ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 2, Issue 4, April 2013*

# Short Text Classification Using kNN Based on Distance Function

Khushbu Khamar

Government Engineering College, Modasa

ABSTRACT- In the present day circumstances nowadays, the scope of short text such as Twitter messages, blogs, chat massages, book and movie summaries, forum, news feeds, and customer review is increasing very drastically. These applications pose a tremendous challenge to the text classifications due to sparseness of the relevant data & lack of similarity between the words. Short text classification is nothing but a process of assigning various input short texts to one or more target categories based on its contents.

Here we compared various algorithms such as Support Vector Machine, Naive Bayes, K-Nearest Neighbor, etc... . Based on this comparison I have selected knn for this. This paper includes various methods to reduce processing time and give good accuracy for testing instances.

Keywords- Short Text, Nearest Neighbor, Euclidean distance, Manhattan distance, Mahalanobis distance, Cosine similarity, k-fold cross validation , Holdout, Loocv, Repeated random sub-sampling

## I. INTRODUCTION

In a recent scenario short text is the modern means of Web communication and publishing, such as Twitter messages, blogs, chat massages, book and movie summaries, forum, news feeds, and customer reviews. Short text is nothing than more than short document which contains few words. For ex. Twitter limits the length of each twit to 14o characters, Facebook status length is limited to 420 characters. To actually verify the importance of Short text classification and recognize its relevance, we need to look at the case study conducted for the same in a well-known news agency 'Reuters". The data of various news articles was classified by applying short-text from the "Reuters data-set" based on the headlines only. Then, had compared the result obtained as above with the results obtained from the older approaches used in the literature that used full text. The evaluation actually shows that the text classification algorithms perform really well in both the set-ups. Also, the same levels of results were obtained with a very nominal loss in the accuracy and the results are obtained by using short-text classification in a lesser amount of time. So, it can be very much inferred that the results available by using short text classification can be available in a very lesser amount of time with the same level of accuracy.

## II. SHORT TEXT CLASSIFIER

### *A. Naive Bayes*

Under various module classifier methods of priory probability & class conditional probability, Naive Bayes method is a kind of module classifier. It is a simple probability classifier which is based on applying the well-known Bayes' theorem with strong (naive) independent bases/assumptions. The major benefit of this classifier is that it only requires a less amount of training data to assume the parameters (means & variances of the variables) which are essential for classification. By analyzing and finding the dependency among various attributes, Naive Bayes is very easy for implementation & computation. Hence, it is used for pre-processing.

### *B. SVM (Support Vector Machines)*

The positive and negative training data sets which are not common for other classification methods, is needed by the SVM. These training sets are required for the SVM to seek for the decision surface which separates the positive from the negative data in the 'n' dimensional space & hence the same is known as the hyper plane. Support vectors are the document representatives which are closest to the decision surface.

The aim of SVM is to find out the best possible classification function in order to differentiate between members of two classes in the training data in a two-class learning task.

### *C. K-Nearest Neighbors (KNN)*

One of the various classifier, 'KNN classifier' is a case based learning algorithm which is based on a distance or similarity function for various pairs of observation such as the Euclidean distance function. It is tried for many applications because of its effectiveness, non-parametric & easy to implementation properties. However, under this method, the classification time is very long & it is difficult to find optimal value of K. Generally, the best alternative of k to be chosen depends on the data. Also, the effect of noise on the classification is reduced by the larger values of k but make boundaries between classes less distinct. By using various heuristic techniques, a good 'k' can be selected. In order to overcome the abovesaid drawback, modify traditional KNN with different K values for different classes rather than fixed value for all classes.

## III. COMPARING KNN, NAÏVE BAYES AND SVM FOR SHORT TEXT CLASSIFICATION

**Data set 1:**

| Category | k-NN | | | Naïve Bayes | | | SVM | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| 1 | 0.8 | 0.77 | 0.82 | 0.66 | 0.89 | 0.76 | 0.63 | 0.77 | 0.85 |
| 2 | 0.6 | 0.8 | 0.72 | 0.5 | 0.2 | 0.2 | 0.33 | 0.2 | 0.1 |
| Avg. | 0.80 | 0.78 | 0.78 | 0.61 | 0.64 | 0.59 | 0.53 | 0.5 | 0.5 |

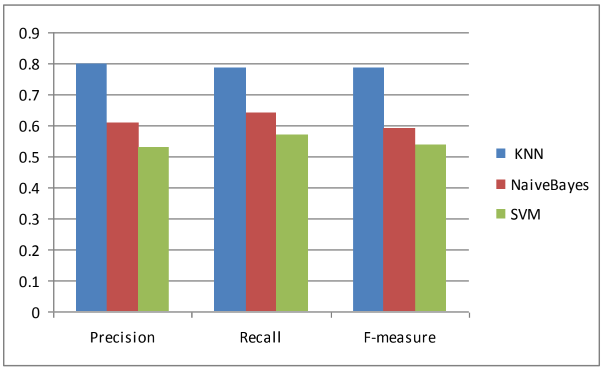

Fig 3.1 Comparison between kNN, NB and SVM

## IV. HOW K-NEAREST NEIGHBOR ALGORITHM WORKS

KNN algorithm is used to classify instances based on nearest training examples in the frame space. KNN algorithm is known as lazy learning algorithm in which function is approximated locally & computations are delayed until classification. A majority of instances is used for classification process. Object is classified into the particular class which has maximum number of nearest instances.

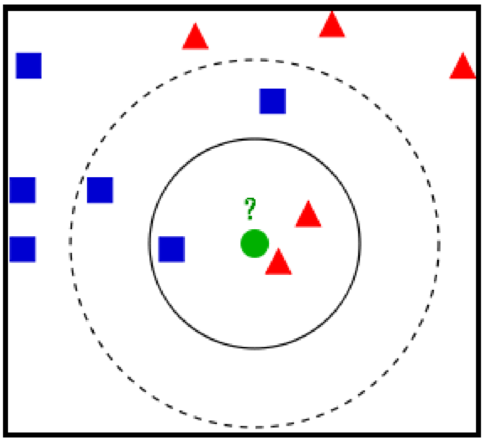

Fig Example of k- nearest neighbor[13]

In above figure the test instance (green circle) should be classified either into blue square class or into red triangle class.

If k = 3 (solid line circle) test object(green circle) is classified into red triangle class because there are 2 triangle instances and only 1 square instance in the inner circle.

If k = 5 (dashed line circle) test object(green circle) is classified into blue square class because there are 3 blue square instances and only 2 red triangle instances in the inner circle.

## V. SCOPE OF KNN

Improvement of KNN can be done on following parameters.[14]

***(1) Distance/similarity Function:*** The distance/similarity function is used for measuring the difference or similarity between two instances is the standard *Euclidean distance.*

***(2) Selection of Value K:*** It represents neighborhood size, which is artificially assigned as an input parameter.

***(3) Calculating Class Probability***: The class probability assumption, based on a simple voting.

## VI. PPLICATIONS OF SHORT TEXT CLASSIFICATION

- Classify the comments on blogs and forums
- Classify the chatter messages
- Classify the headlines of the news article
- Classify the title of research papers
- Classify the subject of email
- Classify the customer review (e.g. movie review, book review, product review etc.)
- In tweeter to classify incoming tweets into different categories (e.g. News, Events, Opinions, Deals, and Private Messages), so that users are not overwhelmed by the raw data.

## VII. EXISTING SYSTEM'S DISTANCE FUNCTION (SIMILARITY FUNCTION)

***Computation*** of distance function in KNN is based on distance between input test instance & training set instances. To compute the distance between instances, the distance/similarity function is important.

**Euclidean distance:** This is distance/similarity function is also called as the 'Pythagorean theorem'.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$$

## VIII. PROPOSED SYSTEM'S DISTANCE FUNCTION

***Pearson Correlation distance:*** This distance function is based on correlation.:

$$\mathrm{Sim(x, y)} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

## IX. PERFORMACE MEASUREMENT:

The classifiers' performances can be analyzed and compared by the measure obtained from the confusion matrix.

CONFUSION MATRIX:

| | Category 1 | Category 2 |
|---|---|---|
| Classified as 1 | True positive | False positive |
| Classified as 2 | False negative | True negative |

Table: Confusion matrix

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}.$$

**Dataset1**

| | IBk | IBk Pearson |
|---|---|---|
| Correctly Classified Instances | 37 | 40 |
| Incorrectly Classified Instances | 20 | 17 |

Table : Dataset1

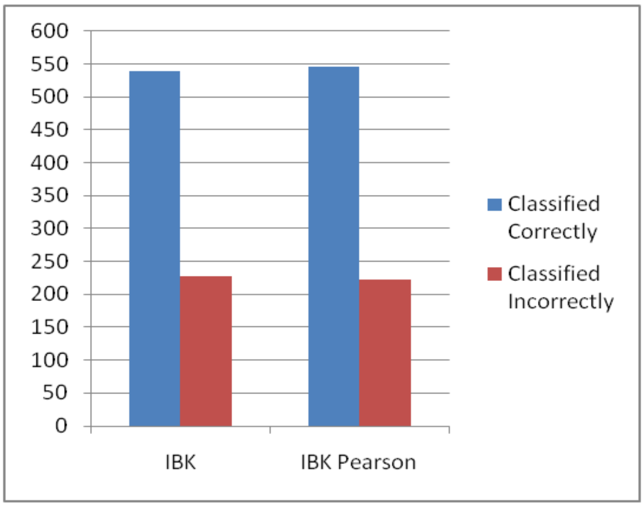

Fig. Comparison between existing system & proposed system using dataset1

**Dataset2**

| | IBk | IBk Pearson |
|---|---|---|
| Correctly Classified Instances | 37 | 40 |
| Incorrectly Classified Instances | 20 | 17 |

Table: Dataset2

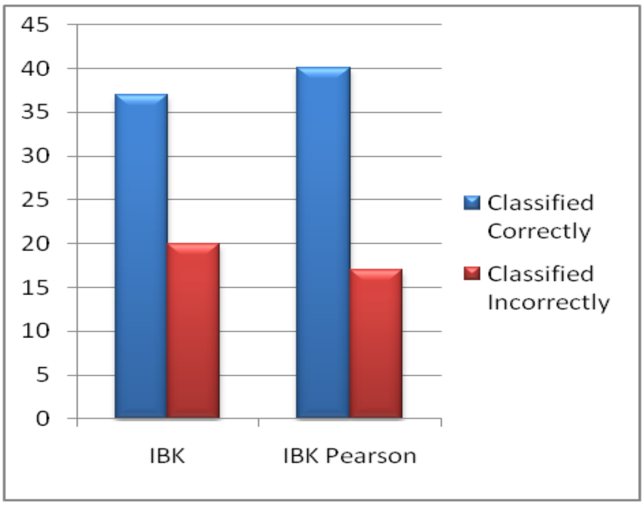

Fig. Comparison between existing system & proposed system using dataset2

In WEKA knn algorithm is named as IBK ,so I here used IBK for traditional kNN & IBK Person for proposed system. Experiment result shows that proposed system gives better result than the IBK using Euclidean distance.

## X. CONCLUSION

Short text classification uses few words for classification. This type of classification takes less time than the use of full text classification. To classify short text k-NN , Naive Bayes & SVM algorithms can be used. Result of section 5.1, concludes that the k-NN gives better accuracy than the other two algorithms. k-NN algorithm depends on distance function and value of k-nearest neighbor. Traditional k-NN uses Euclidean as a distance function but weakness of the Euclidean distance function is that if one of the input attributes has a relatively large range, then it can overpower the other attributes.

## REFERENCES

[1] Free encyclopedia. (2012, May 13). Data Mining [Online]. Available: http://en.wikipedia.org/wiki

[2] Improving Short-Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.5244&rep=rep1&type=pdf

[3] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”,2nd edition, Morgan Kaufmann Publishers, March 2006.

[4] Comparison of Automatic Classifiers’ Performances using Word-based Feature Extraction Techniques in an E-government setting

[5] Weka User Manual www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[6] McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe (2004). Analyzing microarray gene expression data. Wiley.

[7] Aixin Sun, “Short Text Classification Using Very Few Words,” in *Proc. of ACM SIGIR Conference (SIGIR’12)*, Portland, Oregan, USA, 2012.

[8] Improving the performance of k-Nearest neighbor algorithm for the classification of Diabetes dataset with missing vales(2012)

[9] Quantitative distance: “http://www.codeproject.com/Articles/32292/Quantitative-Variable-Distances”

.[10] Learning Distance Function http://www.cs.utexas.edu/~grauman/courses/spring2008/slides/Learning_distance_functions.pdf

[11] Han, Kamber and Pei, Data mining: Concept and techniques, 3rd Edn, Morgan Kaufmann,2000

[12] http://courses.cs.tamu.edu/rgutier/cs790_w02/l8.pdf

[13] k-nearest neighbor algorithm http://en.wikipedia.org/wiki/K-earest_neighbor_algorithm

[14] Laszlo Kosma , “k Nearest Neighbors algorithm (kNN)” (2008) Helsinki University of Technology http://www.lkozma.net/knn2.pdf

[15] Manning C. D. and Schutze H., 1999. Foundations of Statistical Natural Language Processing [M]. Cambridge: MIT Press

[16] Cross-validation_(statistics).htm ,Wikipedia, the free encyclopedia

[17]MathWorkhttp://www.mathworks.in/help/stats/classificationknnclass.html