International Journal of Engineering Research and General Science Volume 3, Issue 4, July-August, 2015
ISSN 2091-2730

SVM BASED IMPROVEMENT IN KNN FOR TEXT CATEGORIZATION

[1]SWATI [2]MEENAKSHI

1A.P, Seth Jai Prakash Mukundlal Institute Of Engineering and Technology

[2]Student, JMIT

ABSTRACT- In today's library science, information and computer science, online text classification or text categorization is a huge complication. [1]With the enormous growth of online information and data, text categorization has become one of the crucial techniques for handling and standardizing text data. Various learning algorithms have been applied on text for categorization. On the basis of accuracy and efficiency KNN (K Nearest Neighbour) algorithm prove itself to be very efficient algorithm as compared to other learning algorithms. The framework of KNN with TF-IDF is studied and some changes need to be done for removing time complexity and improve accuracy so, proposed work is based on using SVM classifier which helps in splitting of training and testing data and take less time from the previous work with iKNN (improved KNN) algorithm which gives less time and more accuracy and overall improve text categorization.

KEYWORDS: categorization, KNN, TF-IDF, SVM, documents.

1. INTRODUCTION

Very tremendous growth in the amount of text data leads to development of different automatic methods purpose to improve the speed and efficiency of automated text classification with textual content. [1]The documents to be classified can contain text, images, music, etc. Each content type requires significant classification methods. This article is based on the textual content documents with special priority on text classification. Text classification is the technique from the main problems of text mining. Text categorization is defined as the term which helps in assigning uncategorized data or text in to fixed or predefined categories. The main objective of text categorization is to assign a category to a new document. [3]Category must be assigned according to their textual content Document or text can reside in multiple. one or no category at all. It is based on supervised machine learning method where documents are framed into VSM (vector space model) where words used as important features. First step is to train the data so that while testing the data results should be effective and efficient. Various different types of classification methods have been applied such as SVM (support vector machine), Naive Bayesian classifier, Decision trees, Entropy, Fuzzy logic, KNN (k- nearest-neighbour) and many more. KNN performance is quite better than other algorithms but still some improvement is required in KNN for reducing its time complexity and improving its accuracy. KNN is a type of lazy learning algorithm; it is based on finding the most similar objects from sample group with the help of euclidean distance. In this paper a framework is established having SVM (support vector machine) as a train/test splitter classifier which helps in training the documents in such a way while testing it requires minimum time. And further KNN is changed to iKNN (improved K Nearest Neighbour) which plays a major role in reducing the time and improving the accuracy of text categorization. This method determines the functionality of framework of SVM and iKNN performing together for effective categorization.

2. RELATED WORK

B. Trstenjaka et al. (2013)**[1]** presented a framework of KNN with TF-IDF for text categorization. This framework was totally based on the quality and speed of classification. It helps in finding similar objects based on the euclidean distance and TF-IDF calculates the weight for each term in each document. Both KNN and TF-IDF embedded together prove good together gave good results and confirmed initial expectations. Framework is performed on several categories of documents and testing is performed. During testing, classification gives accurate results due to KNN algorithm. This combination gives better results but need to upgrade and need to improve the framework for better and high accuracy results.

B. Liu et al. (2012)**[4]** proposed a rough set theory to solve the problem of effective categorization of text data in taxation system. The proposed work based on rough set model, consist of training the data, two-part test, and then finally to classify the new text, by using the training data. The purpose of testing is for comparing the effect of the text categorization system, if the test result is higher than a set categorization accuracy (threshold), the output rules, or the end of operation, re-calculate the weight, take a new feature subset, repeat the process until the results are satisfied. This paper is focused on the large part of text data and analyzes the information and gives the process of text categorization systems.

A. K. Mandal and R. Sen (2014)**[5]** In this paper, four supervised machine learning algorithms revisited including DT (C4.5), NB, SVM, and KNN and compared their classification performances on Bangla text documents. On this aspect, BD corpora was

developed, and then implemented of a tool for feature extraction and selection. The key findings of experiments are summarized as follows:

• On small and well-organized training sets, NB and KNN algorithms prove to be more capable than SVM and DT (C4.5) in categorization of documents. But, for large documents SVM prove to be superior from other classifiers in text categorisation.
• DT (C4.5) takes more time from other three algorithms for training, whereas SVM is fast in learning.
• From the experiment, average F-measures prove that SVM produces the best result followed by KNN, DT (C4.5) and NB.

V. Bijalwan et al. (2014**)[6]** proposed to categorizing the documents with the help of KNN based machine learning approach and then return the most appropriate and relevant documents. Results show that KNN shows the maximum accuracy as compared to the Naive Bayes and Term-Graph. But the shortcoming for KNN is that time complexity is very high but provides a better accuracy than other algorithms. Implementation of Term-Graph with other methods not with the traditional Term-Graph used with AFOPT. This kind of hybrid approach shows a better result than the traditional combination.

[5] *X. Zhoua* (2014**)[7]** proposed a k-means clustering method to collect and choose features for categorization. K means is required to collect several cluster centroids for each class and choose the highest frequency among centroid. K-means gives three steps as follows: first select initial cluster centroid randomly then assign each sample to the nearest centroid and finally update centroid by means of each cluster. Based on selected features compare the methods with original classifiers and finally the accuracy of text categorization, macro-F score, and the running time are tested. Results of k-means are faster than all the original methods and objective is achieved successfully.

3. PROBLEM STATEMENT

This section describes the problem statement as in introduction we discussed about text categorization and its various methods which help in proper and efficient categorization. Categorization is better done by KNN (K Nearest Neighbour) algorithm. Working of KNN is to find most similar objects from a sample group and assign it to a related document. KNN finds the euclidean distance and then assign the closest category to a document. But in section survey we have conclude that KNN has greater time complexity which affects its accuracy. So proposed work mainly focused to reduce its time complexity and improve its efficiency. In this paper a methodology is proposed in which SVM and iKNN work together.

4. PROPOSED WORK

This section describes the whole flow chart of the proposed work. The proposed system can be summarized into three main steps that are integrated to give accurate results: text document representation, classifier construction and performance evaluation.

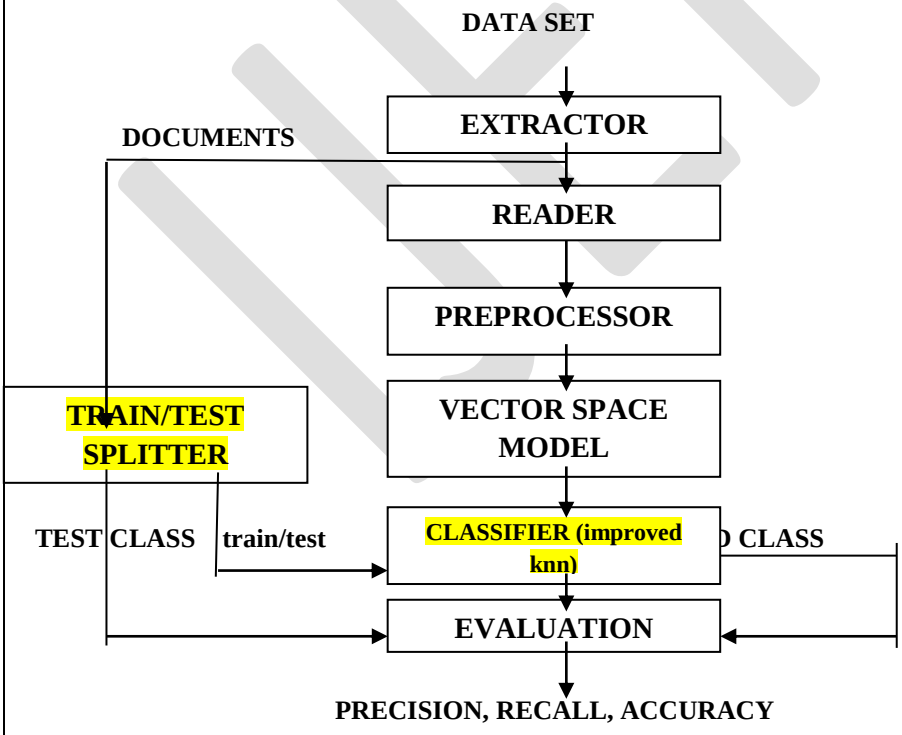


FIGURE 1 THE PROPOSED TEXT CATEGORIZATION SYSTEM FRAMEWORK

4.1 EXTRACTOR & READER

Extractor extracts the data set and read the number of documents, the number of topics and the documents which are related to the specific topics. It is an agnostic content summarization technology that automatically parses news, information, documents into relevant and contextually accurate keyword and key phrase summaries. Reader read input text document and divide the text document into a list of features which are also called (tokens, words, terms or attributes).

4.2 PREPROCESSOR

Pre-processor processed the document words by removing

- Symbols removal
- Stop words removal
- Lower Case Conversion
- Stemming

All symbols are removed in pre processing step and a stop list is a list of commonly repeated features which appear in every text document. [7]The common features such as it, he, she and conjunctions such as and, or, but etc. need to be removed because they do not have effect on the categorization process. Stemming is the process of removing affixes (prefixes and suffixes) from features. It improves the performance of the classifier when the different features are stemmed into single feature. For example: (convert, converts, converted, and converting) stemming remove different suffixes (s, -ed, -ing) to get single feature.

4.3 VECTOR SAPCE MODEL

In vector space model each input text document is represented as a vector and each dimension of this space represents a single feature of that vector and on the basis of frequency of occurrence, weight is assigned to each feature in text document. This representation is called vector space model. In this step, each feature is assigned to an initial weight equal to 1.[8] This weight may increase based on the frequency of each feature in the input text document. Vector space model use feature extraction method which detects and filter only relevant features which are far smaller than actual number of attributes And this process enhances the speed of supervised learning algorithms.

IF-TDF term is used in vector space model for assigning weight to each feature. It determines the relative frequency of words in a specific document. For calculation, TF-IDF method uses two elements:

TF - term frequency of term in document (the number of times a term appears in the document)

IDF- inverse document frequency of term i (the number of documents where the term appears)

[1]Formula for tf and tdf are:-

$$tf(t,d) = 0.5 + \frac{0.5 * f(t,d)}{\max\{f(w,d) : w \in d\}}$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$tfidf(i, d, D) = tf(t, d) * idf(t, D)$$

For creating VSM(vector space model) :

for i = 1 to num docs
  j = 1 to num of unique words

  Vsm (i,j)= (tf(i,j) * log 2(num docs/df(j)));

j=Number of all unique words

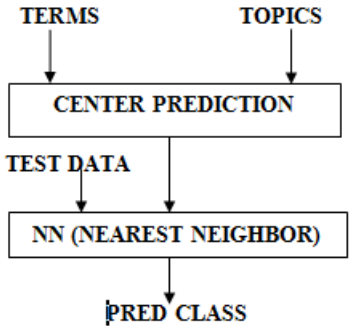
a(0,0)	a(0,1)	a(0,2)	a(0,3)	....	.....	....	....
a(1,0)							
a(2,0)							
a(3,0)							
.....							
.....							
.....							

Fig. 2 Weight matrix.

4.4 TRAIN/TEST SPLITTER

In this paper two train/test splitter are used one random splitter which only works randomly and second SVM which do binary splitting of train and test data by making exclusive classes And the advantage of the training the data in this way is that during testing the data it gives most accurate results from the previous work done. With the classified dataset from document classification, [9] SVM will prepare the model and classifier. Some data is trained and some data is tested. Training data is used for supervised test data. Results can be calculated with the help of test data.

4.6 DOCUMENT CLASSIFIR



4.7 CENTER PREDICTION

Before calculating centre prediction, a vocabulary is formed from the documents. Like, for one category a vocabulary of 30 frequent words are chosen same for category two and same for category three. These 90 words are quite frequent from each category and centre calculation is properly based on the frequent coming words and vocabulary. First 30 words are centre for one category next 30 words are centre for second category and next 30 words are centre for third category.. Training of data is based on the centre prediction. Centres are predicted from intelligent vocabulary which contains top rated terms which reduce dimension and complexity and testing become easier.

VOCABULARY	**Category one n words**	**Category two n words**	**Category three n words**

Where ‘n’ belongs to number of words taken in respective categories according to dataset.

4.8 NEAREST NEIGHBOUR

In this paper centre prediction and nearest neighbour play a major role in proposed work. Testing of data is based on the nearest neighbour. Nearest neighbour is different from KNN (k-nearest-neighbour). KNN works on the principle of calculating centres again and again for each test term but NN works on the principle that it only calculates centre for one time and never update it and it calculates the minimum distance from with the help of euclidean distance. [1]

$$D_{Euclidean}(x, y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$$

Where (x_1, x_2) are coordinates of x and (y_1, y_2) are coordinates of y

5. EXPERIMENTAL SET UP

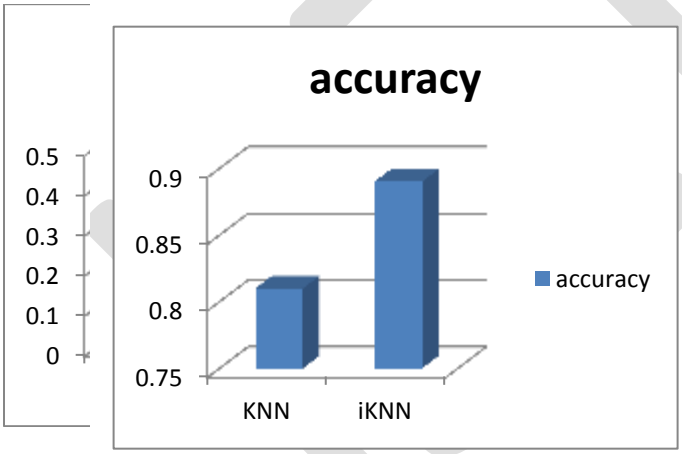
The experiments are carried out using mini- newsgroup dataset from UCI KDD Archive which is an online repository of large data set which encompasses a wide variety of data types, analysis tasks and application areas. Mini newsgroup contains 20 groups of 100 documents each. Our experiment uses 3 newsgroups of 100 documents each.

6. RESULTS

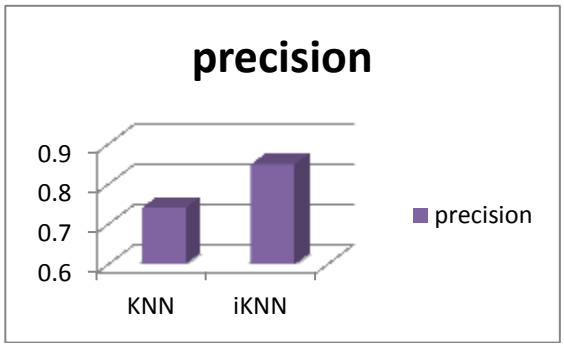
In this section, we investigate the performance of our proposed algorithm iKNN (improved KNN) and compare it with KNN algorithm.

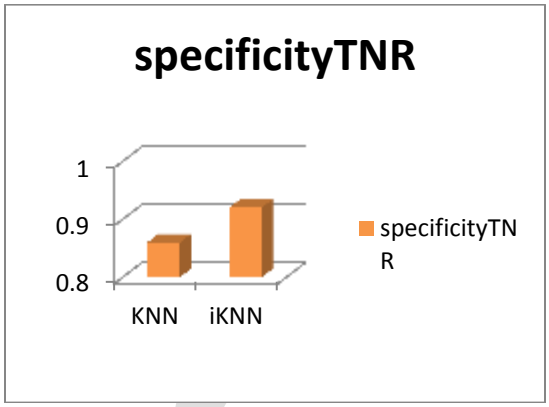
Results	Accuracy	Precision	Specificity	Sensitivity	F-score	Time
KNN	0.84	0.76	0.87	0.77	0.76	0.029
iKNN	0.89	0.81	0.90	0.83	0.81	0.009

$$[13]\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

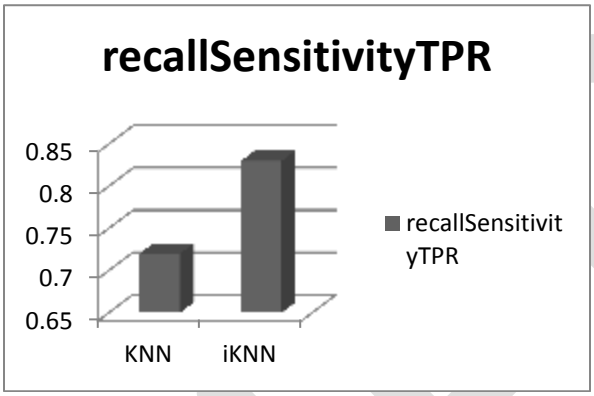


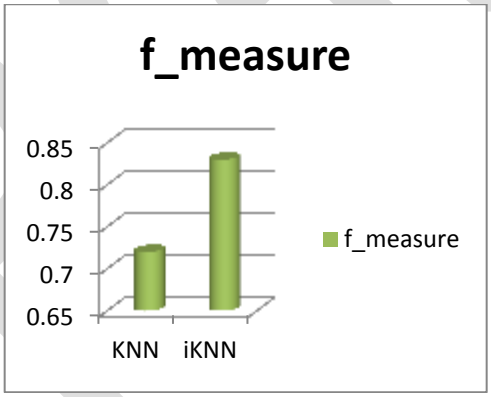
$$\text{Precision} = \frac{\text{Number of correct positive prediction (TP)}}{\text{Number of positive examples (TP+FP)}} \qquad \text{Specificity} = \frac{TN}{TN+FP}$$





$$\text{SENSITIVITY [13]} = \frac{\text{Number of correct positive predictions (TP)}}{\text{Number of correct predictions (TP+FN)}}$$





$$\text{F-Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

From above result we can conclude that iKNN perform better than KNN in terms of accuracy, time complexity, precision, recall and f-measure. Accuracy becomes more efficient from previous algorithm. Time the main issue become resolved from iKNN and sensitivity results are also better than KNN. Calculated results prove that our proposed work is more efficient than previous work.

7. CONCLUSION AND FUTURE WORK

In this paper we present a framework for text classification based on better categorization by using SVM as train/test splitter and iKNN (improved) algorithm instead of KNN. The main motivation for this proposed work is to improve the existing algorithm. Results produced are more efficient and more accurate than existing algorithm. The main factor of time complexity of KNN is reduced with the help of improved algorithm or work. Future work can be proposed for highly accurate results with the help of topic modelling and we can use hybrid models with more effective framework so that results can improve further.

REFERENCES:

[1]B. Trstenjak, S. Mikac and D. Donko, "KNN with TF-IDF Based Framework for Text Categorization," *ELESVIER*, 1356 – 1364, 2014

[2]M. Hrala and P. Kral, "Evaluation of the Document Classification Approaches,"

[3]G. Guo, H. Wang, D. Bell, Y. Bi and K. Greer, "KNN Model-Based Approach in Classification,"

[4]B. Liu, G. Xu, Q. Xu and N. Zhang, "The Research of Tax Text Categorization based on Rough Set," *ELESVIER*, 1683 – 1688, 2012

[5]A. K. Mandal and R. Sen, "Supervised Learning Methods for Bangla Web Document Categorization," *INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE & APPLICATIONS (IJAIA)*, Vol. 5, No. 5, 2014

[6]V. Bijalwan, V. Kumar, P. Kumari and J. Pascual, "KNN based Machine Learning Approach for Text and Document Mining," *INTERNATIONAL JOURNAL OF DATABASE THEORY AND APPLICATION*, Vol.7, No.1, pp.61-70, 2014

[7]X. Zhoua, Y. Hua and L. Guoa, "Text Categorization Based on Clustering Feature Selection," *ELESVIER*, 398 – 405, 2014

[8]A. T. Sadiq and S. M. Abdullah**, "**Hybrid Intelligent Techniques for Text Categorization**,"** *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND INFORMATION TECHNOLOGY (IJACSIT)*, Vol. 2, No. 2, pp. 23-40, 2013

[9]J. Tang, S. Alelyani and H. Liu, "Feature Selection for Classification: A Review,"

[10]Z. Xu, P. Li and Y. Wang, "Text Classifier Based on an Improved SVM Decision Tree," *ELESVIER*, 1986-1991, 2012

[11]Y. Man, "Feature Extension for Short Text Categorization Using Frequent Term Sets," *ELESVIER*, 663 – 670, 2014

[12]Feldman, R.& Sanger, J. (October 2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, USA, New York.

[13]H. Alshalabi, S. Tiun, N. Omar and M. Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization," *ELESVIER*, 748 – 754, 2013

[14]T. S. Zakzouk and H. I. Mathkour, "Comparing text classifiers for sports news," *ELESVIER*, 474 – 480, 2012

[15]Z. Yao and C. Z. Min, "An Optimized NBC Approach in Text Classification," *ELESVIER*, 1910-1914, 2012

[16]H. Drucker, D. Wu and V. N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, Vol. 10, No. 5, 1999