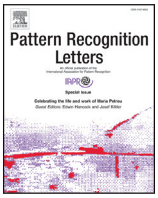
Pattern Recognition Letters 000 (2017) 1–9

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Classification of text documents based on score level fusion approach

S.N. Bharath Bhushan*, Ajit Danti

Department of Computer Applications, Jawaharlal Nehru National College of Engineering, Shivamogga – 577204, India

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Text classification
Text representation
Neural network

ABSTRACT

Text document classification is a well known theme in the field of the information retrieval and text mining. Selection of most desired features in the text document plays a vital role in classification problem. This research article addresses the problem of text classification by considering Sentence–Vector Space Model (S-VSM) and Unigram representation models for the text document. An enhanced S-VSM model will be considered for the constructive representation of text documents. A neural network based representation for text documents is proposed for effective capturing of semantic information of the text data. Two different classifiers are designed based on the two different representation models of the text documents. Score level fusion is applied on two proposed models to find out the overall accuracy of the proposed model. Key contributions of the paper are an enhanced S-VSM model, an interval valued representation model for the proposed S-VSM approach. A word level representation model for semantic information preserving of the text document and score level fusion approach.

1. Introduction

Over the recent years, computerized organization of electronic files is treated as a major study in the domain of computer science [1]. Text files have turn out to be one of the most common vicinity types of statistical repository, mainly due to the elevated reputation of the World Wide Web (WWW) [2]. The main sources of text data generations include websites, newsgroup discussions, forum messages and emails. In a few decades, management of the electronic document based on the content, has received a distinguished reputation in information system domain. This is due to the extended accessibility of text data in the digital form [3].

An important component of text classification system is selection of good quality of document collection, representation model and adaption of suitable classification techniques [4]. At its best, a text file series can be any grouping of text content based files. Algorithmically, solutions of the text mining applications depend on the selection of patterns in the large datasets. The total number of (text files) text documents in such datasets may range from a few thousands to millions. Text documents, is a collection of terms (words), which is difficult to understand and challenging to be interpreted by a classifier. Due to this, the unstructured text data is transformed into machine understandable form.

This article presents the text classification by considering score level fusion techniques. Two different classifiers are designed based on the two different representation models of the text documents. An enhanced S-VSM model will be considered for the constructive representation of text documents. A neural network based representation for text documents will be proposed for effective capturing of semantic information of the text data. Further, two different classification techniques are designed based on the above mentioned different representation models. Classification scores of these two classifiers are fused for the development of novel fusion based technique for text classification problem.

The key contributions of the proposed method are as follows:

- An enhanced S-VSM model is considered for the constructive representation of text documents.
- Developing an interval value representation model for the proposed S-VSM representation techniques.
- Devising a word level representation model for effective representation of the text data.
- Preserving semantic information of the text document.
- Proposing a novel fusion based technique for text document classification.

This article is ordered as follows: a brief introduction of various techniques of text classification is presented in Section 2. A novel fusion based approach for text classification is presented in Section 3. The experimental setup and Result and Discussion are presented in Section 4. Finally, this article will be concluded by drawing conclusions in Section 5.

2. Related work

Generally text files are not appropriate for regular databases as they contain unstructured text data. These text data present a

* Corresponding author.
E-mail address: sn.bharath@gmail.com (S.N. Bharath Bhushan).

http://dx.doi.org/10.1016/j.patrec.2017.05.003
0167-8655/© 2017 Elsevier B.V. All rights reserved.

wide variety of information, which is difficult to understand automatically for supervised learning algorithms. Thus, free running and unstructured text data should be converted to machine understandable format. To address this issue, many algorithms are proposed in literature. Once this conversion of unstructured text data to structured format is achieved, effective representation model is provided to represent the text data since it contributes a lot for accuracy of the proposed approach. Bag of Word (BoW) representation is considered as the primary technique for text document representation. The BoW is considered by generating a vector representation of a document by capturing number of occurrences of each word present in the text document. This methodology of representation text document is known as VSM (Vector Space Model) [5]. Jain and Li [6] provided a binary representation for text documents. The main limitation of binary representation method and VSM method is that these results with the big sparse matrix, which increases the dimension of the feature matrix. Hotho et al [7] presented a representation based on ontology for a text document to capture the semantic (meaning in the document) relationship in between the words existing in the document.

The proposed ontology based representation model has the capability in preserving the class / dictionary information of the document. Cavanar W.B. [8] considered a series of parameters like a character or a word and treated them as n-grams, which is collected from big sentences present in the text document. The main limitation of n-gram model is selecting number of grams (terms) like fixing the N value is the challenging task. Many approaches can be found on classification of text documents in the literature. These approaches include naïve bayes [9–12], nearest neighbor [13–17] decision trees [18], support vector machines [19–22] and neural network [23–25] approaches.

3. Proposed model

This paper presents a classification of text documents using fusion based approach in an uncontrolled environment. Here the problem of text classification is addressed by considering score level fusion techniques.

Two different classifiers are designed based on the different representation models of the text documents. The overall steps involved in the proposed model are shown in the Fig 1.

S-VSM: Text documents, which are typically collection of strings, are not understandable by classification algorithms. Unfortunately, machines cannot understand the words as human beings do, but they need an appropriate representation of the text documents. These text documents, typically collection of strings, need to be transformed into suitable representation for classification algorithms. In order to enable further analysis, the documents have to be mapped onto some representation language like vector space model, n-gram Representation Model.. etc.

Vector space representation model transforms the text document into numerical vector by maintaining the term occurrence count in the document. The main limitation here in this method is that, it generates the high dimension feature matrix. To address this problem a novel S-VSM, which helps in constructing a lower dimensional feature presentation of text documents is designed here.

Let C_j, $j = 1, 2, 3, . . . p$ be the different classes in the database which contains D_m, $m = 1, 2, 3, . . . , q$ documents in each class. Intern each document consists of t_n, $n = 1, 2, 3, . . . , n$ set of terms. The construction of S-VSM is as follows. All the terms t_n from D_m documents are collected and a dictionary Dic_j will be formed to represent the class C_j. The newly formed dictionary Dic_j is subjected for stop word elimination algorithm to remove stop words such as *is, as, was, it, a, an, … etc* from the dictionary Dic_j.

Once the preprocessing is completed on the newly constructed dictionary Dic_j, the dictionary size will be reduced because of the stop word elimination. The same process will be carried out for all the class C_j, hence Dic_j will be constructed for all the remaining p number of classes in the database. Then, S-VSM will be constructed by calculating the probability between terms in dictionary Dic_j and class C_j. The calculation of the probability between terms in dictionary Dic_j to the class C_j is explained below.

Let there be j number of classes/domains in the database which contains q number of training documents. A text preprocessing algorithm is considered to extract terms from each document and stop words are eliminated and then the remaining terms are pooled for construction of dictionary of the respective class. Let n be the number of terms in the dictionary. Each document can be given a term frequency vector representation of dimension n based on frequency of occurrences of each of the n terms in the class. Each dictionary is vectorized using Bayes theorem as explained below.

To estimate the probability of a particular dictionary which is interpreted to a specific class say, Dic_j , $j = 1, 2, \ldots j$, we calculate the posterior probability of the text data interpreted as the specific dictionary is given by the formula

$$P_r\left(Dic_j|d\right) = \frac{P_r\left(Dic_j|t_1, t_2, \ldots, t_n\right)}{n} \tag{1}$$

where t_1, t_2, . . . ,t_n is the term frequency vector representing d.

The posterior probability of word t_i from document d from the dictionary Dic_j is given by

$$P_r\,(Dic_j|w_l\,) = \frac{P_r(w_l\left|Dic_j\right.)\,P_r(Dic_j)}{P_r(t_l)} \tag{2}$$

Where

$$P_r(Dic_j) = \frac{Number\ of\ terms\ in\ Dic_j}{Number\ of\ terms\ in\ training\,set} \tag{3}$$

For normalization of the word t_l, $P_r(t_l)$ is calculated by,

$$P_r(t_l) = \frac{\sum occurrences\ of\ term\ t_l\ in\ all\ dictionaries}{\sum occurrences\ of\ all\ terms\ in\ all\ dictionaries} \tag{4}$$

and

$$P_r(t_l\left|Dic_j\right.) = \frac{Occurrences\ of\ terms\ t_l\ in\ class\ Dic_j}{\sum occurrences\ of\ all\ terms\ in\ class\ Dic_j} \tag{5}$$

On the basis of the Bayes formula, priori probability $P_r(Dic_j)$ value is, the likelihood $P_r(t_l|Dic_j)$ and $P_r(t_l)$ which will be the evidence, along with the posterior probability for each term in the input document is d as $P_r(Dic_j|t_l)$, its posterior probability being annotated to the class Dic_j, $P_r(Dic_j|d)$ can thus be measured using equation.

The posterior probabilities for all words t_l, $1 \le l \le t$ present in the document d with respect to the class Dic_j, $1 \le j \le k$ are calculated and stored in a tabular form as shown in Table 1. In Table 1 every row presents the posterior probability of the term present in the document d, and each column corresponds to a dictionary.

From the obtained posterior probabilities of words in the document d, we calculate the posterior probability of the document d being annotated to the class Dic_j using Eq. 1. Similarly, the posterior probabilities of the document d being annotated to all other classes Dic_j, $1 \le j \le k$ are calculated and presented like a vector which provides the probability of the respective document belonging to each individual dictionary. These k values are used to approximate the document d in form of k level feature space and hence each dictionary will be provided with a k-dimensional vector representation with each dimensional value being crisp as shown below.

$$\left(P_r(Dic_1|d\,),\ P_r(Dic_2|d\,), \ldots, P_r\left(Dic_j|d\,\right)\right) \tag{6}$$

Please cite this article as: S.N. Bharath Bhushan, A. Danti, Classification of text documents based on score level fusion approach, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.05.003