Smart Computing Review

Feature Selection: A literature Review

Vipin Kumar and Sonajharia Minz

School of Computer and Systems Sciences, Jawaharlal Nehru University / New Delhi-110067 / {rt.vipink, sona.minz}@gmail.com

*Corresponding Author: Vipin Kumar

Received March 28, 2013; Revised June, 02; Accepted June 09, 2014; Published June 30, 2014

Abstract: Relevant feature identification has become an essential task to apply data mining algorithms effectively in real-world scenarios. Therefore, many feature selection methods have been proposed to obtain the relevant feature or feature subsets in the literature to achieve their objectives of classification and clustering. This paper introduces the concepts of feature relevance, general procedures, evaluation criteria, and the characteristics of feature selection. A comprehensive overview, categorization, and comparison of existing feature selection methods are also done, and the guidelines are also provided for user to select a feature selection algorithm without knowing the information of each algorithm. We conclude this work with real world applications, challenges, and future research directions of feature selection.

Keywords: feature selection, feature relevance, classification, clustering, real world applications

Introduction

The amount of high-dimensional data that exists and is publically available on the internet has greatly increased in the past few years. Therefore, machine learning methods have difficulty in dealing with the large number of input features, which is posing an interesting challenge for researchers. In order to use machine learning methods effectively, pre-processing of the data is essential. Feature selection is one of the most frequent and important techniques in data pre-processing, and has become an indispensable component of the machine learning process [1]. It is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the process of detecting relevant features and removing irrelevant, redundant, or noisy data. This process speeds up data mining algorithms, improves predictive accuracy, and increases comprehensibility. Irrelevant features are those that provide no useful information, and redundant features provide no more information than the currently selected features. In terms of supervised inductive learning, feature selection gives a set of candidate features using one of the three approaches [2]:

- The specified size of the subset of features that optimizes an evaluation measure
- The smaller size of the subset that satisfies a certain restriction on evaluation measures
- In general, the subset with the best commitment among size and evaluation measure

DOI: 10.6029/smartcr.2014.03.007

Therefore, the correct use of feature selection algorithms for selecting features improves inductive learning, either in term of generalization capacity, learning speed, or reducing the complexity of the induced model.

In the process of feature selection, irrelevant and redundant features or noise in the data may be hinder in many situations, because they are not relevant and important with respect to the class concept such as microarray data analysis [3]. When the number of samples is much less than the features, then machine learning gets particularly difficult, because the search space will be sparsely populated. Therefore, the model will not able to differentiate accurately between noise and relevant data [4]. There are two major approaches to feature selection. The first is Individual Evaluation, and the second is Subset Evaluation. Ranking of the features is known as Individual Evaluation [5]. In Individual Evaluation, the weight of an individual feature is assigned according to its degree of relevance. In Subset Evaluation, candidate feature subsets are constructed using search strategy.

The general procedure for feature selection has four key steps as shown in Figure 1.

- Subset Generation
- Evaluation of Subset
- Stopping Criteria
- Result Validation

Subset generation is a heuristic search in which each state specifies a candidate subset for evaluation in the search space. Two basic issues determine the nature of the *subset generation* process. First, *successor generation* decides the search starting point, which influences the search direction. To decide the search starting points at each state, forward, backward, compound, weighting, and random methods may be considered [7]. Second, *search organization* is responsible for the feature selection process with a specific strategy, such as sequential search, exponential search [9, 10] or random search [11]. A newly generated subset must be evaluated by a certain evaluation criteria. Therefore, many evaluation criteria have been proposed in the literature to determine the goodness of the candidate subset of the features. Base on their dependency on mining algorithms, evaluation criteria can be categorized into groups: independent and dependent criteria [8]. Independent criteria exploit the essential characteristics of the training data without involving any mining algorithms to evaluate the goodness of a feature set or feature. And dependent criteria involve predetermined mining algorithms for feature selection to select features based on the performance of the mining algorithm applied to the selected subset of features. Finally, to stop the selection process, stop criteria must be determined. Feature selection process stops at validation procedure. It is not the part of feature selection process, but feature selection method must be validate by carrying out different tests and comparisons with previously established results or comparison with the results of competing methods using artificial datasets, real world datasets, or both.

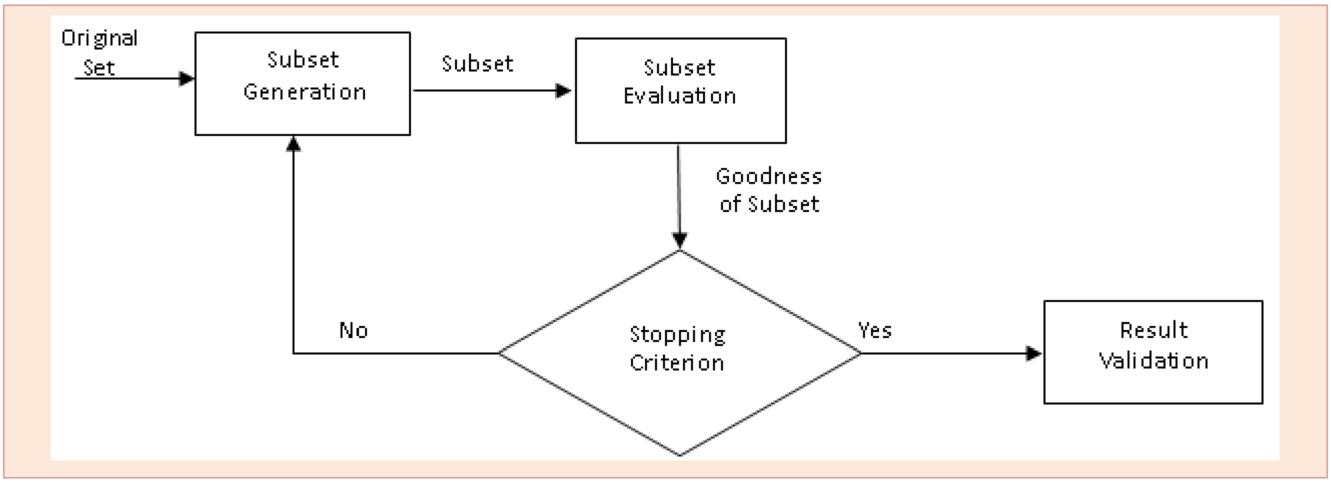

Figure 1. Four key steps for the feature selection process [3]

The relationship between the inductive learning method and feature selection algorithm infers a model. There are three general approaches for feature selection. First, the *Filter Approach* exploits the general characteristics of training data with independent of the mining algorithm [6]. Second, the *Wrapper Approach* explores the relationship between relevance and optimal feature subset selection. It searches for an optimal feature subset adapted to the specific mining algorithm [12]. And third, the *Embedded Approach* is done with a specific learning algorithm that performs feature selection in the process of training.

State of Art

Many, feature selection methods have been proposed in the literature, and their comparative study is a very difficult task. 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. There are many performance measures mentioned in the literature such as accuracy, computer resources, ratio of feature selection, etc. Most researchers agree that there is no so-called "best method" [6]. Therefore, the new feature selection methods are constantly increasing to tackle the specific problem (as mentioned above) with different strategies.

- To ensure a better behavior of feature selection using an ensemble method [84, 85]
- Combining with other techniques such as tree ensemble [86], and feature extraction [87]
- Reinterpreting existing algorithms [88, 89]
- Creating a new method to deal with still-unresolved problems [90, 91]
- To combine several feature selection methods [92, 93]

Many comparative studies of existing feature selection methods have been done in the literature, for example, an experimental study of eight filter methods (using mutual information) is used in 33 datasets [94], and for the text classification problem, 12 feature selection methods are compared [95]. The capability of the survival ReliefF algorithm (sReliefF) and tuned sReliefF approach are evaluated in [96]. Seven filters, two embedded methods, and two wrappers are applied in 11 synthetic datasets (tested by four classifiers), which are used for comparative study of feature selection performances in the presence of irrelevant features, noise in the data, redundancy, and the small ratio between the number of attributes and samples [6]. Related to the high-dimensional dataset (in both samples and attributes), the performance of feature selection methods are studied for the multiple-class problem [90, 97, 98, 99].

In a theoretical perspective, guidelines to select feature selection algorithms are presented, where algorithms are categorized based on three perspectives, namely search organization, evaluation criteria, and data mining tasks. In [2], characterizations of feature selection algorithms are presented with their definitions of feature relevance. In the application perspective, many real-world applications like intrusion detection [100, 101], text categorization [95,102, 109, 110], DNA microarray analysis [103], music information retrieval [104], image retrieval [105], information retrieval [106], customer relationship management [107], Genomic analysis [83, 103] and remote sensing [108] are considered.

Defining Feature Relevance

The optimal feature subset is a subset of all relevant features. Therefore, the relevance of the features must be properly defined according to their relevance. In the literature, features are classified by their relevancy with three qualifiers: irrelevant, weakly relevant, and strongly relevant. A graphical representation is shown in Figure 2 [34]. Many definitions have been proposed to answer a question "relevant to what?" [18]. Therefore, in this section, the definition of the relevance of the feature is presented as suggested in the literature, and the degree of relevance is suggested as well.

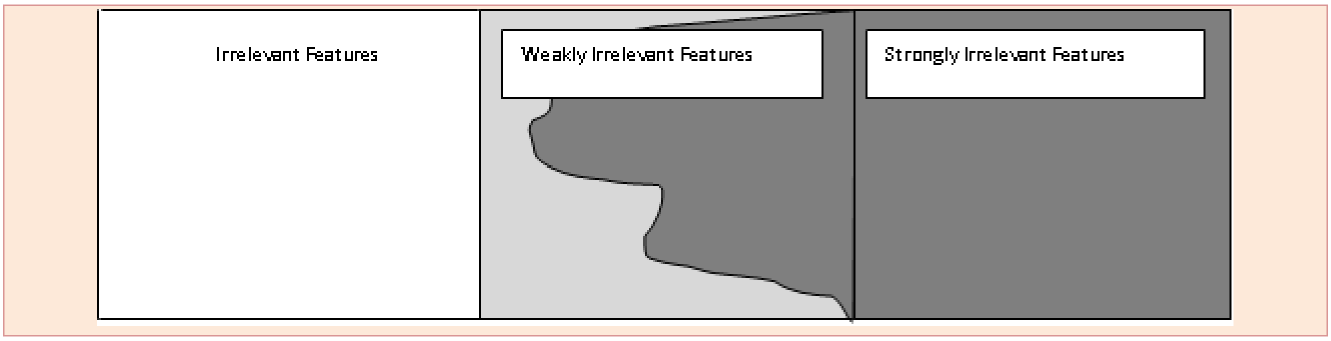

Figure 2. A view of feature relevance [34]

Let a dataset S be composed by $|S|$ instances and be seen as the result of sampling I. The domain of the features $X = \{x_1, x_2, x_3, \ldots, x_n\}$ and instance space is defined as $I = I_1 \times I_2 \times I_3 \times \ldots \times I_m$. P is considered a probability distribution of I. Objective function $C: I \longrightarrow L$ according to its relevance feature, where L is a space of labels.

Definition 1: Relevance to the target [18]

"A feature $x_i \in X$ is relevant to a target concept C if there exists a pair of examples A and B in the instance space such that A and B differ only in their assignment to x_i and $C(A) \neq C(B)$."

Definition 2: Strongly Relevant to the Samples/ Distribution [18]

"A feature $x_i \in X$ is strongly relevant to the sample S if there exists a pair of examples $A, B \in S$ that only differ in their assignment to x_i and $C(A) \neq C(B)$. Or, a feature $x_i \in X$ is strongly relevant to an objective C in distribution P if there exists a pair of examples $A, B \in I$ with $P(A) \neq 0$ and $P(B) \neq 0$ that only differ in their assignment to x_i and $C(A) \neq C(B)$."

Definition 3: Weakly Relevant to the Samples/Distribution [18]

"A feature $x_i \in X$ is weakly relevant to sample S if there exists at least a proper $X^{'} \subset X$ $(x_i \in X^{'})$ where x_i is strongly relevant with respect to S. Or, feature $x_i \in X$ is weakly relevant to objective C in distribution P if there exists at least a proper $X^{'} \subset X$ $(x_i \in X^{'})$ where x_i is strongly relevant with respect to P."

The above definitions focus on which features are relevant. Put another way, we simply want to use relevance as a measure of complexity to show how "complicated" a function is.

Definition 4: Relevance as a Complexity Measure [18]

"Given a sample of data S and a set of concept C, let $r\ (S, C)$ be the number of features relevant using Definition 1 to a concept in C that, out of all those whose error over S is least, has the fewest relevant features."

In another way, we imply optimal performance over S with concept C using the smallest number of features. The above concepts of relevance are independent of the specific learning algorithm. This means that it is not necessary that a given relevant feature is suitable for learning algorithms. Therefore, Caruana and Fritag [19] define the explicit notion of "*incremental usefulness*."

Definition 5: Incremental Usefulness [19]

"Given a sample of data S, a learning algorithm L, and a subset of features X', feature x_i is incrementally useful to L with respect to X' if the accuracy of the hypothesis that L produces using the feature set $\{x_i\} \cup X'$ is better than the accuracy achieved using just the feature subset X′."

Definition 6: Entropy Relevance [20]

"Denoting mutual information $I(x; y) = H(x) - H(x|y)$ with Shannon Entropy (x), the entropy relevance of x to y is defined as $r(x; y) = I(x; y)/H(y)$." Let C be the objective seen as a feature and X be the original set of features, a subset $X' \subset X$ is sufficient if $I(X^{'}; C) = I(X, C)$. For a sufficient subset, it must satisfy $r(X^{'}; C) = r(X, C)$. Therefore, $r(X', C)$ and $r(C; X')$ are jointly maximized.

Feature Selection

Feature selection is the process of selecting relevant features, or a candidate subset of features. The evaluation criteria are used for getting an optimal feature subset. In high-dimensional data (number of samples <<number of features), finding the optimal feature subset is a difficult task [13]. There are many related problems that are shown as NP-hard [12, 14]. The data with N number of features, there exists 2^N candidate subset of features.

Definition 7: Feature Selection

Let the original set of features A and $L(.)$ be an evaluation criterion to be maximized (optimized) and defined as $L: A' \subseteq A \longrightarrow \mathcal{R}$. The candidate subset of features can be considered under the following considerations [15]:

- Let $|A| = m$ & $|A'| = n$, then, $L(A^{'})$ is maximized, where $m > n$ and $A' \subset A$.
- Set a threshold θ such that $L(A^{'}) > \theta$; to find a subset of the feature with the smallest number $(m > n)$.
- Finding the optimization function $L(A^{'})$ with optimal feature subsets $|A'|$.

There is *continuous* feature selection problem in which each feature $a_k \in A$ is assign weights w_k to preserve the theoretical relevance of the features. Binary weight assignment is considered under the *binary* feature selection problem [16, 17]. The optimal feature subset is considered one of the most optimal subsets; therefore, the above definition does not ensure that the optimal feature subset is unique. The optimal feature subset is defined in terms of induced classifier accuracy as follows.

Definition 8: Optimal feature subset

"Let dataset $\mathcal{D}$ be defined by features $\{A_1, A_2, A_{3,\ldots,}A_k\}$ from a distribution $\mathcal{P}$ over the labeled instance space and inducer $\mathcal{L}$. An optimal feature subset, A_{opt}, is a subset of the features such that the accuracy of the induced classifier $\mathcal{C} = \mathcal{L}\,(\mathcal{D})$ is maximal" [12].

There are four basic steps for feature selection, namely subset generation, subset evaluation, stopping criterion, and result validation (shown in Figure 1). Subset generation is a search procedure using the certain search strategy [21]. According to the certain evaluation criterion, a generated subset feature is evaluated with the previous best feature subset. If the new feature subset is better than the previous best feature subset, then the previous best feature subset is replaced by a new feature subset. This process is repeated until some certain stopping criterion is satisfied. After the stopping criterion, the produced optimal feature subset needs to be validated. Validation may be done using synthetic data or real-world data set [3]. A general algorithm for feature selection is shown on Table 1.

Table 1. General algorithm for feature selection

```
INOUTS:
    $X$  : Set of features of a data set having $n$ features
    $SG$ : Successor Generator Operator
    $E$ : Evaluation measure (dependent or independent
    $\theta$ : Stopping Criteria
OUTPUT:
    $X_{opt}$ : Optimal feature set or weighted features

Initialize:
    $X' \coloneqq$ Start_ point($X$);
    $X_{opt} \coloneqq$ {Best of $X'$ using $E$ };
Repeat:
    $X' \coloneqq$ Search_Strategy $(X', SG(E), X)$;
    $X_{opt} \coloneqq$ {Best of $X'$ according to $E$ };
    If $E(X') \geq E(X_{opt})$ or $(E(X') == E(X_{opt})\&|X'| < |X_{opt}|)$
        Then $X_{opt} = X'$;
Until    Stop criteria is not found;
```

The next section describes each basic step of feature selection.

General Procedure of Feature Selection

■ Subset Generation

Subset generation is the process of the heuristic search. The process of subset generation has two basic issues to determine a feature subset, namely *search organization* and *successor generation.*

Search Organization

As we know, for a data set D with N number of features, there exists 2^N number of candidate subsets. Even with moderate N, the search space is exponentially increased, and it is prohibited for exhaustive search. Therefore, many strategies have been proposed in the literature, namely sequential search, exponential search, and random search.

- *Sequential Search*

In sequential search, search selects only one among all successors. It is done in an iterative manner and the number of possible steps is $O(N)$. Sequential search gives completeness, but not an optimal feature subset. In the variation of the greedy hill-climbing approach, many methods have been proposed, namely *sequential forward selection*, *sequential backward elimination*, and *bi-directional selection* [21]. Instead of adding or removing one feature at a time, another is to add or remove K features in one step and remove or add L features in the next step, where $\mathrm{K} > L$ [7]. Sequential search is easy to implement, and the order of the search space is $\mathrm{O}(\mathrm{N}^2)$.

- *Exponential Search*

Exhaustive search is an optimized search that guarantees the best solution. However, optimal searches need not be exhaustive. Different heuristic functions can be used to reduce the search space without tempering the optimal solution.

BRANCH AND BOUND [22] and Beam Search [7] are evaluated for smaller numbers of subsets for an optimal subset. The order of the search space is $O(N^2)$.

- Random Search

Random search starts with randomly selected subset. There are two ways to proceed to get an optimal subset. One, generation of the next subset is completely random manner known as the *Las Vegas algorithm* [23]. The other is sequential search, which includes randomness in the above sequential approach. The concept of randomness is to avoid local optima in the search space. The order of the search space is $O(N^2)$.

Successor Generation

There are five operators considered for successor generations, namely Forward, Backward, Weighing, Random, and Weighting. Let a data set D with X feature sets and $X' \subset X$, where $= \{x_1, x_2, x_3, \ldots, x_n\}$, a feature subset of input dimensions. $E(\mathrm{X})$ denotes the error incurred on the validation, when only the inputs in X are used.

- Forward

This method starts with no feature $\mathrm{X}' = \phi$ and feature $\mathrm{x_i} \in (\mathrm{X} - \mathrm{X}')$ is added to X', which is not yet selected. In each step, $E(X' \cup x_i)$ is evaluated and chooses x_i that causes the least error $F = \arg\min E(X' \cup x_i)$ and if $F < E(X')$ then $X' = X' \cup x_i$. One of the stopping criteria may be used in the following:

- $|X'| = n'$; where n' is pre-decided
- $(F - E(X')) \simeq 0$
- The value of error Fhas exceeded the prefixed error value F_0

This method is not considered the interaction between features. The computational cost of this method is $O(N)$.

- Backward

This method starts with no feature $X' = X$ and feature $x_i \in X'$ is removed to X', which is not yet selected. In each step, $E(X' - x_i)$ is evaluated and chooses x_i to remove that causes the least error $F = \arg\min E(X' - x_i)$ and if $F < E(X')$ then $X' = X' - x_i$. One of the stopping criteria may be used, which is mentioned above [24].

- Compound

The main idea of this method is to apply a K number of consecutive forward steps and L number of consecutive backward steps. Based on $E(X')$, forward or backward steps are selected to discover new interactions among features. The stopping criterion should be $K = L$ or $x_i = x_j$. In the sequential feature selection algorithm it is assured the maximum n number of steps with cost $O(n^{K+L+1})$, where $K \neq L$.

- Random

This method comprises all operators. Those are able to generate a random state in a single step. Other operators are restricted with some criterion such as the number of features or minimizing the error $E(X')$ at each step [2].

- Weighting

This method presents all the features in the solution to the certain degree; where, the search space is continuous. The successor state is a state with different weighting by iteratively sampling the available set of instance [2].

■ Evaluation of Subset

The goods of the newly generated feature subset must be evaluated using certain evaluation criteria. An optimal feature subset generated by one criterion may not be same according to the other evaluation criteria. There are two broadly used evaluation criteria, based on their dependency and independence on the algorithms, which are mentioned below.

Independent Criteria

Basically, a filter model is used for independent criteria feature subset selection. It does not involve any learning algorithm. It exploits the essential characteristics of the training data to evaluate the goodness of the feature subset. There are many independent criteria proposed in the literature, namely *Distance measures* [25], *Information or uncertainty measures* [26], *Probability of error measures* [28], *Dependency measures* [16, 27], *Interclass distance measures*, and *Consistency measures* [11].

- Distance or Divergence Measures

This criterion is known as divergence or discrimination and separability, which computes divergence or probabilistic distance among the class-conditional probability densities. In the two-class problem, a feature $\mathrm{x_i}$ is preferred than $\mathrm{x_j}$ if $\mathrm{x_i}$ makes a greater difference between class-condition probabilities than $\mathrm{x_j}$ [8].

Suppose a space of all probability distributions is S and a divergence on S is a function $D: S \times S \longrightarrow \mathcal{R}$. Then the following condition must be satisfied:

- D(p, q) ≥ 0 for all p, q ∈ S
- D(p, q) = 0 if and only if p = q
- The matrix g (D) is strictly positive-definite everywhere on S

where p and q are the continuous probability distributions. The solution features subset $X' \subset X$ is called a good feature subset, if their divergence among the conditional probabilities is significant. In another way, we can say that the probabilities of a weak relevant feature are very similar. Some measures are shown below [29].

Kullback-Liebler [30]: $D_{KL}(p,q) = \int p(x) \log_2 \left(\frac{p(x)}{q(x)}\right) dx$

Bhattacharya [31]: $D_B(p,q) = \int \sqrt{p(x) \times q(x)}\, dx$

Jeffrey'sdivergence: $D_J(p,q) = \int (p(x) - q(x)) \left(\log_2 \left(\frac{p(x)}{q(x)}\right)\right) dx$

Matusita: $D_M(p,q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$

Kagan's divergence: $D_K(p,q) = \frac{1}{2} \int \frac{(p(x)-q(x))^2}{p(x)} dx$

- *Information or Uncertainty Measures*

This measure is based on the information gain of the features. Information gain of a feature is defined as the difference between the prior uncertainty and expected posterior uncertainty. Information gain is maximal for equal probable classes, and uncertainty is minimal. Shannon entropy is widely used for uncertainty measures and is defined as [32].

$$H(A) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where, A is a discrete probability space.

- *Probability of error measures*

This method is concerned about minimizing the probability of errors. Let us consider two categories of cases such as class w_1 and w_2 that divide into two regions $\mathcal{R}_1$ and $\mathcal{R}_2$ in a possibly non-optimal way. Classification errors may occur in two possible regions $\mathcal{R}_1$ (true state of nature is w_1) and $\mathcal{R}_2$ (true state of nature is w_2). These events are mutually exclusive. The probability of the error is [38].

$$P(error) = \int_{\mathcal{R}_2} p(x|w_1)p(w_1)\, dx + \int_{\mathcal{R}_1} p(x|w_2)p(w_2)\, dx$$

In general, if $p(x|w_1)\, p(w_1) > p(x|w_2)\, p(w_2)$, then it is best to classify x as in $\mathcal{R}_1$, so that the other region will participate in the error integration. This is exactly the concept to achieve the Bayesian decision rule. In a multi-class situation, there are more ways to be wrong than to be right, and it can be computed similarly.

- *Dependency Measures*

This measure is also known as similarity measure and correlation measure. Those features, which are strongly associated with the classes, are preferred for the classification problem. In the context of feature selection for the classification problem, strongly associated features are preferred. The correlation coefficient is a classical evaluation measure [33]. The different approach is to estimate the divergence between unconditional density and class conditional. This kind of purpose may be served by any un-weighted probabilistic distance.

- *Consistency Measure*

This evaluation measure finds the minimum number of features that separate class as consistently as the full set of features. It is greatly dependent on class information [25]. An inconsistency is defined as two instances that have the same feature values but belong to a different class.

- *Interclass Distance Measures*

The concept behind this measure is that different instances of the class are distant in the instance space. Therefore, the metric between classes is sufficient to measure.

$$D(l_i, l_j) = \frac{1}{n_i n_j} \sum_{k_1}^{n_i} \left(\sum_{k_2=k_1+1}^{n_j} d\left(I_{(i,k_1)}, I_{(j,k_2)}\right) \right)$$

and

$$E = \sum_{i=1}^{m} P(l_i) \sum_{j=i+1}^{m} P(l_i)\, D(l_i, l_j)$$

where n_i is the number of instances of the class l_i, and $I_{(i,j)}$ is an instance j of class l_i. The Euclidian distance is usually used for measurements.

Dependent Criteria

Dependent criteria require a predetermined mining algorithm. The performance of the algorithm is used to evaluate the goodness of the feature subset to determine which features are selected. The selected feature subset is best suited to a fixed algorithm. Therefore, the performance of the algorithm is usually better. But it is computationally expensive, because every feature subset estimates accuracy [25, 34]. A wrapper model is used for dependent criteria.

Stopping Criteria

The stopping criterion for the feature selection process must be defined. There are some general stopping criteria:

- Predefined maximum number of iterations or minimum number of features or minimum classification error rate
- The search completes
- Deletion or addition of features to the subset do not produce a significant difference

General Approach for Feature Selection

There are three general approaches for feature selection.

Filter Approach

The filter approach incorporates an independent measure for evaluating features subsets without involving a learning algorithm. This approach is efficient and fast to compute (computationally efficient). However, filter methods can miss features that are not useful by themselves but can be very useful when combined with others. The graphical representation of the filter model is shown in Figure 3.

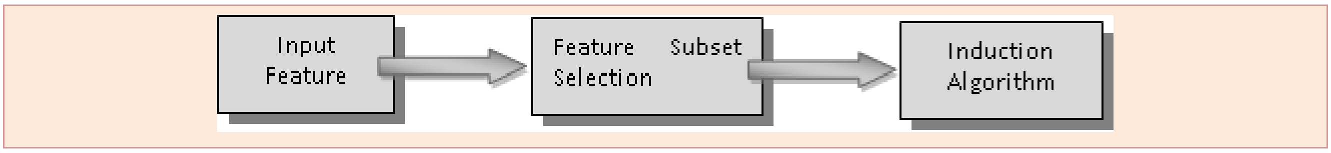

Figure 3. The feature filter model [34]

A general algorithm for a given data set $D = \{X, L\}$ (where X and L are the feature set and labels respectively) is shown in Table 2. The algorithm may start with one of the following subsets of X' such as $X' = \{\phi\}$ or $X' = \{NULL\}$ or $X' \subset X$. Independent measure I_m evaluates the each generated subset X_g and compares it to the previous optimal subset. The search iterates until the stopping criterion θ is not met. Finally, the algorithm outputs the current optimal feature subset X_{opt}. Different filter algorithms may be designed by varying the subset generator and evaluation measure. Many existing algorithms fall under filter approach such as FOCUS [35], ABB [36], relief [37], etc.

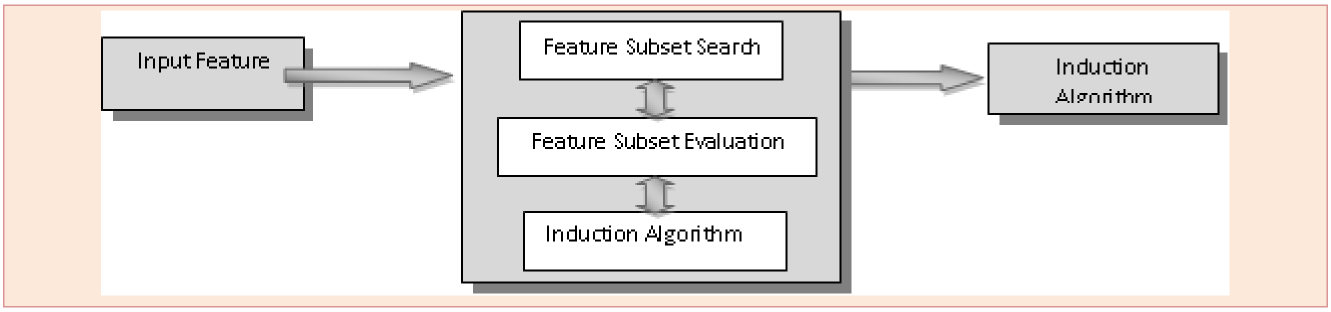

Figure 4. The wrapper model [34]

Table 2. A general filter algorithm

```
INPUT:
$D = \{X, L\}$              // a training data set with $n$ number of features where
                            //  $X = \{f_1, f_2, f_3, \dots, f_n\}$ and $L$ labels
$X'$                        // predefined initial feature subset ($X' \subset X$ or $X' = \{\phi\}$)
$\theta$                    // a stopping criterion
OUTPUT: $X'_{opt}$          // an optimal subset

Begin:
Initialize:
    $X_{opt} = X'$;
    $\varphi_{opt} = E(X', I_m)$;     // evaluate $X'$ by using an independent measure $I_m$
do begin
    $X_g = generate(X)$;   // Subset generation for evaluation
    $\varphi = E(X_g, I_m)$ ;     // $X_g$current subset evaluation by $I_m$
    If ($\varphi > \varphi_{opt}$)
        $\varphi_{opt} = \varphi$;
        $X'_{opt} = X_g$;
repeat (until $\theta$is not reached );
end
return $X'_{opt}$;
end;
```

Table 3. A general wrapper algorithm

```
INPUT:
$D = \{X, L\}$              // a training data set with $n$ number of features where
                            //  $X = \{f_1, f_2, f_3, \dots, f_n\}$ and $L$ labels
$X'$                        // predefined initial feature subset ($X' \subset XorX' = \{\phi\}$)
$\theta$                    // a stopping criterion
OUTPUT:  $X'_{opt}$         // an optimal subset
Begin:
Initialize:
    $X_{opt} = X'$;
    $\varphi_{opt} = E(X', A)$;     // evaluate $X'$ by using mining algorithm $A$
do begin
    $X_g = generate(X)$; // Subset generation for evaluation
    $\varphi = E(X_g, A)$ ;     // $X_g$ current subset evaluation by $A$
    If ($\varphi > \varphi_{opt}$)
        $\varphi_{opt} = \varphi$;
        $X'_{opt} = X_g$;
repeat (until $\theta$is not reached );
end
return $X'_{opt}$;
end;
```

Wrapper Approach

The filter and wrapper approach can only be distinguished by the evaluation criteria. The wrapper approach uses a learning algorithm for subset evaluation. A graphical representation of the wrapper model is shown in Figure 4.

A different wrapper algorithm can be generated by varying the subset generation ($\boldsymbol{X_g}$) and subset evaluation measure $\boldsymbol{A}$ (using dependent criterion). The wrapper approach selects an optimal subset that is best suited to learning algorithm. Therefore, the performance of the wrapper approach is usually better. The wrapper algorithm is shown in Table 3.

Embedded Approach

This approach interacts with learning algorithm at a lower computational cost than the wrapper approach. It also captures feature dependencies. It considers not only relations between one input features and the output feature, but also searches locally for features that allow better local discrimination. It uses the independent criteria to decide the optimal subsets for a known cardinality. And then, the learning algorithm is used to select the final optimal subset among the optimal subsets across different cardinality. A general embedded algorithm is shown in Table 4.

Usually, it starts with an empty set X' using sequential forward selection. For the optimal subset of cardinality k, it searches all possible subsets of cardinality $k+1$ by adding a feature from the remaining subsets. A subset generated at cardinality $k+1$ is evaluated by independent criterion I_m and compared with the previous optimal subset. Then learning algorithm A is applied to the current optimal subset, and performance δ is compared with the performance of the optimal subset at cardinality k. After stopping criterion, it returns a final optimal subset.

Table 4. A general embedded algorithm

```
INPUT:
$D = \{X, L\}$              // a training data set with $n$ number of features where
                            //  $X = \{f_1, f_2, f_3, \dots, f_n\}$ and $L$ labels
$X'$                        // predefined initial feature subset $(X' \subset X or X' = \{\phi\})$
$\theta$                    // a stopping criterion
OUTPUT: $X'_{opt}$          // an optimal subset
Begin:
Initialize:
    $X_{opt} = X'$;
    $\varphi_{opt} = E(X', I_m)$;     // evaluate $X'$ by using independent evaluation measure
    $\delta_{opt} = E(X', A)$;        // evaluate $X'$ by using mining algorithm $A$
    $C_0 = C(X')$;                    // cardinality calculation of $X'$
do begin
    for $k = C_0$+1 to $n$
        for $i = 0$ to $n - k$
            $X_g = X_{opt} \cup \{f_i\}$;   // Subset generation for evaluation with cardinality$k$
            $\varphi = E(X_g, I_m)$ ;       // evaluation the current subset $X_g$ by $I_m$
            If $(\varphi > \varphi_{opt})$
                $\varphi_{opt} = \varphi$;
                $X'_{opt} = X_g$;
            end
     $\delta = E(X'_{opt}, A)$;             // evaluating subset $X'_{opt}$ by $A$ learning algorithm
 If $(\delta > \delta_{opt})$
    $X'_{opt} = X_{opt}$;
    $\delta_{opt} = \delta$;
     else
  break and return $X'_{opt}$
  end
 return $X'_{opt}$;
end
```

Categorization and Characteristics of Feature Selection Algorithms

In the literature, a large number of feature selection algorithms are available. Each algorithm can be different in order to inner mechanism and commonalities. Huan Liu and Lei Yu [8] proposed a three-dimensional categorization framework, shown in Table 5. More algorithms are introduced in Table 5 to strengthen the categorization. Search strategy and evaluation are two dominating factors in the feature selection algorithm. Therefore, mentioned factors are used as two dimensions in the framework. In search strategy, sub-categorization is done - namely complete, sequential, or random corresponding to the data mining task (classification and clustering). Algorithms are categorized as filter, wrapper, and embedded under evaluation criteria. Further categorization of a filter is done in distance, information, dependency and consistency. Wrapper and embedded algorithms are also categorized into predictive accuracy and filter+wrapper, respectively.

There are empty boxes in Table 5. These empty boxes indicate that according to the literature survey no algorithms of that category exist. Therefore, the table also gives an insight for future feature selection algorithms in the respective categories.

Table 5. Categorization of feature selection algorithms based on search strategy, evaluation criteria, and data mining tasks [8]

			Search Strategies			
			Exponential	***Sequential***		***Random***
Evaluation Criteria	***Filter***	Distance	B&B [39], BFF [40], BOBRO [41], OBLIVIIN [42]	Relief [43], Relief [44], RelifS [45], SFS [46], Segen’s [47], SBS [48]		
		Information	MDLM [49], CARDIE [50],	DTM [51], Koller’ [52], FG [11], FCBF [53], BSE [17]	Dash’s [54], SBUD [55],	
		Dependency	Bobrowski’s [49]	CFS [27], RRESET, [56], POE+ACC [57], DVMM [58]	Mitra’s [59],	
		Consistency	FOCUS [35], ABB [36], MIFESI [60], Schlimmer’s [61],	Set Cover [62], WINNOW [63]		LIV [11], QBB [11], LVF [76],
	Wrapper	Predictive Accuracy or Cluster Goodness	BS [7], AMB&B [64], FSLC [65], FSBC [66], CARDIE [50], OBLIVIIN [63],	SBS-SLASH [17], WSFG [28], WSBG [28], BDS [7], PQSS [7], RC [68], SS [69], Queiros’ [17], BSE [17], K2-AS [71], RACE [72], SBS-W [28], SBS-SLASH [17]	AICC [73], FSSEM [74], ELSA [75],	SA [7], RGSS [7] LVW [77], RMHC-PF [78], GA [79], RVE [80],
	Embedded	Filter + Wrapper		BBHFS [82], Xing’s [91]	Dash-Liu’s [81]	
			Classification	***Classification***	***Clustering***	***Classification***
			Data Mining Tasks			

A space of characteristics of feature selection algorithms according to their criteria, namely search organization, generation of successors, and evaluation measure is presented in Figure5 [2].

Application of Feature Selection in Real World

During data collection, many problems are often encountered such as a high dependency of features, too many features, or redundant and irrelevant features. To deal with the mentioned problem, feature selection provides a tool to select a feature subset or feature to learn algorithms effectively. Therefore, in the literature, the applications of feature selection are used frequently in many research areas.

■ Text Categorization

The massive volume of online text data on the Internet such as emails, social sites, and libraries is increasing. Therefore, automatic text categorization and clustering are important tasks. A major problem with text classification or clustering is the high dimensionality of the document features. A moderate size text document may have hundreds of thousands of features. Therefore, feature selection (dimension reduction) is highly enviable for the efficient use of mining algorithms. In the literature, many applications of feature selection techniques are effectively used in the area of text mining. Feature selections using the information Gain Ratio (GR) is used for lyrics [110] and poems [109, 111] for text data classification. Many feature selection techniques are used for feature reduction, then evaluated and compared to the classification problem [1, 2, 3, 34].

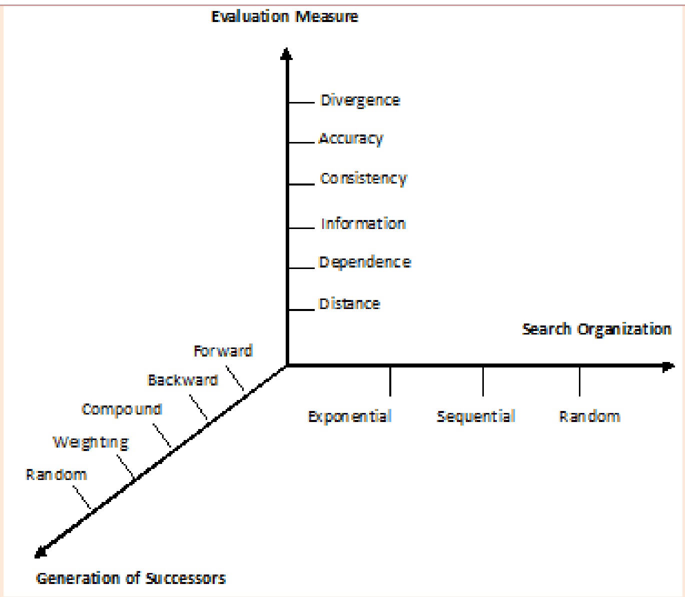

Figure 5. A space of characteristics of feature selection algorithms [2]

Remote Sensing

Feature selection is one of the important tasks in the remote sensing image classification. In paper [116], the challenges and various issues in feature selection and hyper spectral remote sensing image analysis is explained. In [117], pre-processing techniques have been proposed for hyper spectral images in which feature extraction and feature selection have been emphasized as important components in hyper spectral image classification. Feature selection guided by evolutionary algorithms has been proposed, and use a self-adaptive differential evolution for feature subset generation. Generated feature subsets are evaluated by the wrapper method with the help of fuzzy k-nearest neighbor classifier [118]. Shijin Li, Hao Wu, Dingsheng, and Wan Jiali Zhu have developed a hybrid approach for feature selection using support vector machine and genetic algorithm [119]. They have used the wrapper method to select the optimal number of features in order to obtain better accuracy. In [120], a novel technique has been proposed to select a subset of bands from a hyper spectral image to improve the performance of the classification. It utilizes spatial and spectral information simultaneously to improve the discrimination capability of the classifier [9].

Intrusion Detection

In this modern age, information sharing, distribution, or communication is widely done by network-based computer systems. Therefore, the security of the system is an important issue protecting communication networks from intrusion by enemies and criminals. One of the ways to protect communication networks (computer systems) is intrusion detection. Feature selection plays an important role to classifying system activity as legitimate or an intrusion. In [112], data mining techniques and feature selection techniques are used for intrusion detection. In this paper they did a comparative study about techniques, their advantages, and disadvantages. In [113], there is a systematic data mining framework that constructs an intrusion detection model for analyzing audit data. In this work, a large data set is used for an analysis of the frequency patterns. These patterns are guided to select system features for automatic learning using additional statistical and temporal features.

Genomic Analysis

A large quantity of genomic and proteomic data is produced by microarray and mass spectrometry technology for understanding of function of an organism, and the behavior, dynamics, and characteristics of diseases. Tens of thousands of genes are measured in a typical microarray assay and mass spectrometry proteomic profile. Special data analysis is demanded because of the high dimensionality of the microarray data. One of the common ways to handle high dimensionality is identification of the most relevant features in the data. Therefore, in the literature, feature selection has been done successfully on full microarray data. In [114] the Filter, Wrapper, and Embedded methods have been used for feature selection and dimensionality reduction. The techniques covered by them are the most effective for proteomics data and genomic analysis. In [115], comparative studies of 8 feature selection for classification task and their combinations have been done based on gene expression data. It is also shown that classification accuracy can be significantly boosted by a small number of genes by using a feature selection method.

Image Retrieval

Recently, the amount of image collections from military and civilian equipment has increased. To access the images or make use of the information, images should be organized in a way that allows effective browsing, retrieving, and searching. As stated in [121], content-based image retrieval is scalable for the large size of images, but it is also cursed by high dimensionality. Therefore, feature selection is an important task for effective browsing, searching, and retrieval. In [122], content-based image retrieval is proposed that annotates images by their own colors, textures, and shape.

Challenges and Future Direction

Forward vs Backward Selection

In the literature, it is argued that backward elimination is less efficient than forward selection. To defend backward selection, it is said that forward selection finds weaker subset of features, because weaker features are not assessed while subset selection. Moreover, the computational complexity forward feature selection method is less than backward feature selection. Pros of the forward greedy feature selection method are that it is computationally efficient and does not over fit. Cons, errors made in the early stage by forward greedy feature selection method are do not correct later stages. Backward greedy feature selection has corrections of errors by looking at all the models, but it starts with non-over-fit or sparse model. Both methods have their own pros and cons for feature selection. Therefore, in [123], a combination of forward greedy and backward greedy feature selection has been presented that does not over-fit, is computationally efficient, is error corrected by backward greedy step later, and that is made in the early stage in order to trade off. For future research, error correction, over-fitting, and computational efficiency can be considered as features of effective algorithms.

Feature Selection with Large Dimensional Data

Recently, the amount of data collections have increased in the form of text documents, images, videos, and medical data that cause the high dimensionality of the data. Dimensionality N in the range of hundreds is called high-dimensional data [8]. Recently, feature selection has been applied to tens or hundreds of thousands of features [95, 109, 110, 111]. Moreover, feature selection is cursed by high dimensionality [121]. Many feature selection algorithms have higher time complexity about dimensionality N, therefore the scalability of feature selection is a difficult problem. A filter approach has less computational complexity than a wrapper approach, because it uses independent subset evaluation criteria for subset evaluation. A filter approach is more scalable than the wrapper, so is preferred to a wrapper approach for feature selection. In literature, the embedded approach [82] has been proposed to utilize the qualities of the filter and wrapper approach high dimension environment. The embedded method has similar time complexity as the filter approach. To handle the high dimensional data, an efficient correlation-based filter algorithm has been proposed [53]. The inference of the above discussion is that future research must be concentrated on low time complexity with high scalability feature selection algorithms. There is a great research opportunity to develop algorithms using sequential and random search strategies for clustering and classification tasks respectively, as mentioned in Table4.

Subspace Searching and Instance Selection

In clustering, many clusters may exist in different subspaces for small dimensionality with over lappedor non-overlapped dimensions [125]. Subspace searching is not only the feature selection problem. It is finding many subspaces in which feature selection finds one subspace. In literature, many algorithms (subspace clustering) have been developed [126].

Therefore, there is a requirement for efficient subspace search algorithms for clustering. In instance selection, sampling methods have been developed to search for a set of instances that can perform in a focused way [127, 128].

■ Feature Selection with Sparse Data Matrix

A relatively high percentage of variables that do not have actual data are called sparse data. There are two types of sparsity, namely Controlled Sparsity and Random Sparsity. Controlled Sparsity is a range of values of one or more than one dimension that have no data. Random Sparsity, in contrast, is empty values scattered throughout the data variable [124]. In a business context, many individual transactions are recorded in the application such as market basket analysis, direct-mail marketing, insurance, and health care [8]. These types of data collections have a sparse matrix with a large number of attributes. Some other sparse data are commonly available through computer and internet web technology such as HTML, XML, emails, news, and customer reviews. Video stream data is also increasing rapidly with high dimensionality via surveillance cameras, sensors, and web streaming. Feature selection from labelled or unlabelled sparse data is a difficult task, because many feature selection techniques are not suitable for high dimensional sparse data. It is not advised to modify feature selection algorithms for sparse data [8]. Therefore, it is a requirement of future research to develop efficient feature selection algorithms for sparse data.

■ Scalability and Stability of Feature Selection

The scalability of feature selection algorithms is an important issue for online classifiers, because of the rapid growth of the dataset sizes. A large dataset cannot be loaded in the memory for the single data scan. Full dimensionality of the data must be scanned for feature selection. It is very tough to get a feature relevance score without considering sufficient density around each sample. Therefore, the scalability of feature selection algorithms is a big challenge. To solve this problem, some methods have tried to overcome by memorizing only important samples or summaries [129]. More attention is required on the scalability of feature selection algorithms.

The results of classification cannot be trusted if a different set of features are drawn for the same problem in each iteration. That means feature selection algorithms should be very stable (less sensitive). Well-known feature selection algorithms have less stability. Therefore, it is required for developed algorithms with stability and high classification accuracy.

Conclusion

We comprise many definitions of feature relevance, feature selection, and optimal feature subsets. The general procedure of feature selection is described with subset generation, evaluation of subsets, and stopping criteria. Three general approaches of feature selection methods, namely filter, wrapper and embedded methods, are described in detail and their pseudo code is also presented. The categorization and characteristics of feature selection are reviewed, and the interesting facts regarding the advantages and disadvantages of feature selection methods to handle the different characteristics of the real world applications are enumerated. The three dimensional categorization of feature selection algorithms give an insight of future challenges and research directions.

References

[1] A. Kalousis, J. Prados, M. Hilario, “Stability of Feature Selection Algorithms: a study on high dimensional spaces,” *Knowledge and information System*, vol. 12, no. 1, pp. 95-116, 2007. Article (CrossRef Link)

[2] Luis Carlos Molina, Lluís Belanche, Àngela Nebot, “Feature Selection Algorithms: A Survey and Experimental Evaluation,” in *Proc.of ICDM*, pp. 306-313, 2002.

[3] M. Dash, H. Liu, “Feature Selection for Classification,” *Intelligent Data Analysis*, Elsevier, pp. 131-156, 1997.

[4] F. Provost, “Distributed data mining: scaling up and beyond,” *Advances in distributed data mining*, Morgan Kaufmann, San Francisco, 2000.

[5] I. Guyon, A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[6] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, “A review of feature selection methods on synthetic data,” *Knowledge and Information Systems*, vol. 34, no. 3, pp 483-519, Mar. 2013. Article (CrossRef Link)

[7] J. Doak, “An evaluation of feature selection methods and their application to computer security,” *Technical report*,

Davis CA: University of California, Department of Computer Science, 1992.

[8] Huan Liu, Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Transaction on Knowledge and Data Engineering*, 2005.

[9] P. Narendra, K. Fukunaga, "A Branch and Bound Algorithm for Feature Subset Selection," *IEEE Transactions on Computer*, vol. 26, no. 9, pp. 917-922, 1977. Article (CrossRef Link)

[10] J. Pearl, "Heuristics," *Addison-Wesley*, 1983.

[11] H. Liu, H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining," *Kluwer Academic Publishers*, London, GB, 1998.

[12] Ron Kohavi, George H. John, "Wrapper for Feature Subset Selection," *Artificial Intelligence*, Elsevier, pp. 273-324, 1997.

[13] R. Kohavi, G.H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997. Article (CrossRef Link)

[14] A.L. Blum, R.L. Rivest, "Training a 3-node neural networks is NP-complete," *Neural Networks*, vol. 5, pp. 117-127, 1992. Article (CrossRef Link)

[15] M. Kudo, J. Sklansky, "A Comparative Evaluation of medium and large–scale Feature Selectors for Pattern Classifiers," in *Proc. of the 1st International Workshop on Statistical Techniques in Pattern Recognition*, pp. 91-96, 1997.

[16] M. A. Hall, "Correlation–based Feature Selection for Machine Learning," *PhD thesis*, University of Waikato, 1999

[17] R. A. Caruana, D. Freitag, "Greedy Attribute Selection," in *Proc. of the 11th International Conference on Machine Learning*, New Brunswick, NJ, Morgan Kaufmann, pp. 28-36, 1994.

[18] A. L. Blum, P. Langley, "Selection of Relevant Features and Examples in Machine Learning," *Artificial Intelligence on Relevance*, vol. 97, pp. 245-271, 1997.

[19] R. A. Caruana, D. Freitag, "Greedy Attribute Selection," in *Proc. of the 11th Int. Conf. on Machine Learning*, New Brunswick, NJ, Morgan Kaufmann, pp. 28-36, 1994.

[20] K. Wang, D. Bell, F. Murtagh, "Relevance Approach to Feature Subset Selection," *Kluwer Academic Publishers*, pp. 85-97,1998. Article (CrossRef Link)

[21] H. Liu, H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining," *Boston: Kluwer Academic Publishers*, 1998. Article (CrossRef Link)

[22] P. M. Narendra, K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transaction on Computer*, vol. 26, no. 9, pp. 917-922, 1977. Article (CrossRef Link)

[23] G. Brassard, P. Bratley, "Fundamentals of Algorithms," *Prentice Hall*, New Jersey, 1996.

[24] D. Koller, M. Sahami, "Toward Optimal Feature Selection," in *Proc. of the 13th International Conference on Machine Learning*, Bari, IT, Morgan Kaufmann, pp. 284-292, 1996.

[25] H. Almuallim, T.G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features," *Artificial Intelligence*, 69(1-2):279–305, 1994. Article (CrossRef Link)

[26] M. Ben-Bassat, "Pattern recognition and reduction of dimensionality," *Handbook of statistics-II*, North Holland, pp. 773-791, 1982.

[27] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. of the 17th International Conference on Machine Learning*, pp. 359-366, 2000.

[28] P. A. Devijver, J. Kittler, "Pattern Recognition – A Statistical Approach," *Prentice Hall*, London, GB, 1982.

[29] http://en.wikipedia.org/wiki/Divergence_%28statistics%29

[30] S. Kullback, R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79-86, 1951. Article (CrossRef Link)

[31] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, no. 99-109, 1943.

[32] C. E. Shannon, W. Weaver, "The Mathematical Theory of Communication," *University of Illinois Press*, ISBN 0-252-72548-4, 1949.

[33] K. I. Asad, T. Ahmed, Md. S. Rahman, "Movie Popularity Classification based on Inherent Movie Attributes using C4.5, PART and Correlation Coefficient," in *Proc. of IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision* , 2012.

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

[35] H. Almuallim, T. G. Dietterich, "Learning with many irrelevant features," in *Proc. of the Ninth National Conference on Artificial Intelligence*, pp. 547-552, 1991.

[36] H. Liu, H. Motoda, M. Dash, "A monotonic measure for optimal feature selection," in *Proc. of the Tenth European Conference on Machine Learning*, pp. 101-106, 1998.

[37] K. Kira, L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. of the Tenth National Conference on Artificial Intelligence*, pp. 129-134, 1992.

[38] R. O. Duda, P. E. Hart, D. G. Stark, "Pattern Classification," *Willey-India*, Second Edition, 2008.

[39] P. M. Narendra, K. Fukunaga, “A branch and bound algorithm for feature subset selection,” *IEEE Transaction on Computer*, vol. 26, no. 9, pp. 917-922, 1977. Article (CrossRef Link)

[40] L. Xu, P. Yan, T. Chang, “Best first strategy for feature selection,” in *Proc. of the Ninth International Conference on Pattern Recognition*, pp. 706-708, 1988.

[41] L. Bobrowski, “Feature Selection Based on Some Homogeneity Coefficient,” in *Proc. of 9th International Conference on Pattern Recognition*, pp. 544-546, 1988. Article (CrossRef Link)

[42] P. Langley, S. Sage, “Oblivious Decision Trees and Abstract Cases,” in *Working Notes of the AAAI94 Workshop on Case Based Reasoning*, pp. 113–117, 1994.

[43] K. Kira, L. A. Rendell, “The feature selection problem: Traditional methods and a new algorithm,” in *Proc. of the Tenth National Conference on Artificial Intelligence*, pp. 129-134, 1992.

[44] I. Kononenko, “Estimating attributes: Analysis and extension of RELIEF,” in *Proc. of the Sixth European Conference on Machine Learning*, pp. 171-182, 1994.

[45] H. Liu, H. Motoda, L. Yu, “Feature selection with selective sampling,” in *Proc. of the Nineteenth International Conference on Machine Learning*, pp. 395-402, 2002.

[46] P. Pudil, J. Novovicova, “Novel Methods for Subset Selection with Respect to Problem Knowledge,” Liu and Motoda, 2nd Printing, 2001.

[47] J. Segen, “Feature selection and constructive inference,” in *Proc. of the Seventh International Conference on Pattern Recognition*, pp. 1344-1346, 1984.

[48] T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning,” *Springer*, 2001. Article (CrossRef Link)

[49] J. Sheinvald, B. Dom, W. Niblack, “A modelling approach to feature selection,” in *Proc. of the Tenth International Conference on Pattern Recognition*, pp. 535-539, 1990. Article (CrossRef Link)

[50] C. Cardie, “Using Decision Trees to Improve Case–Based Learning,” in *Proc. of the 10th International Conference on Machine Learning*, Amherst, MA, Morgan Kaufmann, pp. 25-32,1993.

[51] C. Cardie, “Using decision trees to improve case-based learning,” in *Proc. of the Tenth International Conference on Machine Learning*, pp. 25-32, 1993.

[52] [52]. D. Koller, M. Sahami, “Toward optimal feature selection,” in *Proc. of the Thirteenth International Conference on Machine Learning*, pp. 284-292, 1996.

[53] L. Yu, H. Liu, “Feature selection for high-dimensional data: a fast correlation-based filter solution,” in *Proc. of the twentieth International Conference on Machine Learning*, pp. 856-863, 2003.

[54] M. Dash, K. Choi, P. Scheuermann, H. Liu, “Feature selection for clustering – a filter solution,” in *Proc. of the Second International Conference on Data Mining*, pp. 115-122, 2002.

[55] M. Dash, H. Liu, J. Yao, “Dimensionality reduction of unsupervised data,” in *Proc. of the Ninth IEEE International Conference on Tools with AI (ICTAI’97),* pp. 532-539, 1997.

[56] A. Miller, “Subset Selection in Regression,” *Chapman & Hall/CRC*, 2^{nd} edition, 2002. Article (CrossRef Link)

[57] A. N. Mucciardi, E. E. Gose, “A comparison of seven techniques for choosing subsets of pattern recognition,” *IEEE Transactions on Computers*, vol. 20, pp.1023-1031, 1971. Article (CrossRef Link)

[58] N. Slonim, G. Bejerano, S. Fine, N. Tishbym, “Discriminative feature selection via multiclass variable memory markov model,” in *Proc. of the Nineteenth International Conference on Machine Learning*, pp. 578-585, 2002.

[59] P. Mitra, C. A. Murthy, S. K. Pal, “Unsupervised feature selection using feature similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301-312, 2002. Article (CrossRef Link)

[60] A. L. Oliveira, A. S. Vincentelli, “Constructive induction using a non-greedy strategy for feature selection,” in *Proc. of the Ninth International Conference on Machine Learning*, pp. 355-360, 1992.

[61] Y. Rui, T. S. Huang, S. Chang, “Image retrieval: Current techniques, promising directions and open issues,” *Visual Communication and Image Representation*, vol. 10, no. 4, pp. 39-62, 1999. Article (CrossRef Link)

[62] M. Dash, “Feature selection via set cover,” in *Proc. of IEEE Knowledge and Data Engineering Exchange Workshop*, pp. 165-171, 1997.

[63] N. Littlestone, “Learning Quickly when Irrelevant Attributes Abound: A New Linear Threshold Algorithm,” *Machine Learning*, vol. 2, pp. 285-318, 1988. Article (CrossRef Link)

[64] I. Foroutan, J. Sklansky, “Feature selection for automatic classification of non-gaussian data,” *IEEE Transactions on Systems, Man, and Cybernatics*, vol. SMC-17, no. 2, pp.187-198, 1987. Article (CrossRef Link)

[65] M. Ichino, J. Sklansky, “Feature selection for linear classifier,” in *Proc. of the Seventh International Confernce on Pattern Recognition*, pp. 124-127, 1984.

[66] M. Ichino, J. Sklansky, “Optimum feature selection by zero-one programming,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-14, no. 5, pp. 737-746, 1984. Article (CrossRef Link)

[67] P. Langley, S. Sage, “Oblivious Decision Trees and Abstract Cases,” in *Working Notes of the AAAI94 Workshop on Case Based Reasoning*, Seattle, AAAI Press, WA, pp. 113-117,1994.

[68] P. Domingos, “Why does bagging work? a bayesian account and its implications,” in *Proc. of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 155-158, 1997.

[69] A. W. Moore, M. S. Lee, “Efficient algorithms for minimizing cross validation error,” in *Proc. of the Eleventh International Conference on Machine Learning*, pp. 190-198, 1994.

[70] C. E. Queiros, E. S. Gelsema, “On feature selection,” in *Proc. of the Seventh International Conference on Pattern Recognition*, pp. 128-130, 1984.

[71] M. Singh, G. M. Provan, “Efficient Learning of Selective Bayesian Network Classifiers,” in *Proc. of the 13th International Conference on Machine Learning*, Morgan Kaufmann, pp. 453-461, 1996.

[72] A. W. Moore, M. S. Lee, “Efficient Algorithms for Minimizing Cross Validation Error,” in *Proc. of the 11th International Conference on Machine Learning*, Morgan Kaufmann, New Brunswick, NJ, pp. 190-198, 1994.

[73] A. K. Jain, D. Zongker, “Feature Selection: Evaluation, Application, and Small Sample Performance,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153-158, 1997. Article (CrossRef Link)

[74] J. G. Dy, C. E. Brodley, “Feature subset selection and order identification for unsupervised learning,” in *Proc. of the Seventeenth International Conference on Machine Learning*, pp. 247-254, 2000.

[75] Y. Kim, W. Street, F. Menczer, “Feature selection for unsupervised learning via evolutionary search,” in *Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 365-369, 2000. Article (CrossRef Link)

[76] H. Liu, R. Setiono, “A probabilistic approach to feature selection - a filter solution,” in *Proc. of the Thirteenth International Conference on Machine Learning*, pp. 319-327, 1996.

[77] H. Liu, R. Setiono, “Feature selection and classification - a probabilistic wrapper approach,” in *Proc. of the Ninth International Conference on Industrial and Engineering Applications of AI and ES*, pp. 419-424, 1996.

[78] D. B. Skalak, “Prototype and feature selection by sampling and random mutation hill climbing algorithms,” in *Proc. of the Eleventh International Conference on Machine Learning*, pp. 293-301, 1994.

[79] [79]. H. Vafaie, I. F. Imam, “Feature selection methods: genetic algorithms vs. greedy-like search,” in *Proc. of the International Conference on Fuzzy and Intelligent Control Systems*, 1994.

[80] D. J. Stracuzzi, P. E. Utgoff, “Randomized variable elimination,” in *Proc. of the Nineteenth International Conference on Machine Learning*, pp. 594-601, 2002.

[81] M. Dash, H. Liu, “Feature selection for clustering,” in *Proc. of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, (PAKDD-2000)*, pp. 110-121, 2000.

[82] S. Das. Filters, “wrappers and a boosting-based hybrid for feature selection,” in *Proc. of the Eighteenth International Conference on Machine Learning*, pp. 74-81, 2001.

[83] E. Xing, M. Jordan, R. Karp, “Feature selection for high-dimensional genomic microarray data,” in *Proc. of the Eighteenth International Conference on Machine Learning*, pp. 601-608, 2001.

[84] Y. Saeys, T. Abeel, Y. Peer, “Robust feature selection using ensemble feature selection techniques,” in *Proc. of the European conference on machine learning and knowledge discovery in databases— part II*, pp 313-325, 2008.

[85] V. Bolon-Canedo, N. Sanchez-Maroño, A. Alonso-Betanzos, “An ensemble of filters and classifiers for microarray data classification,”*Journals Pattern Recognition*, vol. 45, pp. 531-539, 2012. Article (CrossRef Link)

[86] E. Tuv, A. Borisov, G. Runger, “Feature selectionwith ensembles, artificial variables, and redundancy elimination,” *Journal of Machine Learning Research*, vol. 10, pp. 1341-1366, 2009.

[87] I. Vainer, S. Kraus, G. A. Kaminka, H. Slovin, “Obtaining scalable and accurate classification in large scale spatio-temporal domains,” *Knowledge Information Systems*, 2010.

[88] Y. Sun, J. Li, “Iterative RELIEF for feature weighting,” in *Proc. of the 21st international conference on machine learning*, pp 913-920, 2006.

[89] Y. Sun, S. Todorovic, S. Goodison, “A feature selection algorithm capable of handling extremely large data dimensionality,” in *Proc. of the 8th SIAM international conference on data mining*, pp 530-540, 2008.

[90] B. Chidlovskii, L. Lecerf, “Scalable feature selection for multi-class problems,” *Machine Learning and Knowledge Discovery Databases*, vol. 5211, pp. 227-240, 2008. Article (CrossRef Link)

[91] S. Loscalzo, L. Yu, C. Ding, “Consensus group based stable feature selection,” in *Proc. of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, pp 567-576, 2009. Article (CrossRef Link)

[92] Y. Zhang, C. Ding, T. Li, “Gene selection algorithm by combining relief and mrmr,” *BMC Genomics 9(Suppl 2):*S27, 2008. Article (CrossRef Link)

[93] A. El AkadiA,Amine, A. El Ouardighi, D. Aboutajdine, “A two-stage gene selection scheme utilizing MRMR filter and GA wrapper,” *Knowledge Information Systems*, vol. 26, no. 3, pp. 487-500, 2011. Article (CrossRef Link)

[94] H. Liu, L. Liu, H. Zhang, “Feature selection using mutual information: an experimental study,” in *Proc. of the 10th Pacific rim international conference on artificial intelligence: trends in artificial intelligence*, pp 235-246, 2008.

[95] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” *Journal of Machine Learning Research*, vol. 3, pp. 1289-1305, 2003.

[96] L. Beretta, A. Santaniello, “Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets,” *Journal of Biomedical Informatics*, vol. 44, no. 2, pp. 361-369, 2011. Article (CrossRef Link)

[97] J. Hua, W. Tembe, E. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Journal of Pattern Recognition*, vol. 42, no. 3, pp. 409-424, 2009. Article (CrossRef Link)

[98] G. Bontempi, P. E. Meyer, “Causal filter selection in microarray data,” in *Proc. of the 27th international conference on machine learning*, pp 95-102, 2010.

[99] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, “Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation,” *Journal of Machine Learning Research*, vol. 11, pp. 171-234, 2010.

[100] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, “Feature selection and classification in multiple class datasets: an application to KDD Cup 99 dataset,” *Journal of Expert Systems with Applications*, vol. 38, no. 5, pp. 5947-5957, 2011. Article (CrossRef Link)

[101] W. Lee, S. J. Stolfo, K. W. Mok, “Adaptive intrusion detection: a data mining approach,” *Artificial Intelligent Reviews*, vol. 14, no. 6, pp. 533-567, 2000. Article (CrossRef Link)

[102] J. C. Gomez, E. Boiy, M. F. Moens, “Highly discriminative statistical features for email classification,” *Knowledge and Information Systems*, 2011.

[103] L. Yu, H. Liu, “Redundancy based feature selection for microarray data,” in *Proc. of the 10th ACM SIGKDD conference on knowledge discovery and data mining*, pp 737-742, 2004.

[104] P. Saari, T. Eerola, O. Lartillot, “Generalizability and simplicity as criteria in feature selection: application to mood classification in music,” *IEEE Transaction on Audio Speech Language Processing*, vol. 19, no. 6, pp. 1802-1812, 2011. Article (CrossRef Link)

[105] J. G. Dy, C. E. Brodley, A. C. Kak, L. S. Broderick, A. M. Aisen, “Unsupervised feature selection applied to content-based retrieval of lung images,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 373-378, 2003. Article (CrossRef Link)

[106] O. Egozi, E. Gabrilovich, S. Markovitch, “Concept-based feature generation and selection for information retrieval,” in *Proc. of the twenty-third AAAI conference on artificial intelligence*, pp 1132-1137, 2008.

[107] K. S. Ng, H. Liu, “Customer retention via data mining,” *AI Review*, vol. 14, no. 6, pp. 569-590, 2000.

[108] C. De Stefano, F. Fontanella, C. Marrocco, “A GA-Based Feature Selection Algorithm for Remote Sensing Images,” *Applications of Evolutionary Computing* , Lecture Notes in Computer Science, vol. 4974, pp 285-294, 2008.

[109] V. Kumar, S. Minz, “Poem Classification using Machine Learning Approach,” *International Conference on Soft Computing for problem Solving (SocPro 2012)*, Dec. 2012.

[110] V. Kumar, S. Minz, “Mood Classification of Lyrics using SentiWordNet”, in *Proc. of ICCCI-2013*, Jan. 2013.

[111] V. Kumar, S. Minz, “Multi-view Ensemble Learning for Poem Data Classification using SentiWordNet,” in *Proc. of 2nd International Conference on Advanced Computing, Networking, and Informatics (ICACNI-2014)*, 2014.

[112] T. Lappas, K. Pelechrinis, “Data Mining Techniques for (Network) Intrusion Detection Systems,” *Department of Computer Science and Engineering UC Riverside, Riverside CA*, 92521, 2007.

[113] W. Lee, S. J. Stolfo, K. W. Mok, “Adaptive intrusion detection: a data mining approach,” *AI Review*, vol. 14, no. 6, pp. 533-567, 2000.

[114] M. Hauskrecht, R. Pelikan, M. Valko, J. Lyons-Weiler, “Feature Selection and Dimensionality Reduction in Genomics and Proteomics,” *Fundamentals of Data Mining in Genomics and Proteomics*, pp. 149-172, 2007.

[115] H. Abusamra, “A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma,” in *Proc. of 4th International Conference on Computational Systems-Biology and Bioinformatics, Procedia Computer Science*, vol. 23, pp. 5-14,2013.

[116] X. Jia, Bor-Chaen, Melba M. Craford, “Feature mining for hyper spectral image classification,” in *Proc. of the IEEE*, vol. 101, no. 3, Mar. 2013. Article (CrossRef Link)

[117] L. Kauo, “Nonparametric weighted feature extraction for classification” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, May 2004.

[118] A. Ghosh, A. Datta, S. Ghosh, “Self-adaptive differential evolution for feature selection in hyperspectral imagedata,” *Applied Soft Computing*, 2013. Article (CrossRef Link)

[119] S. Li, H. Wu, D. Wan, J. Zhu, “An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine” *Knowledge Based Systems*, vol. 24, pp. 40-48, 2011. Article (CrossRef Link)

[120] Y.-Q. Zhao, L. Zhang, S. G. Kong, “Band-Subset-Based Clustering and Fusion for Hyperspectral Imagery Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, Feb. 2011. Article (CrossRef Link)

[121] T. Hastie, R. Tibshirani, J. Friedman, “The elements of statistical learning,” *Springer*, 2001. Article (CrossRef Link)

[122] Y. Rui, T. S. Huang, S. Chang, “Image Retrieval: Current Techniques, Promising Directions and Open Issues,” *Visual Communication and Image Representation*, vol. 10, no. 4, pp. 39-62, 1999. Article (CrossRef Link)

[123] T. Zhang, “Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations,” *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4689 -4708, 2011. Article (CrossRef Link)

[124] http://docs.oracle.com/cd/A91202_01/901_doc/olap.901/a86720/esdatao6.htm

[125] J. H. Friedman, J. J. Meulman, “Clustering object on subsets of the attributes,” 2002.

[126] L. Parson, E. Haque, H. Liu, “Subspace clustering for high dimensional data: a review,” *SIGKDD Exploration*, vol.

6, no. 1, pp. 90-105, 2005. Article (CrossRef Link)

[127] H. Brighton, C. Mellish, "Advances in instance selection for instance-based learning algorithms," *Data Mining* and Knowledge Discovery, vol. 6, no. 2, pp. 153-172, 2002. Article (CrossRef Link)

[128] T. Reinartz, "A unifying view on instance selection," *Data Mining and Knowledge Discovery*, vol. 6, no. 2, pp. 191-210, 2002. Article (CrossRef Link)

[129] J. Tang, S. Alelyani, H. Liu, "Feature Selection for Classification: A Review," *Data Classification: Algorithms and Applications*, CRC Press, 2013.

[130] S. Alelyani, J. Tang, H. Liu, "Feature Selection for Clustering: A Review," *Data Clustering: Algorithms and Applications*, CRC Press. 2013.

Vipin Kumar received M.Tech (Computer Science and Technology) from Jawaharlal Nehru University, New Delhi, India, in 2013. Presently, he is pursuing Ph.D. (Multi-view Reinforcement Learning for High Dimensional Data Classification) from Jawaharlal Nehru University, New Delhi, India. His research interests include data mining, machine learning, soft computing and classification.

Sonajharia Minz received M.Phil and Ph.D from Jawaharlal Nehru University, New Delhi, India. She is a professor in the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India. Her research interest includes Data Mining, Machine Learning, Soft Computing, Rough Sets and Geo-Spatial-Informatics.