ISSN:0975-9646
Manpreet Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3385-3387

OPTIMIZATION OF TEXT CLASSIFICATION USING SUPERVISED AND UNSUPERVISED LEARNING APPROACH

Manpreet Kaur
CSE Department
Chandigarh University
Gharuan, Mohali

Vijay Kumar
CSE Department
Chandigarh University
Gharuan, Mohali

***ABSTRACT*— Text Classification, also known as text categorization, is the task of automatically allocating unlabeled documents into predefined categories. Text Classification means allocating a document to one or more categories or classes. The ability to accurately perform a classification task depends on the representations of documents to be classified. Text representations transform the textural documents into a compact format. Text Classification plays an important role in information mining, summarization, text recovery and question-answering. It uses several tools from information retrieval (IR) and Machine Learning. Here we are reviewing the effectiveness of different supervised and unsupervised learning approaches in text classification.**

Keywords- Text Mining, Text Classification, Feature Extraction,Term Weighting, Linear SVC, SGD,K-Mean with cosine similarity.

I. INTRODUCTION

With the rapid growth of online information effective retrieval of some particular information is difficult without good indexing and summarization of document content. Text Classification may be the solution to effectively handle and organize such huge text collections. Text Classification is the task of automatically assigning unlabeled documents into predefined categories. It is the process of automatically grouping of documents into some predefined categories. The ability to accurately perform a classification task depends on the representation of documents to be classified. In text Categorization, text representations transform the content of textural document into a compact format so that documents can be recognized and classified by a classifier. A Classifier is a system that repeatedly classifies texts into one of a discrete set of predefined categories. For eg:- for email management one could benefit from a system that classifies incoming messages as important or unimportant. One of the main theme sustaining text mining is transforming text into numerical vectors i.e text representations. Feature selection is a standard procedure for dimensionality reduction. For text classification task an evaluation function is used that is applied in single term for selecting feature subset. After selecting feature subset all terms are sorted and evaluated independently accordingly. The best feature subset is determined by predefined threshold. Document frequency (DF) thresholding, information gain (IG), mutual information (MI), chi-square statistic (CHI) are various commonly used feature selection methods in text classification. In information retrieval documents are generally identified by set of terms or keywords that are collectively used to represent their contents. A document is represented as a vector in the term spaces in Vector Space Model.

$$d = (w1, w2, w3 - - - - - - w|v|)$$

Where |v| = size of vocabulary and,
Between [0,1] the value of w_i represents that how much the term w_i contributes to the semantics of the document. Vector Space Model is one of the mostly used models for text representations. Generally text representations include two types of works: indexing and term weighting. Indexing is done to allocate indexing terms for documents whereas term weighting is done to assign weight to each term of the document which measures the importance of that term. Presently, there are many term weighting methods which are used for text classifications. Text classification has borrowed the term weighting schemes from IR (information retrieval) field, such as tf, tf-idf and its variants. Feature representation is a transformation method that allows documents to be interpreted by classifiers and this method is also called as Term Weighting.

Table1:- Term Weighting Methods

Method	Description
Binary	Boolean Logic Representation 1= Present, 0 = Not Present
TF(Term Frequency)	Frequency of a term in a document i.e no of times the term appears in a document.
DF(Document Frequency)	Frequency of term in collection of documents.

II. LITERATURE SURVEY:-

Wen Zhang et.al [1]:- The purpose of this paper is to study the effectiveness of different indexing methods in text classification. This paper has comparatively studied TF_IDF, LSI and multi-word for text representation. An experimental result has demonstrated that in text classification, LSI has performed very well than other methods in both document collections. Also, while retrieving English documents LSI showed the best performance. This outcome has shown that LSI has both favorable semantic and statistical quality and is different with the claim that LSI cannot produce discriminative power for indexing.

Vishwanath Bijalwan et.al[2]:- In this paper author have first categorize the documents using KNN based machine learning and then return the most relevant documents**.** In this paper author conclude that KNN shows the maximum accuracy as compared to the Naive Bayes and Term-Graph. The disadvantage of KNN classifier is that its time complexity is high but gives a enhanced accuracy than others. In this paper the author rather than implementing the traditional Term-Graph used with AFOPT used Term-Graph with other methods. This hybrid shows a better result than the traditional combination. Finally author made an information retrieval application using Vector Space Model to give the result of the query entered by the client by showing the relevant document.

Tanmay Basu et. al[3]:- Text classification is a difficult task due to its high dimensionality of data. Therefore, efficient method for feature selection is required to improve the performance of text classification. This paper presents a new feature selection method for text classification using a supervised term selection approach. In this paper TS(term significance) a feature selection technique is compared with CHI,IG & MI. The proposed approach derives a similarity score between a term and a class and then ranks the terms according to their scores over all the classes. The experimental results show that the proposed TS can produce better classification accuracy even after removing 90% unique terms.

Youngjoong Ko et. al[4]:- The main purpose of this paper is to improve text classification by efficiently applying class information to a term weighting scheme. The author purposed a new scheme for multi class text classification. Then it was compared to the TF-IDF and previous methods. As a result the proposed scheme utilized class information for term weighting for text classification and performed consistently on the data sets and KNN and SVM classifiers.

Aixin Sun et.ai [5]:- In this paper the author purposed a simple, scalable and non-parametric approach for short text classification. This approach mimics human classification process for a piece of short text like tweets, status updates, and comments. It selects the representative words from a given short text as query words. After that it searches for a set of labeled text those best matches the query words. The author have used four approaches and are evaluated to select the query words: TF, TF.IDF, TF.CLARITY and TF.IDF.CLARITY. Experimental results show that TF.CLARITY performs effectively when three or more words are used in a query whereas TF.IDF.CLARITY performs well when one word is used in a query. The improvement becomes very minor when more than five words are used in a query.

Mengen Chen[6]:- In this paper a new algorithm is proposed for short text classification. The author compares the proposed algorithm with the state of the art baseline over web-snippet data set (one open data set) through two type of classifiers: MaxEnt (Maximum Entropy), SVM(Support Vector Machine). An experimental result shows that proposed algorithm performs better and appreciably reduces the classification errors by 16.68% and 20.25% in the same way.

Svetlana Kiritchenko[7]:- In this paper a learning technique is introduced that decreases the effort needed in applying machine learning. Main Problems in text classification are lack of labeled data and the cost required for labeling the unlabeled data. In this paper Classification is done on E-mail domain with Co-training algorithm that uses unlabeled data along with a small number of labeled examples. In this paper, the author firstly tested SVM classifier on a Labeled edition of unlabeled data and then Naive Bayes classifier is tested. As a result SVM performed very well in comparison with Naive Bayes. Experimental result also shows that the performance of co-training depends on learning method that it uses.

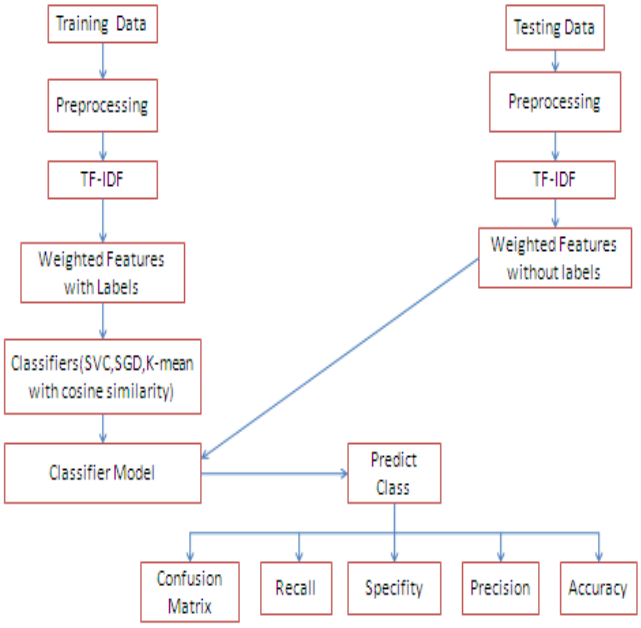
Guansong Pang et. al[8]:- In this Paper author purposed a generalized cluster centroid based classifier(GCCC) to use KNN and Rocchio via a clustering algorithm. In this paper, an algorithm is combined with Rocchio and KNN to make a generalized cluster centriod based model respectively to ensure the scalability and applicability of the GCCs model. Experimental results show that GCCC shows stable and favorable performance than KNN and Rocchio classifier. One drawback of GCCC is that it is more time-consuming than *KNN* and Rocchio.

Samuel Danso et. al[9]:-In this paper, the author has done a relative study for the categorization of verbal autopsy text in three ways i.e. feature representation, effect of reducing features and Machine learning algorithms. The author exhibit that normalized TF and standard TF-IDF achieve comparable performance across different classifiers. Finally author demonstrated the effectiveness of applying semi-supervised feature reduction approach to increase accuracy and SVM (Support Vector Machine) algorithm found to be the best algorithm than other algorithms.

Hugo Larochelle et. al[10]:-In this paper the author used a individual non-linear Classifier (RBM) for the classification. Firstly the classifier RBM (Restricted

Boltzmann Machine) is trained through different strategies and then tested with two classifiers i.e. LOG and NNet. In this paper RBM is compared with two different classifiers on multitask datasets. As a result RBM classifier gives best performance on all datasets than other classifiers.

III. PROPOSED METHODOLOGY



IV. PROPOSED ALGORITHM

Our system consists of three modules:
a) Feature extraction
b) Training
c) Testing

Algorithm (Algorithm for Text Classification)

Input: unstructured text without label
Output: Labeled text
For I = 0 to I = <length (DOC. Taining)
Begin:
Tokenization (DOC [I])
Stop word removal (DOC [I])
TF-IDF (DOC [I])
End
Document with weighted vector
For I= 0 to I<length (DOC.features)
Begin
Put in Linear SVC Classifier
End.
Linear SVC Model
For I= 0 to I <length (Doc .test)
Begin:
Tokenization (DOC [I])
Stop word removal (DOC [I])
TF-IDF (DOC [I])
End
Document test features
Put in Linear SVC model
Check the results in labeled text Precision, Recall, Accuracy, Specifity.

In above algorithm unstructured text is firstly preprocessed and TF-IDF is used to give weight to the text. After preprocessing features are extracted from the text. As features are extracted from the text now the text is converted into a trainable classifier model using SVC classifier. After training, using SVC a model is generated and testing will be done on it. In testing module also, tokenization and features are extracted as before will be extracted and is tested on trained SVC model that whether it predicts class as trained or not.

V. REFERENCES

[1] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, Jordan Pascual. "KNN based Machine Learning Approach for Text and Document Mining", 2014,Vol.7,No.1,pp.61- 70.
[2] Wen Zhang, Taketoshi Yoshida, Xijin Tang. "A Comparative Study of TF*IDF ,LSI and multi words for text classification",2011,Vol.1.
[3] Tanmay Basu, C. A. Murthy, "Effective Text Classification by a Supervised Feature Selection Approach",2008.
[4] Guansong Pang, Shengyi Jiang, " A Generalized Cluster Centroid based classifier for text categorization",2013.
[5] Youngjoong Ko, "A Study of Term Weighting Schemes Using Class Information for Text Classification", Aug 12-16,2012.

[6] Mengen Chen, Xiaoming Jin, Dou Shen, "Short Text Classification Improved by Learning Multi-Granuality Topics", 2010.
[7] Aixin Sun, "Short Classification using very few words", 2012, ACM 978-1-4503-1475-5/12/08.
[8] Svetlana Kiritchenko, Stan Matwin, "Email Classification with Co-training", 2006.
[9] Samuel Danso, Eric Atwell and Owen Johnson, "A Comparative Study of Machine Learning Methods for Verbal Autopsy Classification", 2012.
[10] Hugo Larochelle, Michael Mandel ,Razvan Pascanu, Yoshua Bengio, "Learning Algorithms for the Classification Restricted Boltzmann Machine", Journal of Machine Learning Research 13 (2012) 643-669.
[11] E Leopold, and J Kindermann "Text categorization with support vector machines. how to represent texts in input space? Machine Learning", 2002, 46:423-444.
[12] Y Liu, HT Loh, and A Sun "Imbalanced text classification: A term weighting approach. Expert systems with Applications", 2009, 36:690-701.
[13] M, Lan, CL Tan, J Su, and Y Lu "Supervised and traditional term weighting methods for automatic text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions", 2009, 31:721-735.
[14] M Lan, C. L Tan, H.B Low, and SY Sung "A comprehensive comparative study on term weighting schemes for text categorization with support vector machines" ACM, 2005: 1032-1033.