SURVEY ON EMAIL CLASSIFICATION TECHNIQUES/ALGORITHMS

[1]Prof. Suchita Walke,[2]Miss.Monali Patil

[1]Head of Computer Department,[2]PG Student
[1]HOD (Computer) YTCEM, Mumbai, India
[2]Student of YTGOIFOE, Mumbai, India

Abstract— **In lot of communication e-mail plays important role. E-mail system is used for communication in all type of organizations. It is self-evident that e-mail has become a central means for the discussion of engineering work and sharing of digital assets that define the product and its production process. Engineering communication research has shown that the volume of communication is indicative of progress being made within an engineering project.**

So that e-mail conversations increases as product grows and data in communication also increases. It get difficult to handle the data at emails. So need of classification of emails. Here we have studied different classification techniques which help us to classify the large email data.

Index-TermsEmailclassification, Filtering, Structured and unstructured data, Naïve Bayes

I. INTRODUCTION

Electronic mail, most commonly referred to as email or e-mail since c.1993 is a method of exchanging digital messages from an author to one or more recipients. Modern email operates across the Internet or other computer networks. Some early email systems required that the author and the recipient both be online at the same time, in common with instant messaging. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver, and store messages. Neither the users nor their computers are required to be online simultaneously; they need connect only briefly, typically to a mail server, for as long as it takes to send or receive messages. Historically, the term electronic mail was used generically for any electronic document transmission. For example, several writers in the early 1970s used the term to describe fax document transmission. As a result, it is difficult to find the first citation for the use of the term with the more specific meaning it has today.An Internet email message consists of three components, the message envelope, the message header, and the message body. The message headercontains control information, including, minimally, an originator's email address and one or more recipient addresses. Usually descriptive information is also added, such as a subject header field and a message submission date/time stamp.Email is an information and communications technology. It uses technology to communicate a digital message over the Internet. Users use email differently, based on how they think about it. There are many software platforms available to send and receive. Popular email platforms include Gmail, Hotmail, Yahoo! Mail, Outlook, and many others.Network-based email was initially exchanged on the ARPANET in extensions to the File Transfer Protocol (FTP), but is now carried by the Simple Mail Transfer Protocol (SMTP), first published as Internet standard 10 (RFC 821) in 1982. In the process of transporting email messages between systems, SMTP communicates delivery parameters using a message envelope separate from themessage (header and body) itself.

Email Protocols - Interactions between email servers and clients are governed by email protocols. The three most common email protocols are POP, IMAP and MAPI. Most email software operates under one of these (and many products support more than one). Themost important reason for knowing of their existence? To understand that the correct protocol must be selected, and correctly configured, if you want your email account to work.

POP - POP is the older design, and hails from an era when intermittent connection via modem (dial-up) was the norm. POP allows users to retrieve email when connected, and then act on the retrieved messages without needing to stay "on-line." This is an important benefit when connection charges are expensive.

The basic POP procedure is to retrieve all inbound messages for storage on the client, delete them on server, and then disconnect.Outbound mail is generated on the client, and held for transmission to the email server until the next time the user's connection is active. After it's uploaded, the server forwards the outgoing mail to other email serve until it reaches its final destination.Most POP clients also provide an option to leave copies of email on the server. In this case, messages are only removed from the server when greater than a certain "age" or when they have been explicitly deleted on the client. It's the copies on the client that are considered the "real" ones, however, with those left on the server merely temporary backups.

SMTP - At the risk of overloading you with information, you should know that strictly speaking it's only the incoming mail that is handled by a POP or IMAP protocol. Outgoing mail for both POP and IMAP clients uses the Simple Mail Transfer Protocol (SMTP). When you set up a POP or IMAP email account on email client software, you must specify the name of the (POP or IMAP) mail server computer for incoming mail. You must also specify the name of the (SMTP) server computer for outgoing mail. These names are typically in the same form as Web addresses (e.g., "imap.med.miami.edu"). Depending on the client, there may also be specifications for email directories and searching.

Automated classification of email messages into user-specific folders and information extraction from chronologically ordered email streams have become interesting areas in text learning research. However, the lack of large benchmark collections has been an obstacle for studying the problems and evaluating the solutions. Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders, or identifying SPAM. We will focus on the problem of assigning messages to a user's folders based on that user's foldering strategy. One major consideration in the classification is that of how to represent the messages. Specifically, one must decide which features to use, and how to apply those features to the classification.

It is self-evident that e-mail has become a central means for the discussion of engineering work and sharing of digital assets that define the product and its production process. This is especially the case when teams become larger, increasingly multi-disciplinary and more distributed both spatially and temporally. Delinchant et al. argues that the prominence of e-mail is due to engineering companies offering support for the communication tool and its ubiquity across the engineering domain.

Engineering communication research has shown that the volume of communication is indicative of progress being made within an engineering project. In addition, Dong reveals that almost all successful design teams have high-levels of communication as this helps maintain a shared understanding between the engineers. Although it may seem a positive step to encourage increased communication between engineers, there are a number of limitations of e-mail that need to be addressed both from Personal Information Management and Project Management perspectives.

As per our study we need of email groupings based on users' activities where incoming mails are identified and grouped into appropriate activities and related messages are grouped in the same activity. Email messages are grouped by extracting most frequent words in the content of the message as well as comparing common words with most frequent words in the message to decide which activity the email message belongs.

II. LITERATURE REVIEW

We first provide a brief background for the classification algorithms employed herein before proceeding with a comparison of how they perform when applied to text classification [12] or email classification.

As for the classification phase, different classifiers (such as SVM, NN, KNN and DT) are employed to generate the model. However, this study only focused on using Naïve Bayes[7] to classify the emails. Given the probabilistic characteristic of Naïve Bayes, each training document is vectorized by the trained Naïve Bayes classifier through the calculation of the posterior probability value for each existing.

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

Email classification by Decision Tree (DT)[4]: - Decision trees are statistical data mining technique that express independent attributes and a dependent attributes logically AND in a tree shaped structure. Classification rules, extracted from decision trees, are IF-THEN expressions and all the tests have to succeed if each rule is to be generated. Decision tree usually separates the complex problem into many simple ones and resolves the sub problems through repeatedly using. Decision trees are predictive decision support tools that create mapping from observations to possible consequences. There are number of popular classifiers construct decision trees to generate class models.

Decision tree is introduced in the email classification. Class of representation is used for classify the data in this algorithm. Thus an email-classification tree is a binary tree where each internal node is labeled with a string and each leaf is labelled with a class name. Each email classification tree classifies an input string as follows. 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 labeled string). The class that the input string is classified into is the class at the leaf reached

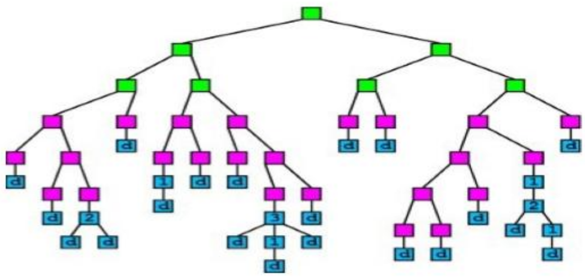

Fig.1 Email classification by Decision Tree (DT)

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 an email-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) Computing probabilities of different possible branches, determining the best split of each node, and selecting optimal combining weights to prune algorithms contained in the decision tree are complicated tasks that require much expertise and experience.
3) 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.
4) Decision trees, while providing easy to view illustrations, can also be unwieldy. Even data that is perfectly divided into classes and uses only simple threshold tests may require a large decision tree. Large trees are not intelligible, and pose presentation difficulties.
5) Drawing decision trees manually usually require several re-draws owing to space constraints at some sections, as there is no foolproof way to predict the number of branches or spears that emit from decisions or sub decisions.
6) Cost of decision tree implementation is very high for good decisions.
7) Analytical area of decision tree is imitated.

Email 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 [1]. 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 [8] algorithm is used forLearning 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.

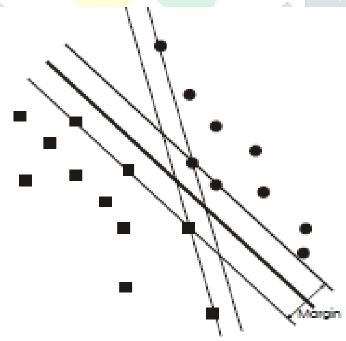

Fig.2 SVM decision margin

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 email 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.

Email classification by Neural Network[3]:-

Email 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.

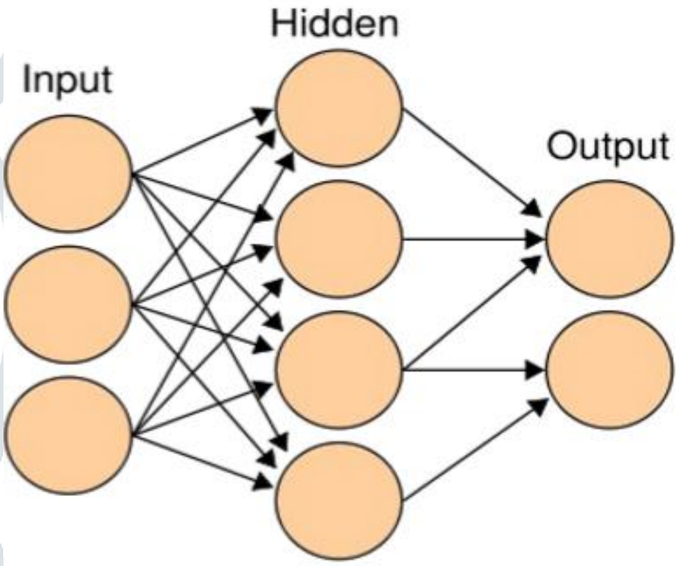

Fig.3 Email classification by Neural Network

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:-

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

Table 1 Comparison Table of Different Classification Techniques

Type	Accuracy	Time to build model	Speed	Precision	Recall	F-measure	Cost
SVM	Low	High	Low	Good	Good	Good	Inexpensive
NN	High	Very Low	Good	Medium	Medium	Medium	Very Expensive
DT	Good	Medium	Medium	Low	Low	Low	Expensive
NB	Very High	Low	High	High	High	High	Expensive

III. EMAIL CLASSIFICATION WITH NAÏVE BAYES[5] (PROPOSED SYSTEM)

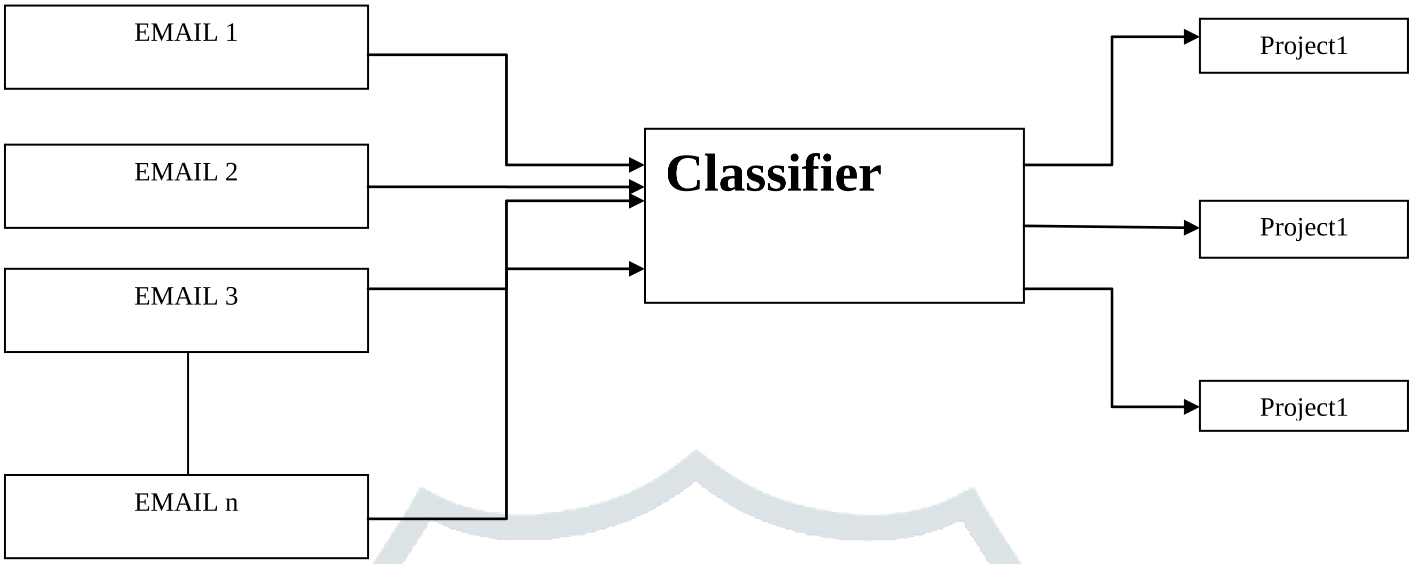

Fig.4 Proposed System

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 empirically.

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

Feature extraction [11]:-

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 first 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 them could be irrelevant or noisy [8]. Dimension reduction is the exclusion of a large number of keywords, base preferably on a statistical criterion, 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 for example 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 email. 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 Mark-up 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 email 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[10]:-

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,

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

Where, n =1, 2, 3, 4…

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

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

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

IV. 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, thousands 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.

V. REFERENCE

[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.

.