Chapter - 1

INTRODUCTION

1.1 Introduction

The world has widely changed in terms of transmitting, acquiring and storing information in the form of unstructured textual data. Managing such vast amount of unstructured data has become a key issue: internet penetration across the world increasingly engage millions of people, thereby increasing pile of data on World Wide Web, making the field text data mining to become the dominant form of retrieving useful information from unstructured data, overtaking conventional traditional data mining techniques. As an academic field of study, text mining might be defined as: Text mining or text data mining, refers to the process of finding interesting or useful patterns, rule, directions, trends, or model from unstructured data, is used to describe the application data mining techniques to automated knowledge discovery from text [HK02]. Text mining has been viewed as natural extension of data mining [Hea03] or sometimes considered as task of adapting the same data mining techniques to domain of large scale textual data [DGS99]. This reflects the fact that the advent of text mining relies on the burgeoning field of data mining to a greater degree.

An important goal of text mining is to extract pattern that can be incorporated into other intelligent applications. Such applications include text categorization, text summarization, text ranking, text retrieval, ranking segmentation, organizing and navigational tool for text. Text categorization is the most explored field for applying data mining techniques because most of the other applications can be cast into task of text categorization. Text Categorization (TC) is a learning task, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents. Text classification task may also be synonymously referred to as text classification, document classification, topic classification, or topic spotting in the literature [Fab02]. The most prevalent technique in research community for solving this problem is rest with machine learning techniques where an automated inductive process incrementally builds a classifier by learning from the pre-classified documents and the characteristics of categories [Mit97].
The discipline of automated Text Categorization is an extremely active area of current research and application. Its applicability is broad, with many potential uses for large businesses as well as individuals. Text categorization is fallen within the paradigm of supervised machine learning techniques. There are many supervised learning algorithms have been developed in past decade such as naïve bayes, decision trees, neural networks, k-Nearest Neighbour and support vector machines, etc. Among these learning algorithms, kNN is the most promising inductive learning algorithm and it has remarkable significance because it has been shown by different researchers that it improves the classification results considerably for text categorization problem.

An empirical evaluation of the text categorization task has been reported [YP97, For03], against quite number of standard benchmark datasets. Most of the empirical evaluations reported on text categorization mention the application of *k*-Nearest neighbour algorithm (*k*-NN) as the most promising classifier for text categorization task. In this thesis, a study has been conducted for improving the effectiveness of *k*-NN based text classifier from the view point of *Dempster-Shafer Theory* [Sha76]. For this purpose, this thesis has proposed an improved evident theoretic *k*-NN for text categorization task and also proposed novel evident theoretic feature selection method for cursing the dimensionality of text categorization. A comprehensive empirical evaluation has also been conducted using state of the art classifiers including the proposed algorithm on benchmark dataset.

1.2 Motivation

In Text Categorization, an important issue faced by the classifier is high dimensionality of the feature space, it has been observed from the literature that performance of the classifier is hurt by redundant and noise present in the feature space [Joa98]. As a result, a pre-processing step known as *dimensionality reduction*, is so warranted to substantially reduce the feature space from \( T \) to \( T' \) (\( T' \gg T \)) a much smaller feature space in terms of number of features. Dimensionality reduction often takes the form of *feature selection*: each term is scored by means of a scoring function that captures its degree of (positive and sometimes also negative) correlation with category \( c_i \), and only the highest scoring terms are used for document representation. With regard to this, a novel evident-theoretic approach to feature selection has been proposed in this thesis to complement the classifier without hurting the performance.
In this thesis, a comprehensive empirical comparison of popular feature selection metrics has been done on standard benchmark datasets which have varying class distributions, number of categories and sizes. While making the comparison, the most widely used feature selection methods such as IG, Chi Square and Odd Ratio have been adapted. Each feature selection method that is examined in this thesis followed a local policy where a unique set of key words are used for each category. In this thesis, an ensemble of feature selection methods has been proposed, which combines various traditional feature selection methods using an evidence theory. The newly proposed ensemble of metric is two-sided metric (i.e. they consider the negative features as well as positive features) and is based on the distributions of a term in both positive class and negative class. The metrics used in the ensemble have shown good performances in prior research using fewer numbers of keywords. This makes the ensemble be precious especially when the practitioner is limited to use a small number of features.

Moreover, a comprehensive empirical evaluation of the text categorization task is also reported [YP97, For03] against quite number of standard benchmark datasets. Most of the empirical evaluations reported on text categorization mention the application of lazy learning method such as k-Nearest neighbours as the most promising technique for text categorization task. As a relatively intuitive algorithm, kNN has a better performance than other methods due to its efficiency and effectiveness [FH51, TP67]. Intuition behind this kNN is very simple and quite straight forward. In essence, kNN makes prediction based on the k training patterns (neighbours) which are closest to the unlabelled (test) pattern.

However, it faces serious challenges when pattern of different classes overlap in some regions in the feature space. In kNN, predicted neighbours as data available to the classifier are imperfect; these neighbours really don’t know exactly to which category the unlabelled test pattern is assigned to. Each neighbour express its opinion about the fact that given instance may belongs to class it holds, this opinion has been represented as a degree of belief by the neighbour and all such belief are then combined using Dempster-Shafer theory of evidence.

Due to this imprecise nature of k neighbours, the classifier faces an uncertainty in classifying the input pattern. Here imprecision is essentially a property of neighbours’ where as uncertainty is the property of classifier. This problem of classifying the pattern on the basis
of its nearest neighbour from the point of view of Dempster–Shafer theory [Sha76] is addressed in this thesis, in which a set of neighbour’s of a pattern to be classified is considered as an item of evidence to support the class membership of pattern. Based on the evidence, basic beliefs are assigned to the subset of training patterns. Such basic belief assignments are obtained for each subset and aggregated using the variant of Dempster-Shafer model called *Transferable belief Model* [SK94]. A comprehensive kNN based classifier framework has been proposed, which provides a platform to compare the performance of various lazy leaning classifier such as majority voting kNN, distance weighted kNN and the proposed improved evident theoretic kNN.

1.3 Need for the Study

Text categorization is a learning problem where the goal is to assign the most relevant categories to unlabeled documents by using a pre-labelled training set of documents. Therefore, the most state-of-the-art machine learning algorithms such as support vector machine (SVM), neural networks, Naive Bayes and k-nearest neighbour algorithms have been constantly used for text categorization research. In the literature, there have been many reported studies in which these learning algorithms are analysed and compared [YP97, Fab02]. Most of these prior researches have shown that kNN is the most sought classifier after SVM in text classification. Therefore, kNN has been adapted as learning algorithm in this study.

Moreover, regardless of the learning algorithm, text classification itself is a very hard problem because of many inherent properties of unstructured textual data such as heterogeneous, noise, redundancy and high dimensionality. Among these attributes, the high dimensionality of the data is really a bottleneck to the classifier. To handle such high dimensionality, feature selection is a core issue in most of the classification problems including domains other than text. Researchers adapted a wrapper method such as genetic search when the dimensionality of the feature space is low. However, the usage of wrapper is not practically feasible if the dimensionality is very large as in text categorization, which leads to situations where feature scoring methods are used instead of wrapper methods. There are many traditional approaches are in practice for feature selection in text categorization. Most of these traditional and popular feature selection metrics have been experimentally evaluated and comprehensive comparison has been done in these studies [YP97, Das01,
However the experimental settings used in these studies have many variations in terms of dataset selection, policy, classifier, etc. For instance, in the study of Yang and Pedersen [YP97] feature selection metrics are evaluated on the Reuters and Ohsumed datasets using the popular feature selection methods. In this study, they adapted kNN and LLSF as major classifiers and they ignored the SVM, which is an important deficiency of their study. In addition to that, they conducted study which merely dealt with feature selection on global policy and no comparison was made with regard to local policy.

A study conducted by Forman [For03], where comprehensive analysis and evaluation of many feature selection metric was done using benchmark datasets. It adapted a local policy instead of global. His study used SVM as base line classifier and includes different types of datasets such as skew datasets, homogenous and heterogeneous. Another significant contribution of this study is the proposal of novel feature selection metric called “Bi-normal Separation” which has been reported most successful feature selection in this study.

An interesting study which compares the existing feature selection is the work carried out by Debole and Sebastiani [DS05]. In this study, the focus is on term weighting where feature selection scores are used to weigh the terms and compare the results of different feature selection metrics with local and global policy. They observed that global policy outperforms local policy. However, they do not reveal a set of key words used in the detailed comparison. Moreover, they relied on Reuters dataset and their finding cannot be generalised to other datasets with varying class sizes and skewness.

However, it is not possible to say anything about the performances of local and global policy using the popular feature selection metrics such as IG, CHI, etc. In addition to the studies which aimed at the comparison of the existing feature selection metrics, there are also some studies which propose novel method of feature selection for text categorization. For instance, Forman [For03] proposes a novel method called Bi-normal Separation which is reportedly performs better than other metric in high-skew datasets.
1.4 Scope and Objectives

Based on the prior reported research work, inherent complexities and challenges of the problem under consideration, this thesis infers the following scope and objectives towards the advancement of knowledge in text categorization research.

1. It proposes \(k\)NN based framework which consist of collection of lazy learning algorithms which perform classification task on wide variety of datasets. The framework also implements various feature selection and feature weighting which effectively capitalizes the information contained in the high dimensional textual data.
2. It proposes an ensemble of evident theoretic feature selection method which combined the evidence in the traditional feature selection methods using theory of evidence.
3. It proposes an improved evident theoretic \(k\)NN based classifier for text categorization. Instantiating the proposed classifier on the standard benchmark datasets leads to considerable improvements in text categorization performance. More specifically, superiority of the proposed approach has been observed in categorizing documents along with the proposed feature selection metric.
4. It formulates a novel approach of analysing the kind of relationship between feature selection methods and the traditional \(k\)NN algorithms empirically, given different benchmark data collections. The explicit and comprehensive comparison based on the proposed framework led to make a clear choice in choosing the appropriate combination of feature selection and inductive learning classifier.
5. It also describe a way to further enhance the performance of feature selection method as well as the classifier on some more benchmark datasets to justify Dempster-Shafer approach over other approaches.

1.5 Methodology

The main motivation of this thesis is to identify and recognise the various inductive learning algorithms and datasets using different configuration settings in order to analyse the practical discriminative power of classifiers and discover the potential configuration pattern which is informative and beneficial for text classification task. In order to find the configuration with optimal performance, this thesis explores various parameters extensively across learning algorithms and datasets. Moreover, to meet the objectives posed by this
research and to provide empirical evidence to the proposed algorithm, methodology of research has been laid out for the experiments in this thesis including the inductive learning algorithms, the benchmark data collections, and text pre-processing and performance evaluation.

A series of experiments have been conducted to investigate various widely-used traditional and the state-of-the-art feature selection methods on various experimental conditions. The purpose of each experiment is to seek answers to the objectives with more general experimental evidence. To accomplish this, a general platform has been built to compare a variety of traditional feature selection methods and classifiers with the proposed feature selection method and classifier using a benchmark corpus. This series of experiments provide deeper insights and a practical guidance on choosing feature selection methods in terms of different learning algorithms and corpora.

1.6 Organization of the Thesis

Chapter 1 introduce the concept of Text Categorization, its potentials and the impact of TC in the present information era, other motivational factors which influenced the research, scope and objectives of this thesis.

Chapter 2 provides a review of techniques for the task of Text Categorization, its peculiar characteristics, taxonomy, applications, and most important, the widely-used approaches to learning text classifiers. This chapter sets the context for the background of this thesis research.

Chapter 3 reviews a mathematical theory of evidence Dempster-Shafer Theory, a powerful tool for representing uncertain knowledge, characteristics, various measures like belief and plausibility and its application to many real world problems. It also reviews the variant of Dempster-Shafer Theory known as Transferable Belief Model. This model is adopted in this thesis as it represents powerful tool for combining measures of evidences.

Chapter 4 explores various pre processing steps in Text Categorization and text representation methods. It mainly focuses on reduction of feature space through Curse of dimensionality and traditional feature weighting methods. It also proposes a novel evident theoretic feature selection method base on TBM for text categorization
Chapter 5 incorporate the theory of evidence to represent evidence in $k$NN based classifier. It also review the existing evidence based $k$NN algorithms and proposes an evident theoretic $k$NN based classifier to improve the effectiveness by handling the various uncertainties associated with the classifier.

Chapter 6 lays out the methodology of research for the experiments in this thesis including the inductive learning algorithms, the benchmark data collections, and text pre-processing and performance evaluation. This chapter provides a detailed description of all the experimental settings in this thesis.

Chapter 7 presents a series of experiments to investigate various widely-used traditional and the state-of-the-art feature selection methods on various experimental conditions. The purpose of this chapter is to seek answers to the three questions with more general experimental evidence. This chapter serves an important role in this thesis since it not only examines the performance of various learning algorithms with more experimental evidence, but also provides us with deeper insights and a practical guidance on choosing feature selection methods in terms of different learning algorithms and corpora.

Chapter 8 summarizes the contributions of this thesis and outlines some possible directions for future research. In this study, the focus has been on the study of text representations and the improvements of inductive learning algorithm using evident theory for Text Categorization.