd} = \sum_{cd} p(c)\log p(c)$ the entropy of selected class c and $E_{ca}=\sum_{ca} p(c|a)\log p(c|a)$ the conditional entropy of attributes of given class [55].

6.3Principal component analysis (PCA)

PCA is not considered a feature selection technique. It is regarded as a feature combination technique [56]. A PCA is performed through following steps:

- Standardization of the data
- Computing the covariance matrix
- Calculating the eigenvectors and Eigen-values of covariance matrix
- Computing the Principal Components (PC)
- Reducing the dimensions of the data set

6.4 The evaluation

Open Source Data Mining suite, WEKA was used to generate the evaluation measurement of ROC values, Precision, Recall, AUC, and F1-Mesures [57]. *Figure 8* shows the general framework for the experimental setup used in this study. Following is a list of major steps, which were taken to complete the study.

- Features were selected through IG, PCA and CS to reduce data dimension
- The model was trained i.e. learning process was performed on sample data
- The model was fine-tuned through validation data. We used K-fold cross validation method to ensure less bias in our model. The value of K used in our study was 10.
- Test dataset was used for unbiased evaluation on final model.
- Evaluation methods i.e. Precision, F-measure, AUC, building time were used to judge the performance of FST

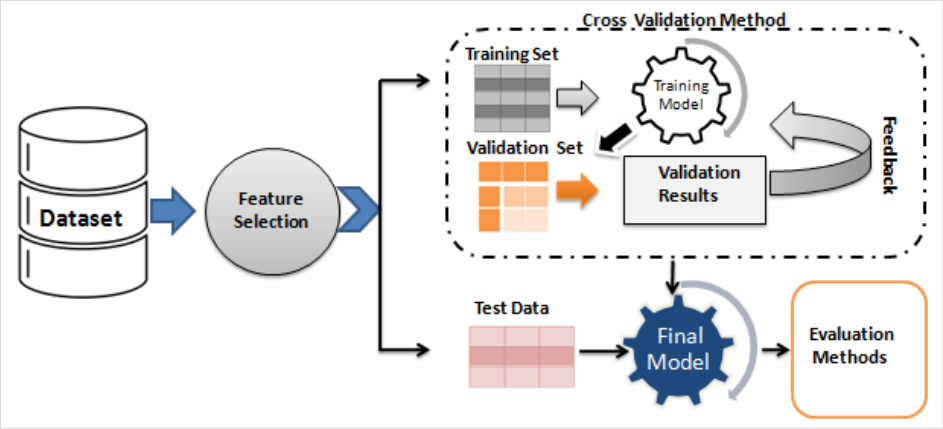

Figure 8 General framework for feature selection & evaluation of results

Experiments were performed on Web spam uk2007 dataset [58,59]. We accomplished experiment on a Windows 7 SP1 machine having configurations Intel® Core 2 Duo CPU 32-bit processor 2.93GHz, 4.00GB RAM. We selected important features for Naïve Bayes (NB) classifier through 10-fold cross validation. For comparative performance comparison of CS, IG, and PCA, we performed 12 experiments with different feature set (10 to 40) on dataset and summarized the results in *Figure 9-13*, and *Table 1*. The Naïve Byes algorithm was selected it provides it provide following benefits:

- Less of amount of training data is required
- Less computational time required to train the data
- Model size is low
- Simplicity

Effectiveness of the NB algorithm and in combination with three feature selection methods on Webspam Uk2007 dataset, we selected the different size features (See *Table 1*). The results presented in terms of Precision, Recall, and F1-Measure, AUC and Model Building time values. In *Table 1*, it can be found out that PCA performance of building time for feature selection is comparatively better, when dealing with large number of features.

In *Table 1*, results indicate that in most cases CS achieved better score in terms of precision, recall, F1-measure and AUC with almost all features. *Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13* depicted the results of IG, SC and PCA in term of above listed metrics. Experimental results also indicate that with CS as feature selection method the ranking performances of Naïve Bayes are more stable than those with IG and PCA as feature selection methods. This is especially true; when the numbers of selected features are moderate.

Table 1 Results of event discovery (Value in bold in each metric column is the best value in the metric)

	No of features	Precision	Recall	F-measure	AUC	Building TIME (MS)
PCA	10	0.901	0.919	0.91	0.53	0.01 sec
PCA	20	0.9	0.914	0.907	0.533	0.03 sec
PCA	30	0.9	0.896	0.898	0.556	0.03 sec
PCA	40	0.898	0.897	0.898	0.554	**0.03 sec**
Chi Square	10	0.901	0.919	**0.91**	0.53	0.02 sec
Chi Square	20	0.9	0.918	**0.909**	0.526	0.04 sec
Chi Square	30	0.899	0.916	**0.908**	0.557	0.05 sec
Chi Square	40	0.9	0.923	**0.911**	0.553	0.05 sec
IG	10	0.901	0.919	0.91	0.53	**0.01 sec**
IG	20	0.9	0.918	0.909	0.526	**0.03 sec**
IG	30	0.899	0.916	0.907	0.556	0.04 sec
IG	40	0.9	0.923	0.911	0.553	0.06 sec

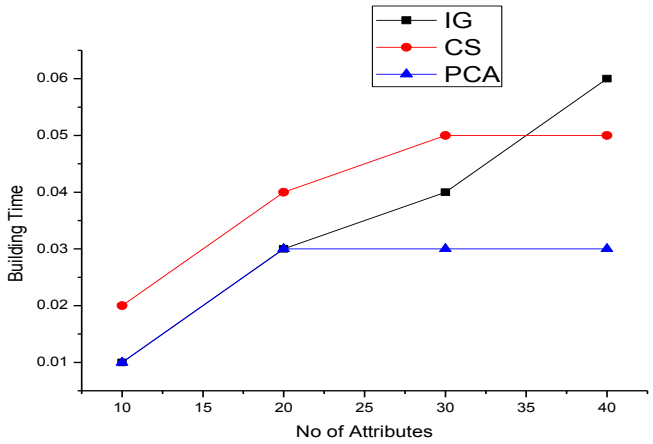

Figure 9 No of attribute vs. building time

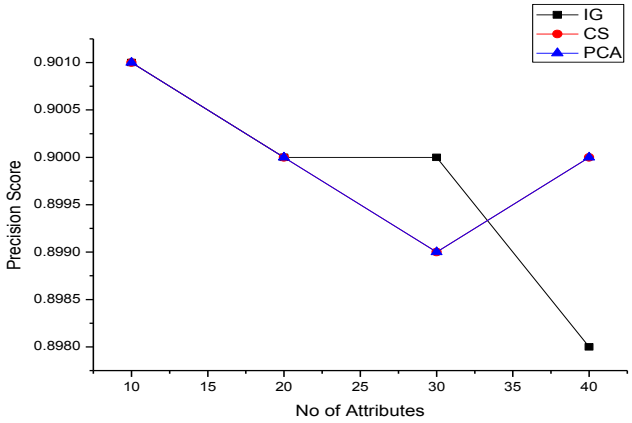

Figure 10 No of attribute vs. precision

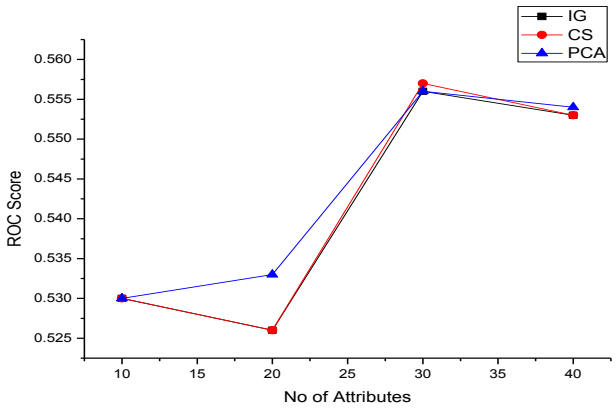

Figure 11 No of attribute vs. AUC score

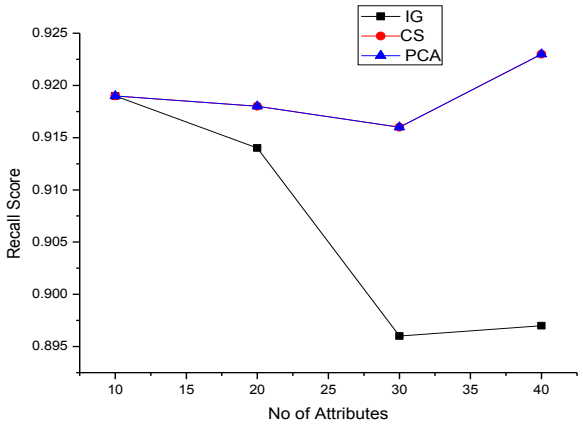

Figure 12 No of attribute vs. recall

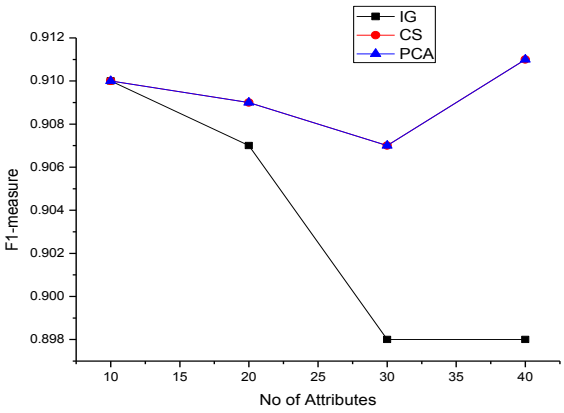

Figure 13 No of attribute vs. F1-measure

7. Conclusion

Feature selection is a well-known issue of many problems in the domain of machine learning and image recognition. For text classification systems, selection of correct features is of vital importance. This is because of the large size of sets of potential features found in different problems. If we take a text document as an example, each word can be considered a feature. Feature selection techniques can help to reduce the number of potentially useful features.

The job of features selection algorithm is to retrieve the most relevant features, or words, in a document. This study has briefly introduced some current work in the area of text classification, followed by a number of common methods for feature selection, and applications of feature selection in machine learning. Experimental results of our study effectively supported the well-known fact that accuracy in terms of recall, F1-measure, precision and AUC can be achieved by selecting the optimum number features.

Features selected through CS, the AUC score are comparatively better than PCA and IG. Down the road, it could be possible to add these methods to encounter to reduce dimensionality, but we must be cautiously attentive to avoid removing features whose variance does not satisfy some threshold values.

Acknowledgment

We would like to offer our deepest gratitude to Professor Dr. Zhu Yan, SWJTU, China. She provided her patient guidance, enthusiastic encouragement, and positive critique of our research work. Our special thanks for the staff at computing laboratory (SIST) for their support and resources provided for the experimental work. In the end, we would like to thank our parents and family members for their continued support and encouragement.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Dasgupta A, Drineas P, Harb B, Josifovski V, Mahoney MW. Feature selection methods for text classification. In proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining 2007 (pp. 230-9).

[2] Shang W, Huang H, Zhu H, Lin Y, Qu Y, Wang Z. A novel feature selection algorithm for text categorization. Expert Systems with Applications. 2007; 33(1):1-5.

[3] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018; 300:70-9.

[4] Khalifi H, Elqadi A, Ghanou Y. Support vector machines for a new hybrid information retrieval system. Procedia Computer Science. 2018; 127:139-45.

[5] Salo F, Nassif AB, Essex A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks. 2019; 148:164-75.

[6] Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A. Feature selection methods and genomic big data: a systematic review. Journal of Big Data. 2019; 6(1).

[7] Ramesh B, Sathiaseelan JG. An advanced multi class instance selection based support vector machine for text classification. Procedia Computer Science. 2015; 57:1124-30.

[8] Caggiano A, Angelone R, Napolitano F, Nele L, Teti R. Dimensionality reduction of sensorial features by principal component analysis for ANN machine learning in tool condition monitoring of CFRP drilling. Procedia CIRP. 2018; 78:307-12.

[9] Gibert D, Mateu C, Planes J. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. Journal of Network and Computer Applications. 2020; 153:102526.

[10] Almuallim H, Dietterich TG. Learning with many irrelevant features. In AAAI 1991 (pp. 547-52).

[11] Swets DL, Weng JJ. Efficient content-based image retrieval using automatic feature selection. In proceedings of international symposium on computer vision-ISCV 1995 (pp. 85-90). IEEE.

[12] Lee W, Stolfo SJ, Mok KW. Adaptive intrusion detection: A data mining approach. Artificial Intelligence Review. 2000; 14(6):533-67.

[13] Forman G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research. 2003:1289-305.

[14] Rill S, Reinel D, Scheidt J, Zicari RV. Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowledge-Based Systems. 2014; 69:24-33.

[15] Idris I, Selamat A. Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing. 2014; 22:11-27.

[16] Uysal AK, Gunal S, Ergin S, Gunal ES. The impact of feature extraction and selection on SMS spam filtering. Elektronika ir Elektrotechnika. 2013; 19(5):67-72.

[17] Zhang C, Wu X, Niu Z, Ding W. Authorship identification from unstructured texts. Knowledge-Based Systems. 2014; 66:99-111.

[18] Saraç E, Özel SA. An ant colony optimization based feature selection for web page classification. The Scientific World Journal. 2014.

[19] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Engineering Journal. 2014; 5(4):1093-113.

[20] Xu S, Chan HK. Forecasting medical device demand with online search queries: a big data and machine learning approach. Procedia Manufacturing. 2019; 39:32-9.

[21] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys (CSUR). 2002; 34(1):1-47.

[22] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97(1-2):273-324.

[23] Donoser M, Wagner S, Bischof H. Context information from search engines for document recognition. Pattern Recognition Letters. 2010; 31(8):750-4.

[24] Guyon I, Gunn S, Nikravesh M, Zadeh LA, editors. Feature extraction: foundations and applications. Springer; 2008.

[25] Ruiz FE, Pérez PS, Bonev BI. Information theory in computer vision and pattern recognition. Springer Science & Business Media; 2009.

[26] Chen J, Huang H, Tian S, Qu Y. Feature selection for text classification with Naïve Bayes. Expert Systems with Applications. 2009; 36(3):5432-5.

[27] Hegde J, Rokseth B. Applications of machine learning methods for engineering risk assessment–A review. Safety Science. 2020; 122:104492.

[28] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In Icml 1997.

[29] Kumar V, Minz S. Feature selection: a literature review. SmartCR. 2014; 4(3):211-29.

[30] Uysal AK. An improved global feature selection scheme for text classification. Expert Systems with Applications. 2016; 43:82-92.

[31] Dash M, Liu H. Feature selection for classification. Intelligent data analysis. 1997; 1(3):131-56.

[32] Wang F, Liang J. An efficient feature selection algorithm for hybrid data. Neurocomputing. 2016; 193:33-41.

[33] Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 2020; 143:106839.

[34] Taşcı Ş, Güngör T. Comparison of text feature selection policies and using an adaptive framework.

Expert Systems with Applications. 2013; 40(12):4871-86.
[35] Chen G, Chen J. A novel wrapper method for feature selection and its applications. Neurocomputing. 2015; 159:219-26.
[36] Liu H, Motoda H, Setiono R, Zhao Z. Feature selection: an ever evolving frontier in data mining. In feature selection in data mining 2010 (pp. 4-13).
[37] Lee SJ, Xu Z, Li T, Yang Y. A novel bagging C4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making. Journal of Biomedical Informatics. 2018; 78:144-55.
[38] Sosa-Cabrera G, García-Torres M, Gómez-Guerrero S, Schaerer CE, Divina F. A multivariate approach to the symmetrical uncertainty measure: application to feature selection problem. Information Sciences. 2019; 494:1-20.
[39] Kumar V, Minz S. Poem classification using machine learning approach. In proceedings of the second international conference on soft computing for problem solving (SocProS 2012), 2014 (pp. 675-82). Springer, New Delhi.
[40] Kumar V, Minz S. Mood classifiaction of lyrics using SentiWordNet. In international conference on computer communication and informatics 2013 (pp. 1-5). IEEE.
[41] Kumar V, Minz S. Multi-view ensemble learning for poem data classification using SentiWordNet. In Advanced Computing, Networking and Informatics 2014:57-66. Springer, Cham.
[42] Jia X, Kuo BC, Crawford MM. Feature mining for hyperspectral image classification. Proceedings of the IEEE. 2013; 101(3):676-97.
[43] Kuo BC, Landgrebe DA. Nonparametric weighted feature extraction for classification. IEEE Transactions on Geoscience and Remote Sensing. 2004; 42(5):1096-105.
[44] Zhao YQ, Zhang L, Kong SG. Band-subset-based clustering and fusion for hyperspectral imagery classification. IEEE Transactions on Geoscience and Remote Sensing. 2010; 49(2):747-56.
[45] Dua M. Attribute selection and ensemble classifier based novel approach to intrusion detection system. Procedia Computer Science. 2020; 167:2191-9.
[46] Mishra S, Panda M. Medical image retrieval using self-organising map on texture features. Future Computing and Informatics Journal. 2018; 3(2):359-70.
[47] Da Silva SF, Ribeiro MX, Neto JD, Traina-Jr C, Traina AJ. Improving the ranking quality of medical image retrieval using a genetic feature selection method. Decision Support Systems. 2011; 51(4):810-20.
[48] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics; 2001.
[49] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering. 2005; 17(4):491-502.
[50] Rodriguez-Galiano VF, Luque-Espinar JA, Chica-Olmo M, Mendes MP. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Science of the Total Environment. 2018; 624:661-72.
[51] Das S. Filters, wrappers and a boosting-based hybrid for feature selection. In Icml 2001: 74-81.
[52] http://docs.oracle.com/cd/A91202_01/901_doc/olap.901/a86720/esdatao6.htm Accessed 10 March 2020.
[53] Fahad A, Tari Z, Khalil I, Habib I, Alnuweiri H. Toward an efficient and scalable feature selection approach for internet traffic classification. Computer Networks. 2013; 57(9):2040-57.
[54] Jiarpakdee J, Tantithamthavorn C, Hassan AE. The impact of correlated metrics on the interpretation of defect models. IEEE Transactions on Software Engineering. 2019.
[55] SL SD, Jaidhar CD. Windows malware detector using convolutional neural network based on visualization images. IEEE Transactions on Emerging Topics in Computing. 2019.
[56] Artoni F, Delorme A, Makeig S. A visual working memory dataset collection with bootstrap Independent Component Analysis for comparison of electroencephalographic preprocessing pipelines. Data in Brief. 2019; 22:787-93.
[57] http://www.cs.waikato.ac.nz/ml/weka. Accessed 10 March 2020.
[58] http://chato.cl/webspam/datasets/. Accessed 10 March 2020.
[59] http://www.cs.waikato.ac.nz/~mhall/thesis.pdf. Accessed 10 March 2020.

Muhammad Iqbal was born in 1972 in Pakistan. He received B.Sc(Hons) and M.Sc degree in Computer Technology from Sindh University, Pakistan and MS in computer Science from SZABIST, Karachi, Pakistan. He is a PhD scholar in School of Information Sciences & Technology (SIST), Southwest Jiaotong University, Sichuan, Chengdu, PR China. His research interests are Network Security, Data Mining, Supervised Machine Learning algorithms and high speed data networks.
Email: miqbal.bukc.edu.pk

Malik Muneeb Abid was born in 1987 in Pakistan. He received B.Sc degree in Civil Engineering from U.E.TTaxila, Pakistan and MS degree in Transportation Engineering from NUST, Pakistan. Since 2013, he has done his PhD from School of Transportation and Logistics, Southwest Jiaotong University, Sichuan, Chengdu, PR China. He is currently working as Assistant Professor and Incharge Civil Engineering Department, at University of Sargodha (UOS). His research interests are Network

Robustness, Transportation network modeling and simulation, Data Mining, Supervised Machine Learning algorithms. He is member of IAROR and PEC.
Email: muneeb.abid@uos.edu.pk

Muhammad Norman Khalid is a Lecturer at Bahria University, Karachi Campus (BUKC), Pakistan. His research interest includes Machine Leering, Network Security and SEO

Email: nomanramzan91@gmail.com

Dr. Amir Manzoor holds a PhD in Management Sciences. He is a graduate of NED University, Pakistan, Lahore University of Management Sciences (LUMS), Pakistan and Bangor University, United Kingdom. He is an HEC-Approved PhD Supervisor, active researcher, and a passionate member of academic community in Pakistan. The professional experience entails a rich blend of both academic and corporate exposure exceeding 20 years with both national and international Organizations. His published research has received 241 citations worldwide. His published books have been adopted by 25 large and reputed Asian and European universities as text/reference book in their curriculum of undergraduate and graduate programs. His research interests include E-commerce, Strategic Management, Enterprise Resource Planning (ERP), Project Management,Supply Chain Management, Data Analysis, and Technology Applications for Competitive Advantage.
Email: amirmanzoor.bukc@bahria.edu.pk