Chapter - 7

EVALUATION OF
EVIDENT THORETIC $k$-NN ALGORITHM

7.1 Introduction

The primary motivation of this chapter is to recognize and experiment with several feature selection methods and inductive learning algorithms using different configuration settings in order to analyze their practical discriminative power and discover the possible most successful configuration pattern which is informative and beneficial for text categorization. The configuration pattern with the optimal performance is extensively analyzed and improved in the following sections.

This chapter also provides a comprehensive experimental research carried out by this thesis. This chapter addresses the objectives posed in Section 1.4. To meet this objective, in this chapter will look into various feature selection methods developed for text categorization task. Then it investigates and choose the best feature selection techniques for $k$NN based text classifications. It also evaluates newly proposed evident theoretic feature selection method and analyse the relationship with the other feature selection methods. Another reason of this chapter is to show the performance in terms of improving the effectiveness of the proposed improved evident theoretic $k$NN based text classifier on wide variety of benchmark datasets.

This chapter is organized as follows. Section 7.2 investigate into various feature selection methods and choose the best feature selection method for $k$NN based classifier. Section 7.3 experiments with the proposed improved evident theoretic $k$NN classifier and verifying the improvement on its effectiveness over other $k$NN variants.

7.2. Exploring the Best Feature selection method for $k$NN based Text Categorization

The prime motive behind this experiment is answering the question “Which is the best feature selection method for kNN based Text Classification”. This experiment also makes a
A comprehensive comparative study on various feature selection methods using kNN based text classifier, since no such comparative study is reported in the literature so far.

In order to do this experiment, this work used a computational framework which it discussed in Section 6.5. The proposed framework implements the all $k$ nearest neighbours under study such as majority voting $k$NN (denoted as $vkNN$), distance weighted $k$NN (denoted as $wkNN$) and newly proposed improved evident theoretic $k$NN (denoted as $ikNN$). In addition to that, the framework also includes the implementation for various feature selection methods such as information Gain, Chi-Square, odd-ratio and also implemented evidence theoretic feature selection method as proposed in Section 4.6. The framework has built-in facility for pre-processing task such as stop word removal, word stemming and document conversion. This frame work is basis for comprehensive comparison of various feature selection methods on wide variety of corpora such as Reuters, WebKB and 20 News Groups using state-of-the-art classifiers. In this experiment all textual documents are represented by the commonly used bag-of-word approach while the remaining background conditions such as data pre-processing and evaluation measure are same except the feature selection technique and the classifiers.

### 7.2.1 Feature Selection Methods

Though there has been many methods proposed in the literature for effectively learning features from the corpora, this work intentionally chooses some sophisticated feature selection methods which have been successful in terms of improving the classifier effectiveness and efficiency. The table 7.1 list the feature selection methods used in experimental settings along with their formulae. All of them are traditional feature selection methods which have been effectively used in text categorization task for long time.

<table>
<thead>
<tr>
<th>Function</th>
<th>Denoted by</th>
<th>Mathematical Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Gain</td>
<td>$IG(t_k, c_i)$</td>
<td>$P(t_k, c_i). \log \frac{P(t_k, c_i)}{P(t_k).P(c_i)} + \frac{P(t_k, c_i). \log \frac{P(t_k, c_i)}{P(t_k).P(c_i)}}{P(t_k).P(c_i).P(c_i)}$</td>
</tr>
<tr>
<td>Chi Square</td>
<td>$X^2(t_k, c_i)$</td>
<td>$g\left(\frac{P(t_k, c_i)P(\bar{c}_k, c_i) - P(t_k, \bar{c}_i)P(\bar{c}_k, c_i)}{P(t_k).P(c_i).P(c_i)}\right)^2$</td>
</tr>
<tr>
<td>Odd Ratio</td>
<td>$OR(t_k, c_i)$</td>
<td>$\frac{P(t_k/c_i).(1 - P(t_k/c_i))}{(1 - P(t_k/c_i)).P(t_k/c_i)}$</td>
</tr>
</tbody>
</table>
The prime purpose of feature selection method is to reduce the term space substantially without hurting the classifier performance. To meet this objective, the above mentioned feature selection methods are chosen because of their success in carefully controlled prior experimental research. Each of the above mentioned feature selection methods have its own rationale, but superiority is measured on the basis of effectiveness it brings about during experimental research.

Information gain is centered on information theory. It actually measures the decrease in the value of entropy when a particular feature is present or absent. [YP97] reported initial application of information gain for text categorization task and had showed the best performance for multi label text categorization task.

In statistic, one can test the independence between two random events using chi-square statistics. In dimensionality reduction, researchers adapted this measure to see independent between category and feature. It has been widely used and has shown effective performance in many text categorization tasks. In the above said mathematical formulae, low value indicates most independence between category and feature, but in order to select a term which has less independence with the category, it select the terms with highest chi-square value. Odd ratio captures the odd of a particular term appearing in the positive class normalised by negative class.

7.2.2 Results and Discussion

To evaluate the effectiveness of newly proposed evident theoretic feature selection method, this work conducted comprehensive experiments on Reuters, WebKB and 20 News Groups corpus using the above mentioned feature selection methods in combination with various kNN based algorithms. In this experiment, it used Reuters 21578 ModApte split, but due to the fact that distribution of these documents is very highly skewed and many of the categories in this collection are left with no training and test documents, it consider only the documents with a single topic and classes which still has at least one training and test examples. The resultant collection is known as R8 by the convention adapted by [Fab02]

In WebKB dataset, there is no standard train/test split reported in the literature and in order to be consistent with previous research reports, two third of the documents were randomly chosen for training and remaining third for testing.
The purpose of each experiment is to select a list of terms with the corresponding scores. The list of terms are simply sorted based on the scores and obtain the top $k$ relevant terms with the highest scores. To evaluate the goodness of each experiment, the $k$ relevant terms are selected and compared to the standard labelled set using the precision, recall and $F_1$ measures. The experiments have been conducted with a wide range of $k$ for each implementation. The range of $k$ has been set from 50 to 1000 features.

In general, the experimental results have shown that the proposed evidence theoretic feature selection method denoted as COM is better than the other traditional feature selection methods such as IG, CHI and OR when precision is strongly preferred over recall. It can be explained that IG while focusing only on positive features, may miss some important negative features. The success of Evident theoretic feature method lies in combining evidence by selecting both positive and negative features considerably from these existing methods. However, for scalability reasons, one is limited to 50-100 features; the best metric that outperformed others is IG.

The classification results for SVM and $k$NN algorithm using proposed methods along with others on Reuters 21578, WebKB and 20 News Groups datasets are also presented. The study also reveals that if the precision is central goal, proposed method beat other methods by a smaller but significant margin.

Figure 7.1 and Figure 7.2 depicts the micro averaged $F_1$ performance of $k$NN and SVM classifiers on Reuter’s corpus respectively. The performance of various feature selection methods at varying number of vocabulary size is observed. In all observations the trends are distinctive that the $F_1$ measure of different feature selection method increases as the number of feature increases. It is quite evident that all feature selection method yield maximum performance at the highest vocabulary size. As for as the proposed method (COM), the best $F_1$ performance of 85.89% is observed at the vocabulary size of 900 by majority voting $k$NN. Similarly, the best $F_1$ performance of 84.14% is observed at the vocabulary size of 900 by Support Vector Machine.

The proposed method has repeatedly shown significant performance than others when the feature set size is larger than 500. Figure 7.3 and Figure 7.4 shows the micro averaged $F_1$ performance of $k$NN and SVM on the WebKB dataset respectively. The performance curve on this dataset is not monotonically increases as opposed to Reuters dataset. The best $F_1$
A measure of 78.08% is obtained at a vocabulary size of 500 by majority voting kNN. Similarly, the best $F_1$ measure of 88.88 is obtained by the proposed method at a vocabulary size of 1000 by Support Vector Machine.

Based on the carefully controlled experiments in these datasets revealed that proposed evident theoretic feature selection method (COM) outperformed other traditional methods for kNN based text categorization tasks. Experimental results strongly recommend that the proposed approach is the most viable for kNN based text categorization task.

![Figure 7.1 Comparison of COM with others in terms of Micro $F_1$ on Reuters for voting k-NN](image)

![Figure 7.2 Comparison of COM with others in terms of Micro $F_1$ on Reuters for SVM](image)
Figure 7.3 Comparison of COM with others in terms of Micro $F_1$ on WebKB for voting $k$-NN

Figure 7.4 Comparison of COM with others in terms of Micro $F_1$ on WebKB for SVM
7.2.3 Further Analysis

Impact of class skewness and vocabulary size is another important criterion to measure the classifier effectiveness. In classification task with high class skew, it could be difficult to obtain good \( F_1 \) measure as induction algorithm often focuses on improving accuracy [For03]. It is observed from this experiment that in low-skew situations, proposed evident theoretic feature selection is the only metric that performed best overall, but one is limited to few number of features, information gain (\( IG \)) is a much better choice. On the contrary, under high class skew, proposed evident theoretic feature selection method performed best by a wide margin as number of feature increases from vocabulary of size 200. Based on the controlled experiments, it is concluded that under low skew, \( IG \) performs best and eventually reaches the best performance as vocabulary size increases, but under high class skew, on the other hand proposed evident theoretic feature selection method performed substantially better than \( IG \).

To achieve reasonable performance gain in terms of \( F_1 \), the size of vocabulary plays an important role. Since the vocabulary size used in two datasets are quite different; varying level of \( F_1 \) measure on these two datasets are observed. For instance, the categories in the Reuters dataset consist of diverse subject matters which involve overlapping of vocabularies. In such situations, in order to improve the classification performance, large vocabulary size is always considered. Hence Reuters requires large vocabulary size. This is not the case in WebKB.

Moreover, the advantage of proposed evident theoretic feature selection method in terms of performance is significant when the number of selected features gradually increases. This suggests that the proposed method might be particularly useful in a situation where the existence of large number of negative examples in the datasets. Overall, the investigation showed that the proposed approach is effective over other traditional feature selection methods and potentially promising for high dimensional datasets.

7.3 Investigating an Evident Theoretic \( k \)NN and its Improvement over \( k \)NN Variants

An evaluation of text categorization performance has always been centred on two main issues; they are computational efficiency and categorization effectiveness. The computational efficiency of the learning algorithms is actually not a concern of researchers
because of the currently available processing ability and memory speed. Therefore, the present work deals with the effectiveness of categorization only since it is believed as more reliable one when it comes to experimentally comparing different learning algorithms.

In single-label text categorization task, researchers measure effectiveness in terms of accuracy, i.e. the percentage of correct classification decisions reported by the learning algorithm. However, in binary as well as multi-label text categorization, accuracy is not generally believed as a good measure. The underlying intuition is that two categories $c_i$ and $\bar{c_i}$ in binary text categorization applications are usually imbalanced, i.e. one contains far more members than the other. In this case, the classifier trivially assigns all documents to the most heavily populated category (i.e. $c_i$), thus the classifier has indeed very high accuracy.

Therefore, this work measures the empirical effectiveness of single-label text categorization task in terms of the classical information retrieval parameters which are discussed in Section 6.4.

7.3.1 Methodology

The objective of this experiment is to seek empirical evidence in favour of newly proposed improved evident theoretic kNN method. Another reason for this experiment is to analyse the kind of relationship between feature selection methods and traditional kNN methods under a more general experimental circumstances. To accomplish this task, it chooses number of widely used feature selection methods and newly proposed evident theoretic feature selection method and investigate the relationship with different learning algorithms such as voting kNN, distance weighted kNN, evident theoretic kNN and SVM on three benchmark datasets.

With regard to the neighbourhood adaption strategy, the proposed improved evident theoretic kNN method (denoted as $i\!k$NN), considered multiple neighbourhoods and each neighbourhood provides a source of evidence supporting the proposition concerning class membership of the query pattern $q$. To compute the neighbourhood for the query pattern $q$, Minkowski family of distance metrics has been adapted as discussed in Section 5.4.

From the observation of the previous experimental setup, it has been learnt that the proposed evident theoretic feature selection method improves the classifier effectiveness substantially. So it is obvious to seek similar result if some other learning algorithms like
distance weighted kNN and Support Vector Machine are included on more general experimental conditions. In order to meet this experiment as comprehensive experimental study, it simply chooses the same set of feature selection methods as used in the previous experimental settings.

The purpose of this experiment is to seek the some sort of improvement over traditional kNN method. It intentionally chooses majority voting kNN (denoted as $v_k$NN) and distance weighted kNN (denoted as $w_k$NN) for comparative study. To compare the observed results with the state-of-art classifiers, it also chooses the SVM as baseline classifiers. The reason behind this selection is SVM has been reportedly proven as effective classifier for TC task.

**7.3.2 Result and Discussions**

The question which has been posed by this thesis is the effectiveness of the proposed approach with respect to classification performance. In order to make a fair comparison among the classifiers, the same controlled settings has been adopted across all experiments. First, each classifier is trained with a chosen data set of documents represented using the standard bag-of-word approach according to the popular feature selection method. Then, the best features which are filtered by the threshold are selected. To make a fair comparison, the thresholds of the various feature selection methods have been set to the same in all experiments. As for as the feature selection is concerned, all feature selection methods including the proposed one as described in the Section 4.6 have been adopted.

The purpose of the experimental research is to assess the effectiveness of the proposed evident theoretic approach in presence of class skewness i.e. different percentage of positive examples versus negative examples on chosen datasets. To assess the effectiveness of the approach with respect to the class skewness, micro-averaging $F_1$ has been adapted.

Various experiments are carried out to verify the validity of the evidence learnt from the heterogeneous distance metric sources as follows. All Experiments have been conducted on four feature selection methods in combination with the proposed improved evident theoretic kNN, denoted as $i_k$NN and other $k$NN variants and SVM on three benchmark corpora. The conducted experiments are as follows.
Experiment 1: This experiment tested the validity of single distance metric by using the majority voting $k$NN ($vk$NN) classification algorithm. It is observed from the experiments that the traditional $k$NN algorithm doesn’t seem to have satisfactory performance in terms of micro-averaged results and hence it is very necessary to improve the effectiveness of the text classifier.

Experiment 2: This experiment is similar to the previous one except that instead of voting $k$NN, it considered distance weighted $k$NN ($wk$NN) algorithm. The results indicated that the distance weighted $k$NN has slight improvement over majority voting $k$NN. This sort of improvement correlates with previous research work reported on this datasets. In both experiments, it used a Euclidian distance metric as a measure of similarity estimation.

Experiment 3: The last experiment was done using an improved evident theoretic algorithm. In order to learn evidence, it considered Minkowski family ($L_1, L_2$ and $L_\infty$) of distance metrics as source of evidence. All experimental results are shown in Figure 7.5, Figure 7.6, Figure 7.7 and Figure 7.8. Each figure presents the performance of the classifier with respect to micro averaged $F_1$ measure on benchmark dataset. Each curve on the figure shows the performance of each classifier as feature set size increases. From the experimental results, it is crystal clear that the micro-averaged $F_1$ measure of $ik$NN has been improved greatly when introducing an ensemble of distance metrics for neighbourhood calculation for each query pattern to be classified, so this experiment has justified that the validity of combining distance metrics on the basis of the effectiveness the classifier brings out. Each figure shows the relationship between $F_1$ values against number of features chosen from the corresponding datasets. It shows clearly that the values of $F_1$ measure of $ik$NN in all three experiments are higher than $vk$NN and $wk$NN. Observed results justified that the improved evidence theoretic $k$NN algorithm is much better choice than the traditional $k$NN algorithms.

Figure 7.5 shows micro-averaged $F_1$ measure results on the Reuters data set using chi-Square feature selection method. The improved evident theoretic $k$NN classifier reaches a maximum of 94.10% at a vocabulary size of 850 words. Note that although an improved evident theoretic $k$NN performs best with a large vocabulary and keeps a rapid increase in performance as the vocabulary size grows, whereas the majority voting $k$NN and distance weighted $k$NN are more even across vocabulary size. Result by evident theoretic $k$NN do an average of 6.2 points higher than other two variants.
Figure 7.5. Performance of $i_k$NN over other variants on Reuters dataset using ChiSquare feature selection.

Figure 7.6 shows micro-averaged $F_1$ measure results on the Reuters data set using information-gain feature selection method. The majority voting $k$NN reaches a maximum of 86.76% with only 150 words, whereas an improved evident theoretic beats the majority voting $k$NN and distance weighted $k$NN at this small vocabulary by showing best performance. Moreover the improved evidence theoretic $k$NN achieves a highest measure of 91.38% at 450 words.

Figure 7.7 and Figure 7.8 shows micro-averaged $F_1$ measure on WebKB data set using chi-square and information-gain feature selection respectively. The same kind of pattern as seen in the Reuters is found on this data set. Here, an improved evidence theoretic $k$NN achieves slightly higher $F_1$ measure both on smaller and larger vocabularies, but on average its best performance is 8.6 points higher than other $k$NN variants.

Figure 7.9 and 7.10 shows micro-averaged $F_1$ measure on 20 News Groups dataset using chi-square and information-gain feature selection respectively. The observed result once again predicted the improvement over other $k$NN variants. The improvement has been again based on micro-averaged $F_1$ measure, for different values of imbalance between positive and negative examples and for different set of features. Unlike other datasets, this is
larger and highly imbalanced. The average performance improvement is slightly higher than voting kNN and distance weighted kNN.

![Figure 7.6. Performance of \(i\)kNN over other variants on Reuters dataset using information gain feature selection.](image1)

![Figure 7.7. Performance of \(i\)kNN over other variants on WebKB dataset using ChiSquare feature selection.](image2)

With regard to the comparison of baseline classifier (Support Vector Machine), precision, recall and \(F_1\) have been calculated for the classifiers and the selected features of varying numbers. Results show that, as for precision is concerned, the improved evident
theoretic $k$NN ($ik$NN) based on evidence theory has reported better results than the one obtained with SVM. On the contrary, as expected, results on recall are not satisfactory than SVM model when number of selected feature are minimal.

Figure 7.8. Performance of $ik$NN over other variants on WebKB dataset using information gain feature selection.

7.3.3 Concluding Remarks

Based on results obtained from the experiments, the superiority of newly proposed an improved evident theoretic $k$NN has been observed over other $k$NN variants irrespective of vocabulary size. The experimental results on these data sets are consistent with best results that were obtained by previous studies. It is 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. The obtained results are consistent with this remark in that the best performance is often achieved with large vocabulary size.
Figure 7.9. Performance of $i_k$NN over other variants on 20 News Groups datasets using chi square feature selection.

Figure 7.10. Performance of $i_k$NN over other variants on 20 News Groups datasets using information gain feature selection.
7.5. Summary

This Chapter investigated the various aspects of evaluating feature selection methods for Text categorization. The effect of class skewness and vocabulary size are analysed with respect to classifier performance on traditional feature selection methods. The performance of the classifier was gradually increased by using the proposed feature selection method, which determines the portion of potential features that are appropriate by learning evidence from other traditional feature selection methods. Feature selection methods are tested on \( k \text{NN} \) as well as on a SVM algorithm. The experimental results on the proposed approach and traditional methods are that 1) the optimal set of features selected by the 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 is evident from the evaluation that the proposed improved evident theoretic \( k \text{NN} \) classifier consistently outperformed other \( k \text{NN} \) variants, which indicates that the superiority of the proposed approach for text categorization Tasks. Moreover, this chapter has analysed the performance of SVM in text categorization tasks. The observed experimental results clearly show that SVM text categorization handles large-scale vectors well without feature scoring for any terms.