**Chapter 2**

**Review of Literature**

Attribute selection has numerous applications in diverse fields of science and engineering. As large and complex datasets become typical and usual, data gathering efforts are increasingly oriented towards extracting informative attributes that reveal the statistic or stochastic patterns in the data.

The main idea of attribute subset selection is to remove redundant or irrelevant attributes from the dataset as they can lead to a reduction of the classification accuracy or clustering quality and to an unnecessary increase of computational cost [Blum and Langley, 1997; Koller and Sahami, 1996]. With dimensionality reduction techniques [Martinez-Uso et.al, 2006] the size of the attribute space can often be decreased strikingly without loosing a lot of information of the original attributes space. An important disadvantage of dimensionality reduction is the fact that the linear combinations of the original attributes are usually not interpretable and the information about how much an original attribute contributes is often lost.

Attribute subset-selection focuses on the reduction of unnecessary attributes or objects in a given dataset to improve the performance of some algorithm that is trying to solve a problem. Whatever be the method, the key aspect in attribute space is the search heuristic as
attribute selection is fundamentally a process of selecting a subset of attributes from a power set of attributes.

It is hard to find a subset of attributes that closely resembles the large attribute set. The key here is to find relevance and the stopping criterion. Ideally all the attributes that contribute to a given context must be selected. This requires codification of context and definition of the stopping criterion [Hall, 2000; Dash and Liu, 2000].

An attribute that is part of the attribute subset used by a learning algorithm is a good attribute to use if it is either a good predictor of the class by itself, or a good predictor of the class when taken together with some other subset of attributes in the set. At the same time, the attribute should not be redundant given the other attributes in the selected attribute set. Attribute Selection needs methods [Zhu et.al, 2009] capable of considering all the available sources of information and of combining them.

Relevance is usually characterized in terms of correlation or mutual information, of which the latter is one of the widely used measures to define dependency of variables.
A good attribute selection algorithm provides:

a) good global search capability that allows for the exploration of new regions of the solution space without getting stuck in local minima

b) rapid convergence to a near optimal solution

c) good local search ability

d) high computational efficiency [Gheyas and Smith, 2009; 2010].

The search procedure that produces candidate attribute subsets for evaluation based on a certain search strategy has been adopted in which each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Then the selected best subset usually needs to be validated by prior knowledge or different tests via synthetic and/or real world datasets.

In order to solve the problem statement given in chapter 1.7, the review of literature has been carried out extensively by considering the following aspects in attribute selection

A. General and hybrid approaches

B. Information gain

C. Clustering methods

The review also highlights the work done in missing values imputation.
2.1 General and hybrid approaches

The review on general and hybrid approaches has been carried out to present a clear idea about the progresses that are carried out in attribute selection with reference to correlation, dimension reduction and hybrid methods.

Langley [Langley, 1994] notes that attribute selection algorithms that search through the space of attribute subsets must address four main issues:

- the starting point of the search,
- the organization of the search,
- the evaluation of attribute subsets and
- the criterion used to terminate the search.

Different algorithms address these issues differently. Although there are many attribute selection algorithms reported in the specialized literature, not all the attributes are taken into account for the mining process.

Algorithms used for selecting attributes prior to concept induction fall into two categories. Wrapper methods wrap the attribute selection around the induction algorithm to be used, using cross-validation to predict the benefits of adding or removing an attribute from the attribute subset used.
Filter methods are general preprocessing algorithms that do not rely on any knowledge of the algorithm to be used. There are strong arguments in favor of both methods. Some of them are effective, but very costly in computational time (e.g., wrappers methods), and others are fast, but less effective in the attribute selection task (e.g., filter methods). Specifically, wrapper methods, although effective in eliminating irrelevant and redundant attributes, are very slow because they apply the mining algorithm many times, changing the number of attributes each time of execution as they follow some search and stop criteria [Kohavi and John, 1997]. Filter methods are more efficient in a way as they use some form of correlation measure between individual attributes and the class [Piramuthu, 1998]. Though they measure the relevance of each isolated attribute, they cannot detect if redundant attributes exist, or if a combination of two (or more) attributes, apparently irrelevant when analyzed independently, are indeed relevant [Molina, Belanche and Nebot, 2002].

Embedded methods [Tuv et al., 2009] use all the variables to generate a model and then analyze the model to infer the importance of the variables. The model gives details of an algorithm using tree-based ensembles to generate a compact subset of non-redundant attributes. Parallel and serial ensembles of trees are combined into a mixed method that can uncover masking and detect attributes of secondary effect.
Janecek et al. [Janecek et al., 2008] presents two classes of attribute reduction strategies: Attribute subset selection and dimensionality reduction are explained. These techniques are used for reducing the attribute space of an attribute set, which is an important component of both supervised and unsupervised classification or regression problems. While in attribute subset selection a subset of the original attributes is extracted, dimensionality reduction in general produces linear combinations of the original attribute set.

The work investigates relationship between several attribute space reduction techniques and the resulting classification accuracy for two very different application areas. On the one hand, they consider e-mail filtering, where the attribute space contains various properties of e-mail messages, and on the other hand, drug discovery problems is also considered, where quantitative representations of molecular structures are encoded in terms of information-preserving descriptor values.

There are three kinds of heuristic search methods [Liu and Kender, 2003] proposed in literature: forward selection, backward elimination and genetic algorithms. Forward selection starts with the empty set and successively adds individual attributes, usually following a variant of a greedy algorithm, terminating when no improvement is possible. However, it can't remove any attributes, and therefore ends up making what amounts to local optimizations to the growing set.
Backward elimination, which does the reverse, starts with the full set of attributes and heuristically subtracts individual attributes. It suffers from a similar problem of local optimization, as removal of an attribute is irrevocable. Genetic algorithm, which permits both the addition and deletion of attributes to a surviving population of evolving subsets of limited cardinality, is more likely to seek a global optimum. But it is computationally costly, and requires a more elaborate definition of algorithm convergence.

A novel approach for attribute selection [Balamurugan and Rajaram, 2009] in high dimensional data using Bayes theorem is discussed. The dependent attributes are the attributes, in which an attribute depends on the other attribute in deciding the value of the class attribute. The proposed work uses dependency methods but, does not use Bayes theorem.

Siddhiq [Siddhiq, 2008] attempts Pearson Correlation measures for attributes selection for static hybrid deduction for malwares. Whereas Hapuarachchi [Hapuarachchi, 2006] establishes the level of Attribute Independence between any two candidate attributes, gives an indication that, the Pearson Correlation is not appropriate for the experiments, because it does not follow a normal distribution. The Pearson Correlation inherently assumes the data to have a normal distribution.
The main drawback of such methods is that they are not able to detect inter-attribute-dependencies, one important example is the XOR-problem. Neither the first nor the second dimension alone helps to determine from which class an example is stemming, only both dimensions together contain enough information about the class membership. The proposed work takes a parallel information gain based approach which is in turn supplemented by clustering and correlation methods to take into account inter-attribute dependencies.

A new attribute selection method [Geng et al., 2007] in which, for each attribute; the value is used to rank the training instances, and define the ranking accuracy in terms of performance measure or loss function as the importance of the attribute. The correlation between the ranking results of two attributes as the similarity between them is considered for attribute selection. The proposed work has a similar aim but it has a multi-stage approach based on ranking, clustering and similarity measures.

A hybrid two stage attribute selection algorithm has been dealt by Vidyavathi and Ravikumar [Vidyavathi and Ravikumar, 2008]. The method uses the Fischer's gain ratio and mutual information. Michalak and Kwasnicka [Michalak and Kwasnicka, 2006] have advocated a correlation based attribute selection strategy to take into account the interdependency between variables.
She et al. [She et al., 2005] present a way of performing attribute selection through class propagation, eliminating the need of join before attribute selection. They propagate a very compact data structure that provides enough information for selecting attributes to each data source, thus allowing attributes to be evaluated locally without looking at any other information.

Deisy et al. [Deisy et al., 2007] compares three methods namely Fast Correlation Based Attribute Selection (FCBF), Multi thread based FCBF attribute selection and Decision Dependent – Decision Independent Correlation (DDC-DIC). These are some approaches that are concerning the relevance of the attributes and the pair wise attributes correlation for redundancy checking in order to improve the prediction accuracy and reduce the computation time. *The proposed work trades-off computation time for better performance. The prediction accuracy in this approach is increased by the unified approach for attribute selection.*

Saeys [Saeys, 2004] contributes to attribute selection by combining filters, wrappers and embedded methods and finally the attribute selection is done using attribute score and performs classification. Law [Law, 2006] proposes the concept of attribute saliency and introduces an expectation-maximization (EM) algorithm for its estimation for clusters.
Loscalzo et al. [Loscalzo et al., 2009] explain that exploiting intrinsic attribute groups in the underlying data distribution will give improved performance. Their work uses an attribute selection framework which approximates intrinsic attribute groups by a set of consensus attribute groups and performs attribute selection and classification in the transformed attribute space.

Guyon et al [Guyon et al., 2003] uses Pearson correlation coefficient for variable ranking by prediction of continuous outcome.

Hall [Hall, 1999] uses Correlation based Attribute Selection (CFS) using three machine learning algorithms namely C4.5 (a decision tree learner), IBL (an instance based learner), and Naive Bayes. Experiments of this approach are carried out on artificial datasets and showed that CFS quickly identifies and screens irrelevant, redundant, and noisy attributes, and identifies relevant attributes as long as their relevance does not strongly depend on other attributes. This approach gives an account of the irrelevant attribute that has been evolved using machine learning algorithm but does not account on the number of attributes that seems to be irrelevant but takes priority in other attribute selection method such as ranking which concludes some attributes that are ranked with high priority. The proposed research work tries to list such outcomes by exactly providing the number of attributes for a particular outcome, using hybrid approach and then aggregating the selected attribute from different subset.
A Parallel approach for attribute selection is presented by De Souza et al, [De Souza et al., 2006] which describe a scalable, effective algorithm to identify groups of correlated attributes. This algorithm can handle non-linear correlations between attributes, and is not restricted to a specific family of mapping functions, such as the set of polynomials. The proposed work uses similar methods, but with different focus and design.

Zhang et al. [Zhang et al., 2008] have advocated an approach involving pair-wise constraints using a filter method. The proposed approach aggregates the pair-wise constraints and information gain measures. The idea is that while inter-relationship clustering is important, it is always necessary to supplement it with measures to find the most important measures first. These are then used as a guide for identifying the optimal subset.

Many attribute selection algorithms assume that the attributes are independent between each other given the class, which can produce models with redundant attributes and/or exclude sets of attributes that are relevant when considered together. Mejia-Lavalle and Morales [Mejia-Lavalle and Morales, 2006] describes a best first search algorithm, called buBF is used for attribute selection. buBF uses a heuristic function based on n-way entropy to capture interdependencies among variables. The proposed approach uses a multistage hybrid approach for identifying various aspects of important
Pairwise selection methods [Bo and Jonasse, 2002] have been mooted to alleviate problems in individual ranking, forward search and backward search.

Michalak and Kwasnicka [Michalak and Kwasnicka, 2006] outlines a correlation based method modeled on a modified pairwise selection strategy which uses mixed univariate and bivariate attribute evaluation based on the correlation between the attribute. The proposed research work uses clustering and complex correlation based systems to expand the scope of subset selection.

Zhang and Ling [Zhang and Ling, 2003] address the theoretical issue of mapping nominal attributes into numerical ones when applying the Naive Bayes learning algorithm. Such mappings can be viewed as a part of discretization in data preparation. They explain that the learnability of Naive Bayes is affected when such mappings are applied. This work helps to understand the influence of numeric mappings on the property of nominal functions and on the learnability of Naive Bayes.

Mutual information based algorithms [Hutter and Zaffalon, 2005] use the attribute to attribute correlation aspects to give good results.
The MIFS [Battit, 1994], its variants MIFS-U [Kwak and Choi, 1994], MIFS-C [Bakus and Kamel, 2003], MIFS-C modified [Bakus and Kamel, 2006] are some well established algorithms in the domain. The proposed research work shares some similarities in the parameter estimation aspects with the above methods.

A forward orthogonal search method [Wei and Billings, 2007] have been mooted for efficient attribute subset search. This method seeks to attributes that are selected in stepwise way, one at a time, by estimating the capability of each specified candidate attribute subset to represent the overall attributes in the measurement space with appropriate correlation function. The method uses an innovative attribute detection, ranking and search procedure. The proposed research work shares similarities in the ranking and search approach.

Difference-similitude attribute selection methods [Skowron and Rauszer, 1992; Ming and Puliu, 2007; Ming et al., 2005; Hao et al., 2003; Keyun, 2000] take not only the difference but also the similitude into account. Usually, forward selection is chosen in DS-based attribute selection. The core attributes are selected at first. Then, the left attributes are selected in the order of their important ranks in similitude matrix or similitude set. The importance is often represented by frequencies of the attributes appearing in the similitude matrix or set. This may cause some attributes to be ignored. If an attribute is very important in classifying but does not appear in
similitude matrix, or appear with a low frequency, the attribute will never be selected. Therefore attributes must be selected according to a combined criterion that takes both the difference and similitude into account. The proposed work is similar in principle in that it uses clustering methods to find the degree of similarity or difference. The importance of an attribute is calculated using the information gain metrics.

A decision tree selects relevant attributes using a tree structure [Quinlan, 1986; Quinlan, 1993; Quinlan and Cameron-Jones, 1995]. The attributes are divided into different subsets and each subset is examined by some data-mining criteria to determine the relevance. The criteria differ from one application to another. The most common tree search strategies are the depth-first and the breadth first tree search [Liu and Motoda, 1988], which give different orders of how the attribute sets are examined. In the depth first tree method, the search starts with a single attribute and adds one attribute for iteration until all attributes are used. On the other hand, the breadth first strategy examines all single attributes first, and then tests on all two-attribute sets and so on. The whole attribute set will be tested at last. These search methods are complete search and they are computationally very expensive. The filter methods are usually criticized because they totally ignore the learning algorithm during the attribute selection process. Building classifiers after determining the attribute set usually limits the performance of the classifiers.
Principal Component Analysis (PCA) has been widely used for multivariate data analysis and dimension reduction [Jolliffe, 2002]. Intuitively, PCA is a process to identify the directions, i.e., principal components (PCs), where the variances of scores (orthogonal projections of data points onto the directions) are maximized and the residual errors are minimized assuming the least square distance.

The stepwise (or greedy) search adds/removes a single attribute to/from the current subset [Kittler, 1978]. It considers local changes to the current attribute subset. Often, a local change is simply the addition or deletion of a single attribute from the subset. The stepwise, which is also called the Sequential Forward Selection (SFS)/Sequential Backward Selection (SBS) is probably the simplest search procedure and is generally sub-optimal and suffers from 'nesting effect'. It means that the attributes that are once selected/deleted cannot be later discarded/re-selected. To overcome this problem, Pudil et al. [Pudil et al., 1994] proposed a method to flexibly add and remove attributes, which they called 'floating search'.

Another famous search approach is based on the Genetic Algorithm. The Genetic Algorithm is a combinatorial search technique based on both random and probabilistic measures. Subsets of attributes are evaluated using a fitness function and then combined via cross-over and mutation operators to produce the next generation of subsets [Yang and Honavar, 1998]. The Genetic Algorithm employs a
population of competing solutions, evolved over time, to converge to an optimal solution. Effectively, the solution space is searched in parallel, which helps in avoiding local optima. Genetic Algorithm based attribute selection solution would typically be a fixed length binary string representing an attribute subset, where the value of each position in the string represents the presence or absence of a particular attribute. The Genetic Algorithm is able to achieve better performance than other conventional methods [Gletsos et al., 2003; Oh et al., 2004].

To deal with the high amount of possible attributes, and often limited amount of samples, the Interpolated Markov Model (IMM) [Salzberg et al., 1998], which used interpolation between different orders of the Markov model to deal with small sample sizes, and a filter method to select only relevant attributes is introduced. In further work, [Delcher et al., 1999] the concept is extended with IMM framework to deal with non-adjacent attribute dependencies, resulting in the interpolated context model (ICM), which crosses a Bayesian decision tree with a filter method to assess attribute relevance. Recently, the avenue of FS techniques for coding potential prediction is further pursued [Saeys et al., 2007]. They combined different measures of coding potential prediction, and then used the Markov blanket multivariate filter approach (MBF) to retain only the relevant ones.
The application of multivariate filter methods ranges from simple bivariate interactions [Bo and Jonassen, 2002] towards more advanced solutions exploring higher order interactions, such as correlation-based attribute selection (CFS) [Wang et al., 2005; Yeoh et al., 2002] and several variants of the Markov blanket filter method [Gevaert et al., 2006; Mamitsuka, 2006; Xing et al., 2001] is carried out. The Minimum Redundancy-Maximum Relevance (MRMR) [Ding and Peng, 2003] and Uncorrelated Shrunken Centroid (USC) [Yeung and Bumgarner, 2003] algorithms are two other solid multivariate filter procedures, highlighting the advantage of using multivariate methods over univariate procedures.

Anytime algorithms [Dean and Boddy, 1988; Horvitz, 1987; Russell and Wefald, 1991; Zilberstein, 1996] offer a trade-off between the solution quality and the computational requirements of the search process. The approach is known under a variety of names, including flexible computation, resource bounded computation, just-in time computing, imprecise computation, design-to-time scheduling or decision-theoretic meta-reasoning. All these methods attempt to find the best set of attributes possible from the operational constraints.

Model building procedure usually deals with attribute selection first, and then with ensemble creation [Ho, 1998; Sharkey, 1996]. However, separation of the issues is not desirable since an attribute subset selected may be optimal for a single classifier, but may not be optimal
for an ensemble. In fact, it may be better off if each member classifier of an ensemble employees different attribute subsets. Since they are inter-related, these two issues need to be addressed at the same time.

Filter approaches are grouped into 5 categories [Legrand and Nicoloyannis, 2005]:

I. Complete methods test all possible attributes subsets. Their computational cost is very high: MDLM [Sheinvald et al., 1990].

II. Heuristic methods have many representatives like Relief, an iterative attribute weight-based algorithm inspired by instance-based learning algorithms, [Kira and Rendell, 1992]. These methods require several accesses to databases.

III. Random methods main representative is LVF, [Liu and Setiono, 1996]. Because of their probabilistic property, the number of selected attributes tends towards the half of the initial attributes number. Like previous methods, these methods require several accesses to databases.

IV. Fast sequential selection method principle is an iterative attribute selection with a single access to databases. In order to have a single data scan, fast correlation measures must be used such as Kendall rank correlation coefficient. This kind of method is represented by MIFS [Battiti, 1994], or the method explained by Lallich and Rakotomalala [Lallich and
Rakotomalala, 2000]. These methods are the fastest and quite efficient.

V. Step-by-step methods use short-sighted criteria to select attributes. This type of method is effective and very rapid particularly for problems with many attributes and objects.

Each approach is characterized by a search procedure to generate the next candidate subset [Langley, 1994] and an evaluation criterion to evaluate the subset under consideration.

There are 4 categories of criteria which measure various attribute specifications:

I. Information measures: these measures determine the information gain: Shannon entropy [Shannon, 1948], gain ratio [Quinlan, 1986]

II. Distance measures: they evaluate the separability of classes: Gini coefficient [Breiman et al., 1984], Mantaras distance measure [De Mantaras, 1991]

III. Dependence measures are the whole correlation or association measures: Tschuprow coefficient [Hart, 1984]

IV. Consistency measures: These measures detect redundant attributes of Zhou [Zhou and Dillon, 1991].
FOCUS [Almuallim and Dietterich, 1991] searches the attribute space by looking at each attribute in isolation, then turn to pairs of attributes, triples, and so on, and stops until it finds the minimal combination of attributes. The minimal attribute subset divides the training data into pure classes.

The RELIEF [Kira and Rendell, 1992] algorithm evaluates the individual attribute according to the evaluation criterion, and thereafter, the best \( n \) attributes are selected. Although both FOCUS and RELIEF use the decision tree induction algorithm after attribute selection, they are naturally not confined to decision tree algorithms, i.e. other induction algorithms can be used instead.

*The proposed study works on a similar principle as FOCUS and RELIEF. The approach uses a combination of the various methods to arrive at a conclusion. The proposed approach builds on the work given by Chen et al. [Chen et al., 2008] in its approach of a combination of multiple information gain and clustering methods for processing. Entropy based and Principle Component Analysis based methods used in this work is similar to the approach proposed by Liu et al. [Liu et al., 2002] and Souza et al. [Souza et al., 2005]. The component analysis method is built in parallel with the entropy based method and the pair wise correlation and clustering approaches are built with variable adaptation methods on top.*
From the above reviewed literature on attribute and hybrid attribute selection methods, it is evident that, there is a need for developing a new hybrid method for attribute selection; which can incorporate attributes obtained from different methods into a single subset by bring out the attributes that are left in one method but are needed for better result on another method. This is only possible when aggregation of attributes from different subsets is performed. Thus the proposed research work brings out the possibilities in aggregating the attributes to obtain better classification accuracy.

2.2 Information gain

The review on information gain approaches has been carried out to give an insight about the various levels and its role on attribute selection.

Information gain (also called cross entropy or mutual information) is a widely used information theoretic measure for the stochastic dependency of discrete random variables [Soofi, 2000]. It is used in learning Bayesian nets [Chow & Liu, 1968; Heckerman, 1998], induce classification trees [Quinlan, 1993], select attributes for classification problems [Duda et al., 2001], i.e. to select a subset of variables. Normally the approaches use a single measure like Shannon's entropy and Pearson's measure for the work.
Given the entropy as a criterion of impurity in a training set $S$, one can define a measure reflecting additional information about $Y$ provided by $X$ that represents the amount by which the entropy of $Y$ decreases. This measure is known as Information Gain or mutual information, and is given by $IG = H(Y) - H(Y | X) = H(X) - H(X | Y)$.

Information Gain is a symmetrical measure. The information gained about $Y$ after observing $X$ is equal to the information gained about $X$ after observing $Y$. A well-known weakness of Information Gain criterion is that it is biased in favor of attributes with more values even when they are not more informative.

Gain Ratio is the non-symmetrical measure

$$GR = \frac{H(Y) + H(X) - H(X, Y)}{H(X)} = \frac{IG}{H(X)}$$

is introduced to compensate for the bias of Information Gain. Here, $H(X, Y)$ is the joint entropy of $X$ and $Y$, which is calculated from the joint probabilities of all combinations of values of $X$ and $Y$. When the variable $Y$ has to be predicted, normalize Information Gain by dividing the entropy of $X$ and vice-versa. Due to this normalization, Gain Ratio values fall always in the range $[0, 1]$. A value of Gain Ratio $= 0$ means that there is no relation between $Y$ and $X$, and Gain Ratio $= 1$ indicates that the knowledge of $X$ completely predicts $Y$. In opposite to Information Gain, the Gain Ratio favors variables with fewer values.
Legrand and Nicoloyannis [Legrand and Nicoloyannis, 2005] in his work use Shannon entropy to measure information gain. Shannon entropy [Hall, 1999], is like the peer category of attribute importance but averages across classes, and the mutual information between the attribute and the class. Jiang [Jiang, 2008] uses Shannon equation for selecting attributes in Ultra-Wideband radar signal analysis.

Chakrabartty and Cauwenberghs [Chakrabartty and Cauwenberghs, 2007] explore the links between maximum entropy based learning techniques and large margin classifiers with extensions to quadratic based impurity functions. This work also used Shannon entropy as a measure for information gain. Within this framework the work introduces Gini Support Vector Machine (GiniSVM), a large margin classifier based on quadratic entropy formulation combined with kernel based quadratic distance.

Baram et al. [Baram et al., 2004] explain a simple maximum entropy criterion that provides effective estimates in realistic settings. The performance of the proposed master algorithm using an ensemble containing two of the best known active-learning algorithms as well as a new algorithm is empirically shown to consistently perform almost as well as and sometimes outperform the best algorithm in the ensemble on a range of classification problems.
MacKay [MacKay, 1992] attempts to measure the information that can
be gained about the unknown target hypothesis using a new labeled
point. It is assumed that given a training set, a probability distribution
over the possible hypotheses is defined. The first measure is how
much the entropy of the hypothesis distribution has decreased using
a new labeled point. A decrease in this entropy can indicate how much
the support of this hypothesis distribution shrinks. The second
measure is the cross-entropy between the hypothesis distribution
before and after the new labeled point is added. An increase in this
cross-entropy can indicate how much the support has changed due to
the new information. It is proved that these two gain measures are
equivalent in the sense that their expectations are equal.

Guyon et al. [Guyon et al., 1996] give methods to measure the
information gain of labeled points in a dataset. The information gain of
a labeled point with respect to a trained probabilistic (soft) classifier is
defined as the probability that the classification of this point is correct.
This probability equals the Shannon's information gain of the class
probability of that point as measured by the probabilistic classifier.
On a given labeled dataset, the information gain of each point can be
measured using the above definition with a leave-one-out estimator.

From the above reviewed literature on information gain methods, it is
emphasized that using information gain is appropriate for attributes
selection.
2.3 Clustering methods

The review on clustering methods has been carried out to give an outline about the role of clustering methods such as distance measure and separability measure for attribute selection.

Borgelt and Nurnberger [Borgelt and Nurnberger, 2004], explain methods to initialize or bias different clustering methods using prior information about the ‘importance’ of a keyword with respect to the specific clusters. This gives a rich set of clustering techniques to improve the clustering performance if prior knowledge is available.

*The proposed approach assumes that prior knowledge is not available.*

Ienco and Meo [Ienco and Meo, 2008] have proposed a hierarchical clustering algorithm based on distance measures. Distance metrics used in the proposed work are similar to this approach. *In the proposed approach clustering is incorporated as one aspect in overall method of attribute selection.*

Ratanamahatana and Gunopulos [Ratanamahatana and Gunopulos, 2003] present an algorithm that uses C4.5 decision trees to select attributes. The purpose is to improve Naive Bayesian learning method. The algorithm uses 10% of a training set to build an initial decision tree. Hruschka et al [Hruschka et al, 2003] introduce a Bayesian approach for attributes selection, where a clustering genetic algorithm is used to find the optimal number of classification rules.
With distance-based clustering, $n$ clusters are constructed by computing a locally optimal solution to minimise the sum of the distances within the data clusters. This is either done by starting from scratch and constructing a new solution or by using a valid cluster solution as a starting point for improvements. A common distance-based algorithm is the K-Means algorithm, which minimises the sum of the distance between each data point and its nearest cluster centre [Selim and Ismail, 1984].

Parallelism in the distance-based clustering methods can be exploited in the outer level, by trying different cluster numbers concurrently, and also in the inner level by computing the distance metrics in parallel.

Hierarchical clustering groups data with a given similarity measurement into a sequence of nested partitions. Two different approaches can be employed. One is to start with each data point as a single cluster and then in each step, merge pairs of points together. This is known as the agglomerative approach. The alternative is to start with all data points in one cluster and then divide one cluster into two clusters in each step. This is the divisive approach. For both methods, $O(N^2)$ algorithms are used. Recent attempts have been made to develop parallel algorithms for hierarchical clustering using several distance metrics in parallel [Olson, 1995].
With the density-based clustering approach, clustering is done by postulating a hidden density model indicating the cluster membership. The data is assumed to be generated from a mixture model with hidden cluster identifiers. The clustering algorithm is then used of finding parameters for individual cluster which maximize the likelihood of the dataset given in the mixture model.

Turner and Ghosh [Turner and Ghosh, 1996] elucidate that the most effective method of achieving classifier independence is by training the members of an ensemble on qualitatively different attribute subsets. In other words, attribute partitioning methods are capable of performing superior to data partitioning methods.

The task of learning attribute subset selection can be obtained by deciding which attributes to use in describing the concept and deciding how to combine those attributes. Thus, the selection of relevant attributes, and the elimination of irrelevant ones, is one of the central problems in machine learning, and many induction algorithms incorporate some approach to address this issue. Recently, it has been noted that the attribute selection stage and classification stage are not independent. Hence, it is recommended to couple attribute selection with effective classification techniques.

From the above reviewed literature on clustering methods, it is clearly shown that clustering methods can also be used in attribute selection.
The proposed attribute selection method builds on the method used by She et al. [She et al., 2005] where a two-stage method is described using information gain and gini index. The present work uses clustering and information gain calculation approach. Also, the work of Legrand and Nicoloyannis [Legrand and Nicoloyannis, 2005] is used as a model for this proposed work. The algorithm uses a three-stage approach aggregating the information from different methods and three metrics and validating the algorithm accordingly. This study uses an approach in which the clustering approach supplemented with information gain based method and the correlation based method incorporates a learning algorithm for the processing.

2.4 MISSING VALUE IMPUTATION

Imputation methods involve replacing missing values with estimated ones based on some information available in the dataset. There are many options varying from naive methods like mean or mode imputation to some more robust methods based on relationships among attributes. This section surveys some widely used imputation methods, although other forms of imputation are available.

2.4.1 MEAN IMPUTATION

This is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean of all known values of that attribute in the class where the instance with
missing attribute belongs. Consider that the value $x_{ij}$ of the $k$-th class, $C_k$, is missing then it will be replaced by

$$x_{ij} = \frac{1}{n_k} \sum_{i: x_i \in C_k} x_{ij}$$

where $n_k$ represents the number of non-missing values in the $j$-th attribute of the $k$-th class. In some studies the overall mean is used, but this does not take into account the sample size of the class where the instance with the missing values belongs to. According to Little and Rubin [Little and Rubin, 2002], the drawbacks of mean imputation are:

a) sample size is overestimated
b) variance is underestimated
c) correlation is negatively biased and
d) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean.

Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. Surprisingly, mean imputation has given good experimental results in datasets used for supervised classification purposes [Chan and Dunn, 1972; Mundfrom and Whitcomb, 1998].

2.4.2 Median Imputation

The mean is affected by the presence of outliers; thus it is natural to use the median instead just to assure robustness. In this case the
missing data for a given attribute is replaced by the median of all known values of that attribute in the class where the instance with the missing attribute belongs. This method is also a recommended choice when the distribution of the values of a given attribute is skewed. Consider that the value \( x_{ij} \) of the \( k \)-th class, \( C_k \), is missing then it will be replaced by

\[
\hat{x}_{ij} = \text{median}_{i': x_{i'j} \in C_k}\{x_{i'j}\}
\]

In case of a missing value in a categorical attribute, mode imputation can be used instead of either mean or median imputation. These imputation methods are applied separately in each attribute containing missing values. It is to be noted that the correlation structure of the data is not being considered in the above methods. The existence of others attributes with similar information (high correlation) or similar predicting power can make the missing data imputation useless or even harmful.

**2.4.3 K-Nearest Neighbor (KNN) Imputation**

In KNN imputation, the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function. The algorithm is as follows:

A. Divide the dataset \( D \) into two parts. Let \( D_m \) be the set containing the instances in which at least one of the attribute is missing.
The remaining instances will complete attribute information form a set called $D_c$.

B. For each vector $x$ in $D_m$:

a) Divide the instance vector into observed and missing parts as $x = [x_o; x_m]$.

b) Calculate the distance between $x_o$ and all the instance vectors from the set $D_c$. Use only those attributes in the instance vectors from the complete set $D_c$, which are observed in the vector $x$.

c) Use $K$ closest instances vectors (K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes, replace the missing value using the mean value of the attribute in the $k$-nearest neighborhood. The median could be used instead of the mean.

The advantages of KNN imputation are:

(i) K-nearest neighbor can predict both qualitative attributes (the most frequent value among the $k$-nearest neighbors) and quantitative attributes (the mean among the $k$ nearest neighbors).

(ii) It does not require creating a predictive model for each attribute with missing data. Actually, the $k$-nearest neighbor algorithm does not create explicit models.

(iii) It can easily treat instances with multiple missing values.
(iv) It takes into consideration the correlation structure of the data.

The disadvantages of KNN imputation are:

(i) The choice of the distance function. It could be Euclidean, Manhattan, Mahalanobis, Pearson, etc.

(ii) The KNN algorithm searches through all the dataset looking for the most similar instances. This is a very time consuming process and it can be very critical in data mining where large databases are analyzed.

(iii) The choice of $k$, the number of neighbors. A similar work is done in [Troyanskaya et.al, 2001], where several numbers are tried and decided to use $k=10$ based on the accuracy of the classifier after the imputation process. The choice of a small $k$ produces deterioration in the performance of the classifier after imputation due to overemphasis of a few dominant instances in the estimation process of the missing values. On the other hand, a neighborhood of large size would include instances that are significantly different from the instance containing missing values hurting their estimation process and therefore the classifier's performance declines. For small datasets, $k$ smaller than 10 can be used.
2.4.4 Imputation using Decision Trees Algorithms

The decision tree building algorithm deals with the problem of missing values. Witten and Frank [Witten and Frank, 2005] outline a solution that involves notationally splitting the instances into pieces, using a numeric weighting method, and sending part of it down each branch. Eventually, the various parts of the instances will reach the leaf node, and the decisions at these leaf nodes must be applied to partial instances. Instead having integer counts, the weights are used. The same weight procedure is used to partition the training set once a splitting attribute has been chosen, to allow recursive application of the decision tree formulation procedure on each daughter nodes. Instances for which the relevant value is missing are notationally splitting the instances into pieces, using a numeric weighting method, and sending part of it down the various branches. Pieces of the instance contribute to decisions at lower nodes in the usual way through the information gain calculation; they may be further split at lower nodes, if the values of other attributes are unknown as well. The expected error reduction, which is called as SDR for Standard Deviation Reduction is calculated by

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} x sd(T_i)$$

Where $T_1, T_2, \ldots$ are the sets that result from splitting the node according to the chosen attributes.
SDR = \frac{m}{|T|} \times \left[ \text{sd}(T) - \sum_{T \in \{L, R\}} \frac{|T|}{|T|} \times \text{sd}(T) \right]

Where \( m \) is the number of instances without missing values for that attribute, and \( T \) is the set of instances that reach this node.

To take an account of missing values, a modification is made to the SDR formula. The formula including the missing value compensation is \( T_L \) and \( T_R \) are sets that result from splitting on this attribute because all tests on attributes are now binary. A threshold is determined for splitting in the usual way, by sorting the instances according to its value and, for each possible split, calculating the SDR according to the preceding formula, choosing the split point that yields the greatest reduction error. Only the instances for which the value of splitting attribute is unknown are used to determine the split point.

Then these instances are divided into two sets \( L \) and \( R \) according to the test then, determine whether the instances in \( L \) or \( R \) have the greater average class value, and calculate the average of these two averages. Then an instance for which this attribute is unknown is placed into \( L \) or \( R \) checking whether its class value exceeds this overall average or not. If it does, it goes into whichever of \( L \) and \( R \) has greater average class value. When the splitting stops, all the missing values will be replaced by the average value of the corresponding attributes of the training instances on reaching the leaves.
It is evident from the above review that missing value should be imputed with some plausible values so as to get better accuracy.

Thus, the proposed work is a hybrid of clustering, information gain, correlation and learning algorithms. While each approach by itself has generated appreciable results, it is perhaps time to use a logical hybrid especially for large datasets with more number of attributes. The proposed work also incorporates suitable missing value imputations to get better accuracy.