The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 8 || Pages || PP 57-67 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

A Survey and Comparative Study of Filter and Wrapper Feature Selection Techniques

Nyararai Mlambo[1], Wilson K. Cheruiyot[2], Michael W. Kimwele[3]
[1]*Department of Information Systems, University of Rwanda, Kigali, Rwanda, Department of Computing, Jomo Kenyatta University of Agriculture and Technology*
[2]*Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Kigali, Rwanda*
[3]*Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya*

----------ABSTRACT----------
Feature selection is considered as a problem of global combinatorial optimization in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data. However, identification of useful features from hundreds or even thousands of related features is not an easy task. Selecting relevant genes from microarray data becomes even more challenging owing to the high dimensionality of features, multiclass categories involved and the usually small sample size. In order to improve the prediction accuracy and to avoid incomprehensibility due to the number of features different feature selection techniques can be implemented. This survey classifies and analyzes different approaches, aiming to not only provide a comprehensive presentation but also discuss challenges and various performance parameters. The techniques are generally classified into three; filter, wrapper and hybrid.

Keywords: *Machine Learning, Feature Selection, Filter, Wrapper, Classification*

Date of Submission: 16 August 2016 Date of Accepted: 27 August 2016

I. INTRODUCTION

Data mining is an inevitable step in knowledge discovery and the knowledge obtained as a result of data mining is used in in many trends; like business and medical use [1]. Sometimes the collected data contain so many attributes (features) of each entity, some of which are irrelevant or redundant. These features not only have no role in in the process of knowledge discovery, but they also increase the complexity and incomprehensibility of the results[1]. Feature Selection is a preprocessing technique which is used to identify the significant attributes, which play a dominant role in the task of classification. This leads to the dimensionality reduction. By applying different approaches features can be reduced. The reduced feature set improves the accuracy of the classification task in comparison of applying the classification task on the original data set [2]. Feature selection does not only reduce the dimensionality of data, it also reduces the computational cost and gains a good classification performance [3]. In many real world problems, feature selection is a must due to the abundance of noisy, irrelevant or misleading features. Different feature selection techniques have been widely used in a wide spectrum of applications, such as genomic analysis, information retrieval and text categorization. Feature selections are labeled as (i) *Relevant:* A relevant feature is one which is related to the minimum cardinality for achieving the high predictive data. (ii) Irrelevant: Irrelevant features do not have any control on the output here the values are generated at random for each data. (iii) *Redundant:* unwanted features occur in the data [6]. Selecting an original subset feature to the relevant one is not an easy task.

The overall literature survey shows that the common classification techniques achieve an accuracy rate of above 70% when applied in the medical field. These techniques; neural networks, Naïve Bayesian classifier, association rules, Decision trees etc, are mostly considered to be easy to understand and implement. However, if the size of data becomes too large then developing a model, generating rules and constructing a tree become a problem. It has also been noted in the literature that there is no so-called "best feature selection method". This research therefore proposes to come up with a novel hybrid feature selection algorithm which will be applied on medical datasets with different classifiers. The research is going to cross validate this with some existing classifiers to measure the accuracy.

Different researchers have come up with different feature selection algorithms with different selection criteria. However, these works have shown that no single criterion is best for all applications, and in fact, different learning algorithms can produce results that can be similar. The success of a given learning algorithm can be determined by a number of factors, among them, the nature of the data used for characterization of the task to be learned. The emergence of new application disciplines, resulting in tasks with large sizes of feature space is proving to be a challenge for existing algorithms. This is because some of them were designed for domains

exploring data of not more than 40 features [5]. Typical examples of the new application fields with high dimensional data are gene selection, where expression levels of many genes are recorded by microarray data but only a few genes can be used for cancer classification, and text categorization.

The feature space in text categorization problems is determined by the vocabularies from the natural language documents whose size is commonly of hundreds of thousands of words [4] and [5]. When the dimensionality of data reduces, the feature selection will improve the performance of learning algorithms and also improves the comprehensibility of the data models. To date, there are so many feature selection algorithms as indicated in the literature, among them[13].

The feature selection algorithms in the literature are diverse and justified by theoretical arguments. In most cases they yield substantially different results even when applied to the same data. [1]noted that these many algorithms available are biased when it comes to dimensionality and none of them stands to be the best for all applications. This therefore makes it difficult to determine the feature selection technique that best suits a new data set in a new application.

II. FEATURE SELECTION

2.1 Introduction

Feature selection is one of the best tools in machine learning. It aims to reduce dimensionality for building comprehensible learning models with good generalization performance. Many feature selection methods have been proposed in the literature. Without knowing the relevant features in advance of the real data set, it is very difficult to find out the effectiveness of the feature selection methods, because data sets may include many challenges such as the huge number of irrelevant and redundant features, noisy data, and high dimensionality in term of features or samples. Therefore, the performance of the feature selection method relies on the performance of the learning method [9]. According to the literature, there are many performance measures such as accuracy, computer resources, ratio of feature selection, etc. Most researchers agree that there is no so-called "best method" [10] as cited in [9]. Therefore, the new feature selection methods are constantly increasing to tackle the specific problem (as mentioned above) with different strategies. The different strategies include reinterpreting existing algorithms, ensuring better behavior of feature selection using hybrid methods, etc. New methods are also needed for still unresolved problems [8]; [7].

2.2 Feature Selection Objectives

The feature selection problem has been studied by the statistics and machine learning communities for many years. It has received more attention recently because of enthusiastic research in data mining. A fundamental problem of machine learning is to approximate the functional relationship f() between an input $X = \{x_1, x_2, \ldots, x_M\}$and an output Y, based on a memory of data points, $\{X_i, Y_i)$, i = 1, 2, …, N, usually the X_i's are the vectors of reals and the Y_i's are real numbers. Sometimes the output Y is not determined by the complete set of the input features $\{x_1, x_2, \ldots, x_M\}$, instead, it is decided only by a subset of them $\{x_{(1)}, x_{(2)}, \ldots, X_{(m)}\}$, where m < M. With sufficient data and time, it is fine to use the input features, including those irrelevant features, to approximate the underlying function between the input and the output. But in practice, there are two problems which may be evoked by the irrelevant features involved in the learning process.

1. The irrelevant input features will induce greater computational cost. For example, using cached kd-trees, locally weighted linear regression's computational expense is $O(m^3 + m^2 \log N)$ for doing a single prediction, where N is the number of data points in memory and m is the number of features used. Apparently, with more features, the computational cost for predictions will increase polynomially; especially when there are a large number of such predictions, the computational cost will increase immensely.
2. The irrelevant input features may lead to overfitting. For example, in the domain of medical diagnosis, our purpose is to infer the relationship between the symptoms and their corresponding diagnosis. If by mistake we include the patient ID number as one input feature, an over-tuned machine learning process may come to the conclusion that the illness is determined by the ID number.

The objectives of feature selection are manifold and different feature selection algorithms may have various objectives to achieve. The advantages of feature selection techniques come at a certain cost, as the search for a subset of relevant features introduces an additional layer of complexity in the modeling task. Given below are some of the common objectives given in the literature:

(i) Find the minimally sized feature subset that is necessary and sufficient to the target concept [11].

(ii) Select a subset of *N* features from a set of *M* features, *N <M*, such that the value of a criterion function is optimized over all subsets of size *N [12]*.

(iii) Choose a subset of features for improving prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features [5].

(iv) Select a small subset such that the resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution given all feature values [5].

2.3 Feature Selection Process

A typical feature selection process consists of four basic steps namely, subset generation, subset evaluation, stopping criterion, and result validation [34]. Subset generation is a search procedure that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied [13];[34]. Feature selection can be found in many areas of data mining such as classification, clustering, association rules, and regression [13]. Figure 1.1 [9] show the four key steps of feature selection.

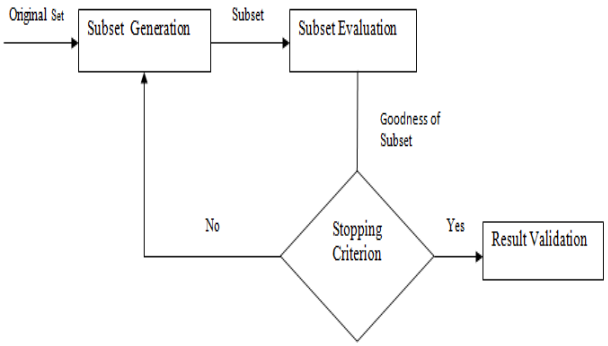

Figure 2.1: Four Key Steps of Feature Selection

2.3.1 Subset Generation

Subset generation is a heuristic search in which each state specifies a candidate subset for evaluation in the search space [9].The nature of the subset generation process is determined by two issues namely; (i) Successor generation- this decides the search starting point, which influences the search direction. Search start point can be an empty set, the full set, or a randomly generated subset [5]. To decide the search starting points at each state, forward, backward compound, weighting, and random methods may be considered. (ii) Search organization, which is responsible for the feature selection process with a specific strategy, such as sequential search, exponential search or random search [9].

Sequential Forward Search (SFS), sequential backward search, and bidirectional search are some variations to the greedy hill climbing method. Algorithms with sequential searches are fast in time complexity of O (N^2) and simple to implement [5]. While SFS starts with an empty set of selected features and each step of the algorithm adds one of the informative features to the set, SBS starts with the full set of features and in each step, one of the redundant ot irrelevant features is omitted. Bidirectional search adds and deletes the features simultaneously [1]. Both SFS and SBS algorithms have the "nesting effect" problem, which means that while a change is considered positive, there is no chance of re-evaluating that feature. Complete search algorithms can also be used for this purpose. Any example is the Best First Search (BFS), which allows backtracking in the search space.

2.3.2 Subset Evaluation

The candidate feature subsets need to be evaluated by some criteria so that the best feature subset can be determined according to the goodness measure. An optimal feature subset generated by one criterion may not be the same according to other evaluation criteria. There are two broadly used evaluation criteria, based on their dependency and independence on the algorithms [9].

Independent Criteria

An independent criterion is typically used in filter algorithm. It tries to measure the intrinsic characteristics of the dataset without involving any mining algorithm. Some popular criteria are probability of error measures, information measures, and dependency measures [33].

Dependent Criteria

A dependent criterion is used by wrapper models. The criterion is measured with a specific mining algorithm. The performance of the mining algorithm is applied to determine the goodness of the feature subset. Usually, a dependent criterion yields better performance than an independent criterion for the predefined mining algorithm. However, the selected feature subset may not be suitable for other mining algorithms, and the computational

cost is often expensive. For classification problems, the predicting accuracy of unseen instances is widely used to select feature subset which yields high testing accuracy[33].

2.3.3 Stopping Criteria

The feature selection process terminates when a stopping criterion is achieved. Some frequentlyused stopping criteria are as follows:

(i) The search is completed.
(ii) Subsequent addition or deletion of any feature does not yield better result.
(iii) A sufficiently good subset is selected.
(iv) Some given bound, i.e. the number of iterations or the number of selected features, is reached.

2.3.3 Result Validation

The prior knowledge of the underlying dataset is often used to directly validate the result of a feature selection process. For a synthetic dataset, the relevant feature subset and irrelevant feature subset is usually known. The former is expected to appear in the resulting feature subset, while the latter is not. Thus we can validate the results by comparing the known relevant and/or irrelevant features with the feature subset produced by the feature selection algorithm. However, in real world applications, such a prior knowledge is usually unknown. Validation of results must occur in an indirect way. A frequently used method is to conduct experiments not only on the selected feature subset, but also the whole feature set. The resulted validation is achieved by comparing the performance of these before-and-after feature selection experiments.

2.4 Feature Selection Algorithms

Feature selection algorithms (FSA) can be classified into different groups according to the subset generation methods, the subset evaluation methods, or data mining tasks. The different algorithms present different conceptual frameworks. Under subset generation methods, the feature selection algorithms can be categorized into four groups: complete search, sequential search, random search, and integral weighting. Under subset evaluation criteria, they can be categorized into three groups: filters, wrappers, and hybrids. Under data mining task criteria, they can be categorized into two groups: supervised learning and unsupervised learning. This research will consider the subset evaluation criteria. The general feature selection algorithms comprise of two categories: the filter and wrapper methods [14] and [15].

A *Filter* method evaluates the relevance of features according to some discriminating criterion that looks at the general characteristics of the data [16]. The results from such a method are usually a ranked list of features, where the features at the top of the list are relevant and the features at the bottom of the list are not so relevant or totally irrelevant. A *Wrapper*, however, evaluates the relevance of features by using a classifier and selects only the most relevant subset of features. Therefore, the results obtained from a Wrapper are different to that of a Filter because it actually selects a subset of the most relevant features rather than list all features in order of relevance[17].

The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model. The filters are efficient because of their independence from learning algorithms, while wrappers can obtain higher classification accuracy with deficiency in generalization and computational cost [18]. The following table, from [9], gives a general algorithm for feature selection.

==========

INPUTS:

X: *Set of features of a data set having n features*
SG: *Successor Generator Operator*
E: *Evaluation measure (dependent or independent)*
O: *Stopping Criteria*

OUTPUT:

X_{opt}: *Optional feature set or weighted features*

==========

Initialize:

X':= Start_point(X);
$X_{opt:=}$ *{Best of X' using E};*

Repeat:

X':= Search_Strategy (X', SG(E), X);

X_{opt}:= {Best of X' according to E};
If E(X') ≥ E(X_{opt}) or (E(X) == E(X_{opt}) & |X'| < |X_{opt}|)
Then X_{opt} = X';
Until *Stop criteria is not found;*
==========

Table 2.1: General algorithm for feature selection

From figure 2.1, present in the list is a function of the evaluation measure which defines the expansion order. Heuristic search algorithms maintain this list of open nodes and the weighting is the value of the heuristic. Sequential algorithms maintain |X'| = 1 whereas random search methods eg evolutionary algorithmsare characterized by |X'| ≥ 1 (the list is the population and the weighting is the fitness value of the individuals).

2.5 Filter Methods

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Generally, filters are less computationally intensive but they produce a feature set which is not tuned to a specific type of model. We will look at some of the common filter methods.

2.5.1 Information Gain

Information gain (IG), also called Kullback-Leibler distance, is the measure of entropy gained due to the operations performed on a given data / random variable. Entropy is essentially the measure of variation in the data, the lesser the variation we have the lesser the entropy is and the greater the data is correlated. In other words, a feature is more important if its IG is larger. The IG treats all features as independent. We will use the concept of normalized information gains $G'_i$for feature f_i.The normalized IG, introduced by Setiono and Liu [19], calculates information gains as follows:
The information contained in the whole training set is;

$$I(S) = -\sum_{J=1}^{K} p(C_J)\log_2 p(C_J) \qquad (1)$$

Where $p(C_j) = n_j/n$ is the fraction of samples X from class C_j, j = 1..k. Continuous features are discretized to compute information associated with a single feature. Let n_{ik}be the number of samples for which features f_i takes a value inside the interval $r_k(f_i)$ and n_{ikj} be the number of such samples X for which X C_j. Information contained in the subset $S_{ik}$of samples with f_i in the interval $r_k(f_i)$ is:

$$I(S_{ik}) = -\sum_{J=1}^{K} p_{ikJ}\log_2 p_{ikJ}; \quad p_{ikJ} = n_{ikJ}/n_{ik} \qquad (2)$$

Summing (or integrating) $I(S_{ik})$ over all M intervals information $E_i$contained in all subsets of feature f_i is computed. The same information may also be computed directly.

$$E_I = \sum_{k=1}^{M} p_{ik} I(S_{ik}); \quad I_I = -\sum_{k=1}^{M} p_{ik}\log_2 p_{ik}; \quad p_{ik} = \frac{n_{ik}}{n} \qquad (3)$$

The information gain and normalized information gains are respectively given as:

$$\mathbf{G_i = I(S) - E_i;\ G'_i = G_i / I_i} \qquad (4)$$

2.5.2 Relief Feature Selection Algorithm

Relief algorithm is a kind of feature weighting algorithm, which gives different weights according to the relevance of features and categories [20]. Its strengths are that it is not dependent on heuristics, runs in low-order polynomial time, and is noise-tolerant and robust to feature interactions, as well as being applicable for binary or continuous data; however, it does not discriminate between redundant features, and low numbers of training instances fool the algorithm. Kira and Rendell [21] proposed the first version of Relief. Its strengths are that it is not dependent on heuristics, runs in low-order polynomial time, and is noise-tolerant and robust to feature interactions, as well as being applicable for binary or continuous data [21].
To date so many variants have been proposed, among them RReliefF[24], HRelief [22], SWRF* [23], etc. The pseudocode for Relief is shown in Table 2.2 below. The weight of an attribute is updated iteratively as follows: A sample is selected from the data, and the nearest neighboring sample that belongs to the same class (*nearest hit*) and the nearest neighboring sample that belongs to the opposite class (*nearest miss*) are identified.

A change in attribute value accompanied by a change in class leads to upweighting of the attribute based on the intuition that the attribute change could be responsible for the class change. On the other hand, a change in attribute value accompanied by no change in class leads to downweighting of the attribute based on the observation that the attribute change had no effect on the class. This procedure of updating the weight of the attribute is performed for a random set of samples in the data or for every sample in the data. The weight updates are then averaged so that the final weight is in the range [−1, 1]. The attribute weight estimated by Relief has a probabilistic interpretation [23].

set W[a] = 0 for each attribute a
for i = 1 to n do
select sample s_i from data at random
find nearest hit s_h and nearest miss s_m
for each attribute a do
$\Delta W_i[a] = diff(a, s_i, s_m) - diff(a, s_i, s_h)$
$W[a] = W[a] + \Delta W_i[a]$
end for
end for
for each attribute a do
$W[a] = W[a] / n$
end for
where $diff(a, s_i, s_j) = 0$, if $s_i[a] = s_j[a]$
$= 1$, if $s_i[a] \neq s_j[a]$

Table 2.2: Pseudocode for Relief Algorithm

2.5.3 Correlation Feature Selection

The correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". The following equation gives the merit of a feature subset S consisting of k features:

$$\mathbf{Merit_{Sk}} = \frac{\mathbf{k\overline{r_{cf}}}}{\sqrt{\mathbf{k + k(k-1)\overline{r_{ff}}}}} \quad (5)$$

Here, $r_{cf}$is the average value of all feature-classification correlations, and $r_{ff}$is the average value of all feature-feature correlations. The CFS criterion is defined as follows:

$$CFS = \max_{S_k}\left[\frac{r_{cf1} + r_{cf2} + \cdots + r_{cfk}}{\sqrt{k + 2(r_{f1f2} + \cdots + r_{fifj} + \cdots + r_{fkf1}}}\right] \quad (6)$$

The $r_{cfk}$and r_{fifj} variables are referred to as correlations. Let xi be the set membership indicator function for feature fi; then the above equation can be rewritten as an optimization problem:

$$\mathrm{CFS} = \max_{x\in\{0,1\}^n}\left[\frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} x_i + \sum_{i\neq j} 2b_{ij}x_i x_j}\right] \quad (7)$$

2.5.4 Symmetrical Uncertainty

Symmetrical uncertainty measures the correlation between pairs of attributes using normalization of information gain. The output of this method results is a feature ranking. The method can be computed as follows (Hernandez-Torruco et al, 2014):

$$U(A,B) = 2 * \frac{MI(A,B)}{\text{Entropy}(A) + \text{Entropy}(B)}$$

$$MI(A,B) = \sum P(A,B)\log_2\frac{P(A,B)}{P(A)\,P(B)}, \quad (8)$$

where P(x) is the marginal probability of feature X, R_A is the range of feature A, and P(A, B) is the joint probability of features A and B.

Most of these methods do not perform feature selection but only feature ranking, they are usually combined with another method when one needs to find out the appropriate number of attributes. Forward selection, backward elimination, bi-directional search, best first search, genetic search and other methods are often used on this task.

2.6 Wrapper Methods

As explained in 2.4, a wrapper evaluates the relevance of features by using a classifier and selects only the most relevant subset of features. Figure 2.2 below shows the main operational differences between Filter and Wrapper feature selection techniques. The search algorithm, induction algorithm, evaluation metric are three components of Feature selection.

The original feature space has N features. The target feature space is a subset of original feature space, including k features selected from N features, k is the number between 1 and N. Since the number of possible feature subsets is the power set of N, search algorithm focus on how to search the feature subsets space to get the target feature subset as soon as possible. In wrapper model, feature subsets selected by search algorithm will pass a classifier to train and test on the given data. This classifier is designed for evaluating the performance of selected feature subset, so we call it induction algorithm. The classification result from wrapper model is compared with the correct label of the data in the evaluation stage. Based on the prediction error, we will decide how to search next or stop search.

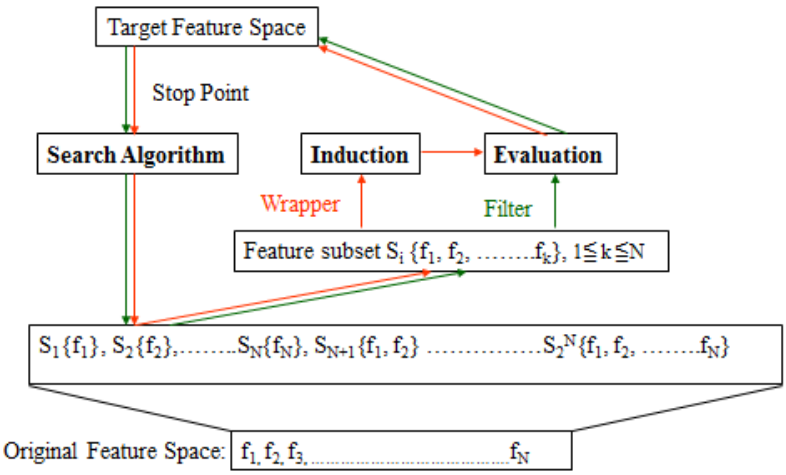

Figure 2.2: Feature Selection Algorithm Design

2.6.1 Heuristic Selection Algorithms

Greedy hill climbing algorithm, branch and bound method, beam search and best first algorithm are the heuristic methods of feature selection problem. Greedy hill climbing algorithm considers all local changes in order to select the relevant features [25]. In this algorithm adding a feature to the selected features and deleting one of them can be considered as local changes. SFS (Sequential Forward Selection) and SBS (Sequential Backward Selection) are two kinds of hill climbing. While SFS starts with empty set of selected features and each step of the algorithm adds one of the informative features to the set, SBS starts will the full set of features and in each step, one of the redundant or irrelevant features is omitted. Another method is bi-directional search; which considers both adding and deleting the features simultaneously. Both SFS and SBS algorithms have the "nesting effect" problem, which means that while a change is considered positive, there is no chance of re-evaluating that feature.

Best first search is another method based on artificial intelligence methods, which allow backtracking in the search space [25]. This algorithm, like greedy hill climbing algorithm, makes use of local changes in the search space. But in contrast to it when the path for reaching the optimum solution is not hopeful, it is possible to backtrack the search space. Given below are some of the common sequential algorithms.

1) Sequential forward selection (SFS)

SFS is the simplest greedy search algorithm. It starts with an empty set, sequentially adding the feature x^+ that maximizes $J(Y_k + x^+)$ when combined with the features Y_k that have already been selected.

1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $X^+ = \arg\max J(Y_k + x)$
   $x \notin Y_k$
3. Update $Y_{k+1} = Y_k + x^+$; $k = k+1$
4. Go to 2

Table 2.3: Algorithm for Sequential Forward Selection

SFS performs best when the optimal subset is small. When the search is near the empty set, a large number of states can be potentially evaluated. Towards the full set, the region examined by SFS is narrower since most features have already selected.

2) Sequential Backward Selection (SBS)

SBS works in the opposite direction of SFS. It starts from the full set, sequentially removing the feature x^- that least reduces the value of the objective function $J(Y - x^-)$. Removing a feature may actually increase the objective function $J(Y_k - x^-) > J(Y_k)$. SBS works best when the optimal feature subset is large, since it spends most of its time visiting large subsets. The main limitation of SBS is its inability to reevaluate the usefulness of a feature after it has been discarded.

1. Start with the full set $Y_0 = X$
2. Remove the worst feature $x^- = \arg\max J(Y_k - x) \quad X \in Y_k$
3. Update $Y_{k+1} = Y_k - x^-$; $k = k + 1$
4. Go to 2

Table 2.4: Algorithm for Sequential Background Selection

3) Plus-L minus-R Selection (LRS)

This algorithm is a generalization of SFS and SBS. If L>R, LRS starts from the empty set and repeatedly adds L features and removes R features and if L<R, LRS starts from the full set and repeatedly removes R features followed by L additions. LRS attempts to compensate for the weaknesses of SFS and SB[25S with some backtracking capabilities. Its main limitation is the lack of a theory to help predict the optimal values of L and R.

1. If L>R then $Y_0 = \{ \emptyset \}$
   Else $Y_0 = X$; go to step 3
2. Repeat L times
   $X^+ = \arg\max J(Y_k + x)$
   $x \notin Y_k$
   $Y_{k+1} = Y_k + x^+$; $k = k + 1$
3. Repeat R times
   $x^- = \arg\max J(Y_k - x)$
   $x \in Y_k$
   $Y_{k+1} = Y =_k - x^-$; $k = k + 1$
4. Go to 2

Table 2.5: Algorithm for LRS

4) Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward (SFBS)

Sequential floating selection is an extension to LRS with flexible backtracking capabilities. Rather than fixing the values of L and R, these floating methods allow those values to be determined from the data. The dimensionality of the subset during the search can be thought to be "floating" up and down. There are two floating methods namely; Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward Selection (SFBS). SFFS starts from the empty set. After each forward step, SFFS performs backward steps as long as the objective function increases. SFBS starts from the full set, after each backward step, it performs forward steps as long as the objective function increases.

1. $Y = \{ \emptyset \}$
2. Select the best feature
   $X^+ = \arg\max J(Y_k + x)$
   $x \notin Y_k$
   $Y_k = Y_k + x^+$; $k = k + 1$
3. Select the worst feature *
   $x^- = \arg\max J (Y_k - x)$
   $x \in Y_k$
4. If $J(Y_k - x^-) > J (Y_k)$ then
   $Y_{k+1} = Y_k - x^-$; $k = k + 1$
   Go to step 3
   Else
   Go to step 2

Table 2.6: SFFS Algorithm (SFBS is analogous)

Heuristic algorithms perform better than complete search methods, but recently meta-heuristic algorithms like Genetic Algorithm (GA), Particle Swarm Intelligence Optimization (PSO) and Ant Colony Optimization (ACO) show more desirable results while comparing time complexities.

2.6.2 Meta-Heuristic Search Algorithms

A Meta-heuristic is formally defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space, learning strategies are used to structure information in order to find efficiently near-optimal solutions. Meta-heuristic algorithms are among these approximate techniques which can be used to solve complex problems. Most widely known Meta-heuristic algorithms are Genetic algorithm (GA), simulated annealing (SA) and Tabu search (TS). Genetic algorithm (GA) emulate the evolutionary process in nature, whereas tabu search (TS) exploits the memory structure in living beings, simulated annealing (SA) imitates the annealing process in crystalline solids [32].

2.6.2.1 Genetic algorithm

Genetic Algorithm is a Meta-heuristic algorithm that aims to find solutions to NP-hard problems. The basic idea of Genetic Algorithms is to first generate an initial population randomly which consist of individual solution to the problem called Chromosomes, and then evolve this population after a number of iterations called Generations. During each generation, each chromosome is evaluated, using some measure of fitness. To create the next generation, new chromosomes, called offspring, are formed by either merging two chromosomes from current generation using a crossover operator or modifying a chromosome using a mutation operator. A new generation is formed by selection, according to the fitness values, some of the parents and offspring, and rejecting others so as to keep the population size constant. Fitter chromosomes have higher probabilities of being selected. After several generations, the algorithms converge to the best chromosome, which hopefully represents the optimum or suboptimal solution to the problem (Said et al., 2014).

2.6.2.2 Tabu Search

Tabu search is the technique that keeps track of the regions of the solution space that have already been searched in order to avoid repeating the search near these areas [8]. It starts from a random initial solution and successively moves to one of the neighbors of the current solution. The difference of tabu search from other Meta-heuristic approaches is based on the notion of tabu list, which is a special short term memory. That is composed of previously visited solutions that include prohibited moves. In fact, short term memory stores only some of the attributes of solutions instead of whole solution. So it gives no permission to revisited solutions and then avoids cycling and being stuck in local optima.

2.6.2.3 Simulated annealing

Simulated Annealing is an early Meta-heuristic algorithm originating from an analogy of how an optimal atom configuration is found in statistical mechanics. It uses temperature as an explicit strategy to guide the search. In Simulated Annealing, the solution space is usually explored by taking random tries. The Simulated Annealing procedure randomly generates a large number of possible solutions, keeping both good and bad solutions.

As the simulation progresses, the requirements for replacing an existing solution or staying in the pool becomes stricter and stricter, mimicking the slow cooling of metallic annealing. Eventually, the process yields a small set of optimal solutions. Simulated Annealing advantage over other methods is its ability to obviate being trapped in local minima [32].

2.7 Hybrid Feature Selection Methods

In recent years a lot of research has been going on hybrid feature selection methods. The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

The literature has shown that there exist, though very few, some hybrid feature selection techniques for classification in medical data mining. Different classifiers, with different stopping criteria have been applied. [26] stated that the choice of the best technique to a specific problem can be decided by experimenting many possibilities based on the measures such as accuracy, speed, robustness, scalability and interpretability. The following are some of the proposed hybrid techniques.

Das [27] proposed a hybrid algorithm that uses boosting and incorporates some of the features of wrapper methods into a fast filter method for feature selection. The empirical results are reported on six real world datasets from the UCI repository, showing that hybrid algorithm is competitive with wrapper methods while being much faster, and scales well to datasets with thousands of features. To determine an optimal feature, [28], a hybrid feature selection method which is a fusion Correlated-based Feature Selection (CFS), Support Vector Machine (SVM) and Genetic Algorithm. The proposed method reduces the computational resource while

maintaining the detection and false positive rate within tolerable range. It also reduces the training time and testing time.

Asha et al [30] proposed a hybrid model to classify the diabetes patients data. The hybrid model encompasses k-means clustering, k-nearest neighbor classification and correlation feature selection. Rahendran et al[31] proposed a method to classify the brain tumor in the CT scan. CT brain images are preprocessed using median filtering process and features are extracted using canny edge detention technique. Frequent patterns from the CT scan images are generated using frequent pattern tree algorithm. The decision tree algorithm is used to classify the medical images for diagnosis. This proposed method proved more accurate than a conventional method. Also for patients' disease classification, [26] proposed a hybrid approach which is a combination of CART decision tree classifier with clustering and feature selection on breast cancer data sets. The effectiveness of the hybrid approach is compared against CART with feature selection, classification with clustering and without feature selection in terms of accuracy. The experimental results demonstrate that the hybrid approach is better than CART with FS and cascading of classification and clustering without FS.

In the same year, [2] presented a novel approach for feature selection by using association and correlation mechanisms. A Two stage hybrid selection algorithm for diagnosing erythemato-squamous diseases was also proposed by [18]. The two-stage algorithms adopt Support Vector Machines as a classification tool and the extended Sequential Forward Search (SFS), Sequential Forward Floating Search (SFFS) and Sequantial Backward Floating Search (SBFS) as search strategies.[29] proposed a supervised feature selection method based on Rough Set Quick Reduct hybridized with Improved Harmony Search

Algorithm. The Rough Set Improved Harmony Search Quick Reduct (RS-IHS-QR) algorithm is a relatively new population-based meta-heuristic optimization algorithm. The proposed algorithm reveals more than 90 % classification accuracy in most of the cases and the time taken to reduce the dataset also decreased than the existing methods.

2.8 Summary

The feature selection algorithms in the literature are diverse and justified by theoretical arguments. In most cases they yield substantially different results even when applied to the same data. It is also noted that these many algorithms available are biased when it comes to dimensionality and none of them stands to be the best for all applications. This therefore makes it difficult to determine the feature selection technique that best suits a new data set in a new application. The available hybrid techniques are still few but they also behave in the same way.

REFERENCES

[1] Mohammad-Reza Feizi-Derakhshi, and Manizheh Ghaemi. " Classifying Different Feature Selection Algorithms Based on the Search Strategies". International Conference on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014) Dubai (UAE).

[2] Rajeswari, K, V. Vaithiyanathan and Shailaja V. Pede. Feature Selection for Classification in Medical Data Mining. International Journal of Emerging Trends and Technology in Medical Data Mining. International Journal of Emerging Trends and Technology 2010.

[3] Hua JP, Tembe WD, Dougherty ER :"Performance of feature selection methods in the classification of high-dimension data". *Pattern Recognit*, 2009, 42:409–424.

[4] Laura Maria Cannas, "A Framework for Feature Selection In High-Dimensional Domains". Dissertation, University of Cagliari, 2012

[5] Tan Feng, "Improving Feature Selection Techniques for Machine Learning". Dissertation,2007, Georgia State University

[6] Kalaiselvi. R, Premadevi. P. Me, Hamsathvani, M. (2014): Survey on Semi- Supervised Feature Selection In Data Mining. International Journal of Computer Science and Information Technologies, Volume 5(6), 2014, 7114-7117.

[7] B. Chidlovskii, L. Lecerf, "Scalable feature selection for multi-class problems," Machine Learning and Knowledge discovery Databases, Vol. 5211, 2013, pp227-240

[8] S. Loscalzo, L. Yu, C. Ding, "Consensus group based stable feature selection," in Proc. of the 5th ACM SIGKDD International Conference on knowledge discovery and Data Mining, 2012, pp567-576

[9] Vipin Kumar and SonajhariaMinz"Feature Selection: A Literature Review. Smart" Computing Review, 2014, Volume 1, No. 3

[10] Veronica Bolon-Canedo, Noelia Sanchez-Marono, Amparo Alonso-Betanzos "A review of feature selection methods on synthetic data," Knowledge and Information Systems,2013, Vol.34, No.3, pp483-519

[11] Shuxin Zhu, Bin Hu, "Hybrid Feature Selection Based on Improved Genetic Algorithm", TELKOMICA, Vol,11, No4, April 2013, pp1725-

[12] Salam SalamehShreem, SalwaniMohdZakree, Ahmad Nazri and MalekAlzaqebah, "Hybridizing Relief, MRMR Filters and GA Wrapper Approaches for GeneSelection', Journal of Theoretical and Applied Information Technology", 2013, Vol. 47 No 3.

[13] RituGanda and Vijay Chahar. "A Comparative Study on Feature Selection Using Data Mining Tools. International Journal of Advanced Research in Computer Science and Software Engineering", Volume 3, Issue 9. 2013.

[14] Kohavi R, John G. "Wrappers for feature selection. *Artif Intell*,1997,1–2:273–324.

[15] Blum A, Langley P. "Selection of relevant features and examples in machine learning". *Artif Intell*,1997,1–2:245–271.

[16] Bhavani, S.D., Rani, T.S., and Bapi, R.S."Feature selection using correlation fractal dimension: Issues and applications in binary classification problem"s. Applied Soft Computing, 8, 2008, 1, 555-563.

[17] Huang, C. J., Yang, D. X., and Chuang, Y.T., "Application of wrapper approach and composite classifier to the stock trend prediction". Expert Systems with Applications, 2008, 34, 4, 2870-2878.
[18] JuanyingXie, Jinhu Lei, Yong Shi and Xiaohui. (2013): Two Stage Hybrid Selection Algorithms for Diagnosing Erythemato-Squamous Diseases. Health Information Science and Systems, 2013, 1:10.
[19] Setiono R, Liu H. , "Improving Backpropagation learning with feature selection. Applied Intelligence: The International Journal of Artifical Intelligence, Neural Networks,and Complex Problem-Solving Technologies , 1996, Vol. **6**, pp129-139
[20] LunGao, Taifu Li, Lizhong Tao, Feng Wen (2014): "Feature Selection Based on Relief Algorithm", Journal of Software, 2014, Vol. 9, No. 2
[21] K. Kira, L. A. Rendell, "Feature selection problem: traditional methods and a new algorithm,"in Artificial Intelligence - AAAI-92, Proceedings Tenth National Conference on, vol.1992, pp. 129-134. 1992
[22] Mhamdi, F: "A New Hybrid Relief Algorithm for Biological Motifs Selection", Bioinformatics, 2013
[23] Mathew E. Stokes, ShyamVisweswaran, "Application of a Spatially-weighted Relief Algorithm for Ranking genetit predictors of diseases", BioData Mining, published online, doi 10.1186/1756-0381-5-20, 2012
[24] Kononenko, I and M, Robnik-Skonja, "Overcomong the Myopia of Inductive Learning Algorithms with RELIEFF". Applied Intelligence 1997,7, 39-55
[25] Mark A. Hall, "Correlation-Based feature selection for Machine Learning", Dissertation, The University of Waikato.1999
[26] Lavanya, D. and K. Usha Rani, "A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks", International Journal of Application or Innovation In Engineering and Management", 2013, Volume 2, Issue1.
[27] Das, S."Filters, wrappers and a Boosting-BasedHybrid for Feature Selection" Proc. 18th International conference on Machine Learning. 2001, Pp74-81,
[28] Sridevi, R. and Chattevelli, R."Genetic Algorithm and Artificial Immune Systems: A combinational approach for network instusion detection", International Conference on Advances in Engineering, Science and Management (ICAESM-2012), pp494-498.
[29] H. Hannah Inbarani, M. Bagyamathi, Ahmad TaheAzar,"A novel hybrid feature selection method based on rough set and improvedharmony search", The Natural Computing Applications Forum, 2015
[30] Asha, T. S. Natarajan and K. N. B, Murthy "A Data Mining Approach to the diagnosis Tuberculosis by Cascading Clustering and Classification" 2011
[31] Rajendran P. and M, Madheswaran, "Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm", Journal of Computing", 2010, Vol.2 Issue 1, ISSN 2151-9617
[32] LI, Yuanhong, "Localised Feature Selection for Unsupervised Learning", Wayne State University. Paper 21.
[33] H. Liu and H. Motoda. "*Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic. 1998