International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Assamese Text Classification using *k* Nearest Neighbor

Moromi Gogoi, Shikhar Kumar Sarma

***Abstract*: Knowledge is the most powerful weapon of a society. And in today's world it is just a click away from the mouse. There is abundance of knowledge and information in the form of newspaper , electronic newspaper ,articles, online journals, webpages , search results etc. And there is a wide range of news from all over the world. But then the choice of news varies from person to person. Some people may prefer sports news to amusement news and some people may prefer political news over sports news and likewise there can be a number of other choices. It completely relies on individual's decision. Document Classification is the process of classifying a document into a number of predefined classes.**
In this paper we have done document classification of Assamese text using k-Nearest Neighbor. We have considered only four classes sports , politics , law and science. Our dataset consists of 200 documents collected from major Assamese newspaper . We have divided our data into 3:1. Majority of our datasets that is 75% data from datasets is used for training and the rest 25% of the datasets is considered for testing.

Keywords* : *Document classification, Assamese Text, k Nearest Neighbor

I. INTRODUCTION

A machine that is intellectually capable as much as human has always triggered curiosity and imagination in the minds of earlier computer scientist who were excited about Artificial Intelligence and Machine Learning. Learning is the ability to improve one's behavior based on experience. It is about building a computer system that automatically improves with experience. Machine Learning explores algorithm that learns from data , builds model from data and this model can be used for different tasks like prediction , decision making. The input to this learning algorithm is training data which represents experience and the output is some expertise which is usually a model or a computer program which can solve subsequent task. Tom Mitchell definition of Machine Learning says that " A computer program is said to learn from experience 'E' , with respect to some class of task 'T' and performance measure 'P' if its performance tasks in 'T' as measured by 'P' improves with experience 'E'.

It is the matter of concern that very few work has been done on assamese text although large number of resources are available for English language on this field. Recently, several related work on Natural Language Processing has been done on Chinese[8] ,Indonesian[5,6], Hindi[3], Arabic[4,7] , English-Hindi[2], Bengali Language [1] and so on. Resources on these languages are also increasing as more and more research are going on in this field. In this paper we have done classification of Assamese text using k Nearest Neighbor . Nearest Neighbor algorithms are among the simplest of all machine learning algorithms. It is a extremely popular algorithm and can perform surprisingly well. It is not only efficient but also easier to implement. The experimental result showed that kNN algorithm performed quite well.

A. k-Nearest Neighbor Algorithm

k-Nearest Neighbor is one of the oldest and most enduring method of classification. In a supervised learning if we have a set of input features (X_1,X_2,….,Xn), a target feature Y, a set pre-classified training examples (X_i,Y_i) where the values for the input features and the target features are given for each example, a new example where only the values for the input features are given. We predict the values for the target feature of the new example.

k Nearest Neighbor is also considered a lazy algorithm or instance based learning. In instance based learning when we get the training examples we don't process them immediately and learn a model instead we just store the examples. We do something when we need to classify an instance . So we don't immediately learn a model. That is why it is also called a lazy algorithm .The algorithm does not come with a model rather when we get the test instance it uses the stored instances in memory in order to find the possible Y.

In order to give a test instance we have to find similar instance or find some neighbouring instances or nearest instances. Instances are subsets of the datasets present. The instance based learning model works on the groups of instances related to the problem. Results are compared against the various instances present .

Manuscript published on November 30, 2019.
*Correspondence Author
Moromi Gogoi*, Computer Science, Dibrugarh University, Dibrugarh,India. Email: moromigogoi@dibru.ac.in
Shikhar Kumar Sarma, Information Technology, Gauhati University, Guwahati, India. Email: sks001@gmail.com

The comparison between two instances can be measured by various similarity metrics depending on the data. Importance is given on the representation of the instances and the similarity measure used for comparing between instances. This is the basic Nearest Neighbor.

In our experiment we have used the Euclidean distance . Suppose each instance has N attribute and between two instances

X_i=(X_{i1} , X_{i2} ,……,X_{iN})

X_j=(X_{j1} , X_{j2} ,……,X_{jN})

We want to find how close they are. Then the Euclidean Distance between X_i, X_j is given by

$$\text{Euc-Dist}(X_i, X_j) = \sqrt{\sum_{m=1}^{N}(X_{im} - X_{jm})^2}$$

We can find the Euclidean distance from a test point to all the point that we have used for training and select that point which has the smallest Euclidean distance.

B. Working of a kNN algorithm

Normally all learning algorithms are divided into two phase. Training phase and testing or Use Phase . Usually a model is learned in a training phase. Given a test instance we have to find similar instance. But in k-Nearest Neighbor we do the following:

- Training time: We don't learn a model in the training phase. We just save the training examples . For more advance implementation the training examples can also be stored in some data structure so that searching through this examples becomes faster and easier.
- At prediction time:
  - We get the test instance x_t and find the k training examples (x_1,y_1), $(x_2,y_2)\ldots(x_k,y_k)$ that are closest to the test example x among all the examples that we have stored in the training phase
  - We have to find that value of x_i corresponding to y_t.
  - We find all the training exanples (x_1 ,y_1) that is closest to X_t .Among all the examples that we have stored in the training phase given x_t we find that value of x_i that is closest to x_t .
  - And then we predict the most frequent class $_{yi}$ as output $_{Yt}$ among those $_{yi}$'s.

This is how a basic 1Nearest Neighbor algorithm works

Now in a more generalised algorithm, instead of finding the single example which is closest to the test example .We find k training examples (x_1,y_1), $(x_2,y_2),\ldots(x_k,y_k)$ that are closest to the test example X_t .The value of k may be equal to 2,3,4… etc. And it works as follows :

- Training time:
  - Save the training examples
- At prediction time:
  - Find the k training examples
  - Predict the most frequent class among those y_i's. That is we predict the most frequent of the majority class from $(y_1,y_2,y_3,\ldots.y_k)$
  -

C. Performance of kNN algorithm

The performance of a kNN algorithm can be improved by the following :

- *Weighting* examples from the neighborhood
- Measuring "*closeness*" by calculating distance between any two instances.
- Finding "close" examples in a large training set *quickly*

1.1.2.1 Choosing "k"

- When k is Large:
  - It is less sensitive to noise (particularly class noise)
  - Provides better probability estimates for discrete classes
  - larger training sets allow larger values of k
- When k is small :
  - It captures fine structure of problem space better
  - When k is small better performance can be obtained when we have a small training set.
- It is necessary to maintain a balance between large and small k depending on size of the datasets.
- As training set approaches infinity, and k grows large, kNN becomes Bayes optimal

II. MODELING AND CLASSIFICATION OF THE ALGORITHM

Let D be a collection of n pre-labeled documents {d_1, d_2, …, d_n} with m classes. Document d_i∈D is represented by a feature vector of the form (w_{i1}, w_{i2}, …, w_{if}), where w_{ij} is the numeric weight for the j-th feature and f is the total number of features.

Typically, each feature corresponds to a word or phrase appearing in the training corpus after the removal of function words, word stemming, and feature selection. In the kNN algorithm as explained earlier, we will use vector space model as the default feature selection criterion, the term frequency combined with inverse document frequency (TFIDF) as the default weighting measure, and the Euclidean Distance as the default similarity metric.

kNN algorithm used to develop the classifier is as follows :

ALGORITHM 1: Train-*k*NN(C,D)

D' ← Preprocess (D)

k ← Select-k(C, D')

return D',k

ALGORITHM 2: Apply-*k*NN(C, D',k,d)

S_k ← ComputeNearestNeighbor(D',k,d) // S_k is the set of d's k Nearest Neighbor

For each $c_j \in C$ // c_j denotes the set of all documents in the class C

do p_j ← $|S_k \cap C_j|/k$ // p_j is an estimate for $P(c_j| S_k) = P(c_j|d)$

return arg $\max_j P_j$

III. EXPERIMENT AND EVALUATION

We have conducted experiments on four classes collected from different Assamese newspapers. All documents for training and testing involve a pre-processing step, which includes tasks of function word removal, word stemming, feature selection, and feature weighting. We used vector space model as the feature selection criterion and the normalized TFIDF as the weighting function in our text categorization prototype system.

Experimental results reported here are based on the "F1 - measure," which is the harmonic mean of precision and recall and is given by,

$$F1 (recall, precision) = \frac{2\times \text{recall x precision}}{\text{Recall + precision}}$$

In the above formula, precision and recall are two standard measures widely used in text classification literature to evaluate an algorithm's effectiveness .

A. Datasets for Experiment

Our dataset consists of documents from three major newspaper "Aamar Asom", "Oxomiya pratidin" and "Khabor". We have considered only four classes sports , politics , arts and science. Our dataset consists of 200 documents collected from major Assamese newspaper . We have divided our data into 3:1. Majority of our datasets that is 75% data from datasets is used for training and the rest 25% of the datasets is considered for testing. Experimental details are shown in following tables.

Several domain specific corpora are developed and available in the research laboratory of NLP, Department of Information Technology, Gauhati University. Our corpus was collected from five daily assamese newspaper 'Asomiya Protidin', 'Khabor', 'Amar Asom', 'Agradoot' and ' Dainik Janambhumi'. The different datasets available are summarized in the table I .

Table- I: Datasets showing number of documents

Newspaper	No. of document	No. of class
Asomiya Protidin	500	10
Khabor	400	10
Amar Asom	200	5
Agradoot	300	9
Dainik Janambhumi	180	4

The different categories of documents available in the laboratory are political science, music , law, religion, sports, report, cartoon, obituaries, news, editorial, horoscope, cookery , arts, publicity , health and politics. Class distribution of different training and testing sets are shown in table II and the experimental results of the measures of precision, recall and F-measure of four classes is shown in table III.

Table II: Class distribution of the sample in the training and testing datasets

Categories	No of training documents	No of test documents	Total no of documents
Sports (ক্রীড়া)	60	20	80
Politics (ৰাজনীতি)	40	15	55
Arts (কলা)	30	10	40
Science (বিজ্ঞান)	20	5	25

Table III:Experimental result of Precision , recall and F- measure of four classes

Class	Precision	Recall	F-measure
Sports	96.85%	95.32%	96.07%
Politics	95.82%	96.22%	96.01%
Business	90.50%	91.20%	90.80%
Science	94.35%	95%	94.67%

IV. CONCLUSION

In this chapter we have done classification using k-NN classifier. K-NN classifier is a instance-based learning algorithm that is based on a distance or similarity function for two different instances , such as the Euclidean distance which we have used in our classification or Cosine similarity measure's .

Every time a new query instance is received for processing, a set of similar, related instances are retrieved from memory, and then this data is used to classify the new query instance. Overall, the entire database is used to predict behavior. A set of data points referred to as neighbors are identified, having a history of agreeing with the target attribute Once a neighborhood of data points is formed, the preferences of neighbors are combined to produce a prediction or top-K recommendation for the active target attribute. These methods are applicable for complex target functions that can be expressed using less complex local approximations.

Unfortunately, with these methods, the cost of classifying a new instance is always high . Moreover because of its effectiveness, non-parametric and easy to implement properties, the classification time is long and it is difficult to find optimal value of k . However the best choice of k rely greatly upon the choice of data ; generally, greater the value of k , less will be the effect of noise on the classification.

REFERENCES

1. Phani, Shanta, Shibamouli Lahiri, and Arindam Biswas, "A machine learning approach for authorship attribution for Bengali blogs." Asian Language Processing (IALP), 2016 International Conference on. IEEE, 2016.
2. Dutta, K., Kaushik, S. and Prakash, N, “Machine learning approach for the classification of demonstrative pronouns for Indirect Anaphora in Hindi News Items”, The Prague Bulletin of Mathematical Linguistics, 95, pp.3350 , Apr 2011
3. Haque, Rejwanul, et al. "English-Hindi transliteration using context-informed PB-SMT: the DCU system for NEWS 2009." Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. Association for Computational Linguistics, 2009.
4. El-Barbary,O. G. El-Barbary, "Arabic news classification using field association words." SCIENCEDOMAIN Int 6.1 (1-9), 2016.
5. Buana, Putu Wira, and I. Ketut Gede Darma , "Combination of k-nearest neighbor and k-means based on term re-weighting for classify Indonesian news." International Journal of Computer Applications 50.11, 2012.
6. Asy'arie, Arni Darliani, and Adi Wahyu Pribadi, "Automatic news articles classification in Indonesian language by using naive bayes classifier method." Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services. ACM, 2009.
7. Kanan, Tarek, and Edward A. Fox. "Automated arabic text classification with P-Stemmer, machine learning, and a tailored news article taxonomy." Journal of the Association for Information Science and Technology 67.11: 26672683, 2016.
8. Xu, Jun, Yu-Xin Ding, and Xiao-Long Wang, "Sentiment classification for Chinese news using machine learning methods." Journal of Chinese Information Processing 21.6 : 95-100, 2007.
9. Gogoi M and Sarma S.K , “Document Classification of Assamese Text Using Naïve Bayes Approach ” International Journal of Computer Trends and Technology (IJCTT) – volume 30 Number 4 – December 2015.

AUTHORS PROFILE

Moromi Gogoi is an Assistant Professor in the Department of Computer Science, DODL, Dibrugarh University , Dibrugarh, Assam. Her area of interest are Natural language Processing, Artificial Intelligence, Text Mining.

Shikar Kr Sarma is a professor in the Department of Information Technology, Gauhati University, Guwahati, Assam. His area of interest are Natural language Processing, Language technology, Artificial Intelligence. Some of the projects under him are Cross Lingual Information Access (Assamese),

Design and Development of Wordnets (Assamese and Bodo), Language Technology Development Project (Assamese and Bodo), Design and Development of Digital Assamese Thesauras.