Chapter 1

Introduction

Data Mining is the process of nontrivial extraction of implicit, previously unknown, and potentially useful information such as knowledge rules, constraints, and regularities from data [Shapiro and Frawley, 2000; Saravanan and Vivekanandan, 2004]. The data may be stored in repositories such as databases, data marts or data warehouses.

Data mining uses techniques (Figure 1A) which are pioneered by fields like artificial intelligence (pattern recognition, neural networks, and fuzzy logic), distributed computing, statistical analysis (rough sets, support vector machines) and database technologies. The idea of data mining is to build computer programs or algorithms that examine through databases automatically, seeking regularities or patterns. Strong patterns, if found, will make accurate predictions on future data.

Figure 1A: Influence on Data mining
Data mining may also be visualized as an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

Data mining is also defined as the process of analyzing a huge amount of data intended to find useful information for decision making [Sumathi and Sivanandam, 2006; Maimon et al, 2005]. Thus, Data mining can help organizations to have useful insights into their businesses from the data collected over the years and take better decisions to achieve new heights. The key aspects of the Data mining are

a) huge amounts of data,

b) process of extraction of the data and

c) the ability of the process to perceive and make explicit, hidden relationships or patterns.

Data mining has gained importance as the scale of data has increased due to the advent of the data processing, computing technology and internet. Data mining can thus forecast future trends and activities to support people [Ni, 2008].
1.1 Steps in Knowledge Discovery in Databases

The various steps involved in knowledge discovery process [Han and Kambar, 2001; Berry and Linoff, 2004] are shown below.

A. Data Sampling: The aim is to select a representative subset of data from a large population of data. For this, the data sources needed for the task on hand are identified from all the available different sources.

The implication of this step is the fact that, unlike many other predictive tasks, data mining activities usually have a specific context. A task for identifying patterns among the population sample for credit card suitability is different from the task of finding patterns for extending loans. Even though the activities may be similar and use the same datasets; even small differences in the context can have a very big impact on the outcome.

B. Data Selection: In this step only those data components (attributes) in different data sources which are needed for the data mining are selected. The attributes can be selected from different data sources and merged together to form a coherent source.

For example, the dataset for credit card eligibility may use different attributes like age, gender, salary & region and may
use the attribute credit card defaulter percentage of the different regions from another data structure.

**C. Data Cleaning:** The data components that are collected may contain errors, missing values, noisy or inconsistent data. The errors may be due to various factors (design and accidental). Attributes need to be pruned for several reasons: The data may be noisy; i.e. contain spurious signals unrelated to the target class; they may be uninformative; e.g. contain mostly one value, or no repeating values; they may be correlated to other variables. The data may be inconsistent; e.g. contain discrepancies in the department code used to categorize items [Yan et.al, 2003].

**D. Data Transformation:** The resultant data even after cleaning needs to be prepared and transformed into forms appropriate for the mining task. Techniques like smoothing, aggregation, normalization etc., may be used for this purpose.

**E. Feature extraction:** This stage pulls out specified and processed data that is significant in the given context. The feature extraction stage identifies attributes that are significant for the current context based on a combination of various methods. The need is for an accurate outcome that correctly
identifies the most significant attributes and removes the redundant and useless attributes from the dataset.

Feature selection and attribute selection are considered to be equivalent terms. The word attribute is used consistently throughout this work to denote feature selection.

A relevant attribute is defined as; the attribute on removal deteriorates the performance or accuracy of the classifier. An irrelevant or redundant attribute is not relevant. Because irrelevant information is cached inside the totality of the attributes, these irrelevant attributes could deteriorate the performance of a classifier that uses all attributes [Dash and Liu, 2003].

**F. Data Mining:** It is an important step in the knowledge discovery process. This step interacts with the user or a knowledge base. Various data mining techniques such as

- Characterization
- Association
- Classification
- Cluster analysis and
- Evaluation and deviation analysis may be applied to discover the interesting patterns.
The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base.

Data mining is classified into two types based on the output obtained from the techniques used, they are

I. Directed

II. Undirected

Directed data mining attempts to explain or categorize some particular target field such as income or response. Undirected data mining attempts to find patterns or similarities among groups of records without the use of a particular target field or a collection of predefined classes.

G. Pattern Evaluation and Knowledge Presentation: To identify the truly interesting patterns representing knowledge based on some interestingness measures because a data mining system has the potential to generate thousands or even millions of patterns or rules. Only a small fraction of the patterns potentially generated from the data mining system may be of interest to any user. A pattern is interesting if:

- it is easily understood by Humans
- valid on new or test data with some degree of certainty
- potentially useful
- novel
A pattern is also interesting, if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge. This step involves visualization, transformation, removing redundant patterns etc., from the patterns that are generated.

**H. Decisions / Use of Discovered Knowledge:** This step helps to make use of the knowledge acquired to take better decisions.

The first five steps in the data mining process are crucial and critical and form a part of the data preprocessing process. This research work concentrates on ‘attribute selection’ mentioned in step E (*Feature Extraction*).

### 1.2 Data Preprocessing

Data preprocessing is very important for any successful data mining process [Liu and Motoda, 1998] as it transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Nearly 80% of mining efforts is spent on data quality [Famili et al., 1997]. There are a number of data processing techniques available [Han and Kamber, 2001]. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations, such as normalization may be applied to improve the accuracy and efficiency
of mining algorithms. Data reduction can reduce the data size by aggregating, eliminating redundant attributes. Figure 1B clearly shows the techniques in data preprocessing and the importance of attribute selection for data mining process.

A dataset can have millions of records and hundreds of columns or attributes. The data cleaning process will be complex with this large amount of data. Attribute selection is very important to reduce time
and effort for further work such as record similarity and elimination process. Attribute selection is applied to reduce the number of attributes in many applications where data has hundreds or thousands of attributes. Attribute selection picks a subset of variables from the initial set of available attributes without using any transformation [Yu and Liu, 2004].

The goal of attribute selection (Figures 1C, 1D) is to identify those attributes relevant to achieve a predefined task, to improve the efficiency of learning algorithms, and to reduce the size of data storage [Xu et al., 2008].

It is difficult to determine the relevant attribute subset before the mining procedure. At practical data mining situations, data miners often face a problem to choose the best attribute subset for a given dataset.
If it contains irrelevant or redundant attributes, it is not possible to get any satisfactory results from mining/machine learning scheme. Irrelevant attributes not only lead to lower performance of the results and time, but also preclude finding potentially existing useful knowledge.

Besides, redundant attributes not only affect the performance of classification task, but influence the readability of the mining result as well as the time constraint. To choose a relevant attribute subset, trial-and-error testing, expertise for the given attribute set, or/and heavy domain knowledge for the given dataset [Abe and Yamaguchi, 2006] are needed.
Attribute extraction in knowledge and data engineering is the process of identifying and removing as much of the irrelevant and redundant information as possible so as to reduce the dataset for the mining process to be effective [Qu et al., 2005]. Attribute selection involves uni-variate or multi-variate evaluation of attributes with respect to the classification accuracy.

Attribute selection is a term for space reduction method which attempts to select the more discriminative attributes from datasets in order to improve classification quality and reduce computational complexity [Wei et al., 2008].

There are three fundamentally different approaches for attribute selection: wrapper, filter and hybrid.

Wrapper approach uses the learning algorithm to test all existing attribute subsets. The filter approach corresponds to a data pre-processing step preceding the learning phase. The wrapper algorithms use the actual classifier to select the optimal attribute subset, while the filter attributes independently select the optimal attribute subset of the classifier [Haindl et al., 2006]. The fundamental difference between the two families is that the wrapper algorithms are related to the learning algorithm whereas the filter phase is completely independent of it.
The disadvantage of filter approach is that the attributes could be correlated among themselves [Ding and Peng, 2003]. On the other hand, wrapper methods tend to find attributes better suited to the pre-determined learning algorithm resulting in better performance. But, it also tends to be computationally more expensive since the classifier must be trained for each candidate subset [Agarwal and Bala, 2007]. It is also possible to combine the filter and wrapper methods to obtain hybrid approaches [Lee et al., 2006]. In recent literature, the term embedded methods [Lal et al., 2006] has been introduced. The inference system has its own Attribute Selection Algorithm (either explicit or implicit) in embedded methods.

In general, the main needs of an attribute selection strategy are: an evaluation function (Figure 1E) to identify and analyze attributes, a search strategy to find the best attribute subset, a stopping criterion to decide when to stop and a validation procedure to check whether the subset is valid [Piramuthu, 2004].

Each newly generated subset needs to be evaluated by an evaluation criterion. The goodness of a subset is always determined by a certain criterion (i.e., an optimal subset selected using one criterion may not be optimal according to another criterion).
Evaluation criteria can be broadly categorized into two groups based on their dependency on mining algorithms that will be finally applied on the selected attribute subset. Three groups of evaluation criteria are discussed below.

I. **Independent criteria**

Typically, an independent criterion is used in algorithms of the filter model. It tries to evaluate the goodness of an attribute or attribute subset by exploiting the intrinsic characteristics of the training data without involving any mining algorithm. Some popular independent criteria are distance measures, information measures, dependency measures, and consistency measures [Almuallim and Dietterich, 1994; Ben-Bassat, 1982; Hall, 2000; Liu and Motoda, 1998].

- **Distance measures** are also known as separability, divergence, or discrimination measures. For a two-class problem, an attribute $X$ is preferred to another attribute $Y$ if $X$ induces a greater difference between the two-class conditional
probabilities than $Y$, because one tries to find the attribute that

can separate the two classes as far as possible. $X$ and $Y$ are

indistinguishable if the difference is zero.

- **Information measures** typically determine the information

gain from an attribute. The information gain from an attribute $X$

is defined as the difference between the prior uncertainty and

expected posterior uncertainty using $X$. Attribute $X$ is preferred

to attribute $Y$ if the information gain from $X$ is greater than that

from $Y$.

- **Dependency measures** are also known as correlation

measures or similarity measures. They measure the ability to

predict the value of one variable from the value of another. In

attribute selection for classification, one looks for how strongly

an attribute is associated with the class. An attribute $X$ is

preferred to another attribute $Y$ if the association between

attribute $X$ and class $C$, the value from separability measure is

higher than the association between $Y$ and $C$. In attribute

selection for clustering, the association between two random

attributes measures the similarity between the two.

- **Consistency measures** are characteristically different from

the above measures because of their heavy reliance on the class

information and the use of the Min-Attributes bias [Almuallim

and Dietterich, 1994] in selecting a subset of attributes. These

measures attempt to find a minimum number of attributes that

separate classes as consistently as the full set of attributes can.
An inconsistency is defined as two instances having the same attribute value but different class labels.

II. Dependent criteria

A dependent criterion used in the wrapper model requires a pre-determined mining algorithm in attribute selection and uses the performance of the mining algorithm applied on the selected subset to determine which attributes are selected. It usually gives superior performance as it finds attributes better suited to the pre-determined mining algorithm, but it also tends to be more computationally expensive, and may not be suitable for other mining algorithms [Blum and Langley, 1997]. For example, in a task of classification, predictive accuracy is widely used as the primary measure. It can be used as a dependent criterion for attribute selection. As attributes are selected by the classifier that uses selected attributes in predicting the class labels of unseen instances. In this case, accuracy is normally high; it is computationally costly to estimate accuracy for every attribute subset [John et al., 1994].

In a task of clustering, the wrapper model of attribute selection tries to evaluate the goodness of an attribute subset by the quality of the clusters resulted from applying the clustering algorithm on the selected subset. There exist a number of heuristic criteria for estimating the quality of clustering results, such as cluster compactness, scatter separability, and maximum likelihood. Recently
much work has been done on developing dependent criteria in attribute selection for clustering [Dash and Liu, 2000; Dy and Brodley, 2000; Kim et.al, 2000].

III. Stopping criteria

A stopping criterion determines when the attribute selection process ends. Some frequently used stopping criteria are:

a) the search completes

b) some given bound is reached, where a bound can be a specified number (minimum number of attributes or maximum number of iterations)

c) subsequent addition (or deletion) of any attribute does not produce a better subset

d) a sufficiently good subset is selected (e.g., a subset may be sufficiently good if its classification error rate is less than the allowable error rate for a given task)

Thus, the data preprocessing stages enables data mining algorithms to be applied easily, improves the effectiveness and the performance of the mining algorithms. It represents the data in understandable format for both human and machines and supports faster data retrieval from databases. It also makes the data suitable for a specific analysis to be performed.
Although attribute subset-selection and individual heuristics have been used to solve specific problems, limited research has been done using attribute subset-selection and hybrid algorithms.

Pair wise attribute selection is recently proposed as a trade-off between selection process complexity and the need to analyze relationships between attributes [Michalak and Kwasnicka, 2006].

1.3 Missing Values

One relevant problem in data quality is the presence of missing data. Many Data mining algorithms handle missing data in rather naive way [Yan et.al, 2003]. Missing data treatment should be carefully done; otherwise bias might be introduced into the knowledge induced. In most cases, dataset attributes are not independent from each other. Thus, through the identification of relationships among attributes, missing values can be determined. Imputation is a term that denotes a procedure that replaces the missing dataset by some plausible values.

Missing values commonly occur in raw datasets and are problematic to model generation in different fields of study. Development of methods to impute missing values could increase the usefulness of these valuable datasets.
Instances containing missing data, though expensive to collect, often go unused. Many computer models are incapable of managing nonexistent values during the calculation of large matrices typical in many scientific applications. Often, the only way to render such valuable and expensive rows usable is to simply replace any missing points with an average or mean value derived from all other rows in the same column.

**Types of Missing Data**

Missing data can be classified into three types (Laird and Rubin, 1987). They are as follows:

I. **Missing Completely At Random (MCAR).** This is the highest level of randomness. It occurs when the probability of an instance having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method can be applied without risk of introducing bias on the data.

II. **Missing At Random (MAR).** When the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of missing data itself.

III. **Not Missing At Random (NMAR).** When the probability of an instance having a missing value for an attribute may depend on the value of that attribute.
Treatment for Missing Values

Treatment method for missing data can be divided into three categories (Laird and Rubin, 1987). The categories are:

I. **Ignoring and discarding data.** There are two main ways to discard data with missing values. The first one is known as *complete case analysis*; it is available in all statistical programs and is the default method in many programs. This method consists of discarding all instances with missing data. The second method is known as *discarding instances and/or attributes*. This method consists of determining the extent of missing data on each instance and attribute, and deleting the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with high degree of missing values. Both, complete case analysis and discarding instances and/or attributes, should be applied only if missing data are MCAR, because missing data that are not MCAR have non-random elements that can bias the result.

II. **Parameter estimation.** Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm can handle parameter estimation in the presence of missing data.
III. **Imputation.** Imputation is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the dataset assist in estimating the missing values. Imputation of missing values can be done by many naive methods like mean, median and some robust methods using C4.5 and K-Means algorithm which are based on relationships among attributes.

1.4 **Background of the Problem**

Knowledge discovery in database (KDD), data mining and database mining are terms used to express the growing field of interest about selecting proper information from database. This field of interest gains more regards due to the increasing availability of collecting and accessing large amount of information. As data repositories grow, there is an increasing need for data mining tools, which are able to glean information from datasets that are not easily understood by traditional observation or experiment. Researchers and practitioners realize that, in order to use the tools effectively for any mining process; an important part to be considered is pre-processing in which data is processed before it is presented to any learning, discovering, visualizing algorithm or data mining.

The term Data Mining generally refers to a process by which accurate and previously unknown information can be extracted from large
volumes of data in a form that can be understood, acted upon, and used for improving decision processes [Apte, 1997]. Data mining is the means used in extracting hidden knowledge from a dataset; this would be knowledge that is not readily obtained by traditional means such as queries or statistical analysis [Roiger and Geatz, 2003].

Data mining process can be broken down into four distinct phases.

a) Decision, whether or not to carry out data mining on a given dataset.

b) Data preparation, preparing the data for analysis.

c) Model building, where the work of building a prediction model is carried out.

d) Interpretation, which is largely carried out by individuals but can be greatly assisted using automated means, such as graphical representation of the results.

Attribute transformation and attribute selection are applied frequently in data pre-processing for real-world applications. Attribute transformation is a process through which a new set of attributes is created.

Assuming the original set $A$ of attributes consists of $a_1, a_2, ..., a_n$, some variants of attribute transformation can be defined below.

- Attribute transformation process can augment the space of attributes by inferring or creating additional attributes. After
attribute construction, it is possible to get additional $m$ attributes $a_{n+1}, a_{n+2}, \ldots, a_{n+m}$. For example, a new attribute $a_k$ ($n < k \leq n + m$) could be constructed by performing a logical operation of $a_i$ and $a_j$ from the original set of attributes.

- Attribute construction process can also extract a set of new attributes from the original attributes through some functional mapping. After attribute extraction, it is possible to get $b_1, b_2, \ldots, b_m$ ($m < n$), $b_i = f_i(a_1, a_2, \ldots, a_n)$, and $f_i$ is a mapping function. For instance for real valued attributes $a_1$ and $a_2$, for every object $x$, it is possible to define $b_1(x) = c_1 a_1(x) + c_2 a_2(x)$ where $c_1$ and $c_2$ are constants.

- Attribute selection is different from attribute transformation, by which no new attributes will be generated, but only a subset of original attributes is selected and the attribute space is reduced.

Traditionally attribute selection has been done in the data preparation stage to reduce the set of data to be considered in the data mining process. Attribute selection methods can work on labeled and unlabeled data [Liu and Yu, 2005].

The given input training set has $n$ attributes. All these attributes may not contribute to the classification process. Also building a classifier considering all the attributes makes the classification technique computationally expensive. Therefore it is required to do the attribute
subset selection and identify those attributes that are significant for the classification process. The attribute subset selection is done by studying the input training set and various techniques have been proposed in the literature. This work proposes a multidimensional hybrid approach to identify the attribute subset.

A goal of attribute selection is to avoid selecting too many or too few attributes than is necessary. If too few attributes are selected, there is a good chance that the information content in this set of attributes is low. On the other hand, if too many (irrelevant) attributes are selected, the effects due to noise present in (most real-world) data may overshadow the information present. Hence, this is a trade-off that must be addressed by any attribute selection method.

Good Attribute Selection methodologies are the need of the hour. The methodologies must model the entire dataset, select attributes not only based on parametric importance, but on pair-wise importance and cluster relationships as well [Yu and Liu, 2004]. The system must be agile in its operation and have learning characteristics so as to be adaptive.

While some of the above characteristics are present piecemeal in many current methodologies, the proposed methodology lays the ground for a unified approach which includes aggregation of information gain, clustering and correlation methods. The proposed unified approach
uses a combination of methods that will account for the different characteristics needed in attribute selection like good information gain, pair-wise correlation, etc. The hybrid method thus captures characteristics that hitherto may have been overlooked by other methods. The missing values in the attributes subset obtained from the hybrid method are imputed using C4.5 and K-Means algorithm so as to improve the classification accuracy.

1.5 BROAD OBJECTIVE

The objective of this research work is to outline an attribute selection algorithm suitable for different quantities of attributes such as nominal attributes, categorical attributes, ordinal attributes and data. The attribute selection needs to be adaptive and have the ability to learn. Also the attribute selection system must use the best of clustering, correlation and ranking systems to form a hybrid system. The algorithm must combine these attributes in a unified and agile manner, after which missing values are imputed.

1.6 SPECIFIC OBJECTIVE

The specific objectives of this work is to propose a hybrid method for attribute selection that

- uses adaptive learning to guide the attribute selection process
- combines information gain, clustering, intelligent search heuristics and
• captures the relationships between attributes using appropriate filter and wrapper methods.

The applicability of the method for a range of datasets is tested before and after imputation of missing values. The results are benchmarked against existing methods.

1.7 Statement of the Problem

The growing abundance of information necessitates the need for appropriate methods for organization and evaluation. Mining data for information and extracting conclusions has been a fertile field of research. However, data mining needs methods to pre-process the data. Attribute selection is an emerging field of interest about selecting proper information from data repositories. The aim of the proposed work is to highlight the need for attribute selection methods in data mining encompassing the best characteristics of the data by aggregating the attributes obtained from different methods. This is carried out so as to select an attribute that seems to be important in one method but not on another method. In recent times there has been interest in developing hybrid attribute selection methods combining the characteristics of various filter and wrapper methods.

i. The proposed method advocates an adaptive aggregation strategy using

a) the gain ratio for candidate attributes,
b) clustering methods to find the distribution of candidate attributes and
c) correlation based strategies which measure the dependencies of different attributes.

ii. The underlying principle of the strategy is that the best individual attributes need not constitute the best sub-set of attributes representing the problem. A given attribute might provide more information when present with certain other attribute(s) when considered by itself.

iii. Unlike many of the current approaches which emphasize one aspect (information gain, clustering and correlation) over the other, the proposed approach advocates a unified approach in a hybrid manner.

iv. The resultant adaptive method has been implemented for the datasets from the UCI Machine Learning Repository and the results are correlated.

v. The missing values on the selected subset of attributes are imputed with C4.5 and K-Means algorithm and the results are correlated for better classification accuracy.

vi. The conclusions validate the algorithm and give out a performance which is better than other comparable methods.
The importance of the making the method adaptive is to emphasize the role of learning process in the system.

1.8 Chapter-wise Summary

The rest of the thesis is organized as follows.

Chapter 2 discusses the comparison of relevant literature to establish the state of art and outline the basis for this work.

Chapter 3 outlines the adaptive hybrid method for attribute selection.

Chapter 4 describes the missing value imputation methods used in this research work and validates the optimal subset of attributes.

Chapter 5 presents the implementation of the system and the relevant results.

Chapter 6 outlines the conclusions of the research work and gives directions for further work.