Text Clustering and Classification Techniques: A Review

Sneh Lata, Mr. Ramesh Loar
Department of Computer Science and Engineering
Rao Pahlad Singh Group of Institutions, Balana, Mohindergarh

Abstract: Text classification is the task of automatically sorting a set of documents into categories from a predefined set. Text Classification is a data mining technique used to predict group membership for data instances within a given dataset. It is used for classifying data into different classes by considering some constrains. Instead of traditional feature selection techniques used for text document classification. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Automated Text categorization and class prediction is important for text categorization to reduce the feature size and to speed up the learning process of classifiers.

Keywords: *Naïve Bayesian, Clustering, Classification*

*****

Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. Highquality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

Text classification is the problem of automatically assigning zero, one or more of a predefined set of labels to a given segment of free text. The labels are to be chosen to reflect the "meaning" of the text. Selecting the appropriate set of labels may be ambiguous even for a human rater. When a machine is to try and mimic the human behavior, the algorithm will have to cope with a large amount of uncertainty coming from various sources. First of all, on a purely lexicographic level, human language is ambiguous per se, including words and word combinations with multiple senses which are disambiguated by the context. More importantly, the definition of meaning of a text is still vaguely defined, and a matter of debate.

One does not want to answer the question whether a computer has "understood" a text, but rather operationally whether it can provide a result which is comparable to what a human would provide (and find useful).

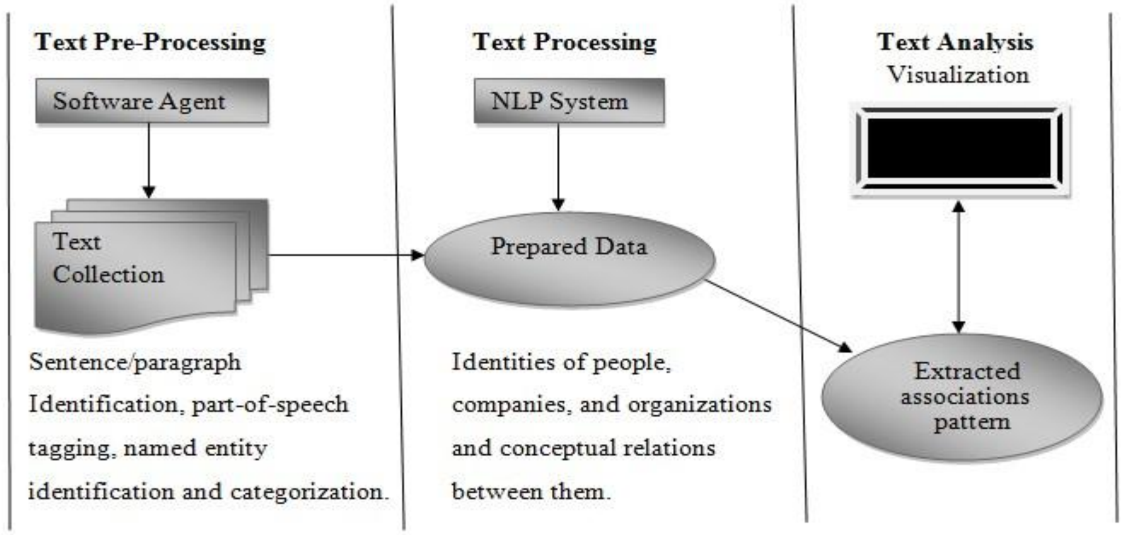

Figure 1.1 Processing in Text Mining

With the increasing availability of text documents in electronic form, it is of great importance to label the contents with a predefined set of thematic categories in an automatic way, what is also known as automated Text Categorization. In last decades, a growing number of advanced machine learning algorithms have been developed to address this challenging task by formulating it as a classification problem. Commonly, an automatic text classifier is built with a learning process from a set of pre labeled documents. Documents need to be represented in a way that is suitable for a general learning process. The most widely used representation is "the bag of words": a document is represented by a vector of features, each of which corresponds to a term or a phrase in a vocabulary collected from a particular data set. The value of each feature element represents the importance of the term in the document, according to a specific feature measurement. A big challenge in text categorization is the learning from high dimensional data.

On one hand, tens and hundreds of thousands terms in a document may lead to a high computational burden for the learning process. On the other hand, some irrelevant and redundant features may hurt predictive performance of classifiers for text categorization.

Literature Review

Several text categorization and classification techniques were proposed in past. In this section I review some of the important existing text categorization and classification techniques using Naïve Bayes as follows:

- Theoretical Framework of Feature Selection
- Selecting The Maximum Discriminative Features
- Text Classification Process
- Classifiers
- A Bayesian Classifier Using Class-Specific Features For Text Categorization
- Improvement In KNN Classifier (Imp-Knn) For Text Categorization
- Feature Selection and Feature Reduction

2.1 Theoretical Framework of Feature Selection [1]

They followed the Information Theory to select feature subsets that had maximum discriminative capacity for distinguishing the samples among two or more classes. They first introduced some concepts on information measures for binary hypothesis testing (also known as "two-class" classification) and present a new divergence measure for multiple hypothesis testing (i.e., for "multi-class" classification).

2.1.1 Divergence Measures for Binary Hypothesis Testing

A Bayesian approach was presented for detecting influential observations using general divergence measures on the posterior distributions. A sampling-based approach using a Gibbs or Metropolis-within-Gibbs method was used to compute the posterior divergence measures. Four specific measures were proposed, which convey the effects of a single observation or covariate on the posterior. The technique was applied to a generalized linear model with binary response data, an over dispersed model and a nonlinear model. The purpose of feature selection was to determine the most informative features which lead to the best prediction performance. Hence, it was natural to select those features that have the maximum discriminative capacity for classification, by minimizing the classification error (i.e., maximizing the KL-divergence or the J-divergence). The J-divergence was only defined for binary hypothesis.

They were next extend the J-divergence for multiple hypothesis testing (i.e., multi-class classification). The measure of discrimination capacity may not hold. The Information Theory to select feature subsets that had maximum discriminative capacity for distinguishing the samples among two or more classes. Some concepts on information measures for binary hypothesis testing (also known as "two-class" classification) and present a new divergence measure for multiple hypothesis testing (i.e., for "multi-class" classification).

2.1.2 Jeffrey's Multi-Hypothesis Divergence

The Jensen-Shannon (JS) divergence had the one that could be used to measure multidistribution divergence, in which the divergences of each individual distribution with a reference distribution were calculated and summed together. Unlike the J-divergence, the measure of discrimination capacity may not hold. In Sawyer presents a variant of Jdivergence with a variance-covariance matrix for multiple comparisons of separate hypotheses. The KL divergence of each detector was the measure of that discriminative capacity for discrimination; the new multi-distribution divergence was able to measure the discrimination capacity over all classes. Note that, since the JMH divergence was the sum of multiple J-divergences, it hold most properties of J-divergence. For example, JMH divergence was almost positive definite.

2.2Selecting The Maximum Discriminative Features [12]

2.2.1 Greedy Feature Selection Approach

Consider a binary (two-class) classification problem first and extend feature selection method to a general multiclass classification problem later. Unlike those existing feature selection methods which compute the score ("importance") of features based on the feature relevance to class, goal was to select the features that offer the maximum discrimination for classification. By doing so, one could expect an improved classification performance for text categorization. For a two-class classification problem, known that the J-divergence indicated the discriminative capacity of discriminating two classes' data under the MAP rule. To examined various values to evaluate the classification performance using those selected features. Hence, it was necessary to assign an importance score to each feature and rank the features. Here, start to propose a greedy approach to rank the features according to their discriminative capacity for naive Bayes. This approach started to determine which feature of the M features produces the maximum JMH-divergence.

The implementation of this greedy feature selection approach based on the maximum Jdivergence for two-class classification. This approach indicated that the discriminative capacity increases when more features was used for classification.

Note that the proposed greedy feature selection algorithm makes a locally optimal choice at each step to approximate the global optimal solution by selecting a feature with the maximum discriminative capacity for classification. The significance of this approach was that it started the best first feature and towards the optimal solution. This greedy approach could be considered as a wrapper approach.

However, unlike those existing wrapper approaches, this greedy approach did not need to evaluate the classification performance on a validation data set through retraining the classifier when a new feature was generated, because a closed form of KL-divergence.

However, this greedy approach still had the computation complexity.

2.3Text Classification Process [20]

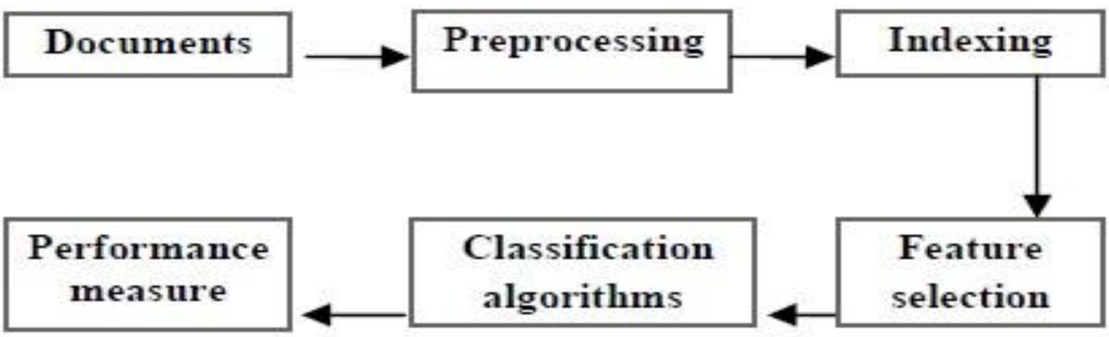

Figure 2.1 Document Classification Process

2.4Classifiers [9]

Classifiers had algorithms for performing the task of classification. Classification was a supervised form of machine learning aiming at identifying the category from a set of categories to which a selected text belongs [1]. It was done on the basis of a predefined training dataset. Various text classifiers had been proposed till date.

2.5Comparison of existing Text Categorization Techniques

Table: Comparison of existing Text Categorization Techniques

Text Categorization Techniques	Concept	Strengths	Weakness
Feature Selection	A novel and efficient feature selection framework based on the InformationTheory.	To rank the features with their discriminative capacity for classification.	Complex methodology, less accuracy as compared to proposed.
Maximum Discriminative Features (A Greedy Feature Selection)	Maximum Discriminative approach for automatic text categorization using class-specific features	Unlike conventional text categorization approaches, proposed method selects a specific feature subset for each class.	Feature selection was required which was not provided. Based on group clustering of records. Feature selection issues provided less accuracy.
Text Classification Process	Proposed combination different classification algorithms for text categorization.	Good detailing was provided for individual techniques. Hybrid technique were first time proposed in systems.	Comparison less. parameters was more to time and it was Complex implement require more Clustering Techniques provided.
Naïve Bayes Classifiers	The hypothesis and routine of two distinctive firstarrange probabilistic classifiers, both of which made the guileless Bayes supposition.	The multinomial model was observed to be consistently superior to the multi variant Bernoulli display.	As compare to proposed Naïve Bayes result was proper. Content oriented with basic advancements in text mining, but no methodologies were provided.
KNN Classifier	Proposed modified KNN classification algorithms for text categorization.	Modified KNN algorithm was less complex as compared to existing one. Nice explanation of points based on content mining with unstructured and semi structed data.	As compared to Naïve Bayes result are proper. Only content information was provided. No methodologies were given.

Conclusion

The Text Classification using analytical approach project proposed a design of the application that can effectively classify text files into appropriate folder depending upon the theme of the file, using the training data to model the classifier. This application automates the text classification process otherwise would take long time doing manually the same task. Text file are appropriately classified using this application. This application allows you to select the test data, training data. A similar concept can be used for different purposes like arrange your computer, classify various documents with various applications and analysethem. Using this system I have improvised the accuracy of system to 92% which provide much better data accuracy and classification for text documents. I have used 20 newsgroup data for simulating our classification algorithm. Modified Naive Bayes Algorithm is used so that more accuracy and less time complexity can be achieved than that of Naïve Bayes algorithm.

References

[1] Bo Tang, “Toward Optimal Feature Selection in Naïve Bayes for Text Categorization,” *IEEE: September 2016.*

[2] Karishma O. Borkar and Prof. Nutan M. Dhande,” A Review on Text Classification Techniques and Algorithms,” *International Journal of Research In Science & Engineering Volume 3 Issue 21st January 2017.*

[3] Karishma Borkar and Prof. Nutan Dhande,” Text Categorization Using Modified Classification Techniques,” *International Journal of Innovative Research In Science, Engineering and Technology, Volume 6, Issue 11, May, 2017.*

[4] Karishma O. Borkar and Prof. Nutan M. Dhande, ”Efficient Text Classification of 20 Newsgroup Dataset using Classification Algorithm,” *International Journal on Recent and Innovation Trends in Computing and Communication, volume 5,Issue 6,June 2017.*

[5] Karishma O. Borkar and Prof. Nutan M.Dhande*,”*.Text Categorization and Class Prediction Using Naïve Bayes Algorithm,” *International Conference on Modern Trends in Engineering Science Technology September, 2017.*

[6] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya,” Pre-processing Techniques for Text Mining - An Overview,” *International Journal of Computer Science & Communication Networks, Vol 5(1), July-2016 ISSN: 2249-5789.*

[7] Haibo He, “A Bayesian Classification Approach Using Class-Specific Features for Text Categorization,” *IEEE: June 2016.*

[8] Senthil Kumar B, Bhavitha Varma E., “A Survey on Text Categorization,” *IJARCCE, Vol. 5, Issue 8, 2016.*

[9] Aditya Jain, Jyoti Mandowara, “Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification,” *IJCA,Volume 6–No.2, March, 2016.*

[10] Shaifali Gupta, Reena Rani, “Improvement in KNN Classifier (imp-KNN) for Text Categorization,” *IJARCSSE-2016.*

[11] Xiaoli Guo, Huiyu Sun, Tiehua Zhou, Ling Wang, Zhaoyang Qu and Jiannan Zang, ”SAW Classification Algorithm For Chinese Text Classification,” *ISSN 2071-1050, 27 February 2015.*

[12] R. Balamurugan, Dr. S. Pushpa, “A Review on Various Text Mining Techniques and Algorithms,” *2nd International Conference on Recent Innovations in Science, Engineering and Management, 22 November, 2015.*

[13] Rajni Jindal, Ruchika Malhotra, Abha Jain,” Techniques for text classification: Literature review and current trends,” *Webology IEEE Volume 12, Number 2, December, 2015.*

[14] Y. Aphinyanaphongs, L. D. Fu, Z. Li, E. R. Peskin, E. Efstathiadis, C. F. Alfieri’s, and A. Statnikov, “A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization,” *Journal of the Association for Information Science and Technology, vol. 65, no. 10, pp. 1964–1987, 2014.*

[15] Antonio Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz, “Rolling out Text Categorization for Language LearningAssessment Supported by Language Technology,” *Springer International Publishing Switzerland, pp. 256–261, 2014.*

[16] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, "Text Mining Methods and Techniques," *International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.*

[17] Sebastian Raschka,” Naive Bayes and Text Classification–Introduction and Theory,” *IEEE, Oct 4, 2014.*

[18] S. Subbaiah, "Extracting Knowledge using Probabilistic Classifier for Text Mining," *Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, IEEE-2013.*

[19] Shuzlina Abdul-Rahman, Sofia Nita Mutalib, Nur Amira Khanafi, Azliza Mohd Ali, "Exploring Feature Selection and Support Vector Machine in Text Categorization,” 16th *International Conference on Computational Science and Engineering, IEEE2013.*

[20] Vandana Korde, C Namrata Mahender, “Text classification and classifiers: A Survey,” *International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.*

[21] M.Sukanyal, S.Biruntha, "Techniques on Text Mining," *International Conference on Advanced Communication Control and Computing Technologies, IEEE-2012.*

[22] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques,” *Morgan Kaufman publishers, San Francisco, Elsevier, 2011, pp. 285-351.*

[23] Nidhi, Vishal Gupta, "Recent Trends in Text Classification Techniques," *International Journal of Computer Applications (0975 – 8887) Volume 35– No.6, December 2011.*

[24] Vaishali Bhujade, N.J.Janwe, "Knowledge discovery in text mining techniques using association rule extraction," *International Conference on Computational Intelligence and Communication Systems, IEEE- 2011.*

[25] William B. Cavnar and John M. Trenkle,” N-Gram-Based Text Categorization,” *IJCSS volume 5, 2010.*