CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

We are living in the information era. Accumulating data is easy and storing it is cheap. It is estimated that every 20 months or so the amount of information in the world doubles. In the same way, tools for use in the various knowledge fields (acquisition, storage, retrieval, maintenance, etc) must develop to combat this growth. Knowledge Discovery in Databases (KDD) (Fayyad et al 1996, Sever 1998) is the process of extracting models and patterns from large databases. Knowledge is only valuable when it can be used efficiently and effectively; therefore knowledge management is increasingly being recognized as a key element in extracting its value.

The term Data Mining (DM) is often used as a synonym for the KDD process, although strictly speaking it is just a step within KDD.

A data mining method (the extraction of hidden predictive information from large databases) is selected depending on the goals of the knowledge discovery task. Data mining, as a multidisciplinary joint effort from databases, machine learning, and statistics, is turning mass of data into human understandable format (Jiawei Han and Micheline Kamber 2004).

Machine learning provides tools by which large quantities of data can be automatically analyzed. Fundamental to machine learning is feature
selection. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining.

Feature selection (FS) is the process of identifying and removing as much of the irrelevant and redundant information. The optimality of a feature subset is measured by an evaluation criterion. The selection of a small number of highly predictive features is used to avoid over fitting the training data. It reduces the number of features, removes irrelevant, redundant, or noisy data, and thereby speeds up data-mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility.

Feature selection (FS) has been a fertile field of research and development since the 1970s in statistical pattern recognition, machine learning, data mining, and widely applied to many fields such as text categorization (George Forman 2003), image retrieval (Dy et al 2003), intrusion detection (Lee et al 2000), and genomic analysis (Xiong et al 2001, Xing 2003). By identifying the most salient features for Feature selection (FS), focuses a learning algorithm on those aspects of the data most useful for analysis and future prediction. The feature selection process is beneficial to a variety of common machine learning algorithms.

In image recognition systems the FS plays the major role to optimize the classification performance (Jelonek and Stefanowski 1997). If the number of features is increased, the classification rate of the classifier decreases after a peak. In melanoma diagnosis, for instance, the clinical accuracy of dermatologists in identifying malignant melanomas is only between 65% and 85%. With the application of FS algorithms, automated skin tumor recognition systems can produce classification accuracies above 95%.
Gene expression of microarrays is a rapidly maturing technology that provides the opportunity to analyze the expression levels of thousands or tens of thousands of genes in a single experiment. A typical classification task is to distinguish between healthy and cancer patients based on their gene expression profile. Feature selection is used (along with some initial filtering) to drastically reduce the size of these datasets which would otherwise have been unsuitable for further processing (Xing 2003, Xiong et al 2001).

The text categorization views documents as a collection of words. Documents are examined, with their constituent keywords extracted and rated according to criteria such as their frequency of occurrence. As the number of keywords extracted is usually in the order of tens of thousands, dimensionality reduction must be performed. This can take the form of simplistic filtering methods such as word stemming or the use of stop-word lists. However, these do not provide enough reduction for use in automated categorizers, so a further feature selection process must take place. Recent applications of FS in this area include web page, bookmark categorization (Jensen and Shen 2001), and spam filtering (George Forman 2003).

The hypothesis explored in this thesis is that feature selection for supervised classification tasks can be accomplished on the basis of information theoretic based interacting features applied to Weather, Lung Cancer, Iris, Breast Cancer Wisconsin, Coil 2000, Splice, Soybean, Wine, Zoo, and US census form UCI Repository data, Reuters-21578 Corpora data set and Ling Spam data.

1.2 BACKGROUND OF THE RESEARCH

There are often many features in KDD, and combinatorially large numbers of feature combinations, to select from. Note that the number of feature subset combinations with m features from a collection of N total features is given by the binomial coefficient C(N, m) = N! / (m!(N-m)!).
features is \(N!/[m!(N-m)!]\). It might be expected that considering a large number of features would increase the likelihood of including enough information to distinguish between classes. Unfortunately, this is not true if the size of the training dataset does not also increase rapidly with each additional feature included. This is the so-called curse of dimensionality (Bellman 1961). A high-dimensional dataset increases the chance that a data-mining algorithm will find spurious patterns that are not valid in general. Most techniques employ some degree of reduction in order to cope with large amounts of data, so an efficient and effective feature selection method is required.

The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In many real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. A detailed review of feature selection techniques devised for classification tasks can be found in (Dash and Liu 1997).

The usefulness of a feature or feature subset is determined by both its relevancy and redundancy. A feature is said to be relevant if it is predictive of the decision feature (or class label), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other. The overall procedure for any feature selection method is given in Figure 1.1.

The generation procedure implements a search method (Langley and Sage 1994b, Siedlecki and Sklansky 1988) that generates subsets of features for evaluation. It may start with no features, all features, a selected
feature set or some random feature subset. Those methods that start with an initial subset usually select these features heuristically beforehand. Features are added (forward selection) or removed (backward elimination) iteratively in the first two cases (Dash and Liu 1997). In the last case, features are either iteratively added or removed or produced randomly thereafter. An alternative selection strategy is to select instances and examine differences in their features.

The evaluation function calculates the suitability of a feature subset produced by the generation procedure and compares this with the previous best candidate, replacing it if found to be better. A stopping criterion is tested for every iteration to determine whether the FS process should continue or not. For example, such a criterion may be to halt the FS process when a certain number of features have been selected based on the generation process.

A typical stopping criterion centered on the evaluation procedure is to halt the process when an optimal subset is reached. Once the stopping criterion has been satisfied, the loop terminates. For use, the resulting subset of features may be validated. Determining subset optimality is a challenging problem. There is always a tradeoff in non-exhaustive techniques between subset minimality and subset suitability - the task is to decide which of these must suffer in order to benefit the other. For some domains (particularly where it is costly or impractical to monitor many features), it is much more desirable to have a smaller, less accurate feature subset.
1.3 RESEARCH MOTIVATION

Feature selection is one of the important techniques in data preprocessing for data mining. The benefits of feature selection for learning can include a reduction in the amount of data needed to achieve learning, improved predictive accuracy, learned knowledge that is more compact and easily understood and reduced execution time.

There are three kinds of Feature selection methods namely, wrapper, filter and hybrid methods. In hybrid method (Lei Yu and Huan Liu 2004) feature selection is integrated into the process of training for a given learning algorithm. Wrappers (Isabelle Guyon and Elisseeff 2003) on the other hand, evaluate attributes by using accuracy estimates provided by the actual target-learning algorithm. The wrapper method is computationally expensive because they are tightly coupled with specified learning algorithm and they are intractable for large-scale problems.

Filters use general characteristics of the training data to evaluate attributes and operate independently of any learning algorithm. Due to its
computational efficiency, the filter methods are very popular to high-dimensional data.

Any feature selection method involves two major steps: generation procedure and evaluation function (Dash and Liu 1997). The optimality of a feature subset is measured by an evaluation criterion. Finding an optimal feature subset is usually difficult, and many problems related to this have been shown to be NP-hard (Blum and Rivest 1992). The various evaluation functions have been introduced namely distance, Gini index, $\chi^2$-test, and dependency metrics (Huan Liu and Lei Yu 2005, Guangzhi Qu et al 2005, Dash and Liu 2003). It is noticeable that among different evaluation criteria the information metric seems to be more comprehensively studied.

The main reason is that information entropy is a good measurement to quantify the uncertainty of feature (Huawen Liu et al 2009). Various researchers Mark Hall (1999), Huan Liu and Lei Yu (2005), Bell and Wang (2000), Mark Last et al (2001), Fleuret (2004), Traina et al (2000), Batti (1994), Al-Ani et al (2003), Lei Yu and Huan Liu (2004), Guangzhi Qu et al (2005), Huang and Chow (2005), Novovicova and Malik (2005), Vanessa Gomez et al (2007) and Gavin Brown (2009) have applied information theoretic criterion as evaluation function. The values of information metric are invariable throughout the selection procedure. However, this cannot accurately represent the relevant degree between features (ie. feature interaction) when the selection procedure continues to work.

The feature interaction was analyzed by Jakuline (2005) for Monk data set, where single feature can be considered as irrelevant based on its correlation with the class, but when combined with other features can result in loss of useful information and thus may cause poor prediction performance. An intrinsic character of feature interaction is its irreducibility. Zheng and Liu
(2007, 2009) proposed searching for interacting features by employing symmetrical uncertainty (SU) as evaluation function for feature relevance and consistency-contribution is used to measure feature interaction.

But Guangzhi Qu et al (2005) argued that SU is not accurate enough to quantify the dependency among features with respect to class labels. Hence in this thesis information theoretic based heuristic value is proposed to find the feature relevance and consistency-contribution is used to measure feature interaction.

This thesis proposes two algorithms namely Information Theoretic based Interact (IT-IN) algorithm, and Dependency Based - Interact Algorithm (DIA) based on information theoretic based heuristic value which concern relevance, redundancy and consistency of the features. Text Categorization (TC also known as Topic Setting or Text Classification) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set.

The three different phases in the life cycle of a TC system are document indexing, classifier learning, and classifier evaluation. Document indexing is one of the most important issues in TC, which includes document representation and a term weighting scheme. For document representation, bag-of-words (Salton and McGill 1983) is the most common way to represent the context of texts. The advantage of this approach is simplicity as only the frequency of word in a document is recorded. Here, for all the predefined categories, the synonyms and prefix words for the category are found and it helps to assign any document to that category based on the synonym or prefix of a term.

Term Frequency (TF) is the simplest measure to weight each term in a text. The drawback of TF is the difficulty in finding the optimal
thresholds (Tokunga 1994). Also TF is known to improve recall but does not improve precision. While TF concerns term occurrence within a text, Inverse Document Frequency (IDF) concerns term occurrence across a collection of texts. The intuitive meaning of IDF is that terms, which rarely occur in a collection of texts, are valuable which improve precision. Thus, the combination of TF and IDF improves recall and precision respectively that gives better performance. All TC researchers use only the product of TF and IDF. A drawback of IDF is that all texts that contain a certain term are treated equally i.e., IDF does not distinguish between one occurrence of a term in a text and many occurrences (Tokunga 1994). The drawback of TF.IDF is that when a new document occurs, recalculation of weighting factors to all documents is needed since it depends on number of documents. Weighted Inverse Document Frequency (WIDF) (Tokunga 1994) overcomes this by weighting term that sums up to one over the collection of texts. WIDF itself improves both the precision and recall. A drawback of WIDF is that when the number of documents becomes large, the terms that have the nearest frequency, have almost equal weight, which makes the learning task more difficult.

To overcome this problem in text categorization this thesis proposes a new term weighting scheme called Modified Inverse Document Frequency (MIDF). The document represented in MIDF is trained using the SVM classifier with linear, polynomial and Radial Basis Function kernels, and Neural Network. The experiments are carried out in Reuters-21578 corpora.

In Spam filtering a major problem is the high dimensionality of the feature or term space. A text domain has several tens of thousands of features. Most of these features are not relevant and not beneficial for Spam classification task. Even some noise features may sharply reduce the
classification accuracy. Furthermore, a high number of features can slow down the classification process or even make some classifiers inapplicable. Hence different term weighting schemes namely TF.IDF, WIDF and MIDF schemes are applied and different feature selection approaches namely FCBF, SU and Information gain are used to reduce the dimension of the email documents. The FCBF algorithm is applied to select features based on feature relevance and redundancy criteria. The different classifiers namely ELM, SVM, Naive Bayes and J4.8 classifiers are used to evaluate the selected feature subset which is used to improve the classification accuracy of spam filter.

1.4 SEQUENCE OF THE RESEARCH WORK

To consider the feature selection, based on the relevance, redundancy, and consistency of the features, the sequence of the research is carried out in the following manner.

i. New two-phase hybrid classifiers by combining correlation based feature selection and classification approaches are proposed for simple rule generation and reduce the computation time. In phase-I, Fast Correlation Based Feature selection (FCBF), Decision Dependent Correlation-Decision Independent Correlation (DDC-DIC) and Mutual Information Based Evaluation Function (MIEF) are applied to generate feature subset based on relevance and redundancy measures. In phase-II, reduced dataset generated from phase-I is applied to either J4.8 or Ant Colony Optimization (ACO) classifier.

ii. The performances of Hybrid classifiers are analyzed for before and after feature selection approaches.
iii. A new two-stage feature selection - Information theoretic based Interact algorithm (IT-IN) is proposed for feature selection.

iv. The performance of the IT-IN is analyzed with other existing feature selection algorithms with ELM (Extreme Learning Machine), SVM (Support Vector Machine), Naive Bayes and J4.8 classifiers.

v. A new three-stage feature selection - Dependency based Interact algorithm (DIA) is proposed for feature selection.

vi. The performance of the DIA is analyzed with other existing feature selection algorithms with ELM, SVM, Naive Bayes and J4.8 Classifiers.

vii. A new term weighting scheme is proposed for text representation and the performance is compared with existing term weighting schemes.

viii. A different term weighting schemes with different feature selection algorithms are applied for e-mail Spam filtering application.

ix. Four different classifiers namely ELM, SVM, J4.8 and Naive Bayes classifiers are considered for Spam Filtering.

1.5 PROPOSED METHODOLOGIES

Performance analysis of the feature selection for classification is obtained by the following methods in this research work.
i. Relevance and Redundancy based feature selection algorithms are analyzed using Ant miner (ACO) and J4.8 for simple rule generation.

ii. New algorithms using information theoretic based Interact algorithm (IT-IN) and Dependency based Interact algorithms (DIA) are proposed to find relevant and irredundant feature subset.

iii. Hash table is implemented for Consistency contribution, which avoid the repeated scanning of data.

iv. The classifications schemes namely ELM, SVM, Naive Bayes and J4.8 are used to find the classification accuracy of the proposed feature selection algorithms.

v. A new term weighting scheme is introduced for text categorization. BPN and SVM classifiers are used to find precision and recall measures.

vi. The new term weighting schemes with different feature selection algorithms are proposed for Spam mail classification. The classification accuracy of Spam filter is analyzed by ELM, SVM, Naive Bayes and J4.8 classifiers.

1.6 RESEARCH CONTRIBUTION

The following are the recent contributions in the area of Feature selection.

i. New initiation of Information theoretic based Interacting algorithm and Dependency based interact algorithm are used
to find relevant, irredudant and interacting features are the new research contribution in this field.

ii. A new ‘term weighting scheme’ is introduced for text categorization, which is used to improve the precision and recall of text data.

iii. A new architecture for Spam mail filtering is a new research contribution in this area.

- A new ‘term weighing scheme’- Modified Inverse Document Frequency (M IDF) is proposed to improve the classification accuracy. The performance of MIDF is compared with the existing TF.IDF and WIDF term weighting schemes with four different classifiers.

- Fast Correlation Based Feature (FCBF) selection approach is applied to reduce the feature space. The performance of FCBF is compared with existing IG and SU.

- The Extreme Learning Machine (ELM) is used for Spam classification; the performance of ELM is compared with the existing Naive Bayes, J4.8 and SVM Classifiers.

- The performance evaluation on the benchmark Spam filtering corpora Ling Spam is conducted for three term weighing schemes (M IDF, TF.IDF and WIDF) with three feature selection approaches (FCBF, IG
and SU) for four classifiers (ELM, SVM, Naive Bayes and J4.8)

iv. The data set such as, categorical, discretised, text and email after feature selection is classified by ELM classifier is another contribution.

1.7 ASSUMPTIONS AND LIMITATIONS

i. The significant drawback of many feature selection algorithms, in the literature is found that the use of user-supplied information is essential to many existing algorithms for feature selection. Some feature selectors require noise levels to be specified by the user beforehand. Some simple feature selection algorithms leave the user to choose their own subset. They require the user to state how many features are to be chosen, or they must supply a threshold that determines when the algorithm should terminate. All of these require the user to make a decision based on their own (possibly faulty) judgment.

ii. It is often difficult to compute the integral in the continuous space based on a limited number of instances, when mutual information is being estimated. For convenience, only discrete features in experiments are considered. Thus, those continuous features were discretized in to nominal ones by the MDL method or Fayyad and Irani method (1993).
1.8 ORGANISATION OF THE THESIS

Successful feature selection algorithms produce relevant, irredundant data, which improve the classification accuracy and reduce the computation time. So, it is important to generate novel, and new feature selection algorithms. The real challenge in this research work is to connect various fields in a clear and concise manner to obtain a well-designed feature selection for categorical, numerical and email data. This chapter gives an overall idea of the research work and its objectives. An overview and step-by-step procedure of the research work, reported in this thesis, is also presented in this chapter.

Chapter 2 presents the existing related works in the literature. It provides the different works on feature selection for machine learning, supervised machine learning algorithms, information theoretic based feature selection algorithms, Text Categorization and Email spam filtering feature selection approaches.

Chapter 3 presents the new two-phase hybrid classifiers by combining Correlation based feature selection and classification approaches. It discusses the performance of ACO and J4.8 before and after feature selection.

Chapter 4 presents the new proposed algorithms namely, Information theoretic based Interact algorithm (IT-IN) and Dependency based Interact Algorithm (DIA). The performances of these algorithms are compared with existing feature selection algorithms. The classifiers schemes such as ELM, SVM, Naive Bayes and J4.8 are considered to evaluate the performance of the proposed algorithms are discussed.
In Chapter 5, architecture of proposed text categorization scheme is introduced. It proposes a new term weighting scheme, and it compares the performance with the existing term weighting schemes. To evaluate this, SVM and BPN classifiers are used to find the precision and recall of text information. The proposed term weighting schemes are also applied to email Spam filtering approach. The various feature selection algorithms along with the new term weighting schemes with ELM, SVM, Naive Bayes and J4.8 classifiers are discussed.

Chapter 6 provides the discussion and conclusions and the scope for further research work.