A HYBRID APPROACH FOR TEXT CLASSIFICATION USING HMM, SVM AND GENETIC ALGORITHM

Aakanksha[1], Er.Dinesh kumar[2]
[1]Student, [2]Head of I.T Dept, D.A.V. Institute of Engg & Tech
Dept. of C.S.E, Jalandhar

Abstract - Text Classification is latest research area in data mining. Data mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, etc. Mining the data means fetching out a piece of data from a huge data block. This paper presents an approach combining Hidden Markov Mode (HMM) and Support Vector Machine (SVM).HMM is used for Feature Extraction and SVM is used for text classification. Three different stages are designed to classify the content of online newspapers such as (a) Text pre-processing (b) HMM based Feature Extraction and (c) SVM for Text Classification.

Keywords-Text Classification, Feature Extraction, Hidden Markov Model, Support Vector Machine.

I. INTRODUCTION

Text Classification classifies the document according to pre-defined categories. It is manually defining set of logical rules that convert expert knowledge on how to classify document under given set of categories. It is widely used because of the availability of the increasing number of the electronic documents from a variety of sources. Text classification (TC) is an important part of text mining. Various news are published daily, it is time consuming process to select the most interesting one. So a method of news-article categorization is required to obtain the relevant information. A news-story categorization system is developed, where a rule base is generated by human expertise. For example automatically label each incoming news story with a topic like “sports”, “politics”, or “art”. A data mining classification task starts with a training set D = (d1….. dn) of documents that are already labelled with a class C1,C2 (e.g. sport, politics). Now Classification Model will assign the correct class to a new document d. This paper suggests that automated text categorization techniques are reaching a level of performance at which they can compete with humans not only in terms of cost-effectiveness and speed, but also in terms of accuracy of classification The rest of the paper is organized as follows: section 2 describes related works, section 3 provides the process of Text classification, section 4 presents basics of Feature Extraction using HMM, and section 5 presents the classifier (SVM), section 6 presents the methodology used.

II. RELATED WORKS

It is essential to have the overview of Feature Extraction using HMM and Feature Classification using SVM.

III. TEXT CLASSIFICATION

The stages of TC are discussing as following points in fig. 1.

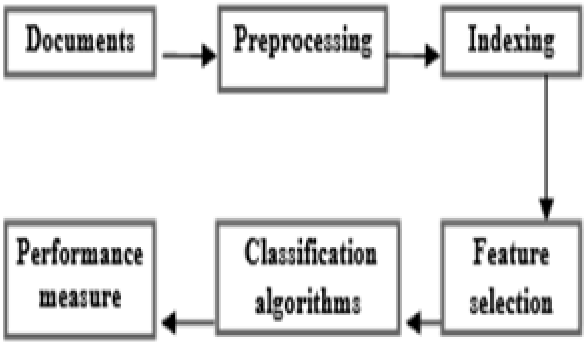

Fig.1 Document Classification Process.

This is first step of classification process in which to collect the different types (format) of document like html, .pdf, .doc, web content etc.

- *Pre-Processing:* The first step of pre-processing which is used to presents the text documents into clear word format.
- *Tokenization:* A document is treated as a string, and then partitioned into a list of tokens.
- *Stemming word:* This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute.
- *Indexing:* The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text.
- *Feature Selection:* After pre-processing and indexing the important step of text classification, is feature selection to create feature vectors.

A. CLASSIFICATION

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Methods used for text classification.

IV. FEATURE EXTRACTION USING HMM

HMM is used for Feature Selection. For text classification words are used as classification features. It is not necessary to use all the words in a text as classification features. By eliminating useless information the classification performance can be improved. The observation is turned to be a probabilistic function (discrete or continuous) of a state instead of an one-to-one correspondence of a state. Each state randomly generates one of M observations (or visible states). To define hidden Markov model, the following probabilities have to be specified: matrix of transition probabilities $A=(a_{ij})$, $a_{ij}= P(s_i \mid s_j)$, matrix of observation probabilities $B=(b_i (v_m))$, $b_i(v_m) = P(v_m \mid s_i)$ and a vector of initial probabilities $\pi=(\pi_i)$, $\pi_i = P(s_i)$. Model is represented by $M=(A, B, \pi)$.
The Feature Extraction using HMM is described in figure 2 [2].

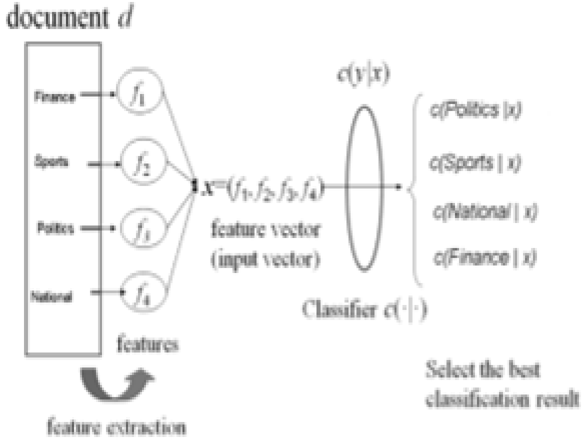

Fig. 2 HMM based Feature Extraction.

After Feature Extraction process, output can be normalized into a new feature vector and a trained SVM is ready to be used for classifying a new text.

A. HMM Assumptions

Markov assumption: the state transition depends only on the origin and destination

Output-independent assumption: all observation frames are dependent on the state that generated them, not on neighbouring observation frames

B. Main issues using HMMs

Evaluation problem: - Given the HMM $M= (A, B,\pi)$ and the observation sequence $O=o_1\ o_2\ ...\ o_K$, calculate the probability that model M has generated sequence O .

Decoding problem: - Given the HMM $M= (A, B,\pi)$ and the observation sequence $O=o_1\ o_2\ ...\ o_{K,}$ calculate the most likely sequence of hidden states s_i that produced this observation sequence O.

Learning problem: - Given some training observation sequences $O=o_1\ o_2\ ...\ o_K$ and general structure of HMM (numbers of hidden and visible states), adjust $M= (A, B,\pi)$ to maximize the probability.

$O=o_1...o_K$ denotes a sequence of observations $o_k \in \{v_1,\dots,v_M\}$.

V. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machines (SVM) A Support Vector Machine is a supervised classification algorithm that has been extensively and successfully used for text classification task. High dimensional input space: When learning text classifiers, one has to deal with large number of features. Since SVM use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces. Most text categorization problems are linearly separable: All categories are linearly separable and so are many of the Reuters Tasks. The idea of SVMs is to find such linear separators. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. With increasing amounts of data being generated by businesses and researchers there is a need for fast, accurate and robust algorithms for data analysis. Improvements in databases technology, computing performance and artificial intelligence have contributed to the development of intelligent data analysis. The primary aim of data mining is to discover patterns in the data that lead to better understanding of the data generating process and to useful predictions. Examples of applications of data mining include detecting fraudulent credit card transactions, character recognition in automated zip code reading, and predicting compound activity in drug discovery.

Benefits of SVM

- High-dimensional input space.
- Few irrelevant features: almost all features contain considerable information. He conjectures that a good classifier should combine many features and that aggressive feature selection may result in a loss of information.
- Document vectors are sparse: despite the high dimensionality of the representation, each of the document vectors contains only a few non-zero elements.
- Most text categorization problems are linearly separable.

A support vector machine (SVM) is for a set of related supervised learning methods, it takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which class.

- A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks.
- A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the

lower the generalization error of the classifier.

- SVM is a machine learning approach. It is used to search a usually very large space of potential hypothesis to determine the one that will best fit the data.
- SVM generates training models for text classification.

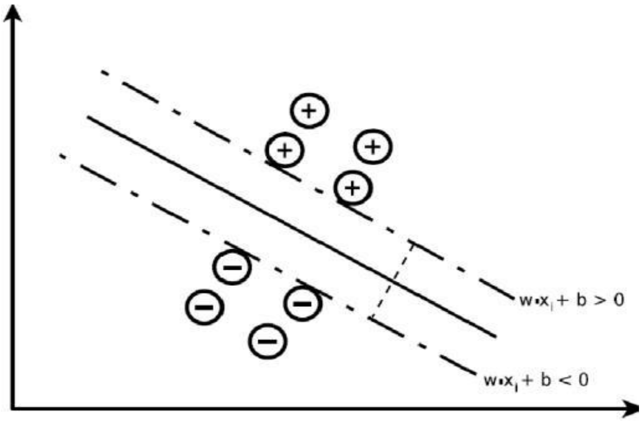

Fig. 3 Support Vector Machine.

VI. METHODOLOGY USED

Step 1 :- (Collect Documents for Database)
Documents are collected by sending request to The Times of India newspaper for providing database of documents and they provided it. A large number of texts is to be prepared, some of which are used for training the system, while others are used for evaluation or testing purpose.

Step 2 :- (Text Preprocessing)
In this present the text document into clear word format. For example Perform Tokenization, stemming word, Removing stop words. In tokenization a document is treated as a string and then partition into list of tokens. In removing stop word remove the stop words for example “the”, “a”, “and”. In stemming words converts different words from into similar canonical form. Statistical analysis of texts is very important, for example text length, the number of texts in each category, and keyword distribution in each text or in all texts.

Step 3:- (Use HMM for Feature Extraction)
After preprocessing the important step of text classification is feature selection. The main idea of feature selection is to select subset of features from the original document. It is performed by using HMM which reads the text as numeric values by creating reduced dimensional matrix. Feature extraction is one of the most important issues in text classification, and it can reduce the text vector space dimension, simplify the calculation, and prevent over-fitting and so on.

Step 4:- (Use SVM for training)
In this Classify documents into predefined categories. For classification RBF kernel of SVM is used for providing better accuracy. SVM trains the document and after training it tests it by comparing it with database for classification.

Step 5:- (Provide Results of Category Classification)
It displays the category of news.

VII. FLOW CHART

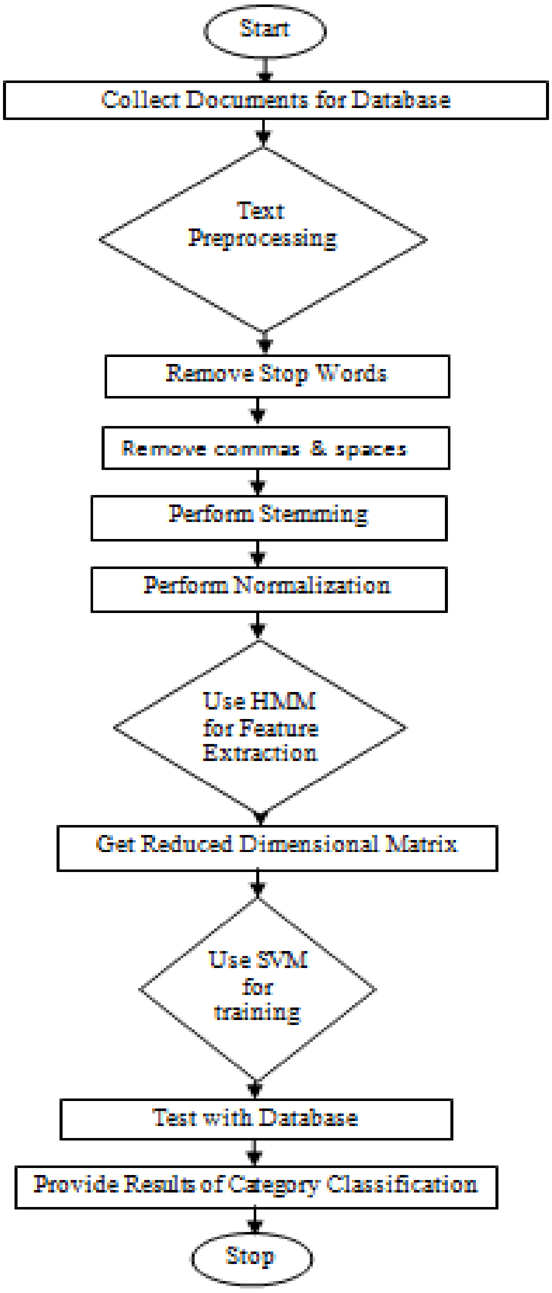

Fig. 4 Steps followed for text classification.

VIII. RESULTS

To evaluate the effectiveness of the new classification method proposed in this paper, we choose a text set, which consists of 10,8000 news, taken from Times of India [46], distributed among three major categories such as Jobs, Business, Religion, Health, Crime, and Politics. In our experiment, 80,000 texts are used as training set, and the rest 28000 texts are used as testing set. The distribution of the test news among various categories is in Table 1

A. Categories with the no of words or texts used for classification.

TABLE 1. Data set of various Categories

Categories/ Source	The Times Of India
Jobs	15000
Business	15000
Politics	20000
Religion	13000
Health	30000
Crime	15000

B. Sub Categories of various categories is listed below.

TABLE 2. Sub Categories of various Categories

Categories	Sub Categories
Jobs	Corporate, Glamour, Mobile jobs, Government jobs, Unemployment
Business	International business, United Kindom business, Local business
Politics	Parliament, International Politics, International political business, Indian politics
Religion	Hindu, Muslim
Health	Normal, Minor disease, Major disease, Critical Situation.
Crime	Crime against women, Reality shows, Police crime,

C. Values of parameters which are used for choosing best among HMM-SVM and HMM-SVM & GA.

TABLE 3. Accuracy of various Categories

Method Used with accuracy/ Parameters Used	HMM-SVM (%)	HMM-SVM & GA (%)
Precision	94	85
Recall	94	90
F-Measure	88	94
Accuracy	75	90

IX. CONCLUSION

The growing use of the textual data which needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. The existing classification methods are compared and contrasted based on various parameters namely criteria used for classification, algorithms adopted and classification time complexities. From the above discussion it is understood that no single representation scheme and classifier can be mentioned as a general model for any application. Different algorithms perform differently depending on data collection. However, to the certain extent SVM with HMM and Genetic Algorithm provides more accuracy and efficiency as compared to text classification using HMM-SVM.

REFERENCES

[1] Korde, Vandana, and C. Namrata Mahender.: "Text Classification and Classifiers: A Survey." International Journal of Artificial Intelligence & Applications (IJAIA) 3.2, 2012, pp-85-99.

[2] Donghui, Chen., "A new text categorization method based on HMM and SVM.", In Computer Engineering and Technology (ICCET), 2010 2nd International Conference on. Vol.7.

[3] Yang, Y., Pedersen J. O., “A Comparative Study on Feature Selection in Text Categorization”,In Proceedings of the 14th International Conference on Machine Learning,1997, pp-412– 420.

[4] XU, Jian-ming, Lei YANG, and Tong-cheng HUANG, "A New Feature Selection Algorithm in Text Categorization." Journal of Shaoyang University (Natural Science Edition) 1, 2008, pp-0-12.

[5] Liu, Zhijie, et al, "Study on SVM compared with the other text classification methods." Education Technology and Computer Science (ETCS), 2010 Second International Workshop on. 2010, Vol. 1.IEEE.

[6] Swati A. Kawathekar1, Dr. Manali M. Kshirsagar2 , “ Movie Review analysis using Rule-Based &Support Vector Machines methods ”, IOSR Journal of Engineering Vol. 2(3),March. 2012, pp: 389-391.