Framework for Text Categorization in Agricultural Domain

Conference Paper · March 2017

CITATIONS
0

READS
91

4 authors:

Sreekumar Biswas
Indian Agricultural Statistics Research Institute
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Rajni Jain
National Institute of Agricultural Economics and Policy Research
75 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Sudeep Marwaha
Indian Agricultural Statistics Research Institute
39 PUBLICATIONS **220** CITATIONS
SEE PROFILE

Alka Arora
Indian Council of Agricultural Research
35 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353
2017 4th International Conference on "Computing for Sustainable Global Development", 01st - 03rd March, 2017
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Framework for Text Categorization in Agricultural Domain

Sreekumar Biswas
Ph. D. Scholar: Discipline of Computer Application
ICAR-IASRI
New Delhi, India
sreekumar.iasri@gmail.com

Rajni Jain
Principal Scientist
ICAR-NIAP
New Delhi, India
rajni@icar.gov.in

Sudeep Marwaha
Principal Scientist: Discipline of Computer Application
ICAR-IASRI
New Delhi, India
sudeep@iasri.res.in

Alka Arora
Principal Scientist: Discipline of Computer Application
ICAR-IASRI
New Delhi, India
alkak@iasri.res.in

***Abstract*—Recent trends in agriculture yields a large number of researches and articles related to it. It has become a daily routine to access these articles by modern day researchers. But due to the fact that these information are in an unstructured form, readers might face difficulty to access these information. Text Categorization, a branch of Text Mining, is a very useful technique to represent these unstructured text in a structured way. In this research, a number of research articles has been categorized using text categorization by applying some popular machine learning algorithms.**

Keywords—Text Categorization; text Mining; Machine Learning; Classifier; Receiver Operating Characteristic.

I. INTRODUCTION

In the modern era, the focus on text categorization is increasing so as to make the information availability easier in a well-structured form. Text categorization, the definition can be stated as the assignment of an unknown document into a class from a set of predefined classes [7]. Previously, it was a manual task, but with the advancement of text mining and the exponential increase in the number of research articles, the application of text document classification algorithms (TDCA) became a popular approach as it diminished the time as well as cost complexity at a rapid rate. We know that text documents suffer from the problem of the Curse of Dimensionality [1]. Therefore, it is very much essential to reduce the dimension of those documents. There are a number of algorithms available for text categorization. Also there are researches available, which compare the effectiveness of the algorithm [2], [6], [10].

Research in agriculture is increasing day-by-day. As a result, the number of research articles are also increasing exponentially. Agriculture, being the backbone of the country, needs the researchers to bring out revolutions, such as the Green Revolution, in terms of food production. For this purpose, the researchers need to consult the research articles day-in and day-out. As mentioned earlier that the information should be easily available, of course in a structured way, there is a need for an effective categorization of the articles apart from the traditional library based categorization system. In this research, an attempt has been made to categorize the articles using some machine learning algorithms.

In the subsequent portions of this paper, the materials and the methodology needed for the research is described.

II. RELATED WORKS IN THE FIELD OF TEXT CATEGORIZATION

The work on text categorization has started a long ago manually, but due to the fact of time and cost, the manual work was shifted to some automated techniques. In this section, a small discussion related to some researches on text categorization is presented.

In [1], Baharudin et. al. showed the important techniques and methodologies involved in text documents classification, focusing mainly on the representation of text and machine learning techniques. They demonstrated a review of the theory and practical methods of document classification and text mining.

In [2], Bhumika et. al. highlighted the important algorithms that can deal with text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved.

In [5], Klassen et. al. demonstrated classification of web pages using keywords from documents as attributes and applying the random forest algorithm. They found that random forest outperforms many other algorithms that were proven to be a good classifier in the context of text categorization.

 ISSN 0973-7529; ISBN 978-93-80544-24-3

In [6], Kotsiantis et. al. studied various classification algorithms and made attempts for improving classification accuracy. They have used techniques based on artificial intelligence (Logic-based techniques, Perceptron-based techniques) and statistics (Bayesian Networks, Instance-based techniques).

In [7], Manevitz and Yousef showed the way of training a simple feed-forward network to filter documents and the method was proved to be superior to some of the modified methods such as Rocchio, Nearest Neighbor, Naive-Bayes, Distance-based Probability and One-Class SVM algorithms.

In [8], Nasa demonstrated the performance of several ML algorithms using WEKA over a large amount of data to test and validate the differences between the classification methods or the algorithms.

III. METHODOLOGY

The methodology has been divided into two sub-sections. The first one being a little description on the data used, and the second one is the workflow of the software used.

A. Data

The data has been collected from Prof. M. S. Swaminathan Library, Indian Agricultural research Institute. It is an arff (Attribute Relation File Format) file and includes titles of research articles from two different domains. The data is summarized in TABLE I.

TABLE I. DESCRIPTION OF THE DATA

Name of the file	Data_Ag.arff
Number of records	44
Source	Prof. M. S. Swaminathan Library, Indian Agricultural research Institute
No of classes	2
No of attributes	142
Type of attributes	Numeric
Missing values	Nil

All the titles are arranged as a single arff file. Then the titles are vectorized to form a word vector model as the software cannot understand string of letters, it understands numeric values only. The further experiment has been conducted using the vectorized data. The detailed information about arff file is given in [11].

B. Software Workflow

WEKA 3.8.0 is used to conduct the experiment. WEKA [3] stands for Waikato Environment for Knowledge Analysis, developed from the University of Waikato, New Zealand. It is an open source machine learning tool licensed under the GNU General Public License. This software contains a collection of visualization tools and algorithms for the task of data analysis and predictive analysis. WEKA supports many data mining tasks, to be more specific, data preprocessing, classification, clustering, regression, association, attribute selection, forecasting and many more tasks including time series analysis. WEKA assumes that the supplied data as one flat file or relation and each data point is described by a fixed number of attributes. WEKA was originally developed for the purpose of processing agricultural data, motivated by the importance of this application area in New Zealand [3]. WEKA is a workbench that provides five interfaces, namely, *Explorer*, *Experimenter*, *KnowledgeFlow*, *Workbench* and *Simple CLI* [12]. This experiment has been conducted in the *KnowledgeFlow* environment. The framework made for the experiment is given on Fig. 1.

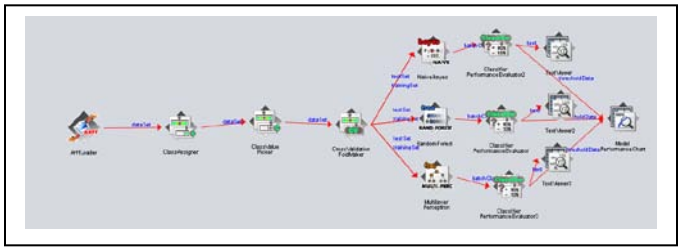

Fig. 1. Framework for the KnowledgeFlow

The first component here is the *ArffLoader*. It is used to load the data. Then the *ClassAssigner*, meant for assigning the class label for the experiment. The *ClassValuePicker* serves the purpose of picking up the label for which the ROC analysis will be taken afterwards. Then comes the *CrossValidationFoldMaker*, here the number of folds for cross validation is adjusted. This experiment is conducted using a 10 fold cross validation. After that, there are three TDCAs, which include Naïve Bayes, Random Forest (RF) [5], [9], [12] and Multilayer Perceptron (MLP). After conducting the experiment, a comparison among these three TDCAs can be obtained as a result. The *ClassifierPerformanceEvaluator* evaluates the performance of the respective TDCA to which it is connected. There are three instances of this component as there are three different TDCAs used in the experiment. Then the component *TextViewer* provides the result of respective TDCA in *PlainText* format. Lastly, there is the final component, *ModelPerformanceChart*. It gives the result of ROC analysis by displaying the ROC curve. There is only one instance of the component because the aim is to compare the performance of the TDCAs together in one plot. A detailed workflow of the interface is given on [11].

IV. Results and Discussions

The experiment was conducted over two domains, Artificial Intelligence and Social Science. The data set contained 48 titles arranged randomly. The vectorized data contained 142 attributes and a 10-fold cross validation was used for the experiment. After conducting the experiment, the result was obtained in terms of statistical as well as ROC curves [4], [8]. The interpretation has a strong statistical inference with support of the ROC analysis. The result is given below:

TABLE II. Naïve Bayes

Sl. No.	Naïve Bayes		
	Statistic	***Number***	***Percentage***
1	Correctly Classified Instances	42	95.4545
2	Incorrectly Classified Instances	2	4.5455
3	Mean absolute error	0.0512	-
4	Root mean squared error	0.2008	-

TABLE III. Random Forest

Sl. No.	Random Forest		
	Statistic	***Number***	***Percentage***
1	Correctly Classified Instances	37	84.0909
2	Incorrectly Classified Instances	7	15.9091
3	Mean absolute error	0.3397	-
4	Root mean squared error	0.3858	-

TABLE IV. Multilayer Perceptron

Sl. No.	Multilayer Perceptron		
	Statistic	***Number***	***Percentage***
1	Correctly Classified Instances	37	84.0909
2	Incorrectly Classified Instances	7	15.9091
3	Mean absolute error	0.2063	-
4	Root mean squared error	0.3459	-

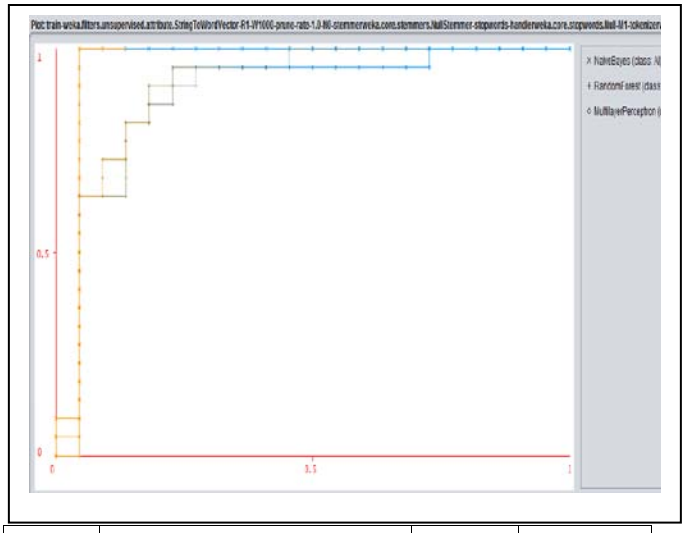

Fig.2 ROCs produced by different TDCAs

The result is very clear that among the three TDCAs that are used for the categorization task, Naïve Bayes has given the best result. It has an accuracy of 95.4545%. Both RF and MLP has the same accuracy, which if of 84.0909%, but if the Root Mean Squared Error (RMSE) is considered, it is clear that MLP is better than RF. The ROC curve obtained from the experiment is given below:

The ROC shows that Naïve Bayes outperforms RF and MLP both. The fact that MLP performs better than RF is clearly visible the ROC.

V. Conclusion and Future Work

The paper presented a framework of text categorization using various TDCAs in WEKA software in agricultural context. In the experiment, Naïve Bayes, , Random Forest and Multilevel Perceptron algorithms have been used to get an idea that which one of the three popular TDCA is suited for categorization in agricultural domain. This research was conducted over real data and the results have been encouraging. Lastly, to conclude among the TDCAs, Naïve Bayes outperforms others in terms of all the measures of performance evaluation. Till now only titles of research papers has been used, in future, abstracts and/or full texts can also be used for measuring the best performing algorithm and the categorization, more precisely, classification of the research articles can have a more easier path.

References

[1] Baharudin, Baharum, Lam Hong Lee, and Khairullah Khan. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology 1.1 (2010): 4-20.

[2] Bhumika, Prof Sukhjit Singh Sehra, and Prof Anand Nayyar. "A review paper on algorithms used for text classification." International Journal of Application or Innovation in Engineering & Management 3.2 (2013): 90-99.

[3] Frank, Eibe, et al. "Weka-a machine learning workbench for data mining." Data mining and knowledge discovery handbook. Springer US, 2009. 1269-1277.

[4] Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning, 45(2), 171-186.

[5] Klassen, Myungsook, and Nikhila Paturi. "Web document classification by keywords using random forests." International Conference on Networked Digital Technologies. Springer Berlin Heidelberg, 2010.

[6] Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques." Artificial Intelligence Review 26.3 (2006): 159-190.

[7] Manevitz, L., & Yousef, M. (2007). One-class document classification via neural networks. Neurocomputing, 70(7), 1466-1481.

[8] Nasa, Chitra. "Evaluation of different classification techniques for web data." International Journal of Computer Applications 52.9 (2012).

[9] Segnini, Armando, and Juanita Joyce Tayou Motchoffo. "Random Forests and Text Mining."

[10] Taneja S, Gupta C, Gureja D and Goyal K. K Nearest-Neighbor Techniques for Data Classification-A Review. ICCIN 2014; Delhi, India.

[11] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[12] Xu, Baoxun, et al. "An improved random forest classifier for text categorization." Journal of Computers 7.12 (2012): 2913-2920.

