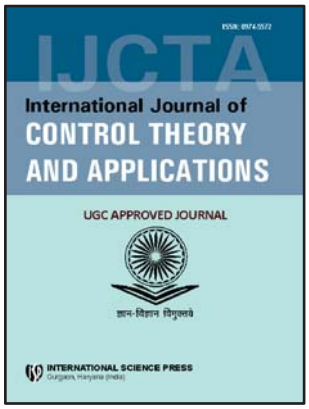


International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 30 • 2017

Naive Bayes Categorization Approach using Class-Specific Features

Kamalesh A.P[a] Arun Rajachandar R[a] and K. Deeba[b]

[a]*Department of Computer Science and Engineering, SRM University (Kattankulathur Campus).*

[b]*Assistant Professor (O.G), Department of Computer Science and Engineering, SRM University (Kattankulathur Campus).*

Abstract: The automaticcategorization of text into prebuilt labels has witnessed a thriving Interest in recent years, due to large number of documents available in digitized format. Hence, there is a need for organizing and ranking data. Efficacy of Naïve Bayes classifier is demonstrated by incorporating it in the Search Engine. This paper deals with datasets related to Bikes. It aids the user of this application to find the bikes that suit their budget and tastes. A single dataset consists of attributes of Bike such as its colour, Price, Model number, Special features and performance parameters. Naive Bayes classifier is used to find similar datasets associated with the input Keyword. After that, the similarly matched datasets are ranked from most similar to the least similar, based on the probability score of each set of attributes associated with the dataset. Therefore, this method is found to be reliable to accomplish the objective of this application.

Keywords: *Machine learning, query classifier, probabilistic data, searching, query processing, ranking*

1. INTRODUCTION

Nowadays, it has been witnessed that myriads of documents are in digitized form, because of which Information Retrieval methodologies have gained an important position.Large number of advanced Machine Learning Algorithms have been deployed to address this challenging task over past few decades. One such process is Text Categorization. Text Categorization assigns one or more classes to a document according to their content. Text Categorization process was in nascent stages during early 60's but it became a major sub domain of Information Systems discipline due to availability of powerful hardware and its application in various fields in the contemporary world.

Naïve Bayes classification technique is based on concept of Bayes' Theorem with an assumption of independence among predictors. The classifier assumes that the presence of particular feature in a class is unrelated to the presence of any other feature. Documents are usually represented by 'bag-of-words'. Naive Bayes model is easy to build and particularly useful for very large Datasets.

This type of learning is called Supervised learning because a supervisor serves as a teacher directing the learning process. The main job of a classifier is to map document to classes. In this paper, Naïve Bayes Classifier is used to assign probability values to the attributes for easy retrieval of documents.

Paper formation is as follows: Section 2 is for Related works and Section 3is for Proposed work. Section 4 contains the conclusion.

2. RELATED WORKS

Bo Tang, Haibo He, Paul M. Baggenstoss and Steven Kay [1] worked on Bayesian classification approach using class specific feature for the purpose of categorizing Text. They built a Bayesian classification rule based on Baggenstoss PDF Projection Theorem to reconstruct PDF's in raw data space from class specific PDF's in low dimensional feature space. The advantage of their method is that it can include existing feature selection criteria easily.

Lucie Skorkovski, Zbynek Zajic, Ludek Muller [2] worked on multi-label text classification of articles where the classifier must decide whether a document does or does not belong to each topic from the existing topic set.Li-Qing Qiu Ru-Yi Zhao Gang Zhou Sheng-Wei Yi [3] worked on improving the algorithm of feature selection. They adopted two step feature selection approach for the TC. Huan Liu and Lei Yu [4] worked on to demonstrate how existing feature selection algorithms can be incorporated into a meta algorithm that can make use of individual algorithms.George Forman [5] worked on Bi-Normal Separation which is a new feature selection metric. His work also explores new evaluation methodology which takes into account the problem of selecting metrics that have best chance of obtaining the best performance on a given Dataset. Y.H. Li and A.K. Jain [6] explored different methods to classify documents from the web. With their adaptive classifier, they were able to achieve accuracy of 83%.

3. PROPOSED APPROACH

The proposed work makes use of Naïve Bayes classifier component in the Search Engine to effectively retrieve required information from a multitude of Datasets. The Datasets used were pre-processed in order to enhance feature extraction.

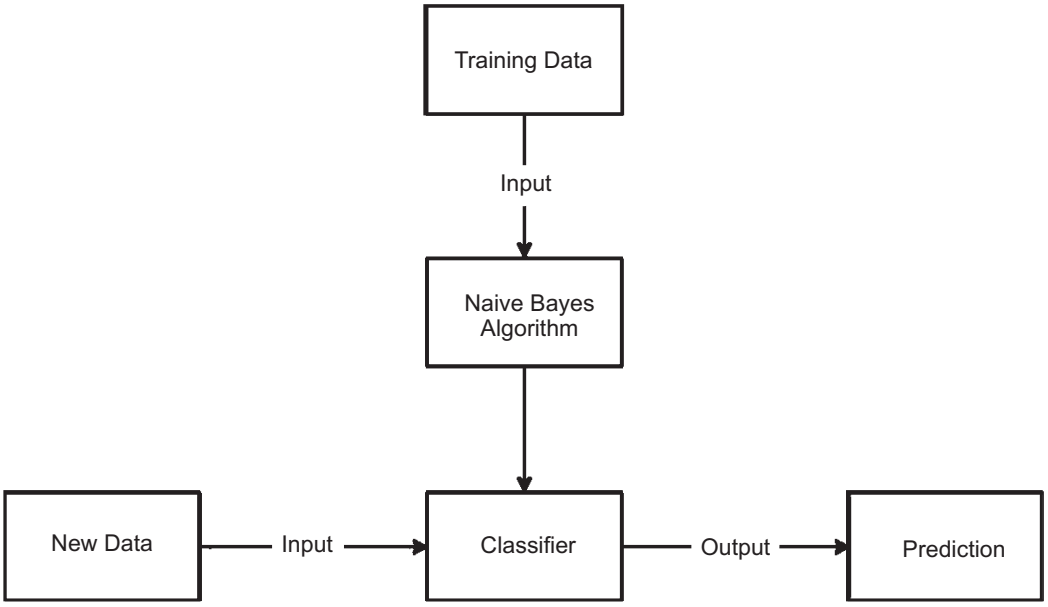


Figure 1

3.1. Pre-processing

Certain pre-processing tasks are usually performed before the datasets in the collection are used for retrieval. The tasks to be performed are hyphens, stemming, handling of digits, stop word removal, and special symbols. HTML tags and special characters in the collected documents are removed. Then the contents of the documents are segmented into sentences. Content words for each sentence is extracted only using nouns.

Stopwords: They are frequently occurring and insignificant words in a language that help construct sentences but do not represent any content of the documents.

Stemming: Stemming refers to the process of reducing words to their stems or roots.

Hyphens: Breaking hyphens are usually applied to deal with inconsistencyof usage.

Handling of Digits: Numbers and terms that contain digits are removed from the Datasets for non-numeric parameter.

3.2. Creating Training Datasets

A single dataset about the Bike contains its General information, it's performance information and feature information. General attributes of the vehicle are its ID, Company name, Model name, Price, Colour, Year of purchase. Performance attributes of the vehicle are the engine power and torque. Feature attributes of the vehicle are the presence or absence of Electric start, Alloys, Tripmeter and Passbeam.

3.3. Indexing

In order to speed up the search of frequently queried keywords in the Search Engine application, an Index is used for the quick retrieval of Information from the Source.

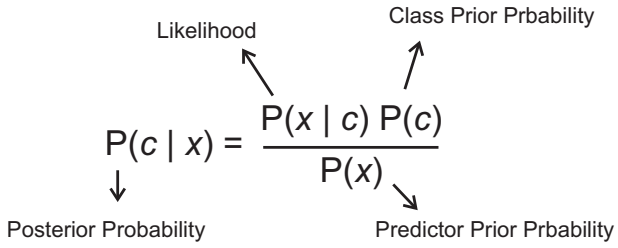
3.4. Naïve Bayes Algorithm:

Input:

1. Documents for a given training datasets

Procedure :

1. Begin
2. Form a reference class which consists of all documents for each class i = 1:N do
   a) Calculate the probability value of each attribute based on a specific criterion using the formula,



$$P(c \mid x) = P(x_1 \mid c) \times P(x_2 \mid c) \times P(x_2 \mid c) \times \ldots \times P(x_n \mid c) \times P(c)$$

   b) In descending order, rank the datasets from the highest similarity to the lowest similarity based on the probability value of each feature associated with the dataset.

   c) Choose the first K datasets that are above the threshold value and display them

Output:

1. Ranked list of similar datasets associated with the inputted keyword is displayed.

3.5. Experimental Results

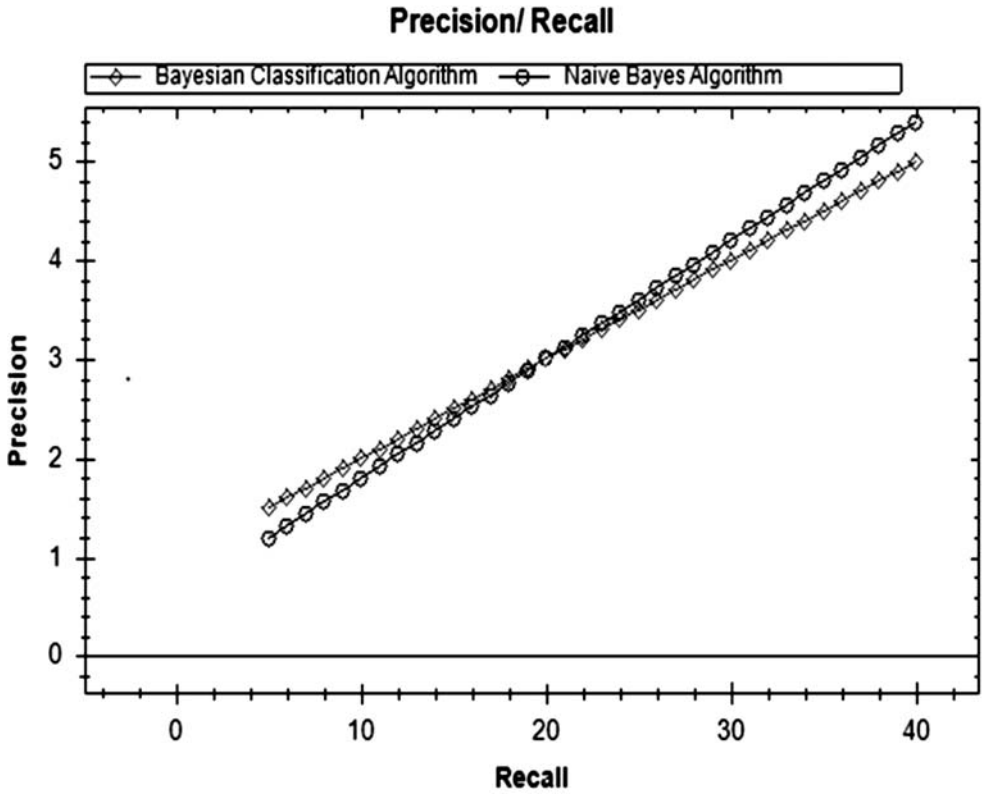


Figure 2

Table 1

Threshold	*Keyword*			*Bayesian Classification*	*Naïve Bayes*
	Name	*Model*	*Price*		
0.6	Hero Honda	Spender	51,608	5/10 (50%)	8/10(80%)
	TVS	Apache	92,925	5/10(50%)	7/10(70%)
	Bajaj	Discover	65,952	7/10(70%)	9/10(90%)
	Suzuki	Access	62,881	4/10(40%)	7/10(70%)
0.7	Hero Honda	Spender	51,608	5/10(50%)	7/10(70%)
	TVS	Apache	92,925	5/10 (50%)	8/10(80%)
	Bajaj	Discover	65,952	8/10(80%)	9/10(90%)
	Suzuki	Access	62,881	6/10(60%)	9/10(90%)

Threshold	*Keyword*			*Bayesian Classification*	*Naïve Bayes*
	Name	*Model*	*Price*		
0.8	Hero Honda	Spender	51,608	4/10(40%)	6/10(60%)
	TVS	Apache	92,925	5/10 (50%)	8/10(80%)
	Bajaj	Discover	65,952	7/10(70%)	9/10(90%)
	Suzuki	Access	62,881	5/10(50%)	6/10(60%)
0.9	Hero Honda	Spender	51,608	5/10 (50%)	8/10(80%)
	TVS	Apache	92,925	4/10(40%)	7/10(70%)
	Bajaj	Discover	65,952	7/10(70%)	9/10(90%)
	Suzuki	Access	62,881	5/10(50%)	7/10(70%)

$$\text{Precision} = |(\{\text{relevant documents}\} \cap \{\text{retrieved documents}\})| / |\{\text{retrieved documents}\}|$$

Precision refers to the closeness of two or more measurements to each other

$$\text{Recall} = =|(\{\text{relevant documents}\} \cap \{\text{retrieved documents}\})| / |\{\text{relevant documents}\}|$$

Recall in information retrieval is the subset of the documents that are related to the query which are proficiently retrieved.

As seen in the graph, the proposed method outperforms the existing method for higher number of Datasets.

4. CONCLUSION AND FUTURE WORK

Identification of appropriate match for queried keyword is a kind of Text Classification problem. This study proposes a method for identification of Bikes (given a keyword from the User) from the Source database. The salient features are extracted from Predefined Datasets and Naïve Bayes classifier is used to find the similar datasets associated with the inputted Keyword (*i.e.* Bikes). As a future work, this can be extended for other vehicular types.

REFERENCES

[1] Bo Tang, Haibo He, Paul M. Baggenstoss, Steven Kay: A Bayesian Classification Approach Using Class-Specific Features for Text Categorization. IEEE Trans. Knowl. Data Eng. 28(6): 1602-1606 (2016)

[2] Lucie Skorkovská and Zbyněk Zajíc and Luděk Müller: Comparison of Score Normalization Methods Applied to Multi-Label Classification. IEEE International Symposium on Signal Processing and Information Technology, Institute of Electrical and Electronics Engineers (IEEE), Noida, India, 2014.

[3] Li-Qing Qiu, Ru-Yi Zhao, Gang Zhou, Sheng-Wei Yi:An Extensive Empirical Study of Feature Selection for Text Categorization. ACIS-ICIS 2008: 312-315.

[4] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502, 2005.

[5] G.Forman, "An extensive empirical study of feature selection metrics for text classification," The Journal of machine learning research, vol. 3, pp. 1289–1305, 2003.

[6] Y. H. Li and A. K. Jain, "Classification of text documents," The Computer Journal, vol. 41, no. 8, pp. 537–546, 1998.