Chapter - 8

CONCLUSION

The focus of this thesis has been the incorporation of Dempster-Shafer theory to single label text classification and has resulted in several new insights. It has presented the study of novel evident theoretic techniques for improving the effectiveness of Text Categorization and has experimentally demonstrated the validity of the proposed approach. Experiments have been performed with the help of framework explicitly devised for dealing with tasks involved in text categorization.

This thesis discussed the following points related to text classification from the viewpoint of Dempster-Shafer Theory.

- How to combine and utilise the evidence concerning the distribution of positive and negative features from the traditional feature selection methods.
- How to devise a novel evidence theoretic $k$-NN based classifier for improving the effectiveness of text categorization by collecting the evidences from the multiple neighbourhoods.
- How to analyse the kind of relationship between feature selection methods and the various lazy learning algorithms empirically, given different benchmark data collections.

In Chapter 2, a review of text classification from machine learning perspective was described. Text Categorization is believed as the most challenging task of text data mining because of many intrinsic properties of textual data. The recent trend in solving such text classification problem is through supervised machine learning paradigm. Two of the reasons for this fact has been the involvement of the Machine Learning community in TC, which has resulted in the use of the state-of-the-art Machine Learning technology in TC applications, and the availability of standard data corpus, which has encouraged research by providing a setting in which different research efforts can be compared to each other, so that the best methods can be discovered.
In Chapter 3, the concept of uncertainty and a method for handling uncertainty in view of mathematical theory of evidence was described. The Bayesian formulation for handling uncertainty of a given problem implicitly assumes a Boolean phenomenon results in over-commitment i.e. the degree of belief which one has in existence of certain hypothesis has a causal effect on the belief in non-existence of the hypothesis. Thus a small degree of belief in hypothesis automatically leads to large degree of belief to the negation of the hypothesis. Thus lack of knowledge leads to a over-committed formulation of the problem. Dempster-Shafer theory of evidence in contrast to the Bayesian theory keeps as much belief in a hypothesis as implied by evidence, thereby avoiding the over-commitment. As such, under evidence theory formulation of the problem, lack of belief does not mean disbelief, leading to an adequate representation of uncertainty. This ability of representing uncertainty has been attracted researchers to implement it in decision making problems where there is a lack of knowledge such as kNN where a set of neighbour’s of a pattern to be classified are imprecise and mislead the classification.

In Chapter 4, many inherent attributes of natural language text such as high dimensionality, noise, heterogeneous made the task of text categorization quite complicated to the researchers. Moreover, standard benchmark datasets used for text categorization often have an enormous amount of noisy and irrelevant features which hurts the performance of the text classifier invariably. Therefore, it was necessary to reduce the dimension of features used in learning by selecting a moderate number of features which contain much information. However, it was difficult for the classifier to learn with a sufficiently high accuracy only with several hundred features. This chapter presented a study on reducing size of the feature space and improving the learning process thru a process called as curse of dimensionality. It also analysed various traditional feature selection methods for text categorization in order to improve the classifier performance. Based on the analysis, it proposed a novel method of learning evidences from traditional methods for feature selection. The theoretical investigation based on evidence theory indicated that the proposed method could improve the performance of text categorization in situations where the features of negative class play a vital role.
In Chapter 5, the most promising classifier in the arsenal of machine learning the $k$ nearest neighbour algorithm was described. It considered the issue of uncertainty in $k$NN and its consequences in the learning process. The way of representing evidence in $k$NN using evidence theory and various types of uncertainty arises in $k$NN due to instance based learning and the methods to tackle those uncertainties were studied. Based on the conceptual framework of Dempster Shafer Theory and $k$NN rule, an improved evident theoretic $k$NN algorithm adopting a simple and efficient neighbourhood selection strategy has been proposed. To classify the pattern, the proposed algorithm considered several neighbourhoods, each of which is a set of neighbours. The neighbours are taken as a single source of evidence supporting the propositions concerning the class membership of the pattern. This evidence is represented as single mass function in order to quantify the uncertainty attached to the class membership of that pattern. A small scale pilot study was conducted using proposed approach on a small scale dataset in order to justify the algorithm effectiveness. It observed the superiority of proposed approach over other $k$NN variants. The experimental results encouraged to do more experiments on real time datasets under carefully controlled settings.

In Chapter 6, the various state-of-the-art inductive learning algorithms adapted in this thesis was described. Moreover, different benchmark datasets and evaluation criteria using different configuration settings was analysed for comprehensive experimental research. In order to analyze the discriminative ability of classifiers and feature selection methods, a novel computational framework was constructed and described the practical advantages of the implementation. The proposed framework is informative and beneficial for text classification task from various perspectives.

In Chapter 7, the various aspects of evaluating feature selection methods for Text categorization were investigated. The generic framework has also been customized to perform experimental comparisons between proposed $k$NN algorithm and other variants, focusing on their capability of improving the effectiveness of text categorization on contrasting the imbalance between positive and negative examples. Experiments performed on Reuters, WebKB and 20 News Groups datasets clearly showed that the superiority of improved evident theoretic $k$NN based classifier is verified and also able to counteract skewness.
The effect of class skewness and vocabulary size were analysed with respect to classifier performance on traditional feature selection methods. The performance of the classifier was gradually increased by the proposed feature selection method, which determines the portion of potential features that are appropriate by learning evidence from traditional feature selection methods. Existing and proposed feature selection methods were tested on kNN as well as on a SVM light algorithm. The experimental results on proposed approach and traditional methods are that 1) the optimal set of features selected by proposed approach had actually combined the features produced by other metrics, and 2) the average performance on select datasets was best when number of features gradually increases.

In addition, a comprehensive empirical evaluation of text classifier has been done on standard datasets using the proposed framework. It was evident from the evaluation that the proposed improved evident theoretic kNN classifier consistently outperformed other kNN variants, which indicates that the superiority of proposed evident theoretic approach for kNN based text categorization tasks. What could one conclude from those results are the following: If one has a sufficient amount of features and it is modeled as a binary-class text classification problem, it is better to use the proposed evident theoretic method for feature selection. If one has only a small amount of features, conventional information gain is effective. Furthermore, in the case of large class skewness in the dataset, evident theoretic feature selection approach is useful for improving the classification performance.

Based on results obtained from the experiments, it was observed that the superiority of newly proposed an improved evident theoretic kNN over other kNN variants irrespective of vocabulary size. The experimental results on these two data sets were consistent with best results that were obtained by previous studies. It was also observed from previous study that smaller vocabulary is sufficient for high performance, but many real world classification tasks do not adhere to this phenomenon due to the fact that a category consists of diverse subject matters with overlapping vocabularies. Such classification tasks manifest the use of larger vocabularies for adequate classifier performance. Obtained results were consistent with this remark in that the best performance is often achieved with large vocabulary size.
Although all of the text categorization tasks introduced in this thesis are flat and single label classifications, it is quite interesting and challenging to work with nested categories which form a hierarchical or network such as the link structure of the World Wide Web. Furthermore, although I assumed only static structure of categories in this thesis, there are many cases of overlapping among categories and updating texts in the categories. I intend to aim my research toward machine learning for the hierarchical structure of texts and online learning for dynamic structures.

As for future directions: (i) from a research perspective, the hierarchical and multi label text categorization could be investigated instead of single label classification; (ii) from an evident theoretic perspective, some other means to aggregate the evidence in the neighbourhood could be adopted and experimented; and (iii) from an experimental perspective, improved evident theoretic $k$NN could be tested on other datasets, such as TREC and MeSH.