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

Employing Information Gain as Feature Selection Method for Classification of Biomedical Text Abstracts

NUR FAATIHAH MD ZUBADI[1], ROZILAWATI DOLLAH @ MD ZAIN*[2]

Department of Information System, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[1]nurfaatihahzubadi26@gmail.com, [2]rozilawati@utm.my

Abstract

Nowadays, due to overwhelming of the amount of online published in biomedical articles, most researchers experiencing difficulties to manage and retrieve the required information. The purpose of this research is to investigate the approaches that could be used for extracting the informative and relevant terms from biomedical literatures. This research focuses on classification of biomedical texts abstracts using Information Gain (IG) as feature selection method to identify and select the meaningful terms or features to be selected. Therefore, we conduct and evaluate several experiments using a subset of OHSUMED dataset that involved features selection using IG for classifying biomedical text abstracts. This research tends to extract the meaningful features from a subset of OHSUMED dataset that can be used for various applications. The findings of this study can be used in search engine and information retrieval research areas.

Keywords: Information Gain, Feature Selection, Text Classification, OHSUMED, Biomedical Text Abstracts

1.0 Introduction

In this recent years, there has been an increasing amount of online biomedical literatures on web. Many researchers have interest to propose methods or techniques to solve issues on difficulties for organizing and retrieving the relevant information. Therefore, we need to select features that only informative and meaningful from biomedical literatures. Furthermore, most knowledge and other unstructured information are difficult to use and to integrate the information. It also demands for effective extract the words to index the articles or documents and make used of the meaningful features.

Recently, there are several approaches that has been proposed by many researchers to identify terms in biomedical literatures due to difficulties for users to find the effectively and efficiently ways for organizing data and retrieving relevant information from the text such as by performing classification. However, big issue of classifying is high dimensionality of data. In order to reduce high dimensionality, feature selection is the best solution to use. As mentioned by Guyon (2003), by employing the feature selection in classification, there are benefits can be gained such as reducing in time and storage.

Therefore, the purpose of this research is to improve the classification accuracy by reducing the high dimensionality features using information gain as feature selection method. Therefore, this research focuses on several activities such as investigate several feature selection methods for reducing high dimensionality data, employ information gain method for features selection process, conduct several classification experiments using a set of features before perform feature selection and a set of features after perform feature selection. In addition, this research also focuses on evaluate the effect of employing informati*on* gain as features selection method for classifying biomedical text abstracts.

This paper is organized as follows: In Section 1, we present the introduction of this paper. Then, Section 2 explains some related works. While in Section 3, we provide the research methodology that used in this research. Next, we discuss the experimental results in Section 4. In section 5, we present the discussion of the results that are obtained in our experiments and finally, Section 6 provides the conclusion of this research.

2.0 Related Work

Due to overwhelming amount of online published in biomedical articles, the uses of classification is being applied. The recent reviews on special journal issues and books show that broadly useful contents and information mining instruments are not appropriate for the biomedical area in light of the fact that it is highly specialized (Simpson & Demner-Fushman, 2012). Therefore, the management of data in biomedical literatures are difficult to retrieve relevant information and organize the material. However, there is used of technique selection to organize it properly.

In addition, we choose biomedical dataset due to the number of patients who suffer these diseases and also awareness among individual to get more information about heart disease because of the difficulties for them to find the relevant articles from web. So that, to improve the classification accuracy and retrieve more relevant information, various classification methods have been proposed. Effective knowledge management is a key element for the success in biomedical field. Moreover, text classification has many issues that being carried out in most of researchers' articles nowadays. Text classification have been used in various fields. For example, text classification also used for Arabic language that has been widely investigated (Al-Anzi et al., 2016). Other than that, many researchers also conduct researches that related to text classification (Chi et al., 2016; Iglesias et al., 2013; Jiang et al., 2016; Onan et al., 2016).

Furthermore, the main challenges of organizing database on biomedical literatures are due to high dimensional data, high dimensional features space and term recognition. Terms recognition occurs during features extraction process. In pattern classification problems, feature extraction is an important step and quality of features in discriminating different classes plays an important role in pattern classification problems. In addition, one of the solution to reduce high dimensional features is using features selection method. Meanwhile, for Banka & Dara (2015), feature selection is helpful as a pre-processing step for reducing dimensionality, removing irrelevant data, improving learning accuracy and enhancing output. In consequence, the proposed method used for simplifying the information to make the users retrieve the features easier in order to find the document about biomedical. Moreover, according to Dasgupta et al. (2007), discovered that the powerful tools that can simplify or speeding up computations is feature selection, and when it implemented suitably it can lead to little loss especially in classification quality. Based on the problem above, the purpose of this research

is to investigate the features selection methods used for selecting informative features from biomedical literature. This research focuses on classification of biomedical texts abstracts using features selection method to identify the meaningful words or features to be extracted.

3.0 Methodology

This section presents the methodology adopted in this research. Figure 1 illustrates the research steps involved in this research which includes collect the dataset, perform text reprocessing, perform feature selection process, calculate feature selection, conduct text classification experiments and analyze the results of classification.

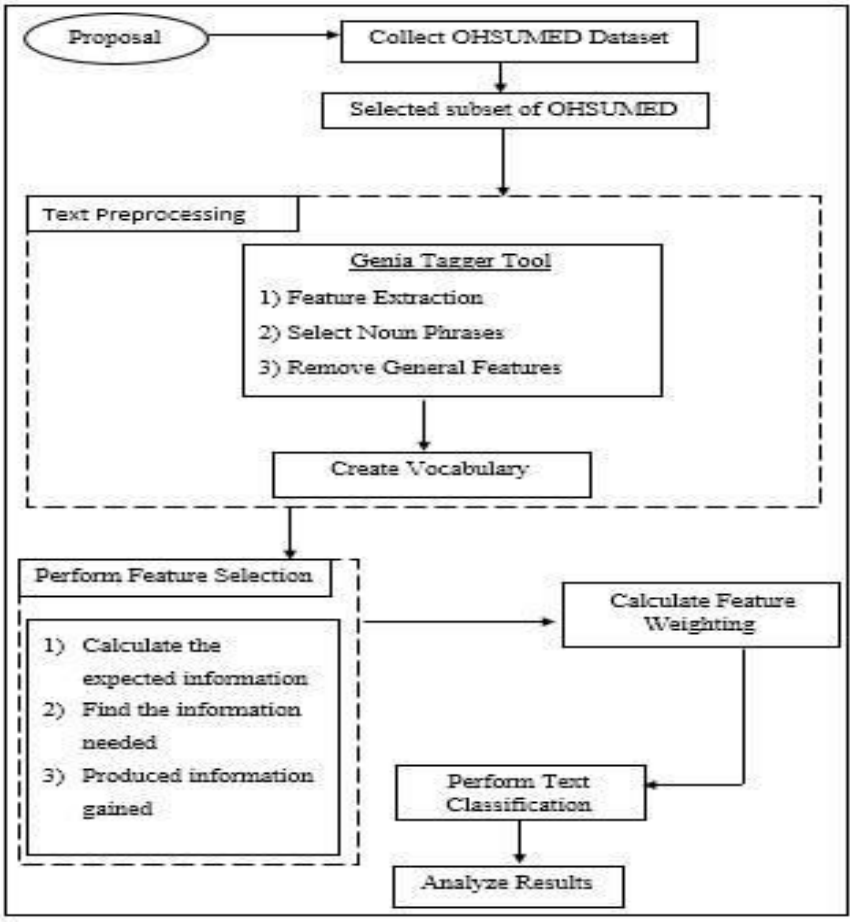

Figure1: Overview of research methodology

4.0 Experimental Results

In this research, we conduct several experiments using two sets of features before and after feature selection process. We perform the classification experiments using LIBSVM. The performance of text classification is measured using the standard information retrieval

measures such as precision, recall and F-measure. Precision can be defined as the number of relevant documents retrieved (true positives) divided by the total number of documents retrieved (true positives + false positives). Recall is the number of relevant documents retrieved (true positives) divided by the total number of relevant documents (true positives + false negatives). While, F-measure is the harmonic mean of the two and it can be weighted towards either precision or recall, but for the purposes of this experiment, we weight them both equally.

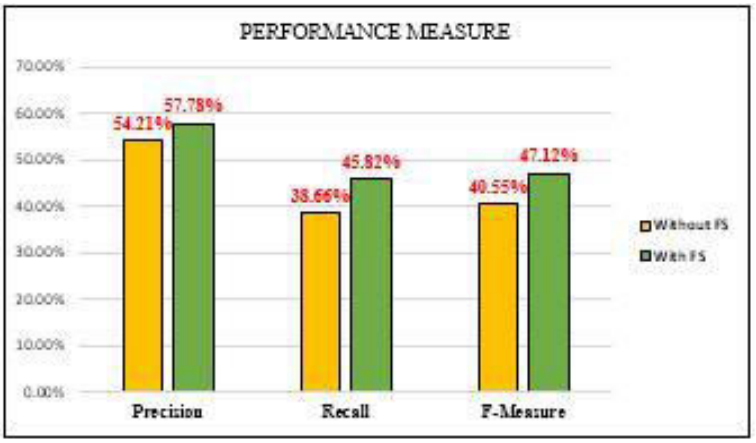

Figure 2: Result of the experiments using Information Gain as feature selection and without using Information Gain

Figure 2 shows the results of classification experiments using 9,313 features before feature selection and 9,283 features after feature selection process. Next, we compare the performance of classification between the result of experiments with Information Gain as feature selection and without Information Gain.

5.0 Discussion

One of the main challenge in performing text classification is managing high dimensionality data. Javed et al (2012) mention high dimensionality data may contains a large number of redundant and irrelevant features that influence the accuracy of text classification. Due to this reason, feature selection has becomes an important task to manage this problem. Therefore, in this research we perform feature selection using Information Gain for selecting the relevant features for text classification. For our experiments, we use 9,283 features that are selected using Information Gain and 9,313 features that are extracted from a subset of OHSUMED dataset.

Based on the results shown in Figure 2, overall, we found the performance of the experiment using feature selection slightly higher than the performance of the experiment without using feature selection method. The results of the experiment using Information Gain show the average value for precision, recall and F-Measure are 57.78%, 45.82% and 47.12%, respectively. While, the results of the experiment without using Information Gain show the

average value for precision, recall and F-Measure are 54.21%, 38.66%, and 40.55% respectively. These results might be cause by the number of features that are reduced only 30 features and not much influencing the result of classification performance.

6.0 Conclusion

In this research, we investigate the effectiveness of employing Information Gain for selecting the relevant features and also reducing the high dimensionality from a subset of OHSUMED dataset for classification purpose. Towards this effort, we conduct several experiments using a set of features that are extracted from the dataset and a set of features that are selected using Information Gain. Then, we evaluate the performance of classification accuracy for both sets of features.

Based on the results that are obtained from the experiments, our research did not much help in the problem of reducing high dimensionality. We found that Information Gain reduce only small number of features that are extracted from the dataset. Even though, only 30 features are reduced after feature selection process, the results that are produced from the experiments show slightly increased the classification accuracy. Therefore, for the future research, we have an interest to enhance this research which focuses on increase the total number of dataset. Furthermore, research may be carried out concerning different feature selection methods.

References

Al-Anzi, F. S., and AbuZeina, D. (2016). Towards an Enhanced Arabic Text Classification Using Cosine Similarity and Latent Semantic Indexing. Journal of King Saud University - Computer and Information Sciences. (2016)

Banka, H., and Dara, S. (2015). A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation. Pattern Recognition Letters, 52, 94-100.

Chi, N.-W., Lin, K.-Y., El-Gohary, N., and Hsieh, S.-H. (2016). Evaluating the strength of text classification categories for supporting construction field inspection. Automation in Construction, 64, pp.78-88.

Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., and Mahoney, M. W. (2007). *Feature selection methods for text classification.* Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, pp.1157-1182.

Iglesias, E. L., Seara Vieira, A., and Borrajo, L. (2013). An HMM-based over-sampling technique to improve text classification. Expert Systems with Applications, 40(18), pp.7184-7192.

Javed, K., Babri, H., and Saeed, M. (2012). Feature selection based on class-dependent densities for high-dimensional binary data. IEEE Transactions on Knowledge and Data Engineering, 24(3), pp.465-477.

Jiang, L., Li, C., Wang, S., and Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52, pp.26-39.

Onan, A., Korukoglu, S., and Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57, pp.232-247.

Simpson, M. S, and Demner-Fushman, D., (2012). Biomedical text mining: A survey of recent progress *Mining text data*: Springer, pp. 465-517.