A Comparative Study of Feature Selection Approaches: 2016-2020

Syed Asim Ali Shah
University Institute of Information Technology, PMAS Arid Agriculture University
Rawalpindi, Pakistan, 46000
Syedasimshah14@gmail.com

Hafiz Muhammad Shabbir
University Institute of Information Technology, PMAS Arid Agriculture University
Rawalpindi, Pakistan, 46000
Shabirahmedz888@gmail.com

Saif Ur Rehman
University Institute of Information Technology, PMAS Arid Agriculture University
Rawalpindi, Pakistan, 46000
Saifi.ur.rehman@gmail.com

Muhammad Waqas
University Institute of Information Technology, PMAS Arid Agriculture University
Rawalpindi, Pakistan
Wmuhammad333@gmail.com

ABSTRACT
Feature Selection (FS) is a dimensionality reduction method that is commonly adopted in the fields of machine learning, pattern recognition, statistics, and data mining. It is a preprocessing course of action universally used for huge volume of data and FS technique aims to select a subset of relevant features from the original set of features according to some criteria. Other than selecting the subset, it also congregates some other purposes, such as dimensionality reduction, compact the amount of data which are required for learning process, progress in predictive accuracy and increasing the constructed models. In literature, comprehensive work exists on the feature selection techniques. The preliminary task of any feature selection method is to reduce the dimensionality of the data and increase the performance of an algorithm. It is a research area of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. In this paper, the basic theme is to provide an overview of the different latest feature selection methods suggested during the years 2016-2020. Furthermore, each of the selected feature selection technique is presented focusing on the methodology adopted, the strengthens and weaknesses. Finally, we have identified the challenges which need to be coped in the newly proposed feature selection approaches.

Keywords Data Mining; Machine learning; Feature Selection, filter methods, wrapper methods, hybrid methods.

1. INTRODUCTION
We are now in the era of big data, where huge amounts of high-dimensional data become ubiquitous in a wide range of fields, such as social media, health care, bioinformatics and online education. The rapid growth of data presents challenges for effective and efficient data management. It is desirable to apply data mining and machine learning techniques to automatically discover knowledge from data of various sorts [1]. Data mining refers to extracting knowledge from a large amount of data, in the other way we can say data mining is the process to discover various types of pattern that are inherited in the data and which are accurate, new and useful.

A feature is an individual measurable property of the process being observed. Feature Selection (FS) is the process of selecting out the most significant features from a given dataset. In many of the cases, FS can enhance the performance of a machine learning model as well. Using a set of features, many machines learning algorithm can perform classification. Real-world data contain a lot of irrelevant, redundant, and noisy features. Removing these features by FS reduces storage and computational cost while avoiding significant loss of information or degradation of learning performance. FS mechanism has the advantages of improving learning performance, increasing computational efficiency, decreasing memory storage, and building better generalization models. Therefore, FS is often preferred in many applications such as text mining and genetic analysis. A generic FS approach is shown in Figure 1.

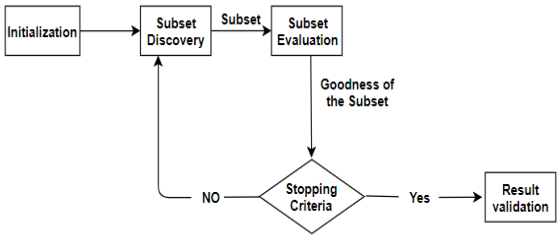

Figure 1. A general FS Process

In Fig.3, Subset generation is essentially a process of heuristic search, with each state in the search space specifying a candidate subset for evaluation. The nature of this process is determined by two basic issues. First, one must decide the search starting point (or points) which in turn influences the search direction. Search may

start with an empty set and successively add features (i.e., forward), or start with a full set and successively remove features (i.e., backward), or start with both ends and add and remove features simultaneously (i.e., bidirectional). Search may also start with a randomly selected subset in order to avoid being trapped into local optima. Second, one must decide a search strategy. For a data set with N features, there exist 2N candidate subsets. This search space is exponentially prohibitive for exhaustive search with even a moderate N. Therefore, different strategies have been explored: complete, sequential, and random search. Each newly generated subset needs to be evaluated by an evaluation criterion. An evaluation criterion can be broadly categorized into two groups based on their dependency on learning algorithms that will finally be applied on the selected feature subset. The one is independent criteria, the other is dependent criteria. Some popular independent criteria are distance measures, information measures, dependency measures, and consistency measures. An independent criterion is used in algorithms of the filter model. A dependent criterion used in the wrapper model requires a predetermined learning algorithm in feature selection and uses the performance of the learning algorithm applied on the selected subset to determine which features are selected.

FS offers a lot benefits as it enhances the prediction performance, understandability, scalability, and generalization capability of the underlying classifier. It also further reduces the computational complexity and storage, provides faster and more cost-effective model, and plays very significant role in knowledge discovery.

In literature, mostly the FS methods can be broadly categorized as wrapper, filter, embedded, Hybrid methods, Figure 2 depicts this categorization.

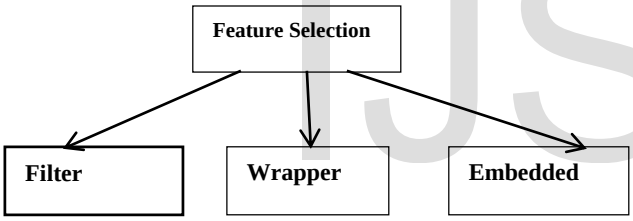

Figure 2. Classification of FS approaches

Wrapper methods rely on the predictive performance of a predefined learning algorithm to evaluate the quality of selected features. Given a specific learning algorithm, a typical wrapper method performs two steps: (1) search for a subset of features; and (2) evaluate the selected features. Wrapper technique repeats (1) and (2) until some stopping criteria are satisfied. The feature set search component first generates a subset of features, and then the learning algorithm acts as a black box to evaluate the quality of these features based on the learning performance. For example, the whole process works iteratively until the highest learning performance is achieved or the desired number of selected features is obtained. Then the feature subset that gives the highest learning performance is returned as the selected features. Unfortunately, a known issue of wrapper methods is that the search space for *d*-features is $\mathbf{2}^{\boldsymbol{d}}$, which is impractical when *d* is very large. Examples of wrapper methods are [17-19]. Figure 3 shows the working principle of the wrapper FS techniques.

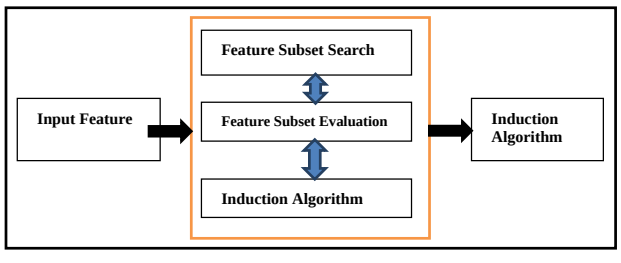

Figure 3. Wrapper FS working principle

The majority of researchers focus on the development of supervised feature selection with filter evaluation framework. Filter methods are independent of any learning algorithms. They rely on characteristics of data to assess feature importance. Filter methods are typically more computationally efficient than wrapper methods. However, due to the lack of a specific learning algorithm guiding the feature selection phase, the selected features may not be optimal for the target learning algorithms. A typical filter method consists of two steps. In the first step, feature importance is ranked according to some feature evaluation criteria. The feature importance evaluation process can be either univariate or multivariate. In the second step of a typical filter method, lowly ranked features are filtered out. Examples of filter methods are [20-24]. A typical diagrammatical representation of filter FS techniques is shown in Figure 4.

The last type of FS approaches is termed as Embedded methods, which are a trade-off between filter and wrapper methods which embed the feature selection into model learning. Thus, such approaches inherit the merits of wrapper and filter methods – (1) they include the interactions with the learning algorithm; and (2) they are far more efficient than the wrapper methods since they do not need to evaluate feature sets iteratively. The most widely used embedded methods are the regularization models that target to fit a learning model by minimizing the fitting errors and forcing feature coefficients to be small (or exact zero) simultaneously. Afterwards, both the regularization model and selected feature sets are returned as the final results. Hybrid methods can be regarded as a combination of multiple feature selection algorithms (e.g., wrapper, filters, and embedded). The main target is to tackle the instability and perturbation issues of many existing feature selection algorithms. Examples of Hybrid methods are [25-27]. For example, for small-sized high-dimensional data, a small perturbation on the training data may result in totally different feature selection results. By aggregating multiple selected feature subsets from different methods together, the results are more robust and hence the credibility of the selected features is enhanced.

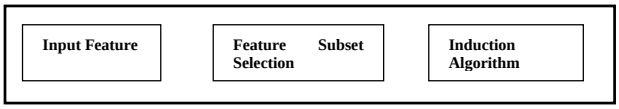

Figure 4. Filter FS Model

Currently, some surveys give a summarization of feature selection algorithms, such as those in Guyon and Elisseeff (2003), Alelyani et al. (2013), Chandrashekar and Sahin (2014), and Tang et al. (2014). Recent survey introduced representative FS algorithms to cover all components, such as J Li et al. (2018). These studies either focus on traditional feature selection algorithms or specific learning tasks like classification and clustering. In this paper, we focus on new FS techniques. FS techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables; hence, offering the advantage of interpretability by a domain expert.

In this survey, we are aim to provide a comprehensive review of literature with regards to feature selection. Moreover, this study

summarizes the FS process, its importance, different types of FS algorithms such as Filter, Wrapper and Hybrid.

This paper is divided into five sections. Section 2 gives a review on various FS approaches especially that have been proposed over the past five years. Section 3 describes the challenges and issues inherent in FS task. The overall discussion with a few recommendations for future directions is presented in the last section 4.

2. FEATURE SELECTION TECHNIQUIES

Since the mid-1990s, hundreds of feature selection algorithms have been proposed. In this section we will discuss each feature selection techniques focusing on the technical aspect of each feature selection method. In this section, the papers are reviewed according to the year of publication. The existing and proposed methods of feature selection are analyzed briefly.

Khoshgoftaar et al. [1] presented a set of seven univariate feature selection techniques; called first order statistic (FOS) based feature selection. In their work, three classifiers Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) are used. SVM is a popular linear discriminant classifier, LR is a simple and effective regression model, and RF is a powerful ensemble based classifier. RF was implemented by the data mining tool Weka. Previous research [29] shows that the optimum number of trees is 100, so they used these numbers. The classifiers evaluate the classification power of seven FOS techniques. Major contribution of their works which tied together those related feature selection techniques into a single family and examines them to one another. To test the learners, 5-Fold cross validation were used. One of the simplest criteria of Significance Analysis of Microarrays (SAM),as in (1) is defined as:

$$\text{SAM} = \frac{\mu p - \mu N}{\sigma^* + \sigma_\circ} \quad (1)$$

Where σ^o represents exchangeability constant and σ^* represents an overall standard deviation. Therefore, calculating σ^* in their experiments by in (2):

$$\boldsymbol{\sigma^*} = \sqrt{\frac{n_T}{n_P\ n_N\ (n_T-2)}\left(\sum_{j=1}^{np}|\boldsymbol{x_j}-\boldsymbol{u_p}|^2 + \sum_{j=1}^{nN}|\boldsymbol{x_j}-\boldsymbol{\mu_n}|^2\right)} \quad (2)$$

As in (2) represents $\sum_{j=1}^{n_p}$ and $\sum_{j=1}^{nN}$, the sum across the instances of the positive and negative class respectively. Where n_T is equal to the total number of instances in the dataset. They used eleven DNA microarray dataset. The result of the individual datasets consists of four runs of five-fold cross-validation. They performed different experiments to validate their techniques. However, majority of the rankers perform well for classification. In all experiments, the SAM techniques showed better as compared to other techniques. The clear worst performer in all techniques was Fold Change Ratio (FCR).

Recently, new feature selection techniques called; query expansion ranking (QER) was proposed [2]. It was based on query expansion term weighting methods from the field of information retrieval. The proposed method especially developed for reducing dimensionality of feature space in sentiment analysis (SA) problems. Moreover, it helps to find more relevant documents for a given query. Then, their method computes and selects features of having lowest scores which was used in classification process. They used four classifiers naïve Bayes multinomial (NBM), SMO, J48, LR and five feature selection methods QER, Chi square (CHI2), information gain (IG), document frequency difference (DFD), and optimal orthogonal centroid (OCFS) to evaluate their results. The limitation of method such as CHI2 method produces high scores for rare features. DFD method is requires an equal or nearly equal number of documents in both classes. The proposed QER method is both language and classifier independent and can select better features than other methods for SA. But, the limitation of QER is that it is only suitable for classifying two classes. For the sake of simplicity, proposed formula (3) to compute features score. In previous study [28], they observed that feature sizes up to 3000 tend to give good classification performance improvement; therefore they choose these feature sizes in our experiments.

$$Score_f = \frac{Pf+qf}{|pf-qf|} \quad (3)$$

Where p_f ratio of positive documents containing feature f and q_f ratio of negative documents containing feature f given as in (4, 5)

$$\text{Pf} = \frac{Df_+^f + 0.5}{N^+ + 1.0} \quad (4)$$

$$\text{qf} = \frac{\text{DP}_-^{\text{f}} + 0.5}{\text{N}^- + 0.5} \quad (5)$$

Where Df_+^f and DP_-^f are the raw counts of documents that contain f in the positive and negative classes, respectively and N^+ and N^- are the numbers of documents in the positive and negative classes, respectively.

The limitation of methods such as CHI2 produces high scores for rare features. Since, rare features are not frequently used in text and thus do not have a big impact for text classification. One limitation of the DFD method is that it requires an equal or nearly equal number of documents in both classes. The contribution of QER is both language and classifier independent and can select better features than other methods for sentiment analysis. But, the limitation of QER is that it is only suitable for classifying two classes. Turkish and English product review datasets were used. They used Weka data mining tool for their experiments. However, they performed different experiments to validate their method. In all experiments, the NBM classifier performed better than other classifiers. And, the proposed QER method was the best performance as compared with other feature selection methods.

In this study, Novel Feature Selection (NFS) algorithm was proposed in [3]. Additionally, the NFS algorithm extracts more relevant features that support for attaining maximum classification accuracy. Hence, the main objective of their work is minimum feature used to predict the possibility of heart disease at its early stages. Data mining techniques were used for prediction of dataset. Next; Receiver Operating Characteristics (ROC) was a graphical plot which illustrates the performance of a binary classifier system which was created by True Positive Rate (TPR) vs. False Positive Rate (FPR). The larger the area under ROC curve, the higher the performance of the algorithm .The formula of FPR (6) and TPR (7) are:

$$\text{FPR=FP/ (FP+TN)} \quad (6)$$

$$\text{TPR=TP/ (TP+FN)} \quad (7)$$

They were collected data record from Cleveland heart disease database. NFS algorithm are applied in datasets with 13 feature and select 6 best feature. They performed different experiment to validate their technique. In all experiments, neural network predict 93% of accuracy and SMO predict 89 % by using NFS. Hence, their study concluded that these approach leads to a superior feature selection process in term of sinking the number of variable required and an increased in classification accuracy for better prediction.

Another NFS technique called; SIP-FS was proposed in [4]. The proposed method improved stability and interpretability without losing predictability. The two main contributions of their work, first generalized correlation rather than mutual information. Second, stability constraint was adopted in SIP-FS to select consistent results of ranking in the case of data variation. Therefore, SIP-FS algorithm selects a reasonable and compact feature subset for data representation efficiently. Their approaches is to focused on the filter methods based on different evaluation measures, such as distance criterion, Separability criterion, correlation coefficient, consistency, and mutual information. Various stability evaluation indexes were only used to evaluate feature selection method rather than improve the stability of the method itself. Therefore, stability constraint is employed in theirs study to obtain robust selection results

$$f_{opt} = argmax(D - R + K \times S) \quad (8)$$

Where S represents existing stability evaluation index. K is a parameter, which balances prediction factor (D−R) and stability factor S. Then, the stability evaluation index can be computed by in (9, 10)

$$S\ (f, F) = \frac{1}{i-1} \sum_{j=1}^{i=1} S(F_f, F_j) \quad (9)$$

$$S\ (F_f, F_j) = \frac{|F_f \cap F_j|}{|F_f \cup F_j|} \quad (10)$$

Where F_f is the union between the selected features, F_j (j = 1, 2... i − 1) represents the selected feature subset, $|F_f \cap F_j|$ and $|F_f \cup F_j|$ represent the intersection and union between feature sets F_f and F_j respectively. Predictability and stability were used in the feature selection criterion of Eq. (8) as well. Four feature selection methods, mRMR, ReliefF, En-mRMR, and En-Relief, and one proposed method SIP-FS are also used for performance comparisons on three publicly available datasets (MIML, NUS-WIDE-LITE, and USGS21). They performed different experiments to validate their techniques. In all experiments, the SIP-FS performed better as compared to other method.

Recently, a feature selection method for mass classification was suggested [5]. The new feature selection method called SRN. Several feature selection methods were investigated, including the F-score, the Relief, the SVM-RFE, the SVM-RFE (mRMR), and the proposed SRN. Major three contributions are as follows: 1) fuzzy c-means (FCM) with a spatial information constraint, which can reduce the labour cost in mass segmentation compared with previous work [30-31]; 2) a new feature selection method was proposed, which can balance the redundancy and relevance in feature selection, and can improve classification accuracy; 3) several feature selection methods are investigated for the mass classification problem. They used normalization technology to develop and modify the redundancy term with its value range to [0, 1]. It select least important features were identified with the criterion in (11) at each iteration is sequentially removed.

$$\gamma_i = \beta|\omega_i| + (1 + \beta)\left[R_{s,f_i} - Q_{s,f_i}\right]$$
$$= \beta|\omega_i| + (1 - \beta)\left[I(c, f_i) - \frac{1}{|s|}\sum_{f_i \in s} NI(f_i, f_s)\right] \quad (11)$$

In addition to, two classification methods LDA with k-NN and SVM with radial basis function (RBF). RBF kernel was defined as in (12)

$$K\ (X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|}{2\sigma^2}\right) \quad (12)$$

Where σ > 0 is a constant that defines the kernel width.

The classification performance is measured by the true positive rate (TPR), the true negative rate (TNR), and the accuracy as in (15):

$$TPR = \frac{TP}{TP + FN} \quad (13)$$

$$TPR = \frac{TN}{TN + Fp} \quad (14)$$

$$ACCURACY = \frac{TP + TN}{TP + FP + TN + FN} \quad (15)$$

Digital Database for Screening Mammography (DDSM) dataset which contains more than 2500 cases of mammograms. They used 804 mammogram images in their experiment. They performed different experiment to validate their method. At the end, their study concluded that the proposed SRN method shows better performance than the other selection methods, with less selected features. The one drawback of proposed method is computation cost is more expensive than other methods.

In another study [6], new feature selection methods called RCFS (ranking with clustering based feature selection) to improve performance of image classification. The previous [32] feature selection methods only used to select relevant features. But consist of redundant features had limitation of them. The new method deals with two feature selection metrics. The first metric covers relevancy and the other metric covers redundancy analysis using clustering such as k mean clustering. In RCFS, symmetric uncertainty(SU) is used to measure the correlation or dependency between the feature and target class as fellow in(16)

$$SU(X, Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)} \quad (16)$$

Where H(X) denote the entropy of a discrete random variable X.

The RCFS algorithm is based on two phases. In the first phase, the irrelevant features are removed by selecting the relevancy measure SU with a threshold value from a given image dataset D. Moreover, the selected relevant features are grouped as K number of clusters in second phase, and then the cluster-representative-features are selected from each cluster using the relevancy measure. Thus, the selected significant features are obtained by combining the cluster-representative-features of each cluster. For evaluating performance of RCFS in term of accuracy and runtime, 3 classifiers were presented and 6 methods compared. Therefore, the new method took less average runtime and produced higher average accuracy. They performed different experiments to validate their method. In all experiments, the RCFS method performed better as compared to other method. They also provide some future directions, for example their work can be extended with different mechanisms for redundancy analysis.

P.Gholami et al. [7] discussed about Data Environment Analysis which is useful technique for determining the efficiency of decision-making units. Next one is Entropy method that is based on weighting criteria to select specific features.Additionally, they used WEKA tool for their experiment. Moreover, MATLAB software was also used to apply their model. Hence, their proposed model worked in following steps. First computed the entropy value of each attribute in different classes. Then separated the datasets according to their classes. Also calculated the entropy of each attribute. Second considered each attribute as a decision making units (DMUs). Third placed input of DMUs equal to 1. Fourth placed output of DMUs equal to entropy value getting from step 1. In fifth step computed efficiency of each attribute. Selected efficient attributes in step 6. Therefore, features selection algorithms applied on same datasets for selecting features in step 7. In last step, comparison of new method was performed with step 7. Also, MCDM methods and data envelopment analysis method also introduced to calculate the efficiency of decision making unit. Hence, the unit matrix formula performed to normalize the decision matrix in (17)

$$E_j = -k\sum_{i=1}^{m}\left[p_{ij} ln_{ij}(p_{ij})\right] \rightarrow \begin{Bmatrix} \forall_j = 1,2,\dots,n \\ k = \frac{1}{\ln(m)} \end{Bmatrix} \qquad (17)$$

Three different datasets performed to validate their model. Moreover, the accuracy of proposed technique compared with different existing techniques. Therefore the accuracy of first and second datasets had greater than third one. In all experiments their proposed technique performed better.

Another hybrid feature selection was explored in [8]. The GADP (genetic algorithm with dynamic parameter setting) used for generating number of subsets of genes. Then ranked the genes according to their occurrences frequencies. While x^2 test also performed to determine threshold for selecting specific genes. Microarray data usually contain many genes and a small number of samples. Most of them are irrelevant or insignificant to a clinical diagnosis. Although many genetic algorithms (GA) had been introduced for the microarray data analysis such as a hybrid method, GA/KNN, to analyse the colon dataset. Next, GA used to generate a large number of subsets of genes. Then the k-nearest neighbour classifier performed to filter the subsets of genes according to classification accuracy. Furthermore the modified kernel Fisher discriminant analysis (KFDA) used to analyse the breast cancer dataset. Another hybrid method deals with the intelligent genetic algorithm (IGA) to generate a number of initial gene subsets and then involved SVM to select better gene subsets from the initial gene subsets. The previous [33-37] methods were only used to cover rank genes. On the other hand the new method aims to determine the number of selected genes. Six datasets were picked for comparison. Then it was found that GADP selected few genes with high prediction accuracy. On the bases of initial feature selection algorithm result, it was showed that GADP method had faster and better than SGA.

In another study [9], two support vector data description. SVDD-RFE (radius-recursive feature elimination) is used to minimize size of boundary observation through radius squared value. Furthermore SVDD-dual objective-RFE is used for obtaining compact description in dual space of SVDD. There are two key elements to the feature selection problem. A criterion function and a subset searching method with a given criterion function. The criterion function was used to measure the discriminating power of a feature subset. The subset searching method used to explore the feature subset space in order to identify the best subset of features that optimizes the given criterion functions.

To find spherically shaped boundary with centre μ and radius R. A variable ζi was introduced to penalize outliers for the largest distance between xi and μ. The formula was given as follows in (18)

$$\min_{R,\mu,\zeta_i} R^2 + C\sum_{i=1}^{N}\varepsilon_i \qquad (18)$$

$$s.t.\|x_i - \mu\|^2 \le R^2 + \varepsilon_i, \varepsilon_i \ge 0, \forall_i$$

Simulated, DARPA and WDBC datasets were taken for comparison. In these datasets false alarm rate was compared with high detection rate. Therefore, the performance of simulated and DARPA was better and most effective to be found than WDBC with small number of features.

Recently, a new feature selection method was derived in [10]. The core idea of this method had to obtain a feature subset FS1 by using an optimal feature selection method. Then to filter the redundant features from FS1 to form the final feature subset. Furthermore the MI based feature selections also introduced that worked on MI theory. MI methods were used to calculate relevance of different words. Then measure the dependence information of a word t_i and a category c. The formula is given as in (19)

$$\begin{cases} MI(t_i) = \sum_{c_k} MI(t_j, c_k), \\ MI(t_j, c_k) = p(c_k, t_k)\log_2\left(\frac{p(c_k, t_k)}{p(c_k)p(t_j)}\right), \end{cases} \qquad (19)$$

Where $p(t_i)$ is the occurrence probability of word t_i and $p(c_k, t_i)$is the probability that a document in category c_k contains t_i.

The aim of text classification is to assigns a predefined category to an unlabelled text document. It has become a very efficient method to manage the vast volumes of digital documents available on the Internet. In recent years, many classifiers were performed to text

classification based on machine learning and statistical theory. Among these classifiers, decision trees, k-nearest neighbours (KNN), neural networks, NB and SVM are the most successful and widely used methods.After comparison, the time complexity of proposed method had slightly higher than CMFSX but lower than MI methods. Next, the proposed method compared with results of datasets such as WE (WebKB), NE (20-Newsgroups) and RE (Routers). Therefore, the execution time and classification accuracy of proposed method had better than others.

Miao et al. [11] proposed MD (maximizing the difference) to analyse customer reviews datasets. With the rapid development of e-commerce, online consumer review, as a new type of word-of-mouth (WOM) information, played an increasingly important role in consumers purchase decisions. Most research papers used the quantitative measures of consumer reviews such as total number of reviews, number of positive and negative reviews, average review stars, etc. Next, defined MD through equation in which symbols, for term j, define p_{jm}= ($p_{jm\,0\ldots}p_{jmS}$) for m= 1 ,…, M, where p_{jms}(s = 0,…, S − 1) represented the probability of term j occurring s times in category m; andp_{jms} the probability of term j occurring S or more times in category m. When M = 2, $MD_j = \left(\sum_{s=0}^{S}|p_{j1s} - p_{j2s}|\right)^{q})^{1/q}$. It had the sum of probability differences for term j occurring s times in category 1 and 2. Usually q is taken equal to 1 or 2 to represent the absolute or square root difference. For M > 2, MD_j had the sum of the above difference over any pair of category

$$MD_j = \sum_{a,b\{1,\ldots,M\},a<b}\left(\sum_{s=0}^{S}|p_{jas} - p_{jbs}|)^q\right)^{1/q} \quad (20)$$

The number of S was chosen according to real situation. In the mining task data was taken in document matrix where document represented by D (row) and term represented by T (column). The previous methods counted only whether T term appeared in document D or not. Hence this is limitation of them. The new method based on this idea that a term which has larger probability difference between categories Cm should have bigger ability to distinguish the document into different groups. The performance of proposed method was compared with boosting, RF and SVM method. Hence, the results show that proposed method is more effective. In addition to, it performed better on unbalanced data than balanced data with respect to simulation and empirical results. In all experiment, the proposed method was found good for mining text.

Wang et al. [12] described new feature selection methods called extreme learning machine (ELM) and Fractional-order Darwinian particle swarm optimization (FODPSO).These method was applied for regression problems. The proposed method was two steps. In first, construct the fitness function by ELM. Secondly, seeking the optimal solutions of fitness functions by FODPSO. ELM method was simple yet effective single hidden layer neural network .It was suitable for feature selection due to its gratifying computational efficiency. FODPSO is an intelligent optimization algorithm which owns good global search ability. Given ELM in (21), ω denoted the weight connecting the input layer and hidden layer. β denoted the weight connecting the hidden layer and output layer. b is the threshold of the hidden layer, and G is the nonlinear piecewise continuous activation .H represents the hidden layer output matrix,X is the input layer, and Y is the expected output .

$$\min \|y - y\| = \min\|y - Hy\|$$
$$\mathrm{H} = \mathrm{G}(\omega X + b) \quad (21)$$

In their work, all of the codes were implemented in MATLAB on a desktop computer with a Pentium eight-core CPU (4 GHz) and 32 GB memory. Seven public datasets for regression problems were adopted. However, it described how many number of feature were found in each dataset. They performed different experiment to validate their techniques. In all experiments, it's concluded that ELM was more efficient, ELM-FODPSO was better.

Recently, a new feature selection method called Peculiar Genes Selection (PGS) was proposed [13]. The proposed method improved classification performances of imbalanced data sets. In the proposed method the feature selection is performed in three steps. In first, Identification of Differentially Expressed Genes according to the experimental conditions .Secondly filters out the features with low discriminative power. In third steps, Select good feature for each class. They presented a supervised approach using the SVM as classifier. As in given (22), where p_i is the predicted probability of success for subject i, $\beta_\circ$ the intercept of the model, βj the fitted parameter and X_{ji} the expression of the j-th gene of subject i .PGS method calculates logistic regressions with the help of their regression and the probability.

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_\circ + \beta_j X_{ji} \quad (22)$$

The proposed method was testing on two microarray datasets. They performed different experiment to validate their method. In all experiments, by using PGS method SVM was dominated classifier achieved 82% accuracy as compared to other. To conclude their, comparison of different methods, performance of SVM was better than others.

Liu et al. [14] proposed a new filter Sequence Forward Search (SFS) feature selection method combined with LW-index called (SFS-LW). LW-index is a statistical index for labelled feature subset. Then, it aims to measure the separation degree between two linear separable classes. In addition to, two classifier are used such as SVM and CBC (centroid based classifier).The main advantage of their method was inexpensive in terms of computation cost due to the linear time complexity of LW-index. It was effective in terms of accuracy due to its high searching and learning capacity by traversing the whole feature space. Thus, the mathematical formula of FD is defined as below in (23)

$$FD_{i|j} = FD_{j|i} = d(v_i, v_j) - (r_i + r_j) \quad (23)$$

Where and v_j are the centroid vectors of cluster C_i and C_j respectively. Supposing X_n represents the sample in C_*, then the calculation of V_* is as follows in (24)

$$V_* = \frac{1}{|C_*|}\sum_{X_n \epsilon C_*} X_n \quad (24)$$

Furthermore, r_i and $r_j$are radii of cluster C_i and C_j , respectively. And r_* is calculated by the average distance between the centroid

of cluster and the K_* most far away instances from the centroid that belong to the cluster. Hence, the measure of r_* is as below in (25)

$$r_* = \frac{1}{K_*}\sum_{K=1}^{K} d(X_{K,}V_*) \quad (25)$$

Nine different datasets were drawn from their experiment. WAPPER, DISR, MIFS, ICAP, RELIEF and CMIM and the proposed SFS-LW were applied in datasets. They performed different experiment to validate their method. Now, the performance of WAPPER was better than as compared to other by using both classifier CBC and SVM. However, by applying CBC classifier and they observed that, the performance of SFS-LW was better than as compared to other in Lung cancer dataset. Finally, it concluded that every method have pros and cons by applying in each dataset. At the end, WAPPER techniques were better and dominated techniques in all other techniques.

In 2012, L. Zhen et al. [15] proposed a new feature selection method called BFS for ML internet traffic classification. BFS focus on two goals: 1) select a feature subset which has balance bias degree, 2) reduce feature to improve the classification accuracy. In addition to, they used classification performance metrics such as g-mean, mauc, accuracy and recall. In (23) H(X) is the information of entropy and N_x is the discrete values

$$\mathrm{RU}(X) = \frac{H(X)}{H_{max(X)}} = \frac{H(X)}{log_{2(min\{N_x,\, m\})}} \quad (26)$$

Where equation (24) $\mathrm{IG}(A_i|C)$ is the conditional IG $\mathrm{H}(A_i)$ and $\mathrm{H}(C)$ are the entropy of $\mathrm{H}(A_i)$ and H(C)

$$\mathrm{SU}(A_i, C) = 2\left[\frac{IG(A_i|C)}{H(A_i)+H(C)}\right] \quad (27)$$

Ten skewed datasets were used for measuring the effectiveness of their approach. The feature chosen was more than five datasets using BFS and FCBF. They used WEKA tool to implement NB and FCBF. They performed different experiment to validate their techniques. In all experiments, it concluded that the proposed BFS method was better than FCBF on ten skewed datasets. Classification accuracy of BFS was 90% using Naïve Bayes.

3. ISSUES AND CHALLENGES

Recently, the popularity of big data presents some challenges for the traditional feature selection task. Meanwhile, some unique characteristics of big data also bring about new opportunities for the feature selection research. In the next few subsections, we will present these challenges of feature selection for big data analytics from the following six aspects. In particular, the challenges of structured features, linked data, multi-source data and multi-view data, streaming data and features are from the perspective of data; while the last two challenges of scalability and stability, are from the performance perspective. In feature selection techniques, there are still significant issues and challenges, which will be discussed here.

3.1 Scalability

In 1989, selecting features from a dataset with more than 20 features was called large-scale feature selection [1]. However, nowadays the number of features in many areas, such as gene analysis, can easily reach thousands or even millions. These increases computational cost and requires advanced search mechanisms, but both of these aspects also have their own issues, so the problem cannot be solved by only increasing computational power. On the other hand, the scalability of feature selection algorithms is a big problem. Usually, they require a sufficient number of samples to obtain, statically, adequate results. It is very hard to observe feature relevance score without considering the density around each sample. Some methods try to overcome this issue by memorizing only samples that are important or a summary. Novel methods and algorithms will become a necessity. To solve large-scale feature selection problems, new approaches are needed, including new search algorithms and new evaluation measures. In conclusion, we believe that the scalability of classification and feature selection methods should be given more attention to keep pace with the growth and fast streaming of the data.

3.2 Computational Cost

Most feature selection methods suffer from the problem of being computationally expensive, since they often involve a large number of evaluations. Filter approaches are generally more efficient than wrapper approaches, but experiments have shown that this is not always true [40]. To reduce the computational cost, two main factors, an efficient Search technique and a fast evaluation measure, need to be considered [42]. Table 1 shows that each feature selection method comparison. Therefore, it is still a challenge to propose efficient and effective approaches to feature selection problems.

3.3 Stability

Algorithms of feature selection for classification are often evaluated through classification accuracy. However, the stability of algorithms is also an important consideration when developing feature selection methods. A motivated example is from bioinformatics, the domain experts would like to see the same or at least similar set of genes, i.e. features to be selected, each time they obtain new samples in the presence of a small amount of perturbation. Otherwise they will not trust the algorithm when they get different sets of features while the datasets are drawn for the same problem. Due to its importance, stability of feature selection has drawn attention of the feature selection community. It is defined as the sensitivity of the selection process to data perturbation in the training set. It is found that well-known feature selection methods can select features with very low stability after perturbation is introduced to the training samples. Developing algorithms of feature selection for classification with high classification accuracy and stability is still challenging.

3.4 Representation

Unfortunately, a known issue of wrapper methods is that the search space for d features is 2^d, which is impractical when d is very large. A good representation scheme can help to reduce the search space size. It in turn helps to design new search mechanisms to improve the search ability. Another issue is that the current representations usually reflect only whether a feature is selected or not, but the feature interaction information is not shown. Feature interaction usually involves a group of features rather than a single feature. If the representation can reflect the selection or removal of groups of features, it may significantly improve the classification

performance. Therefore, a good representation scheme may help users better understand and interpret the obtained solutions.

3.5 Linked Data

Most existing algorithms of feature selection for classification work with generic datasets and always assume that data is independent and identically distributed. With the development of social media, linked data is available where instances contradict the independent and identically distributed assumption. Linked data has become ubiquitous in real world applications such as tweets in Twitter (tweets linked through hyperlinks), social networks in Facebook (people connected by friendships) and biological networks (protein interaction networks). Linked data is patently not independent and identically distributed (i.i.d.), which is among the most enduring and deeply, buried assumptions of traditional machine learning methods [43, 44]. Social media data is intrinsically linked via various types of relations such as user-post relations and user-user relations. Many linked data related learning tasks are proposed such as collective classification [45, 46], and relational clustering [47, 48], but the task of feature selection for linked data is rarely touched. There are many Issues needing further investigation for linked data such as handling noise, incomplete and unlabelled linked social media data.

3.6 Feature Construction

Feature selection does not create new features, as it only selects original features. However, if the original features are not informative enough to achieve promising performance, feature selection may not work well, yet feature construction may work well [49], [50].One of the challenges for feature construction is to decide when feature construction is needed. A measure to estimate the properties of the data might be needed to make such a decision. Meanwhile, feature selection and feature construction can be used together to improve the classification performance and reduce the dimensionality. This can be achieved in three different ways: 1) performing feature selection before feature construction; 2) performing feature construction before feature selection; and 3) simultaneously performing both feature selection and construction [49].

4. CRITICAL ANALYSIS

In our experiment, we have chosen the following three feature selection algorithms such as Correlation feature selection (CFS), Fast correlation based feature (FCBF) and symmetrical uncertainty (SU). The Datasets were selected from Kaggle Website in given TABLE 1. We were selected Heart dataset which contain 14 feature and hepatitis dataset that contained 19 feature .We applied these three algorithms and evaluate results as shown in TABLE 2.

ALGORITHM:

Based on the methodology presented before, we have used the following algorithm, named FCBF (Fast Correlation- Based Filter). [56]

```
Input: S (F1,F2, FN , C) // training data set
δ                // predefined threshold value
Output: Sbest     // an optimal subset
1 begin
2  for i = 1 to N do begin
3    calculate SUi,c for Fi;
4   if (SUi,c ≥ δ)
5     append Fi to S'list ;
6  end;
7  order S'list in descending SUi,c value;
8 Fp = getFirstElement(S'list)
9 do begin
10 Fq =getNextElement(S'list,Fp)
11 if (Fq<> NULL)
12 do begin
13 F 'q = Fq ;
14 if (SUp,q ≥ SUq,c)
15 remove Fq from S'list
16. Fq = getNextElement(S'list F'q);
17. else Fq = getNextElement(S'list, Fq);
18 end until (Fq = = NULL);
19 Fp = getNextElement(S'list,Fp);
20 end until (FP = = NULL);
21 Sbest = S'list ;
22 end;
```

Table 1. Data Used

Dataset	No of Feature
Heart	14
Hepatitis	19
Wine	13

Table 2. Number of Feature Selection

Dataset/Algorithm	CFS	FCBF	SU
Heart	10	11	11
Hepatitis	14	16	13
Wine	11	11	11

5. CONCLUSIONS AND FUTURE WORK

This paper provided a comprehensive survey of feature selection techniques, which covered all the commonly used feature selection algorithms and focused on the key factors, such as representation, search mechanisms, and the performance measures as well as the applications. Important issues and challenges were also discussed.

This survey shows that a variety of feature selection algorithms have recently attracted much attention to address feature selection tasks. A popular approach in GADP, SIP-FS and FODPSO is to improve the representation to simultaneously select features and optimize the classifiers, e.g., SVMs, KNN and NB etc. Different algorithms have their own characteristics; GADP is used for generating number of subsets of genes, FODPSO is an intelligent optimization algorithm which owns good global search ability, SIP-FS algorithm selects a reasonable and compact feature subset for data representation efficiently. To improve their effectiveness and

efficiency, it is necessary to design a cheap evaluation measure according to the specific Representation and the search mechanism of a particular feature selection technique. The proposal of novel approaches may involve methods or measures from different areas, which encourages research across multiple disciplines. In addition, combining feature selection with feature construction can potentially improve the classification performance, whereas combining feature selection with instance selection can potentially improve the efficiency.

REFERENCES

[1] Feature Selection: A Data Perspective J Li, K Cheng, S Wang, F Morstatter… - ACM Computing …, 2018 - dl.acm.or

[2] Techniques', D Dittman, R Wald Machine Learning and 2012 - ieeexplore.ieee.org

[3] "A Novel Feature Selection method for predicting heart diseases with Data Mining Techniques" R Suganya, S Rajaram, AS Abdullah, V Rajendran Asian Journal of Information…,2016

[4] "QER: a new feature selection method for sentiment analysis" T Parlar, SA Özel, F Song - Human-centric Computing and Information, 2018 – Springer

[5] "SIP-FS: a novel feature selection for data representation ." Y Guo, J Ji, H Huo, T Fang, D Li - … on Image and …, 2018 - jivp-eurasipjournals.springeropen

[6] "Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method" X Liu, J Tang - IEEE Systems Journal, 2014 - ieeexplore.ieee.org

[7] "A novel feature selection method for image classification" D Singh, AA Gnana …, 2015

[8] "A novel feature selection method based on an integrated data envelopment analysis and entropy model" SMH Bamakan, P Gholami - Procedia Computer Science, 2014 – Elsevier

[9] "A novel hybrid feature selection method for microarray data analysis." CP Lee, Y Leu - Applied Soft Computing, 2011 – Elsevie

[10] A new feature selection method for one-class classification problems

YS Jeong, IH Kang, MK Jeong… - IEEE Transactions on …, 2012 - ieeexplore.ieee.org

[11] "A new feature selection method for handling redundant information in text classification." Y Wang, L Feng - Frontiers of Information Technology & Electronic …, 2018 – Springer

[12] "A new feature selection method for text categorization of customer reviews." M Liu, X Lu, J Song - Communications in Statistics-Simulation and …, 2016 - Taylor & Francis

[13] "A Novel Feature Selection Method Based on Extreme Learning Machine and Fractional-Order Darwinian PSO"YY Wang, H Zhang, CH Qiu, SR Xia - Computational intelligence and …, 2018 - hindawi.com

[14] "Peculiar Genes Selection: A new features selection method to improve classification performances in imbalanced data sets"F Martina, M Beccuti, G Balbo, F Cordero - PloS one, 2017 - journals.plos.org

[15] "subset Liu, W Wang, Q Zhao, X Shen, M Konan - Pattern Recognition Letters, 2017 – Elsevier

[16] "A new feature selection method for internet traffic classification using ml" L Zhen, L Qiong - Physics Procedia, 2012 - core.ac.

[17] "A Survey on Evolutionary Computation Approaches to Feature Selection." Bing Xue, Member, IEEE, Mengjie Zhang, Senior Member, IEEE, Will N. Browne, Member, IEEE, and Xin Yao, Fellow, IEEE

[18] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003; 3:1157–82

[19] Langley P. Selection of relevant features in machine learning. In: AAAI fall symp relevance; 1994.

[20] Blum AL, Langley P. Selection of relevant features and examples imachine learning. Artif Intell 1997; 97:245–7

[21] "Theoretical and empirical analysis of ReliefF and RReliefF" M Robnik-Šikonja, I Kononenko - Machine learning, 2003 – Springer

[22] Feng Yang and K. Z. Mao. 2011. Robust feature selection for microarray data based on multicriterion fusion. IEEE/ACM Trans. Comput. Biol. Bioinform. 8, 4 (2011), 1080–1092.

[23] Lei Shi, Liang Du, and Yi-Dong Shen. 2014. Robust spectral learning for unsupervised feature selection. In ICDM. 977–982

[24] Hiromasa Arai, Crystal Maung, Ke Xu, and Haim Schweitzer. 2016. Unsupervised feature selection by heuristic search with provable bounds on suboptimality. In AAAI. 666–672.

[25] Salem Alelyani, Jiliang Tang, and Huan Liu. 2013. Feature selection for clustering: A review. Data Clustering: Algorithms and Applications 29 (2013).

[26] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 19 (2007), 2507–2517.

[27] Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed. 2016. Supervised, unsupervised, and semi supervised feature selection: A review on gene selection. IEEE/ACM TCBB 13, 5 (2016), 971–989.

[28] Qiang Shen, Ren Diao, and Pan Su. 2012. Feature selection ensemble. Turing-100 10 (2012), 289–306.

[29] Parlar T, Ozel SA (2016) A new feature selection method for sentiment analysis of Turkish reviews. In: International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, Sinaia, pp 1–6.

[30] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," Tools with Artificial Intelligence, IEEE International Conference on, pp. 310–317, 2007.

[31] K. S. Chuang, H. L. Tzeng, S. Chen, J. Wu, and T. J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," Comput. Med. Imag. Graph., vol. 30, no. 1, pp. 9–15, Jan. 2006.

[32] J. Tang and X. Liu, "Classification of breast mass in mammography with an improved level set segmentation by combining morphological features and texture features," in Multi-Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies. New York, NY, USA: Springer-Verlag, 2011, pp. 119–135.

[33] G. Forman, J Mach Learn Res., 2003, p.1289

[34] H.L. Huang, F.L. Chang, ESVM: evolutionary support vector machine for automatic feature selection and classification of microarray data, Bio Systems 90 (2007) 516–528

[35] J.H. Cho, D. Lee, J.H. Park, I.B. Lee, New gene selection method for classification of cancer subtypes considering within-class variation, FEBS Lett. 551 (2003) 3–7.

[36] Guyon, J. Westion, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422

[37] L. Wang, F. Chu, W. Xie, Accurate cancer classification using expressions of very few genes, IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (1) (2007) 40–53.

[38] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, D. Geman, Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (2005) 3896–3904

[39] W. Siedlecki and J. Sklansky, "A note on genetic algorithms forlarge-scale feature selection," Pattern Recognit. Lett., vol. 10, no. 5,pp. 335–347, 1989.

[40] B. Xue, "Particle swarm optimisation for feature selection," Ph.D.dissertation, School Eng. Comput. Sci., Victoria Univ. Wellington,Wellington, New Zealand, 2014.

[41] E. Amaldi and V. Kann, "On the approximability of minimizing non zero variables or unsatisfied relations in linear systems," Theor.Comput. Sci., vol. 209, nos. 1–2, pp. 237–260, 1998.

[42] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, nos. 1–4, pp. 131–156, 1997.

[43] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In International Conference on Machine Learning, pages 259–266, 2002.

[44] B. Taskar, P. Abbeel, M.F. Wong, and D. Koller. Label and link prediction in relational data. In Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data. Citeseer, 2003

[45] S.A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. The Journal of Machine Learning Research, 8:935–983, 2007.

[46] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. AI magazine, 29(3):93, 2008.

[47] B. Long, Z.M. Zhang, X. Wu, and P.S. Yu. Spectral clustering for multi-type relational data. In Proceedings of the 23rd international conference on Machine learning, pages 585–592. ACM, 2006.

[48] B. Long, Z.M. Zhang, and P.S. Yu. A probabilistic framework for relational clustering. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 470–479. ACM, 2007.

[49] A. S. U. Kamath, K. De Jong, and A. Shehu, "Effective automated feature construction and selection for classification of biological sequences," PLoS One, vol. 9, no. 7, 2014, Art. ID e99982.

[50] K. Neshatian, "Feature manipulation with genetic programming," Ph.D. dissertation, Dept. Comput. Sci., Victoria Univ. Wellington, Wellington, New Zealand, 2010.

[51] Iqbal, M., Rehman, S. U., Gillani, S., & Asghar, S. (2015). An empirical evaluation of feature selection methods. In Improving Knowledge Discovery through the Integration of Data Mining Techniques (pp. 233-258). IGI Global

[52] Jameel, S., & Rehman, S. U. (2018). An optimal feature selection method using a modified wrapper-based ant colony optimisation. Journal of the National Science Foundation of Sri Lanka, 46(2)

[53] Iqbal, M., & ur Rehman, S. (2016). Association Rule Mining Using Computational Intelligence Technique. International Journal of Computer Science and Information Security, 14(12), 416

[54] G. H. John, R. Kohavi, K. Pfleger, "Irrelevant feature and the subset selection problem," in Proc. of the Eleventh International Conferenc on Machine Learning, pp. 121-129, 1994.