UTM Computing Proceedings
Innovation in Computing Technology and Applications
Volume: I | Year: 2016 | ISBN: 978-967-0194-81-3

Evaluating Entropy as Feature Selection Method for Biomedical Text Classification

***UMI NAZIHAH ZAMARI*[1], *ROZILAWATI DOLLAH @ MD ZAIN**[2]

Department of Information System, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[1]*zihazam28@gmail.com* , [2]*rozilawati@utm.my*

Abstract

As the amount of digital biomedical literature grows, the task of organizing and retrieving becomes a challenging task. This is due to difficulties for the most researchers in searching and retrieving the required information from the Internet. The application of text classification on biomedical literature would be one of the solution for this problem. However, one of the issue in text classification is managing the high dimensionality of data. Therefore, the purpose of this research is to select the relevant features for classifying the biomedical literature. In this research, we focus on employing of the Entropy as feature selection method for reducing the high dimensionality of data. Towards this effort, we conduct several classification experiments using a subset of OHSUMED dataset in order to evaluate the effectiveness of employing Entropy for text classification. The results demonstrate that the macro-average for precision, recall and F measure are improved by employing entropy method to select the relevant features for classification purpose.

Keywords: Biomedical Text Abstracts, Entropy, Feature Selection, OHSUMED, Text Classification

1.0 Introduction

As a result of the rapidly increasing number of online biomedical literature, biologists and researchers need a conducive system to help them for searching relevant information related to their interest. Classification is one of the effective ways to organize the ample amount of online biomedical literature. Along with the growing of technology, many researches propose different techniques to overcome this problem, for example text mining (Dalianis *et al.,* 2011; Fu *et al*., 2015), classification (Colace *et al*., 2014; Ishii *et al*., 2010), feature selection (Dessì and Pes, 2015; Shu *et al.*, 2014) and so forth.

In order to help researchers to gain information related to the biomedical literature, data mining techniques could be employed. Yoo *et al.* (2012) stated that data mining could unwrap new biomedical and healthcare as it could help to indicate any fraud in health insurance, as well as in classifying the level of health of the patients. However, Vlachos *et al.* (2015) stated that the big amount of data available nowadays leads to the difficulty to retrieve relevant data.

According to Tsai *et al*. (2014), text classification is connected with the construction of a model via the learning from training example in order to automatically classify the biomedical abstracts. However, the amount of the required storage and cost of the model learning become

bigger when the size of the biomedical abstracts increases swiftly. Since there is an increase of data in digital forms, the needs of feature selection method to be employed for reducing the number of features that used for classification also increases. Zhang *et al*. (2016) mentioned that feature selection involves not only reduction number of features but also the choice of the features, which means either selecting the modeling or analyst tool, or eliminating the features based on their advantage for the analysis. In addition, Ghareb *et al*. (2016) stated that there are two purposes of feature selection. First, it applies a classifier in order to decrease the size of effective vocabulary, and second it increases classification accuracy by removing noise features.

Therefore, the purpose of this research is to investigate the feature selection methods for selecting the informative and relevant features for classifying the biomedical literature. In this research, we focus on employing of the Entropy as feature selection method for reducing the high dimensionality of data. Finally, we compare and evaluate the effectiveness of employing Entropy method as feature selection method for classification of biomedical text classification.

This paper is organized as follows: In Section 1, we present the introduction of this paper. Then, in Section 2, we discuss more details of some related works. While in Section 3, we describe the overview of the research methodology that used in this research. Next, we provide the experimental results in Section 4. In section 5, we analyze the results from the research conducted. Finally, Chapter 6 states the conclusion and future works in enhancing this research.

2.0 Related Work

Researchers tend to face a lot of problems in order to organize large amount of online biomedical literature and they also face difficulties to retrieve the required information in their field of study. With the overwhelming online articles in biomedical domain, one of the popular approaches suggested for organizing large volume of documents is text mining. Romero *et al.* (2011) mentioned that the efficient and effective literature mining techniques are very important for the use of gathering and deploying the knowledge in this field. Even though there are lot of approaches used to overcome the issues in managing biomedical literature, but not all problems can be fully solved. However, Li *et al.* (2011) stated that text classification has been recognized as one of the main approaches in text mining research that could be used to classify the large volume of text documents.

Text classification is one of the important and challenging tasks in text mining research. According to Uysal & Gunal (2014), preprocessing is the main element in a typical text classification framework. It is very important as the number of electronic documents in worldwide is growing. Nevertheless, Colace *et al*. (2014) mentioned that in order to achieve better accuracy of text classification, there still is restriction that needs to be faced due to the fact that human labeling is enormously time-consuming. Other than that, high dimension also becomes one of the challenges in performing text classification. Normally, large volume of data is portrayed by their high dimensionality (Perthame *et al.,* 2015). In addition, Pedrycz *et al.* (2009) stated that to perform effective classification for high dimensionality of biomedical data, we need to apply feature selection method.

Feature selection is the process of selecting a subset of relevant feature that representing each training and testing document for classification purpose. It mostly acts as a

filter and is also useful as a part of data analysis process. According to Dessi & Pes (2015), as the number of machine learning and data mining techniques increase, feature selection has become an important step to make the analysis more convenient and to extract useful knowledge about a given domain. As mentioned by Ishii *et al.* (2010), considering that the large number of biomedical articles is being digitalized, tools for mining the part of those articles are required. There are few features selection methods that can be used to select the relevant terms or features as keywords or index for classifying biomedical text abstracts such as Information Gain, Entropy, Mutual Information, Odd Ratio and so forth.

3.0 Methodology

There are four phases to complete in order to gain output for this research. Figure 1 illustrates the proposed research methodology for this research.

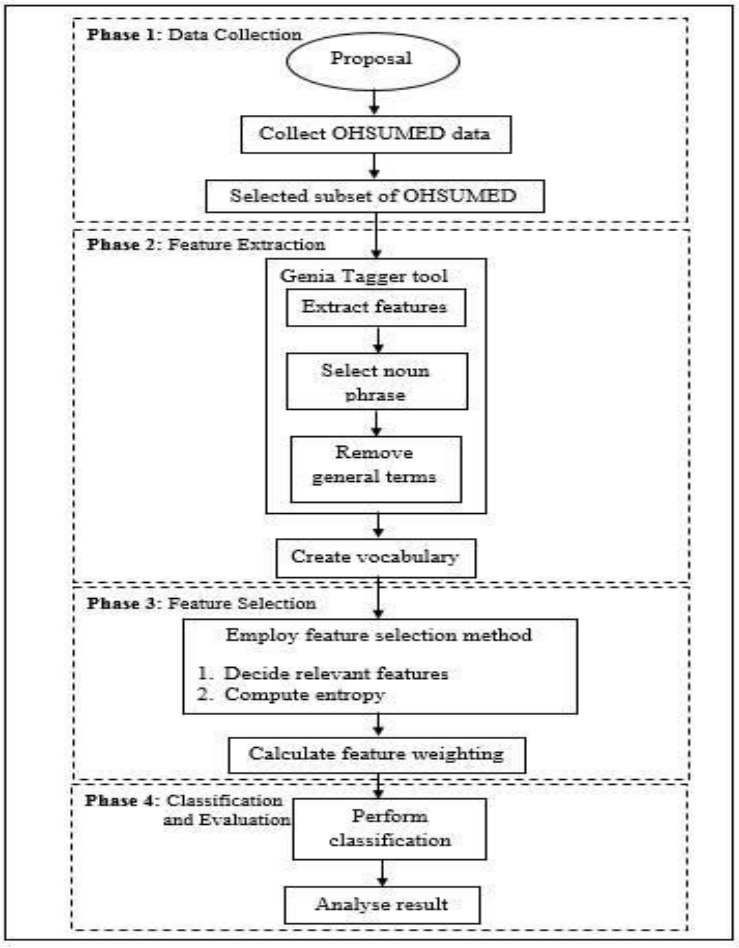

Figure 1: Purposed research framework

In phase 1, we perform data collection. Therefore, we collect 5,768 biomedical text abstracts from OHSUMED dataset. Then, we divide this dataset into two parts, which are training dataset and testing dataset. Phase 2 involves feature extraction process, where features are tagging and chunking using Genia Tagger tool in order to produce a set of features from each document in a subset of OHSUMED dataset. In phase 3, we employ Entropy as feature selection method for reducing the number of features which represents the dataset. Next, we perform feature weighting using term frequency-inverse document frequency (TF-IDF). During phase 4, we perform text classification using LIBSVM and evaluate the result that are collected.

4.0 Experimental Results

In this research, we perform text classification using LIBSVM. Then, the performance of text classification is measured using the standard information retrieval measures such as precision, recall and F-measure. Precision deals with the number of properly recognized items as a percentage if the number of items identified. In other words, it processes how many of the relevant documents retrieved (true positives) divided by the total number of documents retrieved (true positives + false positives). We refer to the Equation (1) for calculating the precision in our experiments.

$$Precision = tp/(tp + fp) \quad (1)$$

where tp and fp are the numbers of true positive and false positive, respectively.

While, Velupillai *et al*. (2009) defined the recall as to evaluate of the ability of a prediction model to select the occurrence of certain class from a dataset. It is also known as sensitivity which related to the true positive rate. We calculate the recall using the Equation (2) below;

$$Recall = Sensitivity = tp/(tp + fn) \quad (2)$$

where tp and fp are the numbers of true positive and false negative, respectively.

F-measure as defined by Sautot *et al*. (2015) is the weighted the harmonic mean of the two, precision and recall and it can be weighted towards either precision or recall, but for the purposes of this experiment, we weight them both equally. The F-measure is formulated as follow;

$$F = 2 \times \frac{(precision.recall)}{(precision+recall)} \quad (3)$$

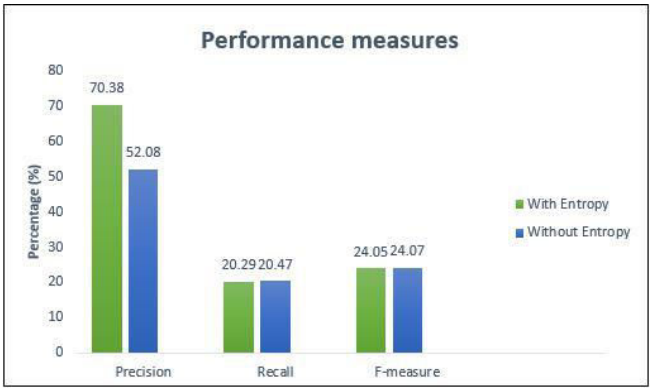

Figure 2: The performance of text classification for experiments using Entropy as feature selection method and for experiments without using feature selection method

Figure 2 illustrates the results of text classification experiments using two different set of features. For the experiments that not employ feature selection method, we use 3,599 features that are extracted from Genia Tagger tool. While, for the experiments that employ feature selection method, we use only 3,392 features that are selected by Entropy method.

5.0 Discussion

In this section, we discuss about the results of text classification using two different set of features. First set of features are extracted using Genia Tagger tool and second set of features are extracted using Genia Tagger tool and then, selected using Entropy method. In our experiments, we use 3,392 features that are extracted using Genia Tagger tool and then, selected by Entropy method and 3,599 features that are extracted from Genia Tagger tool.

As stated in Figure 2, generally we found the performance of precision for the experiments using feature selection method outperform the performance precision of the experiments without using feature selection method. The results show the average for precision with Entropy method is 70.38%, almost 18% higher than the average precision without Entropy method, which 52.08%. However, both averages for recall and F-measure for the classification experiments with Entropy method is slightly lower than both averages for recall and F-measure for the classification experiments without Entropy method. Based on the results obtained in our experiments, the average value for recall with Entropy method is 20.29%, which is 0.18% lower than the average value for recall without Entropy method is 20.47%, while average value for F-measure with Entropy method is 24.05%, whereby 0.02% lower than average value for F-measure without Entropy method is 24.07%. These results demonstrate that, even only 207 features are reduces during feature selection process, the performance of text classification would be increased. In addition, the use of different number of testing documents and training documents in our experiments also might influence the performance of classification accuracy.

6.0 Conclusion

Text classification has been recognized as one of the main approaches used to classify the digital data. However, the high volume text documents influences the performance of text classification due to high dimensionality problem. Therefore, we attempt to investigate the

effectiveness of employing feature selection method for reducing the number of features for classification purpose. In this research, we use Entropy method to select the relevant features for classifying biomedical text abstracts. Then, we perform classification using LIBSVM.

Based on the results obtained in the experiments, we conclude that employing Entropy as feature selection method for biomedical text classification helps reducing the dimension of terms in biomedical text abstracts. Even though only 207 features are reduced after feature selection process, the results that are collected from the experiments show slightly increased the performance of classification accuracy.

Generally, this research is contributing in information retrieval field which relates to search engine. Choosing the correct feature selection method would affects the selecting of relevant features, as it will helps reduce the high dimensional of data. By reducing the number of relevant features could reduce the computational time consuming and helps to increase the result of classification accuracy. Thus, would increases the performance of retrieve relevant information.

For the future work we have an interest to extend this research by focusing on the increase the total number of dataset and also use other sources of dataset in biomedical domain, such as MEDLINE, etc. In addition, we also could explore and employ many other feature selection methods to select the relevant features for text classification purpose.

References

Colace, F., De Santo, M., Greco, L., & Napoletano, P. (2014). Text classification using a few labeled examples. *Computers in Human Behavior, 30*(0), pp.689-697. doi:http://dx.doi.org/10.1016/j.chb.2013.07.043

Dalianis, H., Hassel, M., & Velupillai, S. (2011). Louhi 2010: Special issue on Text and Data Mining of Health Documents. *Journal of Biomedical Semantics, 2*(3), pp.1-4. doi:10.1186/2041-1480-2-S3-I1

Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications, 42*(10), pp. 4632-4642. doi:http://dx.doi.org/10.1016/j.eswa.2015.01.069

Fu, X., Batista-Navarro, R., Rak, R., & Ananiadou, S. (2015). Supporting the annotation of chronic obstructive pulmonary disease (COPD) phenotypes with text mining workflows. *Journal of Biomedical Semantics, 6*(1), pp. 1-11. doi:10.1186/s13326-015-0004-6

Ghareb, A. S., Bakar, A. A., & Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications, 49*, pp. 31-47. doi:http://dx.doi.org/10.1016/j.eswa.2015.12.004

Ishii, N., Koike, A., Yamamoto, Y., & Takagi, T. (2010). Figure classification in biomedical literature to elucidate disease mechanisms, based on pathways. *Artificial Intelligence in Medicine, 49*(3), pp.135-143. doi:http://dx.doi.org/10.1016/j.artmed.2010.04.005

Li, W., Miao, D., & Wang, W. (2011). Two-level hierarchical combination method for text classification. *Expert Systems with Applications, 38*(3), pp. 2030-2039. doi:http://dx.doi.org/10.1016/j.eswa.2010.07.139

Pedrycz, W., Lee, D. J., & Pizzi, N. J. (2009). Representation and classification of high-dimensional biomedical spectral data. *Pattern Analysis and Applications, 13*(4),pp. 423-436. doi:10.1007/s10044-009-0170-1

Perthame, É., Friguet, C., & Causeur, D. (2015). Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, pp.1-14. doi:10.1007/s11222-015-9569-2

Romero, R., Iglesias, E. L., Borrajo, L., & Marey, C. M. R. (2011). Using Dictionaries for Biomedical Text Classification. In M. Rocha, J. C. Rodríguez, F. Fdez-Riverola, & A. Valencia (Eds.), *5th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2011)*, Vol. 93, pp. 365-372: Springer Berlin Heidelberg.

Shu, W., Shen, H., Sang, Y., Li, Y., & Wu, J. (2014). A New Evaluation Function for Entropy-Based Feature Selection from Incomplete Data. In V. Tseng, T. Ho, Z.-H. Zhou, A. P. Chen, & H.-Y. Kao (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 8444, pp. 98-109): Springer International Publishing.3wqa4

Tsai, C.-F., Chen, Z.-Y., & Ke, S.-W. (2014). Evolutionary instance selection for text classification. *Journal of Systems and Software, 90*(0), pp.104-113. doi:http://dx.doi.org/10.1016/j.jss.2013.12.034

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*(1), pp.104-112. doi:http://dx.doi.org/10.1016/j.ipm.2013.08.006

Vlachos, M., Freris, N., & Kyrillidis, A. (2015). Compressive mining: fast and optimal data mining in the compressed domain. *The VLDB Journal, 24*(1), pp.1-24. doi:10.1007/s00778-014-0360-3

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., & Hua, L. (2012). Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of Medical Systems, 36*(4), pp.2431-2448. doi:10.1007/s10916-011-9710-5

Zhang, X., Mei, C., Chen, D., & Li, J. (2016). Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognition, 56*, pp.1-15. doi:http://dx.doi.org/10.1016/j.patcog.2016.02.013