Machine learning approach for categorical document mining

P. Sarma¹ and H. Deka²
¹,² Gauhati University, Guwahati, India
¹ parismita.sarma@gmail.com, ²hrishikesh.deka1991@gmail.com

Abstract. The word text mining or document mining means automatic classification of text documents to different categories depending on their content. At present days this area of research is attracting many researchers due to increasing use of electronic documents in everyday life. In this paper authors have proposed a document mining method using Bisecting K-means algorithm, KNN classifier and Decision tree. There are many machine learning classification algorithms used for information retrieval. But most of them have very high computational complexity. So a method Bisecting K – Means clustering algorithm instead of normal K-means algorithm is used here and this approach trivially reduces the number of comparisons compared to others. The researchers also used decision tree at last to obtain the sub categories more accurately. After analysis it is found that the combination of Bisecting K Means and KNN classifier enhance the accuracy of the categorization. Accuracy of this proposed system is shown for each category at Result and discussion section.

Keywords: Text Mining, TF-IDF, NLP, K Nearest Neighbor, K means Clustering

1. Introduction

Document is an evidence of information. In day to day life varieties of documents like structured, semi structured and totally unstructured are found. Among them Govt. electronics documents, resumes, news paper items, digital documents are most important sources of information. Knowledge mining from huge dataset and classification of these documents are two evolving research areas now a days. KDT means Knowledge Discovery from Text, is a distinctive and explicit way of discovering knowledge. Though text mining is similar to data mining but difference exists as data mining deals with structured documents whereas in text mining knowledge is extracted from un structured or semi structured documents[1]. Automatic document classification has many practical applications. Manual document classification is a tedious work when the dataset is very large. So the authors tried here to classify digital document using machine learning approach and this approach is explained here with methodology applied. For any document classification a categorised and labelled dataset is the most important primary need. Depending upon the contents of a document it is observed that it may belong to more than one category. Text summarization is a mostly cited problem in NLP and machine learning. Among many uses document demonstration, creation and valuation of the classifier are most important. Text mining is just like web searching. Text mining works with the help of algorithms in search of structured data values from unstructured ones. Some quantitative methods are used to do data analysis. The primary intention of text mining is meaningful data extraction from text and manage various primary operations on data like information retrieval, summarization, categorization of documents depending on supervised and unsupervised approach.
2. Literature review

To get ideas about some recent studies on text mining and to explore some sound techniques a number of studies have been done and a few of them are discussed below.

One paper by Bijalwan V et al. [2] we studied and came to know that most of the time machine learning methods based on KNN approach is used for classifying documents into respective categories. They have used Reuter – 21578 dataset for their experiment and identified different keywords and taken out them from the SGML file. A vector space was generated from the documents and its keywords. They had used TF-IDF to prepare the vector space and SVM classifier for classification of the documents.

In the next paper by Khan A et al. [3], different feature extraction machine learning techniques are elaborately discussed. According to them proper feature selection and extraction of those features are very important part of document clustering. Pre processing consists of tokenization, stemming and stop word removal. Some weighting schemes are taken for features selection. Authors of this paper mentioned about some mostly used machine learning algorithms such as KNN, Decision Tree, SVM, Bayesian classifier, Latent Semantic Analysis, Genetic Algorithm etc.

Lamba A et al. [4] discuss about use of KNN algorithm in document clustering. KNN classifier is an essentially used approach for document clustering but researchers are facing some problems when using it in document identification. This algorithm has drawback of high computational cost, huge memory requirement, strictly equal weighted feature etc. To remove these drawbacks two modified versions of KNN algorithms are used. One version tries to improve the performance by changing the result influencing factors like value of initial clusters k. The 2\textsuperscript{nd} category discusses about the hybrid edition of KNN that infuses evolutionary computing.

According to the authors of paper [5], K means clustering algorithm and its variants show better result than other techniques like hierarchical clustering. It happens due to vast difference in the time complexities of hierarchical and K means clustering. K-means clustering is linear whereas hierarchical is quadratic. Bisecting K-means which is a modification of traditional k- means clustering algorithm out performs the original one.

3. Proposed model

In this work text categorization is focused on three basic steps: Pre processing, Vector presentation and Classification. Here a clustering method called Bisecting K Means clustering algorithm is used according to which the beginning of the source dataset is clustered. This initial clustering approach helped in better classification in later stage. The output of Bisecting K-Means clustering algorithm was used by KNN (K Nearest Neighbour) classification algorithm and were able to classify the documents correctly. Figure 1 shows the block diagram of proposed approach. The block diagram clearly shows the prominent phases of the methodology we adopted. An overview of the method used is given below.
The first phase is pre processing. The documents in the set is made standardized using pre-processing technique. Tokenization is the first work, documents in the data set are normally in .txt format. This format is converted to a set of tokens where each token represents one word. Word is a continuous set of alphabet without any delimiters. It is a repetitive process and will be applied to each document, the output is one token per line.

3.1. Stop word removal

Redundant words that carries very little information are called as Stop words. They have insignificant role in document classification hence can be removed from the set. They are almost present in all the sentences, for example determinant (a, an, the), WH question words (Who, What, When ), prepositions etc. Removal of these stop words does not degrade the performance of the classification and we have removed all stop words from our data set. A list which was prepared earlier used to identify the stop words and we have removed them in the tokenization phase itself. This pre processing work reduces the dimensionality of vector space[6]. Eventually efficiency of the classification will be improved due to stop word removing.

3.2. Term frequency-inverse document frequency

Vector space consists of a number of selected features. After extracting relevant features vector space is built for further processing. TF-IDF (Term frequency – Inverse Document Frequency) method is used in our case to select the appropriate features. It is a term weighting method used for text classification. Uniqueness of a term determines the TF-IDF value. Actually TF-IDF is product of Term frequency and Inverse document frequency.

\[ \text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t, d) \]
Equation 1 above shows the mathematical representation that calculates TF-IDF value for term ‘t’ in document ‘d’ [6].

3.3. Vector Space Model
It is a two dimensional vector of M X N size. Text documents are represented as a vector of terms of the document. The terms used for the vector formation must be unique. As mentioned M represents total concerned documents of the vector whereas N represents total number of unique terms in that M documents. A value \( w_{ij} \) in the vector space corresponds to TF-IDF value of term \( T_j \) in the document \( D_i \) [7].

3.4. Clustering
One difficulty frequently arises at the time of clustering due to high dimensionality of the vector space and it puts down the performance of the classification process. That is why researchers have used clustering before classification of the input data set. Clustering reduces the dimension of the vector space and increases the efficiency of classification. From earlier study it is understood that K-Means is an efficient clustering algorithm for text document [8]. But it results in poor clustering on huge dataset, in our case this original K-Means algorithm will degrade the performance when number of features become very large. Moreover number of clusters should be predefined which is not a good approach for document clustering. To overcome these pitfalls we have used Bisecting K-Means clustering algorithm. This clustering shows better result in huge dataset and improved time complexity than general K-Means clustering algorithm. The Bisecting K-Means algorithm iteratively selects and bisects each leaf cluster until \( k \) clusters are found. Following are the steps for Bisecting K-Means clustering algorithm [5].

1. Input the documents as cluster.
2. Select a cluster to split.
3. Take ‘k’ as 2 and apply K-means to selected cluster.
4. Repeat step 2 and 3 until desired number of clusters are received.

3.5. Classification
K Nearest Neighbour (KNN) supervised classification algorithm is used for document classification. According to this algorithm, based on class of K nearest documents in document space, category of given (query) document is identified.

If \( X \) is an query document, the classification of this document according to KNN will be as given below [9] [10].

If \( C_1, C_2, C_3, \ldots, C_j \) are \( j \) numbers of training categories.

Total training samples are \( N \). After requisite pre- processing (tokenization, stop word removal), a \( m \)-dimensional feature vector is obtained.

TF-IDF method is applied to assign weights to each term.
1. From the training samples \( (X_1, X_2, X_3, \ldots, X_n) \), we have obtained a same dimensioned feature vector.
2. Similarity distances between query document \( X \) and training samples are calculated. For example similarity between \( X \) and \( i^{th} \) document will be calculated as follows (given at formula (2)):

\[
SIM(X, d_i) = \frac{\sum_{j=1}^{m} x_j d_{ij}}{\sqrt{\left(\sum_{j=1}^{m} x_j^2\right) \left(\sum_{j=1}^{m} d_{ij}^2\right)}}
\]  

3. \( K \) numbers of samples are chosen from \( N \) similarities of \( SIM(X, d_i) \), \( i = 1, 2, \ldots, N \) and name as KNN collection of \( X \). After that probability of \( X \) fit in to each group and it will be calculated using the formula (3).
Here \( y(d_i, C_j) \) is a category attribute function, which satisfies

\[
(3) \quad P(X, C_j) = \sum_{d_i \in \mathbf{K}_{NN}} SIM(X, d_i) \cdot y(d_i, C_j)
\]

\[
(4) \quad y(d_i, C_j) = \begin{cases} 1, & d_i \in C_j \\ 0, & d_i \notin C_j \end{cases}
\]

Query document ‘X’ should belong to largest of \( P(X, C_j) \) values, that is \( C_j \) category

3.6. Decision tree

Another classification technique called Decision tree is used to classifying a document into its respective subcategory. Decision tree algorithm gives good performance with noisy data. There are various decision tree algorithms among which ID3, C4.5 and C5 are frequently used. Attributes are chosen depending upon highest value of information gain at individual level[11]. Leaf nodes corresponds to document’s category and branches represents the parameters which directs the query document to the particular category. We have shown our output with decision tree in result and discussions section. In the training set \( S(s_1, s_2,..., s_n) \), each element \( s_i \) is \( (x_{i1}, x_{i2},..., x_{im}) \), a \( m \) dimensional vector where \( i \) stands for category of \( S_i \), \( X_i \) represents values of attribute of the sample[12].

4. Results and discussion

4.1. Dataset

The training set is obtained from the BBC news website. It has data distributed among five different classes: Business, Entertainment, Politics, Sport and Tech. Each class contains multiple documents. Each of these documents is in plain text format. A total of 2225 documents of the five different classes present in the dataset[13].

4.2. Work performed

Pre-processing, clustering and classification are three main steps of a document classification process. As mentioned earlier pre-processing phase consists of various steps like tokenization, stop word removal and creation of a vector space from the filtered terms by using some term weighting methods such as TF-IDF.

4.3. Tokenization and stop word removal

From each class, 100 documents are taken to create the training set. After tokenization a series of tokens are found that are stored in a new file with one token in a line. Same process is repeated for each document. Figure 2 shows a sample of tokenized terms. Stop words are those which are irrelevant to the classification process. We have used “bbc stop list.txt” and “smart.txt” to remove such stop words. Figure 3 shows a sample of stop lists that we had.

\[
\text{Figure 2. An example of extracted tokens} \quad \text{Figure 3. An example of stop words used}
\]

4.4. Vector space model

The filtered terms will be used to build a vector space model. A term weighting method will be used on the unique filtered terms to build the vector space. A total of 13,824 unique terms are found from
the dataset and a vector space of size $500 \times 13,824$ is created for 500 documents of the training set having 13,824 unique terms.

4.5. Clustering
The cluster centroids from K-Means clustering are stored in a file. After clustering, each document in the vector space is assigned to one of the five clusters. From these five clusters, five cluster centroids will be calculated. The classification process will use these cluster centroids as input data. Next Bisecting K-Means clustering is used, in each step of the Bisecting K-Means cluster, two clusters are created from the remaining documents in the vector space. A cluster similarity measure is used to select one of the two clusters as final cluster. The remaining cluster will be again divided into two clusters. This process is continued until the desired number of clusters are not obtained. The required number of clusters is set to five for Bisecting K-means algorithm.

4.6. Classification
From each class, 30 documents are used to create a test set for the classifier. The documents that are used to create the test set are not a part of the training set. The KNN classifier is used to determine the class of each of these documents. Figure 4 below shows correctly categorized documents using K-means cluster centroids and Figure 5 shows documents of each class that are used for evaluation using the Bisecting K-Means cluster centroids.

![Figure 4. Total documents and correct output using K-Means cluster centroids](image)

![Figure 5. Total Documents and correct output using Bisecting K Means cluster centroids](image)
4.7. Precision, recall and F-measure of KNN classification

KNN classification algorithm is used and Precision, Recall and F measure are calculated to evaluate the performance of the classification algorithm.

| Document class | Precision | Recall | F-Measure |
|----------------|-----------|--------|-----------|
| Business       | 0.8333    | 0.8064 | 0.8196    |
| Entertainment  | 0.7667    | 0.7667 | 0.7667    |
| Politics       | 0.7333    | 0.7097 | 0.7213    |
| Sports         | 0.8333    | 0.9259 | 0.8772    |
| Tech           | 0.7333    | 0.7097 | 0.7213    |
| **Average**    | **0.7799**| **0.7837**| **0.7818**|

Table 2. The evaluated result of the classification process using Bisecting K-Means cluster centroids

| Document class | Precision | Recall | F-Measure |
|----------------|-----------|--------|-----------|
| Business       | 0.8667    | 0.8667 | 0.8667    |
| Entertainment  | 0.8333    | 0.8333 | 0.8333    |
| Politics       | 0.7333    | 0.7097 | 0.7213    |
| Sports         | 0.90      | 0.9310 | 0.9152    |
| Tech           | 0.6667    | 0.6667 | 0.6667    |
| **Average**    | **0.80**  | **0.8015**| **0.8007**|

From the Table 1 and Table 2, we see that average F-Measure for KNN classification algorithm increases while using the cluster centroids of the Bisecting K-Means clustering.

4.8. Decision tree classification

For the documents that are categorized as sports class, a further classification of those documents into five sub-categories are carried out. For the Sports class the sub categories considered are Football, Rugby, Athletics, Tennis and Cricket. 50 documents of each of these subcategories are used to create the training set of 250 documents. From these documents, a vector space of the weighted unique terms is created. The terms are weighted using TF-IDF scheme.

The KNN classification algorithm categorizes 30 documents that belong to the sports class. From figure 5 it is seen that 28 documents out of 30 are correctly categorized into sports class using Bisecting K-Means cluster centroids. Next these sports documents are categorized by the Decision tree classifier into one of the five sub-categories. The result in table 3 shows that 24 out of these 30 documents are correctly categorized by the Decision tree. It is observed from table 3 that accuracy rates are quite good for most of the sub-categories. We are able to get an average accuracy rate 80%.
Table 3. The evaluation result by the Decision tree classifier of Sports class

| Document class | Documents Evaluated | Correctly Identified | Accuracy   |
|----------------|---------------------|----------------------|------------|
| Athletics      | 6                   | 5                    | 0.8333     |
| Cricket        | 6                   | 5                    | 0.8333     |
| Football       | 6                   | 6                    | 1.0000     |
| Rugby          | 6                   | 4                    | 0.6667     |
| Tennis         | 6                   | 4                    | 0.6667     |
| Total          | 30                  | 24                   | 0.8000     |

5. Conclusion and future work

In this paper various methods and techniques for an accurate and effective text categorization have been discussed. A clustering algorithm is used along with the classification technique to increase the efficiency of the text mining process. A Decision Tree algorithm is also used to find out the sub categories of the documents.

In future, some stemmers can be used in pre-processing steps to find out only the root words of the terms in the training set. This will further reduce the dimensionality of the feature space and thus, complexity of the process may be reduced. Again, some steps may be taken to remove noisy features as the accuracy of the KNN classification process is greatly reduced in presence of noisy features. The relation between the terms is not considered by the authors. Each term is taken as individual one. But there may be some terms which carry some meaning while they are considered together such as some phrases. Hence, other than individual terms, combination of terms may also be considered. This may increase the accuracy of the classifier.

References

[1] Desai A 2015 A review on knowledge discovery using text classification techniques in text mining, International Journal of Computer Applications, vol 111(6)
[2] Bijalwan V, Kumar V, Kumari P and Pascual J 2014 KNN based Machine Learning Approach for Text and Document Mining, International Journal of Database Theory and Application, vol. 7(1), pp. 61-70
[3] Khan A, Baharudin B, Lee LH, Khan K 2010 A review of Machine Learning algorithms for text documents classifications, Journal of advances in information technology, vol. 1(1)
[4] Lamba A, Kumar D 2016 Survey on KNN and its variants, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5(5)
[5] Steinbach M, Karypis G, Kumar V 2000 A Comparison of Document Clustering Techniques, In KDD Workshop on Text Mining, Department of Computer Science /Army HPC Research Center, University of Minnesota
[6] Vijayarani S, Ilamathi J, Nithya Pre-processing Techniques for Text Mining An Overview, International Journal of Computer Science & Communication Networks, vol. 5(1)
[7] Trstenjak B, Mikac S, Donko D 2013 KNN with TF-IDF Based Framework for Text Categorization, 24t DAAAM International Symposium on Intelligent Manufacturing and Automation
[8] Reddy C K and Vinzamuri B 2013 A Survey of Partitional and Hierarchical Clustering Algorithms, Data Clustering: Algorithms and Applications, p 24
[9] Yong Z, Youwen L and Shixiong X 2009 An Improved KNN Text Classification Algorithm
Based on Clustering, *Journal of computers*, vol. 4(3)

[10] Buana PW, Jannet S D R M, Putra I K G D 2012 Combination of K-Nearest Neighbor and K-Means based on Term ReB weighting for Classify Indonesian News, *International Journal of Computer Applications*, Vol 50(11)

[11] Li YH and Jain AK 1998 *Classification of Text Documents*, The Computer Journal, 41(8), pp.537-546

[12] Bhumika, Sehra S S, Nayyar A 2013 A review paper on algorithms used for text classification, *International Journal of Application or Innovation in Engineering & Management (IJAIME)*, vol. 2(3)

[13] Deka H and Sarma P 2017 Machine Learning Approach for Text and Document Mining, *International Journal of Computer Science Engineering (IJCSE)*, ISSN: 2319-7323, Vol. 6(5)