Data Mining: Document Classification using Naive Bayes Classifier

Ekta Jadon
Patel Group of Institution Indore
Ralamandal Indore (M.P.)

Roopesh Sharma
Patel Group of Institution Indore
Ralamandal Indore (M.P.)

ABSTRACT
In data mining, classification is the way to splits the data into several dependent and independent regions and each region refer as a class. There are different kinds of classifier uses to accomplish classification task. Moreover classification is bounded in case of classifying of text documents. The motives of the work which a present in the article is to evaluate multiclass document classification and to learn achieve accuracy of classification in the case of text documents. Naive Bayes approach is used to deal with the problem of document classification via a deceptively simplistic model. The Naive Bayes approach is applied in Flat (linear) and hierarchical manner for improving the efficiency of classification model. It has been found that Hierarchical Classification technique is more effective than Flat classification. It also performs better in case of multi-label document classification. In contrast to retrospect we observe significant increase in the generation of data each day. And hence with the advent of smarter technologies, data is required to be classified and sorted before framing out decisions from it. There are so many techniques available for classifying documents into various categories or labels. Data mining is the process of non-trivial extraction of novel, implicit, and actionable knowledge from large data sets.

Keywords
Data Mining, Mining Techniques, Classification, Document Classification, Naïve Bayes Classifier.

1. INTRODUCTION
There has been a significant increase observed in the generation of data each day. And hence with the advent of smarter technologies, data is required to be classified and sorted before framing out decisions from it. There are so many techniques available for classifying documents into various categories or labels. Data mining is the process of non-trivial extraction of novel, implicit, and actionable knowledge from large data sets [1]. Popularly referred to as Knowledge Discovery in Databases (KDD). It is the automated extraction of patterns representing knowledge implicitly stored in large databases, data warehouses and other massive information repositories. Standard data mining methods may be integrated with information retrieval techniques and the construction or use of hierarchies specifically for text data as well as discipline oriented term categorization systems (such as in chemistry, medicine, law, or economics) [2]. Text databases are databases that contain word descriptions for objects. These word descriptions are usually not simple keywords but rather long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

Classification is a data mining technique used to predict group membership for data instances. It is used to build structures from examples of past decisions that can be used to make decisions for unseen or future cases. Various classification techniques are used for web page classification, data classification etc. Coming on to Text categorization, it is the task of automated assigning of texts to predefined categories based on their content by learning models of categorized collections of documents. Text categorization is the primary requirement of Text Retrieval systems. As the amount of online text increases, the demand for text categorization for the analysis and management of text is increasing. Though the text is cheap, it is expensive to get the information that, to which class a text belongs to. This information can be obtained from automatic categorization of text at low cost, but building the classifier itself is expensive because it require a lot of human effort or it must be trained from texts which have themselves been manually classified. The task is usually performed in two stages: 1) the training phase and 2) the testing phase. During the training phase, sample documents are provided to the document classifier for each predefined category. The classifier uses machine learning algorithms to learn a class prediction model based on these labeled documents. In the testing phase, unlabelled documents are provided to the classifier, which applies its classification model to determine the categories or classes of the unseen documents. This training-testing approach makes the process of document classification a supervised learning task where unlabeled documents are categorized into known categories. Seeing the importance of text mining numerous text mining applications today use some form of text classification and this has fueled extensive research in the area. Efficient training and application and building understandable classifiers, are continuing fields of text classification research. Document (email) filtering and routing is a very important application in large corporate settings. Spam filtering is perhaps the most common application that impacts all of us, with Bayesian or rule-based spam filtering being necessary component in all client and server mail software. Web directories are an invaluable source of well categorized information on a broad variety of topics on the web, and though manually created for now, there are many applications using them for better information presentation and navigation.

2. LITERATURE SURVEY

2.1 Background
Data mining is the process of applying machine learning techniques for automatically or semi automatically analyzing and extracting knowledge from stored data. It can also be defined as technology which enables data analysis, exploration and visualization of very large databases using high level of abstraction. Data mining also known as Knowledge-Discovery in Databases, is the process of automatically searching large volumes of data for useful patterns that might otherwise be unknown. Some other definitions of data mining are: “The nontrivial extraction of implicit, previously unknown, and potentially useful information from data’’. “The science of extracting useful

information from large data sets or databases". Data mining has been used to extract interesting patterns or features of information from large amounts of data [3]. It includes analyzing the data (pre-processing of data), finding relevant frequent patterns and summarizing data (post-processing of results).Data mining [4,9] tasks are mainly divided into two categories.

- Predictive tasks: The main objective is to predict the value of an attribute based on already known values of other attributes.
- Descriptive tasks: The main objective is to derive patterns that lend information in order to derive the underlying relationships in data. Based on the main techniques used in data mining, Predictive Modeling refers to the process of building a model for the target variable as a function of other variables. The main two types of predictive modeling are Classification and Regression.

2.2 Related Study

The automated document classification may follow these approaches:

- **Rule Based Classification:** Here, the user groups the documents together, decide on categories, and formulates the rules that define those categories; these rules are actually query phrases [5]. Then a matching operator is applied to classify the documents. This approach is very accurate for small document sets. Results are always based on what user defines, since user write the rules. But, defining rules can be tedious for large document sets with many categories. As the document set grows, user may need to write correspondingly more rules.
- **Machine Learning Based Approach:** Here, the machine is trained using a set of sample documents that are already classified into the classes (training data) and as it learns it hence automatically create classifiers based on this data [6]. On one hand it shows a high predictive performance but on the other side we might require an effective training data set.
- **Supervised classification:** Supervised classification which requires an external mechanism (such as human feedback) to provide information on correct classification for documents [7, 8, 10]. Nonsupervised classification (also called document clustering) where the classification must be done entirely without reference to external information.

Naive Bayes Theorem: Easy calculation, rapid classification, independent properties, continuous property values difficult to be deal.

3. PROBLEM DEFINATION AND PRAPOSED SOLUTION

Document organization is the task of structuring documents into folders, which may be hierarchical or flat. Supervised algorithms assume that the category structure or hierarchy of a text database is already known. They require a training set of labeled documents and return a function that maps documents to the predefined class labels. Knowing the category structure in advance and generation of correctly labeled training set is very challenging. Manual document classification is known to be an expensive and time-consuming task.

When a document can belong to more than one class, it is called multi-labeled. Multi-labeled classification is a harder problem than just choosing one out of many classes. In addition to estimating the classes a document belongs to, the additional problem is to determine how many classes are relevant for the document in hand. Machine learning approaches to classification suggest the automatic construction of classifiers using induction over pre-classified sample documents. This thesis extends the approach to classify an unknown sample across multiple folders (classes).The main aim of this study is to compare and evaluate the performance of supervised technique for text document organization. For this purpose the traditional method of building a single classifier for all the classification work known as linear or flat classification is used and for improving the classification's performance hierarchical classification is done. A document is classified in top down fashion from root to leaf. The hope in such a hierarchical organization is that different features will be active in different parts of the hierarchy, and it should be possible to build high performance classifiers using such a hierarchical class structure. The problem statement for a document classifier defines the problem being solved by the classifier. It consists of two aspects: the document space and the set of document classes. The document space defines the range of input document samples. The training samples and the test samples are drawn from the document space. The set of document classes defines the possible outputs produced by the classifier and is used to label document samples.

4. PRAPOSED METHADOLOGY AND ARCHITECTURE

The work is implemented and demonstrates in Java. The Java programming language is ideal for large scale projects, it simplifies the development process. The Java I/O package and File handling is used in this project. Memory usage could be at a medium level, because the data has to be modeled in order to work with it efficiently.

4.1 Architecture

Figure 1 a simple architecture of Document classification systems is presented. There is a pool of documents corpus which represents the content at hand that can either be stored on disk containing the training and unlabeled data. There are standard preprocessing steps applied to this document corpus, followed by an appropriate choice of Vocabulary creation, and labeling systems. Classification models are chosen to operate on train-validation-test splits, and classifiers are learned, stored and performance of classification models is evaluated.

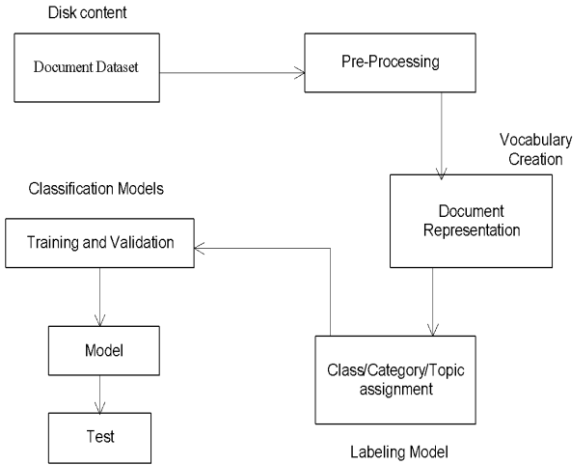

Figure 1: The standard document classification setup

5. IMPLEMENTATION AND TESTING

5.1 Implementation

The steps for preprocessing and classifying a new document can be summarized as follows:

- Remove periods, commas, punctuation, stop words. Collect words that have occurrence frequency more than once in the document. We called this collection of words as vocabulary.
- View the frequent words as word sets by matching the words which are in the vocabulary as well as training set documents.
- Search for matching word set(s) or its subset(containing items more than one) in the list of word sets collected from training data with that of subset(s) (containing items more than one) of frequent word set of new document.
- Collect the corresponding probability values of matched word set(s) for each target class.
- Calculate the probability values for each target class from Naïve Bayes categorization theorem.

5.2 Experimental Results

The work of classifying a new document depends on the word sets generated from training documents. So the number of training documents is important in formation of word sets used to determine the class of a new document. The greater number of word sets from training documents reduces the possibility of failure to classify a new document. In this project both the approaches implemented as discussed previously. The experiments are done with the datasets mentioned above in conjunction with the Naive Bayes classifier learning algorithm. For performance evaluation, the Accuracy, Precision and Recall metrics are used that were presented in the previous sections. The results obtained on performing the experiments are:

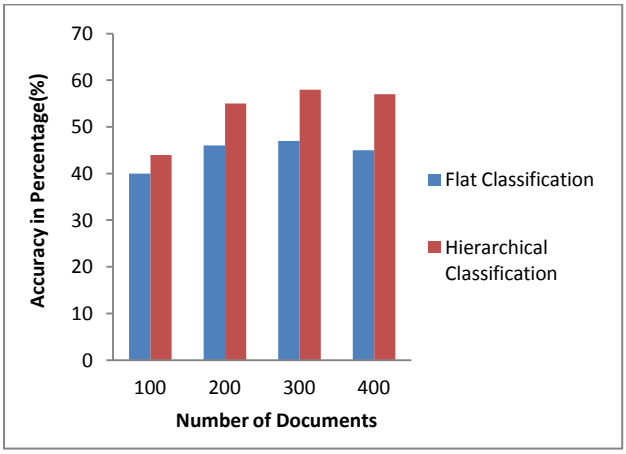

Figure 2: Accuracy in percentage for Different Number of Documents.

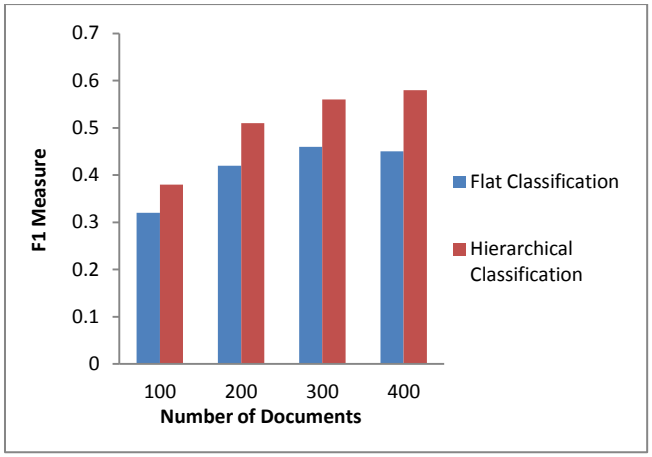

Figure 3: F1 Measure for Different Number of Documents.

6. CONCLUSION AND FUTURE WORK

In the recent years, the wide variety of available information services for use with smart phones and portable mobiles (tablets) devices has been provided a technique that dynamically classifies the quality of the sorted data. A document classifier is essential tool to classify various type documents being generated in the Big Data era. A very popular type of document classification scheme is the naïve Bayes classifier. The naïve Bayes scheme is based on performance classification which varies extensively depend on the method of extraction used in the documents. In last two decades, various researchers have experimented with the task of Text Document Classification using machine learning algorithms. The main Aim of all this work is to improve the efficiency and accuracy of classifier. The Naïve Bayes we have used performs well with even large datasets. Generating hierarchy of the available training classes and then applying classifier model can improve classification performance in most cases. It increases the performance of the classifier even for multi label classification in the field of multi class text classification. But the further research is needed to build statistically significant and meaningful hierarchy. Even for efficient text classification it is required to get strong hierarchy information which needs further investigation. Combining different classification approaches instead of single one along with hierarchic structure of classes also provide avenue for future search. Hence a solution is needed which confirms the selection ofa consistent cluster head, which can handle extreme traffic and maintain stability of cluster head.

7. REFERENCES

[1] Shweta Joshi. "Categorizing the Document Using Multi Class Classification in Data Mining", 2011 International Conference on Computational Intelligence and Communication Networks, 10/2011Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Nigam, Ayan, et al. "Classifying the bugs using multi-class semi supervised support vector machine." Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on. IEEE, 2012.

[3] Ponce, Julio, Alberto Hernndez, Alberto Ochoa, Felipe Padilla, Alejandro Padilla, Francisco lvarez, and Eunice Ponce de Le. "Data Mining in Web Applications", Data Mining and Knowledge Discovery in Real Life Applications, 2009.

[4] Survey of Classification Techniques in Data Mining, Thair Nu Phyu, Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2009, Vol. IIMECS 2009, March 18 - 20, 2009, Hong Kong.

[5] Alexandrin Popescul, Lyle H. Ungar, Steve Lawrence, David M. Pennock, Statistical relational learning for document mining. In Proceedings of IEEE International Conference on Data Mining (ICDM-2003), 2003, pages 275–282.

[6] S. B. Kim, H. C. Rim, D. S. Yook, H. S. Lim, Effective Methods for Improving Naïve Bayes Text Classifiers, In Proceeding of the 7th Pacific Rim International Conference on Artificial Intelligence, 2002, Volume, 2417.

[7] Yang Y., Liu X., A re-examination of text categorization methods. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR'99), 1999, pp. 42-49, ACM Press.

[8] Nigam, B., Ahirwal, P., Salve, S., & Vamney, S. (2011). Document classification using expectation maximization with semi supervised learning. arXiv preprint arXiv:1112.2028.

[9] Senkamalavalli, R, and T Bhuvaneshwari. "Data mining techniques for CRM", International Conference on Information Communication and Embedded Systems (ICICES2014), 2014.

[10] Jain, Rishabh, et al. "Performance evaluation of PSVM using various combination of kernel function for intrusion detection system." International Journal of Modeling and Optimization 2.5 (2012): 613.