Improving K-Nearest Neighbor Efficacy for FarsiText Classification

Mohammad Hossein Elahimanesh[1], **BehrouzMinaei-Bidgoli**[2], **Hossein Malekinezhad**[3]

[1]Islamic Azad University, Qazvin Branch, Qazvin, Iran, Computer Research Center of Islamic Sciences, Qom, Iran

[2]Iran University of Science and Technology, Tehran, Iran, Computer Research Center of Islamic Sciences, Qom, Iran

[3]Islamic Azad University, Naragh Branch, Naragh, Iran, Computer Research Center of Islamic Sciences, Qom, Iran

E-mail: {elahimanesh, bminaei, hmalekinejad}@noornet.net

Abstract

One of the common processes in the field of text mining is text classification.Because of the complex nature of Farsi language, words with separate parts and combined verbs, the most of text classification systems are not applicable to Farsi texts.K-Nearest Neighbors (KNN) is one of the most popular used methods for text classification and presents good performance in experiments on different datasets. A method to improve the classification performance of KNN is proposed in this paper. Effects of removing or maintaining stop words, applying N-Grams with different lengths are also studied.For this study, a portion of a standard Farsi corpus called Hamshahri1 and articles of some archived newspapers are used. As the results indicate, classification efficiency improves by applying this approach especially when eight-grams indexing method and removing stop words are applied. Using N-grams with lengths more than 3 characters, presented very encouraging results for Farsi text classification.The Results of classification using our method are compared with the results obtained by mentioned related works.

Keywords: Text classification, N-grams of characters, K-nearest neighbor

1. Introduction

Classification is one of the central issues in information systems dealing with text data. Text classification is a supervised learning task of assigning natural language text documents to one or more predefined categories or classes according to their contents. With the rapid growth of electronic text documents on the Internet and corporate intranets, as a potential tool for better finding, filtering, and managing these resources, text categorization has gained more and more attention in recent years. While it is a classical problem in the field of information retrieval for a half century, it has recently attracted an increasing amount of attention due to the ever expanding amount of text documents available in digital form. While many researchers apply various machine learning algorithms like Naive Bayes, Nearest Neighbor, Neural Networks, Rule Induction and Support Vector Machines in promoting the effectiveness of text classifiers, few people systematically compare and statistically analyze the impact of different text representations on the generalization accuracy of text categorization systems. As a simple definition, text classification detects the class of a new text based on a sufficient history of tagged and classified texts. There are a variety of methods used in different applications for text classification. We can divide the classification approaches into Contextual and non-contextual approaches. The contextual are language-dependent. But the no-contextual can divided into statistical, like Support Vector Machine (SVM) and Naïve Bayes Classifier, and non-statistical ones that based on examples induction like Decision Tree (DT) and K Nearest Neighbors (KNN). Statistical approaches are widely used in text mining applications.Context based methods essentially make use of contextual information to improve the Classification performance. Latent Semantic Analysis and Lexical Units are widely used for extracting contextual information.

Because of the complex nature of Farsi language, words with separate parts and combined verbs, the most of text classification systems are not applicable to Farsi texts. Previous works deal with to automated Farsi text classification is limited to next few works. Arabsorkhi and Feili developed a Farsi text classifier using of Bayesian model (2006). Basiri et al. presented a comparison between KNN and fuzzy KNN approaches for Farsi text classification based on information gain and document frequency feature selection (2008). Bina et al. developed a Farsi text classifier using n-grams and KNN (2008). Pilevar et al.provided a Farsi text classification system using the Learning Vector Quantization network. In this method, each class is presented by an essence vector called the codebook. These vectors are placed in the feature space in a manner that decision boundaries are approximated by the k-nearest neighbor (KNN) rule (2009). Maghsoodi and Homayounpour have used SVM classifier based on extending the feature vector applying words extracted from a thesaurus. This method has improved classifier performance when training dataset is unbalanced and not comprehensive for some classes (2011).

k-Nearest Neighbor (KNN) is one of the most popular

algorithms for pattern recognition. Many researchers have found that the KNN algorithm accomplishes very good performance in their experiments on different data sets. The traditional KNN text classification algorithm has three limitations: (i) calculation complexity due to the usage of all the training samples for classification, (ii) the performance is solely dependent on the training set, and (iii) there is no weight difference between samples. To overcome third limitation, an improved version of KNN is proposed in this paper.

The rest of this paper is organized as follows. Our approach for text classification is detailed in section 2. Section 3 describes some conducted experiments to demonstrate the suitability of theproposed approach and finally section 4 concludes this paper.

2. Proposed Method

The method that is used for converting documents to numerical vectors and the type of classifier that is applied, are two affecting factors in text classification systems. Proposed method in this paper investigates the effects of using N-grams of characters for converting documents to numerical vectors on text classification performance. Also the effects of improved K-nearest neighbour classifier have been studied.

2.1 Improved KNN Classifier

KNN algorithm uses neighbors of a document to determine its class. Labels of K nearest neighbors to the document are chosen among all labels. Finally, the class with most number of neighbors wins. Unbalanced distribution of training documents in different categories affects this classifier. A class with more training documents has more chance to win. We use the following coefficient to adjust the chance between categories with different numbers of samples:

$$Wj = \frac{N}{Lj * M}$$

whereL_j is the number of samples in j^{th} class, M is the number of classes and sum of the training samples is N, $(d_1, d_2, .., d_N)$. Finally, we use following formulas to calculate probability of sample X belong to each class:

$$P(X, Cj) = Wj. \sum_{i=1}^{N} SIM(X, d_i). y(d_i, C_j)$$

Where

$$y(d_i, C_j) = \begin{cases} 1, & d_i \in C_j \\ 0, & \text{Otherwise} \end{cases}$$

$$SIM(X, d_i) = Dice(X, d_i) = \frac{2 \times |X \cap d_i|}{|X| + |d_i|}$$

We calculate the similarity between document d_i and test sample X, $SIM(X, d_i)$, using Dice similarity measure.Where $|X \cap d_i|$is the number of common N-grams between test sample X and training sample d_i.

Sample X belongs to the class which has the largest $P(X, Cj)$.

2.2 Converting Text Documents into Numerical Vectors

In this paper, we used N-grams of characters and N-grams of words to convert text documents into numerical vectors. Using N-grams consist of characters is a common way to represent text documents as numerical vectors. In this technique, each text document divides into slices with length N of adjacent characters. Vector corresponding to each document contains a list of non-iterative N-Grams with the number of iterations of each N-gram. In previous studies, the common lengths of N-grams were between 2 and 5 but in this paper we investigate the effect of N-grams with various lengths, between 2 and 10, on the efficacy of classifier.Table1 illustrates the tri-gram vector of sentence: "باران (شدیدی) بارید."and we used dash instead of space in tri-gram vectors.

3. Implementation and Results

A variety of experiments were conducted to test the performance of proposed method. Accuracy of the classifier is measured by a 5-fold cross-validation procedure. Essentially, the corpus is divided into 5 mutually exclusive partitions and the algorithm is run once for each partition. Each time a different partition is used as the test set and the other 4 partitions are used as the training set. The results of the 5 runs are then averaged. We first explain the training and test data used in the experiments and then present the obtained results.

3.1 Training Dataset

For this study, a portion of a standard Farsi corpus called Hamshahri1 and articles of some archived newspapers are used.The Hamshahri1 corpus consists of more than 100000 text documents covering 82 different categories. We choose 4000 text documents covering 7 categories: politics (625), social (146), economy (484), sports (484), technology (152), exterior (626) and affairs (283).

3.2 Evaluation Measures

In the text classification, the most commonly used performance measures are precision, recall and F-measure. Precision on a category is the number of correct assignments to this category and recall on a category signifies the rate of correct classified documents to this category among the total number of documents belonging to this category. There is a trade-off between precision and recall of a system. The F-measure is the harmonic mean of precision and recall and takes into account effects of both precision and recall measures. To evaluate the overall performance over the different categories, micro and macro averaging can be used. In macro averaging the average of precision or recall is compared over all categories. Macro averaging gives the same importance to all the categories. On the other hand micro averaging considers the number of documents in each category and compute the average in proportion to these numbers. It gives the same importance to all the documents. When the

corpus has unbalanced distribution of documents into categories, by using macro averaging, classifier deficiency in classifying a category with fewer documents is emphasized. Since an imbalanced corpus is being dealt with, it seems more reasonable to use micro averaging.

3.3 Experimental Results

In this section, the experimental results are presented. The experiments consist of evaluating classifier performance when character n-grams are used to represent documents. We evaluated classifier performance on Hamshahri1 corpus and the effects of removing or maintaining stop words, applying N-Grams with different lengths on classifier performance have been analysed. Finally, comparison between proposed method and other related works for Farsi text classification is presented.

In first experiment we have applied improved KNN classifier on the data. Fig.1 shows the performance of improved KNN classifier in comparison with traditional KNN classifier.

In the second experiment we have used various N-grams lengths in preprocessing and then we have applied improved KNN algorithm on the data. Fig.2 shows the best performance that can be achieved is in the case that we choose N-grams with length 8.

tri-gram	.ید	رید	اری	-با	ب(	ی)-	(دی	یدی	دید	شدی	(شد	ش)-	ـن(	انـ	ران	ارا	بار
Count	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

Table 1: Tri-grams without removing spaces, non-letters and punctuation marks

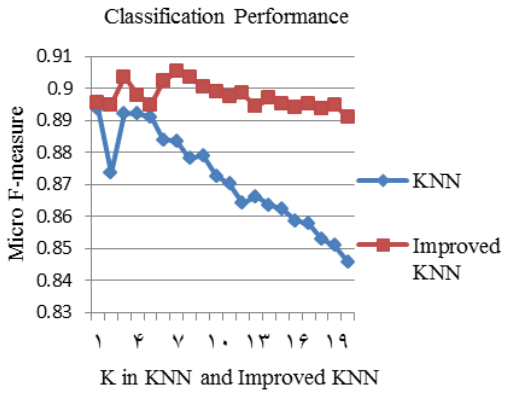

Figure 1: Evaluation of proposed method versustraditional KNN

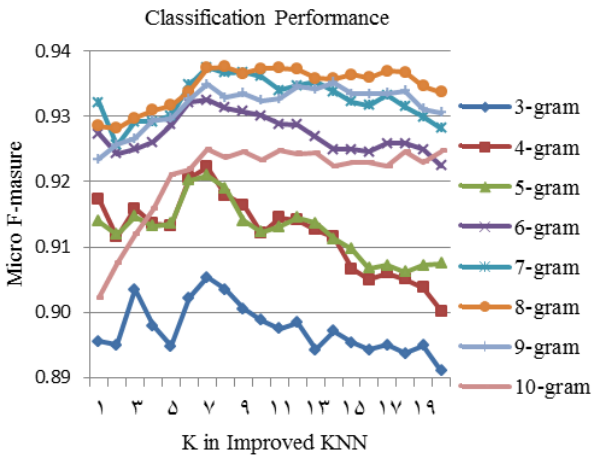

Figure 2: Effect of different N-grams lengths

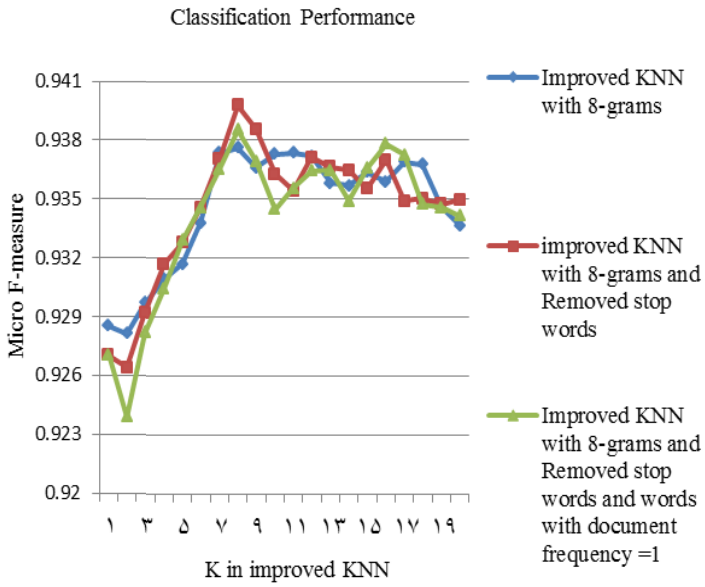

Figure 3: Effects of removing stop words and words with low document frequency

Removing stop words and word with low document frequency is a common technique for improving text classification performance. In the third experiment, we have evaluated the effects of removing stop words and words with document frequency of 1. The results have been shown in Fig.3. Removing words with document frequency of 1 can cause undesirable effect in this method.

In the next experiment, we have used a new datasets consists of 10 categories: economy (209), politic (180), sport (141), theology (148), medicine (110), art (219), agriculture (200), chemistry (130), mathematics (106) and sociology (214). Table 2 is a comparison between

Method	Micro Precision	Micro Recall	Micro F-Measure
Improved KNN	0.92	0.91	0.91
SVM using Thesaurus	0.88	0.90	0.89

Table 2: Improved KNN algorithm compared to SVM using Thesaurus

results of text classification using the improved KNN algorithm and best results for Farsi text classification, to This point, obtained in (Maghsoodi, Homayounpoor, 2011) using the SVM algorithm on the same datasets.

4. CONCLUSIONS

The KNN classifier is one of the most popular neighborhood classifier in pattern recognition. However, it has limitations such as: great calculation complexity, fully dependent on training set, and no weight difference between each class. The proposed approach in this paper aims to enhance the classification performance of KNN classifier for Farsi text classification. We improved the KNN text classifier by inserting a factor to the KNN formula for considering the effects of unbalanced training datasets and used of N-grams with lengths more than 3 characters in text preprocessing. This paper presented the results of classifying Farsi texts using N-gram and word frequency statistics employing a similarity measure called "Dice". As the results indicate, this approach improves the KNN algorithm especially when 8-grams indexing method and removing stop words are applied.

We reported 94% performance for text classification by selecting K=8. We demonstrated that the use of N-grams with lengths more than 3 characters and remaining non-letters characters and spaces in Farsi text preprocessing can improves the KNN classification results.

5. Acknowledgements

The authors would like to thank Noor Text Mining Research Institute of Computer Research Center of Islamic Sciences for supporting this work.

6. References

Arabsorkhi, M., Feili, H. (2006).Using Bayesian model to Persian text classification.In Proceedings of the Second Workshop on Persian Language and Computer, pp. 245--249, [in Persian].

Basiri, M.E., Nemati, S.,Aqaee, N. (2008).Comparing KNN and FKNN algorithms in Farsi text classification based on information gain and document frequency feature selection. In Proceedings of the 13th International Computer Conference of Computer Society of Iran, pp. 383--406, [in Persian].

Bina, B., Ahmadi, M., Rahgozar, M.(2008).Farsi text classification using n-grams and KNN algorithm: A comparative study. In Proceedings of the 4th International Conference on Data Mining, pp. 385--390.

Pilevar, M.T., Feili, H.,Soltani, M. (2009).Classification of Persian textual documents using learning vector quantization. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering , pp. 1--6.

Maghsoodi, N. and Homayounpoor, M. (2011).Using Thesaurus to Improve Multiclass Text Classification. Part II, LNCS 6609, pp. 244--253.