A Survey on Different Techniques of Text Categorization

Bhavna Rani
M.Tech. Student
Department of Computer Science & Engineering
Kurukshetra University, Kurukshetra, India

Abstract— In the Constitution of India, a provision is made for each of the Indian states to choose their own official language for communicating at the state level for official purpose. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. So the Classification of text documents based on languages is essential. The objective of the work is the representation and categorization of Indian language text documents using text mining techniques. Several text mining techniques such as naive Bayes classifier, k-Nearest-Neighbor classifier and decision tree for text categorization have been used. This paper describes various techniques used for semantic text classification. Text classification (Also called Text Categorization) is one of the important research issues in the field of text mining. Due to the rapid increase in addition of text documents on the web or internet, the text classification became a serious issue to retrieve the desired text from the huge amount of data placed in unstructured form on the internet. Categorization is a process of objects and ideas are differentiated, recognized and understood. For some specific purpose, the categorization implies the objects are grouped into categories. The text classification acts as a key function to organize and deal with million of documents. This paper covers different classification techniques along with their advantages and limitations.

Key words: Support Vector Machine, KNN (K-Nearest Tokens, Lemmatization or Stemming, Stop words, Zipf's Law, Bayes Classifier, K-Neighbor Classifier, Decision Tree, Precision (p), Recall (r), F-Measure

I. INTRODUCTION

Data mining is the main area when dealing with structured data in databases. Text mining refers to the process of analyzing and detecting knowledge in unstructured data in the form of text. The main problem in text mining is that the data in text form is written using grammatical rules to make it readable by humans, So to be able to analyze the text, it first needs to be preprocessed.

There are two fundamental approaches to analyse the text. First, Text mining employs Natural Language Processing (NLP) to extract meaning from text using algorithms. This approach cans be very successful but it has limitations. Second, a different approach using statistical methods is becoming increasingly popular and the techniques are improving steadily [1].

Text Classification [2]tasks can be divided into two sorts: supervised document classification where some external mechanism like human feedback provides information on correct classification for documents or to define classes for the classifier and unsupervised document classification which is also known as document clustering, where is no need of any external reference, the classification system do not have any predefined classes. The Text Classification task has an another task called semi-supervised document classification, means some documents are already labeled by the external mechanism.

II. OBJECTIVE

India is the home of different languages. Each state in India has its own official language. The objective of this work is to classify the documents based on language, using supervised learning algorithm. In future, these categorized documents can be used for summarization.

III. TEXT REPRESENTATION

The key objective of data preparation is to transform text into a numerical format such as vector space model. To mine text, we first need to process it into a form that data-mining algorithms can use. The Pre-processing steps are shown in figure 1 standard format.

A. Collecting Documents

The work resource for creating this corpus is the World Wide Web itself. The main problem with this approach to document collection is that the data may be of uncertain quality and require extensive cleansing before use.

B. Document Standardization

Once the documents are collected, it is common to find them in a variety of different formats, depending on how the documents were generated. The documents should be processed with miner modification, to convert them to a

C. Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. The tokenization process is language

IV. SUPPORT VECTOR MACHINES

SVM is a supervised learning technique used for both the classification and the regression. The standard SVM predicts for each given input, which of two possible classes it should be in [7].SVM learning systems trained with a learning algorithm from optimization theory that implement a learning bias determine from statistical learning theory.SVM is one of the best technique for the classification because of

- High dimensional input space
- Few dense concept vector
- Sparse document vectors
- Most text categorization techniques are linearly separable

Support vector machines are best for text classification because it provides both empirical and theoretical evidence. SVM makes the text classification very easier by eliminating the need of feature selection. It avoids catastrophic failure, show better performance in all

experiments while it is not in the conventional methods. SVM has the property of robustness[7].

Lukui Shi et al. combined SVM with the nonlinear dimensionality reduction method for the text categorization [7]. From the two similar text documents to classify text document many classification algorithms are used. In this algorithm to represent the similarity between the documents the geodesic distance is used. According to this algorithm geodesic distance is computed at first of all the documents and after that the high dimensional that is mapped into a low-dimensional space by using ISOMAP algorithm[8].

A. Advantages

1) SVM are less susceptible to over fitting. With excellent classification accuracy the SVM can handle large feature space[2]. For the testing and training sets it gives the best results of classification. SVM is a faster classification technique (very less time consuming technique).
2) SVM are the universal learners. These are independent of the dimensionality of the feature space.

B. Limitations

1) Its implementation is very complex and in the text collection with the number of documents it cannot scale well.

V. KNN(K- NEAREST NEIGHBOUR)

The k-nearest algorithm based on learning by correlation and closest training examples. KNN is a valid, simple and easy technique for the classification. It is also known as lazy learning or instance based learning. In the KNN technique firstly select an arbitrary data point from the vector space model which acts as a initial seed cluster. When it enters the training phase, then based on the Euclidean Distance whose centre is located at the nearest distance the training documents assigned to a cluster. Numeric attributes which are n-dimensional describe the training samples. To the mean of their currently acquired data points the cluster samples are adjusted repeatedly adjusted .when the given unknown samples close to the centroid.(In this each document is represented by nodes. The distance is calculated between the labeled documents and the unlabeled documents .If there are the number of labeled nodes are ‘n’ then the complexity of finding the label for unlabeled nodes is ‘n*log k’[2]).The KNN classification algorithm tries to find the K-nearest neighbor of the data points (text documents) and determine its class label by using the majority of votes[8].

A. Advantages

1) It is valid, easy and simple to implement. For the noisy training data it is very robust and it requires only two parameters.
2) KNN only storing the training examples, no any type of learning is done by KNN.KNN classification is also beneficial for multi classes as its classification decision is based on small group of similar text documents.

B. Limitations

1) KNN classification is a very time consuming technique .For the large samples and high dimensions, it is impossible to implement KNN algorithm.
2) For the nearest-neighbour classification cost becomes very high. With the increase in the training documents the classifier grows.

VI. DECISION TREE

Decision Tree ID3, C4.5, C5.0 algorithms are the classical algorithms having strong learning ability, high classifying speed and simple construction[12].In practical applications these algorithms are no satisfied. Decision Tree structure uses a top down approach. It uses the greedy learning approach.

There are many applications of decision tree in the real life. Some application areas are Business, Intrusion Detection, Energy Modeling, E-Commerce, Image Processing, Medicine, Industry, Intelligent Vehicles, Remote Sensing and web applications etc[11].

1) Decision tree is a classifier in the form of a tree structure
   – Decision node: specifies a test on a single attribute
   – Leaf node: indicates the value of the target attribute
   – Arc/edge: split of one attribute
   – Path: a disjunction of test to make the final decision
2) Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node.

VII. SELF-ORGANIZING MAP(SOM)

The Self-Organizing Map is an unsupervised learning technique of text categorization. It is a type of neural network. It is used to produce the similarity graph of input data. Cheng Hua Li and Soon Choel Park proposed two types of text classification neural networks[12],back-propagation neural network (BPNN) and multi-output perceptron learning(MOLP) and then a novel algorithm is proposed for improving back-propagation neural network. This proposed algorithm is able to overcome the limitations of the traditional back-propagation neural network which are slow training speed and property of local minima. Three methods are compared for the training time and the tested and the performance. By using the precision, recall and F-measure this algorithm is able to achieve the high categorization effectiveness.

Richard Freeman et al.[13] have investigated the use of SOM technique for the document categorization. They presented a growing and hierarchical method by using a series of one-dimensional maps. All the documents were represented by using vector space model. To organize the given set of documents, dynamically growing one-dimensional SOM were allocated [13].The produced hierarchically structured maps were visualized as a hierarchical tree. The result was come out in the form of a set of clustered documents. Yan Yu et al.[14] have proposed a new document clustering technique for text categorization, which is based on 1D-SOM.In this technique the results were obtained by calculating the distance between every two more similar prototype to the input

vector(MSPs).The result showed that it is simple and easy relative with 2D-SOM by using 1D-SOM.

A. Advantages

1) The data is easily interpreted and understood. It requires less amount of data to be trained.

B. Limitations

1) For the development of meaningful clusters, it requires necessary and sufficient data.

VIII. GENETIC ALGORITHM (GA)

Genetic Algorithm is a searching technique which is used for both solving problems and modeling evolutionary systems. It is heuristic in nature. It computes the population of solutions while the other heuristic methods give only single solution in their iterations[15].The GA involves chromosomes (population),reproduction, fitness and evolution. The working steps of algorithm are described as follows:

– Step 1: Generate a random population of n chromosomes which are suitable solutions.
– Step 2: Establish a method to evaluate the fitness f(x) of each chromosome x in the population
– Step 3: Create a new population by repeating the following steps until the new population is Complete.

1) Selection: select from the population according to some fitness scheme.
2) Crossover: New offspring formed by a crossover with the parents.
3) Mutation: With a mutation probability mutate new offspring at each locus (position in chromosome).

– Use the newly generated population for a further run of algorithm.

A. Advantages

Crossover of any Genetic Algorithm is a crucial aspect, but it may seem that with a high fitness function it will dramatically change parents so that they may no longer be fit.GA provides a good chance to find the better solution by creating new variants.

Algorithm-s	Fundament-al Techniques Used	Strengths	Weaknesses
Naive Bayes	Supervised learning Probability based	Simple Robust Wide applicability	Susceptible Depends upon bayes theorem
SVM	Supervised learning Straight line based Classification	No need to train the features	Long training Time
KNN	supervised Learning Neuron based Classification	modify	Performance Degrades with noisy trained data
Decision Tree	Supervised Graph based classification	simple & rule based approach Handle continual & categorical variable Handle missing data	Computationally expansive to calculate the information gain
BPNN	Supervised learning	modify	Slow training speed
SOM	Unsupervised learning	Validation measure, Scalability	Difficult to process for very large scale implementation of DR and PR
Genetic Algorithm	Based on the principles of biological evolution	Reduce the high feature dimension, improved accuracy	Cannot solve certain optimization problem

Table 1: Comparison of Conventional Text Classification Methods

B. Limitations

Genetic Algorithm cannot solve certain optimization problem. This occurs only due to the bad chromosome blocks having poorly known fitness functions.

IX. CONCLUSION

This paper covers the various text classification techniques, their advantages and limitations. All the different techniques are applied for the extraction of feature of the text which can be low dimensional and also can be high dimensional. This survey paper describes the overview of different types of text classification methods for the retrieval of desired data from a huge collection of unstructured data. Some classification techniques are less time consuming and faster for extracting features and easy to implement for large samples (e.g. Decision Tree) while some are more time consuming and their implementation is impossible for the large samples (e.g. KNN).All the different techniques has their own property for the retrieve the text. The research should be still continued for the efficient feature selection of the text and on classification of different types of text in different areas. To improve the text classification tasks various other semantic based machine learning methods can be proposed in future which can be able to overcome the limitations of previously proposed classification techniques.

REFERENCES

[1] B.Mahalakshmi, Dr.K.Duraiswamy, "An Overview of categorization Techniques," International journal of modern Engineering Research (IJMER), Vol.2,Sep.-Oct.2012.
[2] P.Bhargavi and Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, VOL.9, No.8, 2009.
[3] M.Aly, "Survey on Multiclass Classification Methods," Neural Networks, 2005.
[4] M. E. Ruiz and P. Srinivasan, "Automatic Text Categorization Using Neural Networks," School of library and Information Science,Vol.8.
[5] W.and B. Yu, "Text categorization based on combination of modified back propagation neural network and latent semantic analysis," Neural Comput & Applic, Springer Link, Vol. 18, No.8, pp.875–881, 2009.

[6] T. Joachims,'' Text categorization with Support Vector Machines: Learning with many relevant features'', university at Dortmund, Germany.
[7] L. Shi, J. Zhang, E. Liu, and P. He," Text Classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines,'' Third International Conference on Natural Computation, IEEE Xplore,Vol.1,pp.674-677,2007.
[8] C. Wan R. Pan and J. Li, "Bi-Weighting Domain Adaptation for Cross-Language Text Classification," in Proceedings of the twenty second International Joint Conference on Artificial Intelligence, August 1,2010..
[9] N. Remeikis, I. Skucas and V.Melninkaite," A combined Neural Network and Decision Tree Approach for Text categorization," information Systems Development,springer,pp.173-184,2005.
[10] D.D. Lewis and M. Ringuette," A comparison of two learning algorithms for text categorization, Symposium on Document Analysis and Information Retrieval,"1994.
[11] Mr. B. R Patel, Mr. K..K Rana," A Survey on Decision Tree Algorithm For Classification,"Volume 2, Issue 1,ISSN: 2321-9939, 2014.
[12] C. Hua Li and S. Choel Park, "Text Categorization Based on Artificial Neural Networks, "Neural Information Processing, Springer, Vol.4234, pp.302-311, 2006.
[13] R.Freeman,Hujun Yin and N. M.Allinson,"Self-Organizing Maps for Tree View Based Hierarchical Document Clustering,"IEEEXplore,pp.1906-1911,2002.
[14] Y. Yu,Pilian He,Y. Bai and Z. Yang, "A Document Clustering Method Based on One-Dimensional SOM," Seventh IEEE/ACIS International Conference on Computer and Information Science,pp.295-300,2008.
[15] A. Ganatra, Y P Kosta, G. Panchal and C. Gajjar, " Initial Classification Through Back Propagation In a Neural Network Following Optimization Through GA to Evaluate the Fitness of an Algorithm," International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

