ABSTRACT

This work deals with the problem of improving the effectiveness of $k$NN based Text Categorization (TC) task from the viewpoint of Dempster-Shafer Theory. It incorporates theory of evidence into binary text categorization problem in two stages. In first stage, this work adapts the theory of evidence on feature selection to achieve reduction on dimensionality of term space (DR). With regard to this, it proposes an ensemble of feature selection method which combines various traditional feature selection methods such as information gain, odd ratio, and chi square using an evidence theory. This newly proposed ensemble of metric is two-sided metric (i.e. it considers 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 property makes the proposed ensemble be valuable when the practitioner is limited to use a small number of features.

In second stage, this work proposes an improved evident theoretic $k$NN based classifier for text categorization task. As a relatively intuitive algorithm, $k$NN has a better performance than other methods due to its efficiency and effectiveness. Intuition behind this $k$NN is very simple and quite straight forward. In essence, $k$NN makes prediction based on the $k$ training patterns (neighbors) which are closest to the unlabelled query pattern. However, it faces serious challenges when pattern of different classes overlap in some regions in the feature space. In $k$NN, predicted neighbors as data available to the classifier are imperfect; these neighbors really don’t know exactly to which category the unlabelled test pattern is assigned to. Each neighbor 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 neighbor and all such belief are then combined using Dempster-Shafer theory of evidence.

Due to this imprecise nature of $k$ neighbors, the classifier faces an uncertainty in classifying the input pattern. Here imprecision is essentially a property of neighbors’ where as uncertainty is the property of classifier. A set of neighbor’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*.

Moreover, It also propose a comprehensive $k$NN based classifier framework which provides a platform to compare the performance of various lazy leaning classifier such as majority voting $k$NN, distance weighted $k$NN and our proposed improved evident theoretic $k$NN. A comparative study has been conducted in order to evaluate the proposed classifier relative to the state-of-art classifier like Support Vector Machine (SVM). The experiments have been conducted using three benchmark datasets. The experiments concerning feature selection have been conducted using information, gain, odd ratio, chi square and proposed method. The results showed that proposed feature selection method outperforms other traditional methods in terms of micro-averaged $F_1$ measure, in low class-skew datasets, while information gain is superior over others for moderately high class-skew datasets. Furthermore, the results indicated that proposed feature selection improves the performance over all categories when feature set size gradually increases and efficiently balances the bias of some datasets. Additionally, the proposed combination process shows a significant improvement in performance over all categories across datasets.

With respect to the proposed evident theoretic $k$NN based classifier, this work conducted carefully controlled study using majority voting $k$NN, distance weighted $k$NN, and support vector machine (SVM). Experimental research observed the superiority of improved evident theoretic $k$NN over other $k$NN variants. The experimental results also revealed that the improved evident theoretic $k$NN is quit competent with support vector machine as well in term of improving the effectiveness of text categorization task.