Chapter - 2

Background and Review of Literature
CHAPTER - 2

BACKGROUND AND REVIEW OF LITERATURE

2.1 Introduction

In real situation, data often consists of many features. But the entire features are not necessary to get the target concept. Only few features may be used for getting the target concept. Apart from that, dataset are flooded with more of redundant and irrelevant features. Redundant features are those which do not have any information. This is otherwise termed as unnecessary features in the dataset. Irrelevant features are also present in the dataset which provide no useful information in any context. In machine learning and data mining operation, all the features in the dataset are not needed for target finding task. Thus, Feature Selection is termed as the process of selecting a subset of relevant features or non-redundant features for target finding operation. It helps to reduce the computation complexity of learning and prediction algorithms. It saves cost by way of selecting the appropriate features instead of all in the dataset. In common, feature selection can enhance the prediction accuracy by selecting appropriate features in the dataset. Its main objectives include providing a better understanding of data distribution as well as more accurate prediction [Gue, 2000] [Kos, 1996]. Thus, Feature selection technique is used to reduce the number of features in the dataset. Irrelevant and incomplete features in the dataset may have negative effects on a prediction task and in
addition to the increase of computational complexity. So, the importance of feature selection has got popular in data mining and machine learning areas. Now a days, irrespective of the discipline, the feature selection techniques are being adapted to preprocess the data in order to improve the model performance [Lis, 2009]. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information [Han, 2006] [Liu, 2008] [Bij, 2010].

2.2 General Feature Selection Procedures

Feature Selection procedure includes four important key steps; subset generation, subset evaluation, stopping criterion and result validation [Boy, 2011], which are shown in Figure 2.1.

![Figure 2.1 General Feature Selection Procedure](image-url)

**Figure 2.1 General Feature Selection Procedure**
Figure 2.1 explains the flow of common methodology adopted in the feature selection. Feature space is reduced to a subset of features which are evaluated based on the criterion. Finally these features are validated by the validating measures. The steps followed for general feature selection are elucidated in the following sub sections.

2.2.1 Subset Generation

Subset Generation is a search process that generates the candidate feature subset using certain search strategy. The process has two basic issues, namely search direction and search strategy. Firstly, a starting point must be selected which in turn influences the search direction. The search directions are divided into forward search, backward search and bi-directional search. The search process starts with an empty set and adds the features progressively one by one (forward search) or starts with full sets and removes the features one by one (backward search) or starts with both ends and adds and removes the features simultaneously (bi-directional search). Secondly, a search strategy must be decided. The search strategies are broadly categorized into three namely, complete search, sequential search and random search [Lad, 2011].

For $n$ features in a dataset, there are $2^n$ possible subsets. In case of this scenario, a thorough or complete (exhaustive) search of this space is practically not possible. The most realistic strategy involves some search strategy to reduce the search space. Different approaches applied to generate subsets in feature selection algorithms are briefly described here.
**Complete Search**

In complete search approach, all $2^n$ subsets of features are taken for consideration. The best possible or optimal subset ought to be found. The clear disadvantage of this method is in the computational complexity of the search, $O(2^n)$. Though the search is exhaustive, but there is no guarantee of an optimal subset. This means that not all $2^n$ have to be evaluated in order to guarantee an optimal subset.

**Heuristic Search**

The search through the space of subsets is guided by an heuristic algorithm in a way to avoid a complete search. However, since only a fraction of the search space is considered in the search, there are no guarantees that the optimal subset will be found. Several heuristics have been used for this purpose.

**Random Search**

Algorithms that apply on random search approach generate a new subset randomly at each iteration. Even though the search space remains $2^n$, the exact number of subsets that are considered by the algorithm is controlled by the number of iterations. As a result, the performance of the search process will depend on the resources available.

**2.2.2 Subset Evaluation**

In subset evaluation, an evaluation criterion is used to evaluate each newly generated subset. The evaluation criterion is used to determine the
goodness of the subset (i.e., an optimal subset selected using one criterion may not be optimal according to another criterion). The evaluation criteria are divided into Independent, Dependent and Hybrid criteria [Lad, 2011].

Selecting a final subset of features involves picking the best subset according to some evaluation measure. The evaluation method will set a value to each subset based on its ability to distinguish the different target classes. As a result, the subset with best evaluation measure should be able to generate highly accurate classification models. Different evaluation methods have been used for feature selection. These methods can be broadly grouped into four categories.

**Based on Distance**

These measures are based on the assumption that instances of a different class are distant in the instance space. The Euclidian Distance is used by a number of algorithms to compute the distance between instances.

**Based on Information Gain**

First introduced by Quinlan in his classification algorithm ID3 [Qui, 1986], information gain refers to the measure of how well a given feature separates instances according to their target classification. Such a statistical measure can be used to compare and consequently select features. Entropy, which measures the (im) purity of a collection of instances, is often used to characterize information gain.
Based on Dependency

These methods are based on the rationale that good subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Pearson’s correlation coefficient is an example of a measure used to determine the degree of correlation between a subset and the target class, while the uncertainty coefficient and symmetrical uncertainty coefficient [Pre,1988] and the information gain ratio [Qui, 1986] can be used to determine the feature-feature and feature-class dependencies.

Based on Consistency

In order to evaluate a given subset of features, its inconsistency rate is calculated by considering only the features of this subset. This rate refers to the number of instance pairs with same feature values but belonging to different classes. This evaluation scheme uses the Min-Features bias when selecting the best subset of features. Consequently, these measures find out the minimally sized subset that satisfies the acceptable inconsistency rate.

Based on Classifier Accuracy

The classifier created from a given subset of features is used as an evaluation function.

2.2.3 Stopping Criterion

It is used to stop the feature selection process. The feature selection process may stop under one of the following criteria [Lad, 2011].
1. A predefined number of features is selected,
2. Predefined number of iterations is reached,
3. In case, addition (or deletion) of a feature fails to produce a better subset,
4. An optimal subset according to the evaluation criterion is obtained.

### 2.2.4 Validation

The validation process is used to measure the resultant subset using the prior knowledge about the data. In some applications, the relevant features are known beforehand, a comparison is done between the known set of features with the selected features [Lad, 2011]. However, in most real-world applications, the prior knowledge about the data is not available. In such case, the validation task is performed by an indirect method. For example, the *classifier error rate test* is used as an indirect method to validate the selected features. The error rate on the full set of features and the same on the selected set of features are compared to find the goodness of the feature subsets.

### 2.3 Objectives of Feature selection

Classification or clustering problems in data mining often have a large number of features in the dataset, but not all of them are useful for classification or clustering. Irrelevant and redundant features may even reduce the performance. Feature selection aims to choose a small number of relevant features to achieve similar or even better classification performance than using...
all features. It has two main objectives of maximizing the classification performance and minimizing the number of features [Xue, 2012].

The following are the main objectives of feature subset selection:

1. To reduce the number of features
2. To avoid the irrelevant features
3. To ignore the redundant features
4. To increase the accuracy

2.4 Types and different approaches of feature selection

A large number of algorithms have already been proposed for the feature selection issues. These algorithms are varied from its functionality based on

1) the search strategy they use to determine the right subset of features
2) how each subset is evaluated, feature selection algorithms

There are two types of feature selection algorithms namely supervised and unsupervised. Supervised feature selection algorithms rely on measures that take into account the class information. A well known measure is information gain, which is widely used in feature selection [Das, 1997]. For feature selection in unsupervised learning, learning algorithms are designed to find natural grouping of the examples in the feature space. Thus feature selection in unsupervised learning aims to find a good subset of features that forms high quality of clusters for a given number of clusters [Dyb, 2004] [Liu, 2008].

Feature selection techniques can be divided in three main categories. They are,
1) Filter approach [Yul, 2003]

2) Wrapper approach [Koh, 1997]

3) Hybrid approach

2.4.1 Filter Approach

In the filter approach, the feature selection is performed as a pre-processing step in classifying the data. Here, the selection process is continued independently to improve the classification accuracy in the machine learning algorithm. In filter approach in order to evaluate a feature or a subset of features, it applies an evaluation function that measures the discriminating ability of the feature or the subset to differentiate class labels. In practice, different evaluation functions are used by different algorithms. Filters are generally much less computationally expensive than wrapper and hybrid algorithms. However, they may suffer from low performance if the evaluation criterion does not match the classifier well.

Figure 2.2 Filter Approach in Feature Selection
Figure 2.2 depicts the functionality of the filter approach in the feature selection. It encompasses the feature selection and its evaluation in one component. In the filter approach, the individual features are analysed and evaluated. Based on this evaluation, the features are selected for the data mining task such as classification.

2.4.2 Wrapper Approach

In contrast to Filter approach, a Wrapper approach algorithm [Joh, 1994] uses the learning algorithm as an integral part of the selection process. John et al. [Joh, 1994] observed that the idea behind Wrappers come from the fact that the optimal subset of features depends on the specific bias of the learning system. Therefore, the selection of features should consider the characteristics of the classifier. Then, in order to evaluate subsets, wrappers use the classifier error rate induced by the learning algorithms as its evaluation function. This aspect of wrappers results in higher accuracy performance for subset selection than simple filters. However, since wrappers have to train a classifier for each subset evaluation, they are often much more time consuming.

![Figure 2.3 Wrapper Approach in Feature Selection](image)
Figure 2.3 shows functionality of wrapper feature selection process that encompasses three in one single component. Feature subset search, evaluation and learning algorithm are performed under a single unit.

2.4.3 Hybrid Approach

The term “hybrid” refers to the fact that two different evaluation methods are used, a filter-type of evaluation and hybrid-type evaluation. In this approach the feature set is evaluated using both independent measure and a data mining algorithm. The independent measure is used to choose the best subset for a given cardinality and the data mining algorithm selects the finest subset among the best subsets across diverse cardinalities [Vee, 2010]. Figure 2.4 shows the Hybrid approach.

![Figure 2.4 Hybrid Approach in Feature Selection](image-url)

2.5 Dimensionality Reduction

Dimensionality reduction [Bar, 2005] is an important technique in various fields such as Data Mining, Machine Learning, Pattern Recognition, Image Retrieval and Text mining etc. Various real world applications in data mining usually have a high dimensional data. In order to handle the data...
adequately, its dimensionality needs to be reduced. The main objective of dimensionality reduction is to transform the high dimensional data samples into the low dimensional space such that the intrinsic information contained in the data is preserved. Once the dimensionality gets reduced, it helps to improve the robustness of the classifier and it reduces the computational complexity. Figure 2.5 shows the dimensionality reduction process. Generally, dimensionality reduction is performed using various techniques such as, Principal Component Analysis, Principal Feature Analysis, Fisher Criterion, Factor Analysis and Classical Scaling.

![Figure 2.5 Dimensionality Reduction Process](image)

**2.6 Feature Transformation**

Feature transformation is a process through which a new set of features is created. The variants of feature transformation are feature construction and feature extraction. Both are sometimes called feature discovery. Assuming the original set consists of $A_1, A_2, ..., A_n$ features, these variants are defined below. Feature construction is a process that discovers missing information about the relationships between features and augments the space of features by inferring or creating additional features [Mat, 1991] [Wne, 1994] [Tho, 1992]. After feature construction, the set may have additional $m$ features $A_{n+1}, A_{n+2}, ..., A_{n+m}$. 

---

*Feature Selection Techniques to Enhance Classification Accuracy in Data Mining* 18
For example, a new feature $A_k$ ($n < k \leq n + m$) could be constructed by performing a logical operation on $A_i$ and $A_j$ from the original set.

Feature extraction is a process that extracts a set of new features from the original features through some functional mapping [Wys, 1980]. After feature extraction, the set may have $B_1, B_2, ..., B_m$ ($m < n$), $B_i = F_i(A_1, A_2, ..., A_n)$, and $F_i$ functions. For instance, $B_1 = c_1A_1 + c_2A_2$ where $c_1$ and $c_2$ are coefficients. Subset selection is different from feature transformation wherein no new features will be generated instead only a subset of original features is selected and thus the feature space is reduced [Das, 1997] [Lan, 1994]. As to feature transformation, feature construction often expands the feature space, whereas feature extraction usually reduces the feature space.

Feature transformation and subset selection are not two totally independent issues. For example, feature construction and subset selection can be viewed as two sides of the representation problem. Features can be considered as a representation language. In some cases where this language contains more features than necessary, subset selection helps to simplify the language; in other cases where this language is not sufficient to describe the problem, feature construction helps enrich the language. It is common that some constructed features are not useful at all. Subset selection can then remove these useless features. It is also common to see the combined use of feature extraction and subset selection.
2.7 Fundamentals of Relief Algorithm based Feature Estimation

The Relief algorithm is simple and efficient. This research concentrates on Relief based feature selection algorithm. This section discusses the fundamental concept behind Relief algorithm. The Pseudo code of Relief algorithm is given as follows.

Relief Algorithm

Set all weights \( W[A] = 0 \)

For \( i = 1 \) to \( m \) do

Begin

Randomly select an instance \( R \);

Find the nearest hit \( H \) and nearest miss \( M \);

For \( A = 1 \) to all attributes do

\[
W[A] = W[A] - \frac{\text{diff}(A, R, H)}{m} + \frac{\text{diff}(A, R, M)}{m};
\]

End;

Where,

\( A \) – Attribute

\( R \) – Instance

\( H \) – Nearest Hit of Instances

In the pseudo code, the function \( \text{diff}() \) calculates the difference between values of attributes \( [A] \) for two instances. The difference between two values is either 1 for discrete values or 0 for continuous values. Value 1 means that the
attribute values are different between two instances. If the difference is 0, values are equal. For continuous attributes, the difference is the actual difference. The difference in value is normalized between the intervals 0 and 1. The normalization with \( m \) values assures that weights are in the interval of \([-1, 1]\).

The Function difference \( \text{diff}() \) in the Relief algorithm is used for calculating the distance between instances to find the nearest neighbours. The total distance is sum of distance of overall attributes. The default distance function for measuring the distance in the Relief algorithm is Manhattan distance. This distance function helps to find the nearest neighbour of instances.

The difference function for nominal and numerical attribute values are as follows:

**Nominal Attribute**

\[
\text{Diff}(A, I_1, I_2) = \begin{cases} 
0 ; & \text{Value}(A, I_1) = \text{Value}(A, I_2) \\
1 ; & \text{Otherwise}
\end{cases} \quad \ldots (2.1)
\]

**Numerical Attribute**

\[
\text{Diff}(A, I_1, I_2) = \left( \frac{\text{Value}(A, I_1) - \text{Value}(A, I_2)}{\max(A) - \min(A)} \right) \quad \ldots (2.2)
\]

Where, \( A \) - Attribute

\( I_1, I_2 \) - Instances

### 2.8 Principles of Mean-Variance based Relief Algorithm

Mean-Variance based Relief algorithm is one of the variant of Relief feature selection algorithm. The principle behind the Mean-Variance based Relief algorithm is to estimate the quality of attributes according to how well
their values distinguish between instances that are near to each other. This algorithm weights each feature according to its relevance to the classification task. Initially all weights are set to zero and then updated iteratively. In each iteration, this algorithm chooses a random instance \( i \) in the dataset and estimates how well each feature value of this instance distinguishes between instances close to \( i \). In this process two groups of instances are selected such that some closest instances belonging to the same class and some belonging to a different class. With these instances, Relief will iteratively update the weight of each feature according to how well this feature differentiates between data from different classes and at the same time it is recognizing data from the same class. At the end, a certain number of features with the highest weights are selected. A threshold may be used in such a way that only the features with weights above this value are selected.

The basic principle of Relief algorithm is to find two nearest neighbors such as nearest hit and nearest miss. The nearest miss and nearest hit are termed as M and H respectively. Mean-variance based algorithm selects the instances randomly and looks for nearest neighbor. If the instances fall in the same class, it is termed as nearest hit [H]. Otherwise, it is termed nearest ms [M]. Then the weight is estimated by using the \( \text{diff}(\cdot) \) function. Two parameters are passed to this function such as instances X and near hit [H] or near miss [M]. Mean value of weight for feature is taken in this algorithm. This algorithm uses the Manhattan distance for selecting near hit and near miss for instances. The pseudo code for Mean Variance based Relief is shown below.
Pseudocode for mean, variance based Relief

Set the weight $W[A] := 0.0$

For $i := 1$ to $m$ do begin

Begin

Randomly select an instance $X_i$;

Find nearest hit $[H]$ and nearest miss $[M]$;

For $A := 1$ to $m$ do

\[ W[A] := W[A] - \frac{\text{diff}(X_i, H)^2}{m} + \frac{\text{diff}(X_i, M)^2}{m}; \]

End;

2.9 Supervised Learning

Data mining methods take two forms of inductive-learning methods, namely, Supervised Learning and Unsupervised Learning. In supervised learning, the class label of the training data is already known. The training data consist of pairs of input objects and desired outputs. The output of the function can be a continuous value or can predict a class label. The Supervised Learning methods attempt to identify the relationship between the independent variable and a dependent variable. The identified relationship is referred as a model. Generally, the model explains the hidden phenomenon which is present in the dataset and this is used to predict the value of the target attribute.

2.9.1 Naive Bayes Algorithm

The Naive Bayes is a simple probabilistic classifier based on Bayes’ theorem with strong (Naïve) independence assumption. It is used for estimating
the probability of each class value during classification and prediction. In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The conditional probability of the selected class is computed by the following equation (2.3)

\[ P(C_i | X_{test}) = \frac{P(C_i)}{P(X_{test})} \prod_{m=1}^{n} P(x_m | C_i) \]  

...(2.3)

where,

- \( C_i \) is the \( i^{th} \) class,
- \( X_{test} \) is a test data,
- \( x_m \) is the value of the \( m^{th} \) feature in the \( X_{test} \) data.

### 2.9.2 J48 Classifier

In data mining, decision tree is one of the most widely used inductive learning methods to classify the given input data. It is originally implemented from Decision Theory (DT) and Statistics. The DT models are used to examine the data and induce the tree and their rule is used to make predictions. The main goal of the DT is to classify the data into discrete groups which make a strong separation in the values of the dependent variable. The Decision Tree includes various types of algorithms such as ID3. The J48 method is an algorithm for decision tree generation and an extension of ID3.
The J48 algorithm classifies a new data item by creating a decision tree based on the attribute values of the available training data. In that, the initial feature is chosen as the root of the tree. Generally, the training sample set is larger in size. Hence, the generated decision tree contains a number of branches and layers. In addition, irrelevant and noisy data present in training set could lead to some abnormal branches. To overcome the above issue, the decision tree must be pruned. The J48 algorithm provides two pruning methods, namely, Subtree Replacement; where the nodes in the tree are replaced with a leaf and Subtree Raising; where a node is moved upwards towards the root of the tree, replacing other nodes along the way. The information gain is calculated for each feature. The features are segmented based on the information gain and the best features which have high information gain are selected.

2.10 Related Works on Feature Selection

This section describes the related works of feature selection on filter approach.

Lei Yu and Huan Liu [Lei, 2003] introduced a feature selection algorithm called Fast Correlation Based Filtering (FCBF) approach using the predominant correlation concept. In this algorithm, the feature subset was selected based on Symmetric Uncertainty and F-correlation measure. In this approach, redundant and irrelevant features were removed and it showed improvement in the classification accuracy.
Lei Yu et al. [Lei, 2004] identified that there was a need for explicit redundant analysis in feature selection. Hence, a new framework for efficient feature selection was defined through relevance and redundant analysis. This framework divided the relevant analysis and redundant analysis. A new feature selection algorithm was also implemented and it was verified against various learning algorithms to extract the best feature set.

Xiubo Geng et al. [Xiu, 2007] investigated the existing feature selection algorithms for ranking models. They found that there was a striking difference between the ranking and classification. Hence, a feature selection algorithm was proposed. In this algorithm, two ranking models, Ranking SVM and RankNet were used to extract the best feature subset.

Chung-Jui Tu et al. [Chu, 2007] proposed a feature selection algorithm using Particle Swarm Optimization (PSO) and Support Vector Machines (SVMs). The PSO was used to select the best feature subset. The SVM with the one-versus-rest method was used to evaluate the fitness function of PSO. The algorithm was validated with various classification problems.

Noelia Sanchez-Marono et al. [Noe, 2007] studied the four filter methods for feature selection. In their study, they made a comparison among the four filter methods to find the best filter method. Based on the study, they proposed a hybrid filter algorithm using best filter methods. To find the best filter method, a comparison was made among the four filter methods which are ReliefF, Correlation based Feature Selection (CBFS), Fast Correlated Based Filter (FCBF) and INTERACT.
Antonio Arauzo-Azofra et al. [Ant, 2008] defined a feature selection method, namely, Consistent-Based feature selection. It was an useful measure for various feature selection methods. Hence, the proposed method achieved similar accuracy result, than the wrapper approach and it also attained higher feature reduction.

Appavu et al. [App, 2009] proposed a feature selection algorithm using association rule mining and Information gain. The A priori algorithm was used to find the relevant attributes. Information Gain was used to remove the irrelevant and redundant features in the dataset. The result of the algorithm showed that there was no improvement in the classification accuracy.

Huanjing Wang et al. [Hua, 2010] introduced a feature selection method using the filter–based ranking techniques. The proposed technique was called Threshold Based Feature Selection (TBFS). Each attribute’s value was normalized between 0 and 1 using the F-measure and the independent attribute was paired individually with the class attribute. This technique was useful to find the smaller subset of features and it showed an improvement in the classification accuracy.

Athanasios et al. [Ath, 2010] proposed a feature selection algorithm using the Correlation-Based filter approach called Relevance, Redundancy and Complementarity Trade-off (RRCT). In this algorithm, linear correlation coefficient was used to remove the irrelevant, redundant and noisy features. Gaussian distribution method was used to obtain the best feature subset.
Yuxuan Sun et al. [Yux, 2011] studied the existing RELIEF algorithm for feature selection using feature weight estimation. From the study, they found that RELIEF algorithm was unstable and that lead to poor accuracy of expected results. To overcome the problem, a feature selection algorithm was proposed based on Mean-Variance Model.

The Mean Variance Model was used to revise the feature weight estimation method according to the original RELIEIF algorithm. In this algorithm, the mean and the variance of the discrimination among instances were considered as the criterion of feature weight estimation. The algorithm was validated through an experimental study and the result indicated that the feature subsets generated by the proposed algorithm had a better performance.

Qinbao Song et al. [Qin, 2011] developed a Clustering-Based feature subset selection algorithm for high dimensional data. The Graph-theoretic method was used to divide the features into clusters. The features that were strongly related to the target class were selected as the best feature subsets. In this, they treated each cluster as a single feature. Hence, the dimensionality was drastically reduced. The algorithm was compared with various existing algorithms and it showed a minimum improvement in the prediction accuracy and classification performance.

Debahuti Mishra and Barnali Sahu [Deb, 2011] proposed a model for feature selection using Signal to Noise Ratio (SNR) ranking to enhance the predictive accuracy. In this algorithm, they proposed two approaches for
selecting the best features. In the first approach, K-means clustering and SNR ranking were used to get the top ranked features. In the second approach, SNR ranking was used to obtain the best features. The two models were validated through different classifiers by conducting experiment. They found that the performance of the learning algorithms decreased the classification accuracy.

Tingquan Deng et al. [Tin, 2011] introduced a notation called Knowledge Granularity. The knowledge granularity was used to find the relationship between conditional attributes and decision attributes. An evaluation function was used to measure the significance of conditional attributes. They developed an optimized algorithm for feature selection based on the evaluation function. The algorithm was validated and it showed improvement in the classification accuracy.

Myo Khaing and Nang Saing Moon Kham [Myo, 2011] studied the various feature selection algorithms using Multiple Correspondence Analysis (MCA) and found many disadvantages in the same. Hence, they proposed a feature selection algorithm called Modified - Multiple Correspondence Analysis (M-MCA). From the experiment conducted, they claimed that the result of the experiment gave better performance compared to the existing algorithm using simple MCA.

Boyang Li et al. [Boy, 2011] designed a feature selection model based on correlation analysis and SVM ranking method. In this algorithm, the correlation based clustering was used to group the feature into some clusters. Influence Qualities were calculated for each of the feature in the cluster. Using
the feature sensitivity in the SVM, the best features were obtained. This model was tested with some real dataset and they stated the result which showed improvement in the classification accuracy compared to the existing algorithms.

Danyang Cao et al. [Dan, 2012] analyzed the discrimination feature selection algorithm using Information Theory, which did not consider the discriminant and continuous features of the dataset. From the analysis, they proposed a feature selection algorithm. In this algorithm, an entropy breakpoint concept was introduced. This algorithm was validated with various real-world dataset. The result of the experiment specified that the algorithm had a high computational complexity with very low prediction accuracy.

Chinna Gopi et al. [Chi, 2012] proposed an algorithm to solve optimization problem in Feature Selection. This algorithm was used to find the best features using Greedy Search method and Greedy Search Loss of ranking method. The algorithm was validated with public dataset. This algorithm worked well in generated optimized feature subset but, the computational cost was higher than the existing algorithms.

2.10.1 Related Works on Feature Selection on Agriculture and Biological Datasets

Data mining and its various methodologies are used for industrial, commercial, and scientific purposes [Ebr, 2010] [Ebs, 2010]. Recently, agricultural and biological research studies have used various data mining
techniques for analyzing large dataset and creating useful classification patterns in dataset. The novelty and advancement in feature selection on data mining technologies are able to bring more fruitful results in different discipline [His 2006], [Ami, 2010]. Data mining tasks generally involve hundreds and thousands of attributes. The major portion of time in model building process involves examining the variables to be included in the model. Feature selection allows the features set to be reduced in size and it is creating a more manageable set of attributes for modeling [Liu, 2008]. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information [Han, 2006] [Liu, 2008] [Bij, 2010]. When the feature space is complex and the data distribution patterns are not uniform, the use of feature selection method allows analyzing the more complex data compared to other statistical techniques [Dru, 2002] [Gau, 2006].

Ehsan Bijanzadeh [Ehs, 2010] has reported that the supervised feature selection algorithm was applied to determine the most important features contributing to wheat grain yield. Four hundred seventy two fields (as records) from different parts of Iran which were different in 21 characteristics (features) were selected for feature selection analysis.

Selection of the wide range of features, including location, genotype, irrigation regime, fertilizers, soil textures, physiological attitudes and morphological
characters, provided the opportunity of precise simultaneous study of a large number of factors in wheat grain yield topic by hand of data mining. These features included culture type, location, soil texture, 1000 kernel weight, nitrogen supply, irrigation regime, biological yield and organic content of the soil, the amount of rainfall, genotype, plant height, and spike number per unit area. Interestingly, growing season length and plant density were the second most important features for wheat grain yield. Based on the feature selection model, culture type, as dry land farming or irrigated, severely affected wheat grain yield. The soil pH had a marginal effect on wheat grain yield. The results of this investigation demonstrated that feature classification using feature selection algorithms might be a suitable option for determining the important features contributing to wheat grain yield, providing a comprehensive view about these traits. This is the first report in identifying the most important factors on wheat grain yield from many fields using feature selection model.

2.10.2 Related Works on Feature Selection by Relief and its Extension Algorithms

Relief is considered as one of the most successful algorithms for assessing the quality of features due to its simplicity and effectiveness [Die, 1997]. The Relief algorithms are a family of attribute weighting algorithms that can efficiently identify associations between attributes (e.g., SNPs) and the class (e.g., disease status) even if the attributes have nonlinear interactions.
(e.g., epistatic) without significant main effects [Kir, 1992] [Die, 1997]. RELIEF was extended to handle noisy and missing data [Kon, 1994].

An iterative RELIEF (I-RELIEF) algorithm is used to alleviate the deficiencies of RELIEF by exploring the framework of the Expectation-Maximization algorithm. I-Relief was introduced to support to multiclass settings by using a new multiclass margin definition [Yij, 2007].

Abdolhossein Sarrafzadeh et al., [Abd, 2012] studied the effects of reducing the number of features and selecting the most effective subset of features in the context of content-based image classification and retrieval of objects using Relief F algorithm. Their experimental result shows that employing Relief F on Coil-20 image dataset improves the speed and accuracy.

Yan Wei et al., [Yaw, 2011] have briefed that in the feature set of complex products which are in high dimension, the set usually contains useful information, irrelevant information and redundant information. However, the former is usually buried in the latter two. Therefore, the recognition of the most useful information in the original dataset, which is defined as the identification of Critical-To-Quality features, becomes a key process in the field of quality control. The traditional methods include QFD, Taguchi loss function and Decision tree, etc. However, almost none of them can deal with the high dimensional quality feature set with both accuracy and easiness.

Casey S Greene et al., [Cas, 2009] have proposed the SURF algorithm. They reported that SURF's ability to detect interactions in this domain is
significantly greater than that of ReliefF. Similarly SURF, in combination with the TuRF strategy significantly outperforms TuRF alone for single nucleotide polymorphisms [SNP] selection under an epistasis model.

Matthew E Stokes et al., [Mat, 2012] have proposed and developed a new spatially weighted variation of Relief called Sigmoid Weighted ReliefF Star (SWRF*), and applied it to synthetic SNP data. When compared to ReliefF and SURF*, which are two algorithms that have been applied to SNP data for identifying interactions, SWRF* had significantly greater power. They reported that the new Relief algorithm called SWRF* that had greater ability to identify interacting genetic variants in synthetic data compared to existing Relief algorithms.

Fan Wenbing et al., [Fan, 2012] have proposed an adaptive Relief (A-Relief) algorithm to alleviate the deficiencies of Relief by dividing the instance set adaptively. According to them, A-Relief has performed better in image dataset.

Yuxuan SUN et al., [Yux, 2011] devised a new strategy on the Relief algorithm. They briefed the defects of Relief algorithm. According to them, as Relief algorithm selects the instances randomly, the feature weight estimation is uncertain. So the randomicity and the uncertainty of the instances used for calculating the feature weight vector in the Relief algorithm, the results lead to poor evaluation accuracy. To overcome this issue, a novel feature selection algorithm based on Mean-Variance model is proposed by them. This algorithm
takes both the mean and the variance of the discrimination among instances into account as the criterion of feature weight estimation, which makes the result more stable and accurate. Based on real seismic signals of ground targets, experiment results indicate that the subsets of feature generated by proposed algorithm have better performance.

Blessie E.C et al., [Ble, 2011] proposed a new algorithm called Relief-Disc. It works based on Discretization. Discretization partitions features into finite set of adjacent intervals. Instead of using random sampling for selecting the instance, they have suggested to take instance from each interval which reduces the computational complexity and maintains the quality of features. Also, there was no need of user input for sample size parameter. Experimental results showed that the performance of the new algorithm is better when compared with the existing Relief algorithm. According to them, Relief-Disc performed better than Relief.

2.11 Chapter Summary

In this chapter, the basic terminologies related to the thesis were discussed. In real world, various data mining applications need to handle larger volume of data. The performance of the data analysis is affected due to high dimensionality of data. The high dimensionality problem is overcome by the technique called feature selection. The basic concept of feature selection and its procedures are discussed. The feature selection approaches are categorized into
filter, wrapper and hybrid. The overview of dimensionality reduction and its various techniques were discussed. The learning methods, namely, supervised, unsupervised and semi supervised are presented. Among these, only the supervised learning methods are considered to improve the classification accuracy. Supervised learning methods and the two supervised algorithms, namely Naive Bayes and J48 are discussed. Finally, the literatures for feature selection expressed the following assessments:

1. Selecting the relevant features using different approaches.
2. Selecting the appropriate features with feature weight as well as relevant to the target.
3. Selecting relevant features by feature weight or ranking mechanism.

The next chapter presents the proposed methodology for the research work.