Classification of Text Documents -Performance Analysis at Micro and Macro levels

Aprajeeta Singh

Computer Science and System Engineering, Andhra University, AP, India

aprajeeta.singh4157@gmail.com

Abstract

Text mining is the process of exploring and analysing large amounts of unstructured text data aided by software that can identify concepts, patterns, topics, keywords and other attributes in the data. It's also known as text analytic. Text mining is also the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data and finally evaluation and interpretation of the output. The effectiveness of the proposed machine learning techniques for text mining is analysed by experimentation at micro and macro levels on the Reuters-21578 collection dataset with three classifiers models namely KNN, Decision trees and Support Vector Machine.

Keywords: Machine learning, micro, macro, classification

1. Introduction

The purpose of Text Mining is to process unstructured textual data to extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various machine learning algorithms. As we know that the information available on internet is increasing day by day, and there is growing need to classify these documents and measure their performance based on micro and macro levels. The main techniques is the representation of text documents and methods for estimating the classification quality i.e. performance.

Classification of text documents based on machine learning is important because the composition and content of document is changing permanently. Techniques that we use in text document classification makes easy to construct classification models based on training set and also use them to predict a class or a set of classes for new documents. If this approach is properly implemented then the classification quality is nearly equal to that performed by human. In the text classification and performance analysis support vector machine and kNN proved to be more effective. Text documents are usually classified using bag of words model.

The overall structure of this research paper is as follows. Section 2 presents related work Proposed approach is given in Section 3. Section $ discusses about the implementation of the proposed model and revels the results. Finally, Section 5 gives conclusions along with future scope for the work.

2. Related work

Until the late ‘80s the most popular approach to text classification was knowledge engineering, in which rules must be manually defined by a knowledge engineer with the aid of a domain expert, if the set of categories is changed or updated then these two professionals must intervene again. And also if the classifier is moved to a completely new set of classification then the different expert needs to intervene and work start from scratch. In the ‘90s this approach lost popularity especially in research area in favour of machine learning, in which a general inductive process automatically builds a text classifier by learning from a set of pre-classified documents. The advantage of machine learning is accuracy comparable to that achieved

by human experts, and also considerable savings in term of expert manpower, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories. The advantage of the machine learning over the knowledge engineering are evident. The engineering efforts goes towards the construction not of a classifier, but of an automatic builder of classifiers.

3. Proposed Method:

a) Pre-Processing

- Before applying machine learning algorithms, we have to represent and weight every document with respect to the set of textual features. Below transformations are applied
    - Remove punctuation
    - Remove digits
    - Remove extra white spaces
    - Remove stop words (e.g. the, and, for, is)
    - Conversion to lower case
- Also, we produce a weighted version of the TDM by term frequency-inverse document (tf-idf). This weighted version takes into account how often a term is used in the entire corpus as well as in the single document.
- Frequent terms can also be found (those which appears in at least a specifies number of documents. Representation is normalized.

b) Document Classification

We define our testing and training subsets to make sure that we do not evaluate with the documents that the system has learnt from. These methods are only available when training data is tagged/labelled. Currently we are using 3 models.

- Decision tree
- Support Vector Machine
- kNN

c) Evaluating the Models

Measuring the quality of a classifier is a necessary step in order to improve it. Main metrics are

- **Precision:** Number of documents correctly assigned to a category out of total number of documents predicted.
- **Recall:** Number of documents correctly assigned to the category out of the total number of documents in such category.
- **F1:** It is basically the harmonic mean of precision and recall.

We are basically evaluate in multi-class and multi-label environments, the method becomes slightly more complicated because the quality metrics have to be either shown as per category, or globally aggregated. There are two main aggregation approaches.

- **Micro-average:** Every assignment9document, label) has the same importance. Common categories have more effect over the aggregate quality than smaller ones.
- **Macro-**average: The quality for each category is calculated independently and their average us reported. All categories are equally important.

4. Experiments and Results:

4.1 Dataset

Dataset used for this research is Reuters-21578 collection which is publically available. The document in the Reuters collection were collected from Reuters newswire in 1987. This dataset contains structured information about newswire articles that can be assigned to several classes, therefore making this multi-label problem. It has a highly skewed distribution of documents over categories, where a large proportion of documents belong to few topics. The collection consists of 21,578 documents, including documents without topics and typographical errors. For this reason, a subset and split of the collection, referred to as "*ModApte*", is traditionally used. This dataset includes 9,603 documents for training and 3,299 for testing. This split assigns documents from April 7, 1987 and before to the training set, and documents from April 8, 1987 and after to the test set.

Additional step is to focus on the categories that have at least one document in the training and test set. After this, the dataset has 90 categories with 7769 of training documents and 3019 of test documents. The main reason why we will focus on this collection is that it is one of the most classic collection from text classification and it will allow us to compare our results with a large set of previously published results for several algorithms while being able to run it in our laptops.

4.2 Results:

To evaluate the performance of our method, we used precision and recall and F1 measures as shown in table(1) and table(2). Precision and recall decrease when the number of training sample decreases. For evaluating performance average categories we used micro-averaging and micro averaging.

	SVM	Decision Tree	kNN (k=3)
Precision	0.9455	0.9091	0.4624
Recall	0.8013	0.4303	0.3494
F1-measure	0.8674	0.5841	0.398

Table 1: Micro-averaging Table

	SVM	Decision Tree	kNN(k=3)
Precision	0.6493	0.1296	0.4899
Recall	0.3948	0.0552	0.1774
F1-measure	0.4665	0.0692	0.2319

Table 2: Macro-Averaging Table

5. Conclusion:

With the increasing number of documents arriving in internet, there has been explosion of online documents. Text categorization, the assignment of free text documents to one or more predefined categories based on their content, is an important component in many information management tasks. Typical text classification process consists of: preprocessing, indexing, dimensionality reduction and classification. In this paper three classification methods have been used: Support vector machine (SVM), Decision tree and kNN. All of them performed reasonably well, kNN performs well among all of them.

6. Future Work:

Several things can be done going forward to strengthen the analysis, including:

- More ML algorithms (e.g. a neural network, multinomial logistic regression, etc.) can be implemented
- LSA (Latent Semantic Analysis) can be implemented.

7. References:

[1] V.V. Gulin and A.B. Frolov “On the Classification of Text Documents Taking into Account Their Structural Features, 2016

[2] M.Sreelatha, M.Shashi, MP Teja, M.Rajashekar, K.Sasank, "Intrusion prevention by image based authentication techniques", Proceedings of the International Conference on Recent Trends in Information Technology, ICRTIT-2011

[3] S.B.Rahaman, M.Shashi, "Sequential Mining equips e-health with knowledge for managing Diabetes", Proceedings of the 4th International Conference on New Trends in Information Science and Service Science (NISS), 2010

[4] Saritha Vemulapalli, M.Shashi, " Design and Implementation of an Effective Web Server Log Pre-processing System", Proceedings of International Conference on Information System Design and Intelligent Applications, INDIA-2012

[5] Fabrizio Sebastiani “Machine Learning in Automated Text Categorization, 2001

[6] Sam Scott and Stan Matwin “Feature Engineering For Text Classification”, 1999

[7] V.V. Gulin”A comparative analysis of text document classification maethods”, 2011

[8] Jeonghee Yi and Neel Sundaresan “A classifier for semi-structure documents”, 2000