INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

SURVEY ON TEXT CLASSIFICATION TECHNIQUES/ALGORITHMS

MR. JITENDRA P. PATIL[1], DR. J. W. BAKAL[2]

1.PG Student, Terna Engg. College, Nerul, Navi Mumbai, India.

2.SSJCOE, Sonarpada, Dombivli(E) Mumbai, India

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: With the explosive growth of the textual information from the electronic documents and World Wide Web, proper classification of such enormous amount of information into our needs is a critical step towards the business success. Recently, numerous research activities have been conducted in the field of document classification, particularly applying in spam filtering, emails categorization, website classification, formation of knowledge repositories, and ontology mapping. Document classification is a growing interest in the research of text mining. Correctly identifying the documents into particular category is still presenting challenge because of large and vast amount of features in the dataset. In regards to the existing classifying approaches, Naïve Bayes is potentially good at serving as a document classification model due to its simplicity. The aim of this is to highlight the performance of employing Naïve Bayes in document classification.

Keywords: Text classification, Filtering, Structured and unstructured data, Naïve Bayes.

PAPER-QR CODE

Corresponding Author: MR. JITENDRA P. PATIL

Access Online On:

www.ijpret.com

How to Cite This Article:

Jitendra Patil, IJPRET, 2014; Volume 2 (9): 920-934

INTRODUCTION

The term “Text document” refers to written, printed, or online document that presents or communicates narrative or tabulated data in the form of an article, letter, memorandum, report, etc. The Text expresses a vast range of information, but encodes the information in a form that is difficult to decipher automatically. The information available in structured and unstructured form. Unstructured means data that does not reside in fixed locations. The term generally refers to free-form text, which is ubiquitous. Data that resides in fixed fields within a record or file is referred as a structured data. Relational databases and spreadsheets are examples of structured data.

Unstructured information refers to computerized information that either does not have a data model or has one that is not easily usable by a computer program. The term distinguishes such information from data stored in fielded form in databases or annotated in documents. However, data mining deals with structured data, whereas text presents special characteristics and is unstructured.

How these documents can be properly annotated, presented and classified? Is a question of concern? Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important.

In data mining, Machine learning is often used for Prediction or Classification. Prediction means extracting information from data and using it to predict future trends and behavior patterns. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes. Classification involves finding rule that partition the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classifications analyze the training data set and construct a model based on the class label. The goal of classification is to build a set of models that can correctly predict the class of the different objects.

Machine learning is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets.

Structured Data - Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets. Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).

Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data.

Structured data is often managed using Structured Query Language (SQL) – a programming language created for managing and querying data in relational database management systems. Originally developed by IBM in the early 1970s and later developed commercially by Relational Software, Inc. (now Oracle Corporation).

Structured data was a huge improvement over strictly paper-based unstructured systems, but life doesn't always fit into neat little boxes.

Unstructured Data - Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpage's, pdf files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Software that creates machine-process able structure exploits the linguistic, auditory, and visual structure inherent in all forms of human communication. Algorithms can infer this inherent structure from text, for instance, by examining word morphology, sentence syntax, and other small- and large-scale patterns. Unstructured information can then be enriched and tagged to address ambiguities and relevancy-based techniques then used to facilitate search and discovery. Examples of "unstructured data" may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects (e.g. in files or documents ...) that themselves have structure and are thus a mix of structured and unstructured data, but collectively this is still referred to as "unstructured data".

Fig.1 Unstructured data

Text mining algorithms

An increasing amount of information becomes available in the form of electronic documents. There is need to intelligently process such texts makes to understand depth of knowledge from text. The lacking depth of knowledge understanding methods such as information extraction (IE) is useful.

Text classification by Decision Tree (DT):

Decision tree is introduced in the text classification. Class of representation is used for classify the data in this algorithm. Thus a text-classification tree is a binary tree where each internal node is labelled with a string and each leaf is labelled with a class name. Each text classification tree classifies an input string. An input string determines a unique path from the root to a leaf: at each internal node the right (respectively left) edge to a child is taken if the input string contains the string labelled at that internal node as a substring (respectively does not contain the labelled string). The class that the input string is classified into is the class at the leaf reached.

Advantages:-

1. The algorithm is robust for classification noise contained in the sample

2. The algorithm does not need any natural language processing technique.

3. The algorithm constructs a text-classification tree in a top down manner started from the root node inductively from the given sample (Top-Down Induction of Decision Tree).

4. Due to tree structure decision is taken fastly.

Disadvantages:-

1. Decision trees are easy to use compared to other decision-making models, but preparing decision trees, especially large ones with many branches, are complex and time-consuming affairs.

2. Decision trees moreover, examine only a single field at a time, leading to rectangular classification boxes. This may not correspond well with the actual distribution of records in the decision space.

3. Cost of decision tree implementation is very high for good decisions.

4. Analytical area of decision tree is imitated.

Text classification by SVM:

The dimension of the text data is huge for the text documents are usually represented with the vector space model. Thus, it is greatly time-consuming to perform existed text categorization methods. Moreover, it is almost unimaginable to store and enquire high-dimensional text data. To improve the executing efficiency of classification methods, they present a classification algorithm based on nonlinear dimensionality reduction techniques and support vector machines. In the procedure, the ISOMAP algorithm is firstly executed to reduce the dimension of the high-dimensional text data.

Then the low-dimensional data are classified with a multi-class classifier based single-class SVM. Experimental results demonstrate that the executing efficiency of categorization methods is greatly improved after decreasing the dimension of the text data without loss of the classification accuracy.

After pre-processing and transformations, a machine learning algorithm is used for Learning how to classify documents, i.e. creating a model for input-output mappings. A linear model is a model that uses the linear combination of feature-values. Positive/negative discrimination is based on the sign of this linear combination.

SVMs are a generally applicable tool for machine learning. Suppose we are given with training examples **x***i*, and the target values yi_{-1,1}. SVM searches for a separating hyper plane, which separates positive and negative examples from each other with maximal margin, in other words, the distance of the decision surface and the closest example is maximal.

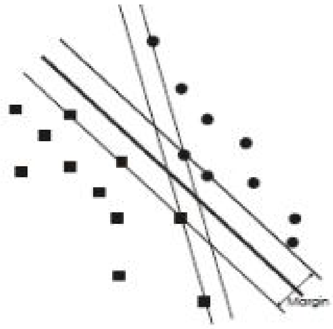

Fig.2. SVM decision margin

Advantages:-

1. Supported vector machine combines nonlinear dimensionality reduction techniques for text classification.

2. In this method, high-dimensional text data are firstly mapped into a low dimensional space with ISOMAP due to this time cost of training and testing of the classifier is greatly reduced.

3. It decreases the dimension without a loss of classification accuracy.

4. Decreases the memory requirements.

5. It can take high-dimensional input space.

6. Most text categorization problems are linearly separable.

Disadvantages:-

1. A text categorization system may have lots of parameters. So need to consider all of them. Often it is not clear, how to set them, it heavily depends on the nature of the problem. Make distinct set of parameter is time consuming process.

2. To lower the dimensionality need to dimensionality reduction techniques, which is also time consuming technique?

3. Due dimension reduction there may be chance of data loss.

4. This is fully depending on vector space that we decides, if parameter get wrong then accuracy of the system also get affected.

Text classification by Neural Network:-

Web document classification is implemented by using wavelet neural network. The structure of web classification mining system based on wavelet neural network is given. With the ability of strong nonlinear function approach and pattern classification and fast convergence of wavelet neural network, the classification mining method can truly classify the web text information.

The neural network is a high nonlinearity dynamics system, and the method of searching problem generally uses the gradient descent method and the random search method. The error back propagation BP network based on the gradient descent method is a new technique in recent years, its ability to approach nonlinear function has been proved in theory also have been validated in actual applications. But the BP network has some problems such as converge to local minimum and slow converge speed. Wavelet neural network is new kinds of network based on the wavelet transform theory and the artificial neural network. It utilizes the good localize character of the wavelet transformation and combines the self-learning function of the neural network. So it overcomes the disadvantages of BP network and has the ability of strong self adaption learning and nonlinear function approach. Meanwhile the wavelet neural network has the simple implementation process and fast convergence rate.

Advantages:-

1. This classification method shows the results feasible and effective.

2. Wavelet neural network can enhance the converging speed and the classification accuracy to a great extent.

3. It does follow pattern classification technique so that time consumption and accuracy is more.

4. It has both the advantages of wavelet analysis and neural network.

5. Speed of classification is fast.

Disadvantages:-

- Neural network requires training to their nodes. So that it is time consuming process.
- Due pre-requisite training cost of implementation also increases.
- It works on the trained patterns so that it does not work for other requirement. It needs training to node to get that.

Text mining With Naïve bayes (Proposed system)

The document representation is the pre-processing process that is used to reduce the complexity of the documents and make them easier to handle, which needs to be transformed from the full text version to a document vector. Text representation is the important aspect in documents categorization that denotes the mapping of a document into a compact form of its content. A major characteristic of the text classification problem is the extremely high dimensionality of text data, so the number of potential features often exceeds the number of training documents. The documents have to be transformed from the full text version to a document vector which describes the contents of the document.

Text classification is an important component in many informational management tasks, however with the explosive growth of the web data, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired. Dimensionality reduction (DR) is a very important step in text categorization, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy, the current literature shows that lot of work are in progress in the pre-processing and DR, and many models and techniques have been proposed. DR techniques can classify into Feature Extraction (FE) approaches and feature Selection (FS), as discussed bellow.

- Feature Extraction
- Feature selection
- Semantic and ontology based document representation
- Learning algorithm

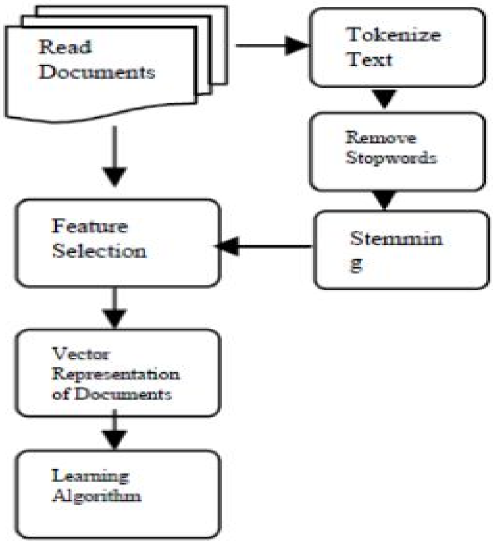

Fig.3. Proposed system

Feature extraction:-

The process of feature extraction is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming. Feature Extraction is fist step of pre processing which is used to presents the text documents into clear word format. Removing stops words and stemming words is the pre-processing tasks. The documents in text classification are represented by a great amount of feature and most of then could be irrelevant or noisy [8]. Dimension reduction is the exclusion of a large number of keywords, base preferably on a statistical criterision, to create a low dimension vector. Dimension Reduction techniques have attached much attention recently science effective dimension reduction make the learning task such as classification more efficient and save more storage space. Commonly the steps taken please for the feature extractions are: Tokenization: A document is treated as a string and then partitioned into a list of tokens. Removing stop words: Stop words such as “the”, “a”, “and”… etc are frequently occurring, so the insignificant words need to be removed. Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form eg. Connection to connect, computing to compute etc.

Feature selection:-

After feature extraction the important step in pre-processing of text classification, is feature selection to construct vector space or bag of words, which improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics. The main idea of FS is to select subset of feature from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Hence feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

There are mainly two types of feature selection methods in machine learning; wrappers and filters. Wrapper are much more time consuming especially when the number of features is high. As opposed to wrappers, filters perform feature selection independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class. We need to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weight each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories. Some of the recent literature shows that works are in progress for the efficient selection of the feature selection to optimize the classification process. A new feature selection algorithm is presented in, that is based on ant colony optimization to improve the text categorization.

We in developed a new feature scaling method, called class–dependent–feature–weighting (CDFW) using naive Bayes (NB) classifier. Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index. A good feature selection metric should consider problem domain and algorithm characteristics. The authors in focus on document representation and demonstrate that the choice of document representation has a profound impact on the quality of the classifier. In the authors present significantly more efficient indexing and classification of large document repositories, e.g. to support information retrieval over all enterprise file servers with frequent file updates.

Semantic and ontology based document representation:-

Ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. It is used to reason about the objects within that domain. Ontology is the explicit and abstract model representation of already defined finite sets of terms and concept, involved in knowledge management, knowledge engineering, and intelligent information integration.Web Ontology Language (OWL) is the ontology support language derived from America DAPRA Agent Markup Language (DAML). Ontology has

been proposed for handling semantically heterogeneity when extracting informational from various text sources such as internet. Machine learning algorithms automatically builds a classifier by learning the characteristics of the categories from a set of classified documents, and then uses the classifier to classify documents into predefined categories. However, these machine learning methods have some drawbacks:

(1) In order to train classifier, human must collect large number of training text term, the process is very laborious. If the predefined categories changed, these methods must collect a new set of training text terms.

(2) Most of these traditional methods haven't considered the semantic relations between words. So, it is difficult to improve the accuracy of these classification methods.

(3) The issue of translatability, between one natural language into another natural language, identifies the types of issues that machine understanding systems are facing.

These type of issues are discussed in the literature, some of these issues may be addressed if we have machine readable ontology, and that's why this is a potential area for research. During the text mining process, ontology can be used to provide expert, background knowledge about a domain. In the author concentrates on the automatic classification of incoming news using hierarchical news ontology, based on this classification on one hand, and on the users' profiles on the other hand. A novel ontology-based automatic classification and ranking method is represented in where Web document is characterized by a set of weighted terms, categories is represented by ontology. In the author presented an approach towards mining ontology from natural language. In the author presented a novel text categorization method based on ontological knowledge that does not require a training set. An automatic document classifier system based on Ontology and the Naïve Bayes Classifier is proposed in Ontology have shown their usefulness in application areas such as knowledge management, bioinformatics, e-learning, intelligent information integration, information brokering and natural-language processing and the positional and challenging area for text categorization.

Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and Proper Nouns. Using statistics-backed technology, these words are then compared to your taxonomy (categories) and grouped according to relevance. According to the statistical techniques are not sufficient for the text mining. Better classification will be performed when consider the semantic under consideration, so the semantically representation of text and web document is the key challenge for the documents classification, knowledge and trend detection.

Learning Algorithm:-

Bayesian Theorem:-

Bayesian theorem is data mining algorithm. The Bayesian belief network was first introduced by Cooper and Herskovits (1992). Bayesian belief networks are statistical techniques in data mining. Bayesian networks are very effective for modeling situations where some information is already known and incoming data is unsure or partially unavailable. The goal of using Bayes rules is to correctly predict the value of designated discrete class variable given a vector of predictors or attributes. In 1993, Sam maes et al has been suggested BN for credit card fraud detection. For the purpose of fraud detection, two Bayesian networks hypothesis for describing the behavior of user are constructed. First,

Bayesian network is constructed to model behavior that has been assumed the user is fraudulent and second model under the assumption that the user is a legitimate. Bayesian networks allow the integration of expert knowledge, which we used to initially set up the models Bayesian Network needs training of data to operate and require high processing speed. BN is more accurate and much faster than neural network, but BBNs are slower when applied to new instances.

Bayes' Theorem is a theorem of probability theory originally stated by the Reverend Thomas Bayes. It has been used in a wide variety of contexts, ranging from marine biology to the development of "Bayesian" spam blockers for email systems. In the philosophy of science, it has been used to try to clarify the relationship between theory and evidence. Many insights in the philosophy of science involving confirmation, falsification, the relation between science and pseudo since, and other topics can be made more precise, and sometimes extended or corrected, by using Bayes' Theorem. These pages will introduce the theorem and its use in the philosophy of science.

Following are the formulas used –

Baye's Formula –

Let$\boldsymbol{B_1}$, $\boldsymbol{B_2}$, $\boldsymbol{B_3}$,……… $\boldsymbol{B_n}$ be a partition of Ω (space) such that P($\boldsymbol{B_n}$) ≠ 0 for any

n = 1, 2, 3… and let P (A) ≠ 0.Then,

$$\boldsymbol{P(A|B_n)} = \frac{\mathbf{P(B_n|A)P(B_n)}}{\sum \mathbf{P(B_n|A)P(B_n)}}$$

Where, n =1, 2, 3, 4…

Naïve Bayes:-

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". An overview of statistical classifiers is given in the article on Pattern recognition.

In simple terms, a naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive

Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of Naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

NB is a popular method for document classification due to its computational efficiency and relatively good predictive performance. NB is very closely related to the simple centroid-based classifier and compares the two methods empirically.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Mathematically it is represented as –

n = 1, 2, 3... and let P (A) ≠ 0.Then,

$$\boldsymbol{P(A|B_n)} = \frac{\mathbf{P(B_n|A)P(B_n)}}{\sum \mathbf{P(B_n|A)P(B_n)}}$$

By using Bayesian network and feature variable Naïve Bayes classifies the data with following formula

$$\mathrm{classify}(f_1,\dots,f_n) = \underset{c}{\mathrm{argmax}}\ p(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c).$$

Here f1...fn are the features set, we can calculate feature probability overclass for 1 to n scope.

Fig. 4. Process

CONCLUSION

Naïve Bayes classifier has been discussed as the best document classifier, which satisfies the literature result, through the implementation of different feature selection and classifier.

There are many words in the documents, therefore when we captured the terms from these documents, thousand of terms are found. However, there are some terms that are usefulness and uninteresting to the results, it is then important to discover and interpret which features are useful and critical.

REFERENCES

1. Text Categorization and Support Vector Machines, István Pilászy, Department of Measurement and Information Systems, Budapest University of Technology and Economics.

2. Text Categorization and Support Vector Machines: Learning with many relevant features, Thorsten Joachims, University Dortmund, Germany.

3. NTC (Neural Text Categorizer): Neural Network for Text Categorization, Taeho Jo School of Information Technology & Engineering, Ottawa University, Ontario, Canada, Vol 2, issue 2, April 2010.

4. Categorization of Genomics text based on Decision tree, Rocio Guillen, California university.

5. Is Naïve Bayes a Good Classifier for Document Classification?, S.L. Ting, W.H. Ip, Albert H.C. Tsang, Vol. 5, No. 3, July, 2011

6. Naive Bayes for Text Classification with Unbalanced Classes, Eibe Frank1 and Remco R. Bouckaert1,2,

7. Naïve Bayes, http://www.wikipedia.com/Naive %20Bayes

8. Aurangzeb Khan , Baharum B. Bahurdin, Khairullah Khan, An Overview of E-Documents Classification, 2009 International Conference on Machine Learning and Computing.

9. Fabrizio sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys

10. Bayesian Theorem, http://www.wikipedia.com/bayesian_theorem

11. George Forman, Evan Kirshenbaum, Extremely Fast Text Feature Extraction for Classification and Indexing, HP Laboratories

12. Pingpeng Yuan, Yuqin Chen, Hai Jin, Li Huang, MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification, IEEE International Workshop on Semantic Computing and Systems

13. Eibe Frank and Remco R. Bouckaert, Naive Bayes for Text Classification with Unbalanced Classes.