Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Review of feature extraction approaches on biomedical text classification

Rozilawati Dollah [1,*], Tiara Izrinda Jafni [1], Haslina Hashim [1], Mohd Shahizan Othman [1], Abd Wahid Rasib [2]

[1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[2]Program of Geoinformation, Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 26 September 2019
Received in revised form
10 January 2020
Accepted 12 January 2020

Keywords:
Biomedical literature
Feature extraction
Feature selection
Text classification
Text mining

ABSTRACT

The overcoming volume of online biomedical literature causes congestion of data and difficulties in organizing these documents and also to retrieve the required documents from the database, especially in the Medline database. One of the solutions to surpass the overwhelming of documents is to apply classification. However, each document must be represented by a set of terminology or feature vectors. The identification of terminology or feature from biomedical literature is one of the most important and challenging tasks in text classification. This is due to a large number of new features and entities that appear in the biomedical domain. In addition, combining sets of features from different terminological resources leads to naming conflicts such as homonymous use of names and terminological ambiguities. Therefore, the purpose of this research is to investigate and evaluate the effective ways for extracting the relevant and meaningful features in order to increase the classification accuracy and improve the performance of web searches. Towards this effort, we conduct several classification experiments to evaluate and compare the effectiveness of feature extraction approaches for extracting the relevant and informative features from the biomedical literature. For our experiments, we use two different sets of features, which are a set of features that are extracted using the Genia tagger tool and set of features that are extracted by medical experts from Pusat Perubatan Universiti Kebangsaan Malaysia (PPUKM). The results show the performance of classification using features that are extracted by medical experts outperform the performance of classification using the Genia Tagger tool when applying feature selection method.

1. Introduction

Nowadays, the volume and growth rate of online biomedical literature creates new challenges for the researchers. MEDLINE (Sampson et al., 2016) is the primary source of medical literature, which consists of over 23 million online entries with a growth rate of over 800 thousand new citations every year. MEDLINE documents are manually categorized under 22,568 MeSH category names by experts from the National Library of Medicine (NLM). The accessibility of the extensive biomedical online collections presents new challenges to organize and retrieve the relevant documents from MEDLINE. Therefore, text classification could be one of the solutions to overcome these problems.

Text classification is one of the challenging research topics due to the need to organize and categorize the growing number of electronic documents worldwide. Text classification can help users to effectively handle and exploit useful information hidden in large-scale documents (Wang et al., 2016). In addition, Cohen (2006) mentioned that automated document classification could be a valuable tool for biomedical tasks that involve large amounts of text. Nowadays, text classification has been successfully applied to various domains such as topic detection, spam e-mailing filtering, SMS spam filtering (El-Alfy and AlHasan, 2016), web page classification (Sabbah et al., 2016; Selamat and Omatu, 2004) and author identification.

A conventional text classification framework consists of preprocessing, feature extraction, feature selection, and classification stages. The preprocessing stages usually comprise of tasks, such as stemming, stop word removal, tokenization, and

* Corresponding Author.
Email Address: rozilawati@utm.my (R. Dollah)
https://doi.org/10.21833/ijaas.2020.04.001
Corresponding author's ORCID profile:
https://orcid.org/0000-0001-6007-1749

lowercase conversion (Uysal and Gunal, 2014). The feature extraction stages generally utilize the vector space model (Salton et al., 1975) that makes use of the bag-of-words approach (Joachims, 1997). Finally, the feature selection stage typically uses the filter method such as document frequency (Azam and Yao, 2012; Yang and Pedersen, 1997), mutual information (Tang et al., 2019; Al-Angari et al., 2016; Liu et al., 2009), information gain (Mendez et al., 2019; Lee and Lee, 2006), chi-square (Asdaghi and Soleimani, 2019; Chen and Chen, 2011) and Odds Ratio (Raza and Qamar, 2016; Feng et al., 2015).

Many text classification approaches (Saqib et al., 2019; Wang et al., 2016; Thaoroijam, 2014; Aljaber et al., 2011; Fang et al., 2011; 2008; Hliaoutakis et al., 2009; Zhang et al., 2008; Li et al., 2007; Chen et al., 2006; Cohen, 2006; Couto et al., 2004; Kamruzzaman et al., 2005) were proposed for improving the results of classification accuracy and retrieving the relevant documents from database. However, the main problem while performing text classification is managing high dimensional data (Rizaldy and Santoso, 2017; Chandrashekar and Sahin, 2014; Khalid et al., 2014; Javed et al., 2012; Maji and Paul, 2011; Wei and Billings, 2007). According to Javed et al. (2012), high-dimensional data may contain a large number of redundant and irrelevant words or features that worsen the performance of a learning algorithm. One of the effective ways to reduce the high dimensionality of data is by performing feature selection.

Feature selection has become an essential and challenging task in which to analyze and select useful knowledge about a given domain. Traditionally, feature selection research has focused on removing irrelevant and redundant features as much as possible (Mirończuk and Protasiewicz, 2018; Maji and Paul, 2011). Recently, some researchers have focused on methods for effectively handling high dimensional datasets (Vinh et al., 2016; Chandrashekar and Sahin, 2014; Khalid et al., 2014). In addition, Vinh et al. (2016) stated that effective feature selection could improve performance while reducing the computational cost of the learning system. While, Dadaneh et al. (2016) mentioned that feature selection is one of the most important fields in pattern recognition, which aims to pick a subset of relevant and informative features from an original feature set. Many other researchers study on feature selection approaches to handle high dimensionality problem (Ghareb et al., 2016; Hernández-Pereira et al., 2016; Tutkan et al., 2016; Vinh et al., 2016; Feng et al., 2015; Pinheiro et al., 2015; Chandrashekar and Sahin, 2014; Inbarani et al., 2015; Khalid et al., 2014; Rehman et al., 2015; Javed et al., 2012; Maldonado and Weber, 2009; Wei and Billings, 2007). Although many feature selection approaches have been proposed and have been employed in various domains, there are still some issues, especially in retrieving the relevant documents.

Therefore, in this research, we investigate several feature selection methods or techniques that could be employed for classification. Most of the research papers that implemented the Odds Ratio produced better results. For example, Raza and Qamar (2016) presented a comparison of using feature selection methods towards large datasets such as Gisette, Isolate, Musk-2, UjlindoorLoc, Egg-Eye-style and Internet advertisement for classification purpose. They found that the use of Odds Ratio as a feature selection method produced high accuracy compared to other feature selection methods. In other research, Ding et al. (2016) proposed a classification method for predicting PH proteins and their distribution in a host cell. The use of feature selection method has been seen to improve the result of their research.

While Feng et al. (2015) performed a comparison among few feature selection methods such as Information Gain, Chi-squared and Odds ratio for classifying MPH-20 and 20 Newsgroups datasets. However, their results show the Odds Ratio and Information Gain outperformed Chi-square for 20 Newsgroups dataset. In other similar research, Tutkan et al. (2016) proposed a new feature selection method named Meaning Based Feature Selection (MBFS). Then, they compared the performance of their proposed feature selection method with other methods such as Information Gain, Chi-squared, Odds ratio. They found that the Odds ratio outperforms other feature selection methods.

Banerjee and Biswas (2012) made a comparison between the Mantel-Haenszel estimator and profile maximum likelihood (PMLE) for estimating the common Odds Ratio. Those estimators converge to the true value of the common Odds Ratio. The result shows that Odds Ratio leads to better performance. In addition, Gregory et al. (2008) used Odds Ratio to calculate cancer incidence from the AIC-minimizing, and their result shows the value that been selected from Odds Ratio leads to the highest performance.

Odds Ratio would be used as a feature selection technique to evaluate the effectiveness of biomedical text classification. Thus, this research focuses on how to execute Odds Ratio as a feature selection method for selecting the relevant and informative features from the candidate features that are extracted using the Genia Tagger tool and also the candidate features that are extracted by medical experts from Pusat Perubatan Universiti Kebangsaan Malaysia (PPUKM). This paper is divided into 5 sections. In Section 2, we describe the details of our methodology for conducting this research. Section 3 contains the experiments and in Section 4, the discussion of the classification results is stated. Finally, we conclude this paper in Section 5.

2. Methodology

This research is conducted based on the methodology as shown in Fig. 1. The details of the methodology are explained in details as follows:

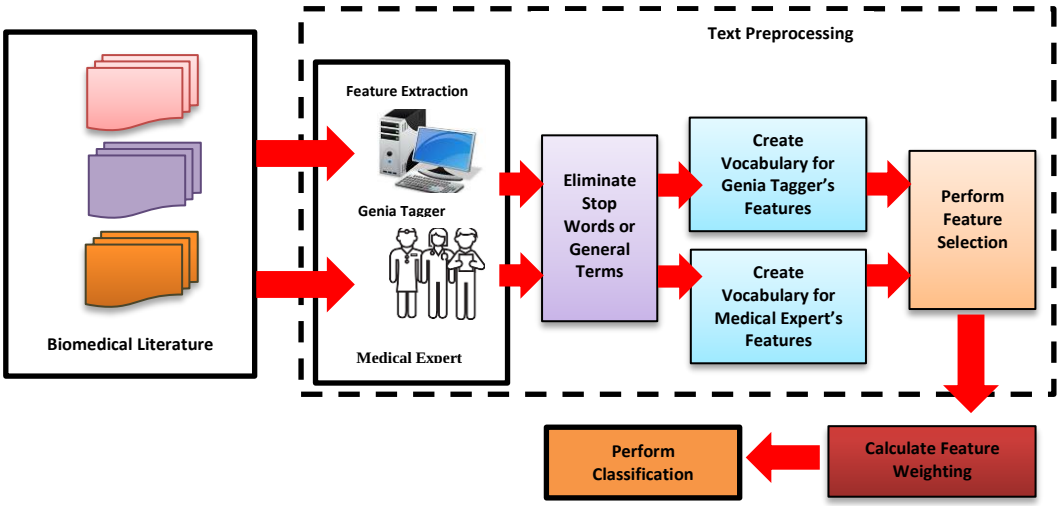

Fig. 1: A research methodology of the study

A. Data collection: In this research, we use the Ohsumed (2005) dataset. The Ohsumed (2005) dataset is a subset of the MEDLINE database, which is bibliographic database literature preserved in the National Library of Medicine. The Ohsumed (2005) data collection contains medical abstracts from Medical Subject Headings (MeSH) categories from the year 1987 to 1991. This dataset contains more than 350,000 documents however the total number of documents that are used in this research is around 5,215 biomedical text abstracts.
B. Text preprocessing: The purpose of text preprocessing is to extract and select the relevant and informative features or terms from the Ohsumed (2005) dataset for representing and indexing the document for text classification. Abstract or text usually holds a large number of unwanted, noise, and uninformative parts such as scripts, HTML tags and stop words. Keeping these unwanted parts will add to the high complexity of the problem. It causes the classification to be more complex and challenging since each word in the text is connected to each other. Eliminating the noisy data will solve the problem of the data being improperly preprocessed.

Typically, text preprocessing involves several steps such as tokenization, stop word elimination, expending abbreviation, stemming and finally feature selection (Al-Angari et al., 2016). However, Uysal and Gunal (2014) mentioned in their research, that a standard text classification framework consists of preprocessing, feature extraction and classification stages. Nevertheless, in this research, the text preprocessing process consists of feature extraction, eliminate stop words and general terms, create a vocabulary and apply feature selection method.

2.1. Perform feature extraction

The purpose of feature extraction is to produce a list of unique features from the dataset. In this research, we perform feature extraction using the GENIA tagger tool. GENIA tagger analyzes English sentences and outputs the base forms, part-of-speech (POS) tagging, phrase chunking and named entity tagging. GENIA tagger may detect the type of entities genes like DNA, RNA and protein name. In addition, this tagger is specifically modified for biomedical text such as the MEDLINE dataset. In this research, the sentences in each abstract assigned or tagged into all chunk types like a noun phrase, verb phrase, adjective, conjunction and etc. Fig. 2 shows the example of biomedical text abstracts from the Ohsumed (2005) dataset, meanwhile Fig. 3 illustrates the example of output after POS tagging and phrase chunking processes, whereby it still contains a few general terms and stop words such as adjectives, verb, conjunction and etc.

In contrast, we also perform the feature extraction process by cardiologist experts from Pusat Perubatan Universiti Kebangsaan Malaysia (PPUKM). In this research, the cardiologist experts identify and extract all the significant medical terms related to heart disease. For both feature extraction approaches, all the stop words and general terms are removed.

2.1.1. Eliminate stop words and general terms

In this phase, all stop words such as adjectives, conjunction and general terms or features from the training and testing documents are removed. While only noun and verb phrases are chosen from each abstract. The primary purpose of removing stop words and the general terms process is to eliminate the noise data.

2.1.2. Create a vocabulary list

Create a vocabulary process is the process of gather all the terms and words in all documents. So through the compilation, we can see the number of times the term has repeated itself in the dataset. The vocabulary list is required to perform the feature selection process. In our experiments, each document must be represented by a set of feature vectors. For that purpose, we create a list of unique

terms or features that are extracted from the Ohsumed (2005) dataset.

2.1.3. Perform feature selection

Feature selection is one of the most feasible solutions to reduce the dimensionality of the datasets by selecting the most informative features and still retains sufficient information for the classification task. Feature selection has many advantages, such as avoiding over-fitting, facilitating data visualization, reducing storage requirements and reducing training time. In this research, the purpose of applying the feature selection method is to reduce the dimensionality of data. This is because not all features are informative and would affect the classification performance.

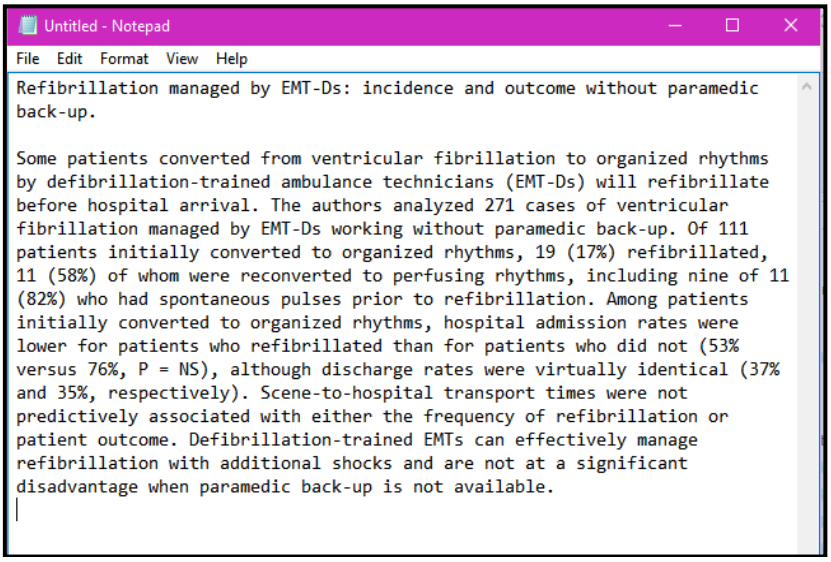

Fig. 2: An example of biomedical text abstracts from Ohsumed (2005) dataset

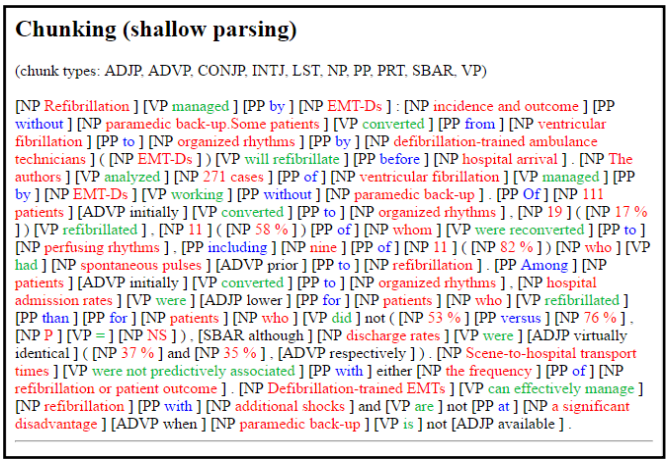

Fig. 3: An example of output after POS tagging and phrase chunking processes

Feature selection research has focused on eliminating redundant and irrelevant features as many as possible. Irrelevant features supply no useful information in any context and redundant features are those which provide no more information than the currently selected features. In this research, we select Odds Ratio as a feature selection technique. Odds Ratio evaluates whether the odds of a specific event or outcome is the same for two groups. Odds Ratio is using a simple equation as follows:

$$Odds\ Ratio = \frac{\frac{A}{C}}{\frac{B}{D}} = \frac{AD}{BC}.$$

where, A is the number of exposed cases; B is the number of exposed non-cases; C is the number of unexposed cases; D is the number of unexposed non-cases.

The list of the terms or features in the vocabulary file will be sorted and categorized into some range based on their frequency. The frequency of the terms will be used to calculate the standard error, 95% confidence interval and Odds Ratio result gained using Equation (1). The value of the Odds Ratio for this research is from 0.6 to 2.1. Thus, all terms in the abstracts that have the Odds Ratio values in the range of 0.6 to 2.1 are selected.

As a result of the features that are extracted using the Genia Tagger tool, 1,061 terms are selected over 25,837 terms, which cover 4.11% over the whole abstracts. While 958 terms of 30,028 terms are selected, which is 3.19% of the whole abstracts for the features that are extracted by medical experts.

2.2. Calculate feature weighting

We compute the feature weighting for each training and testing document. For our experiments, we use 4,643 documents that contain the selected features and their frequency. Meanwhile, for testing documents, we use 1,217 documents with the selected features and their frequency. Thus, we calculate each feature weight for both training and testing documents using the Term Frequency–Inverse Document Frequency (TF-IDF) equation as follows;

$$w_{i,d} = tf_{i,d} \log\left(\frac{n}{df_i}\right);\ \frac{n+1}{df_i+1},$$

where, term frequency $tf_{i,d}$ is the frequency of term i occurs in document j and d = 1, ..., m, document frequency df_i is the total number of documents that contain the term ,i and n is the total number of documents.

2.3. Perform text classification

For our experiments, we perform text classification using Library Support Vector Machine (LIBSVM). We conduct several experiments for text classification using a set of features that are extracted using the Genia Tagger tool and also set of features that are extracted by medical experts from PPUKM. Then, we compare the performance of both sets of features based on the precision, recall, and F-measure produced in the experiments. In addition, we also conduct the experiments using all features without performing the feature selection process for both sets of features.

The performance of text classification is measured using the standard information retrieval measures in terms of precision, recall, and F-measure using the following equation;

$$Precision = \frac{TP}{(TP+FP)}$$
$$Recall = \frac{TP}{(TP+FN)}$$
$$F - Measure = \frac{2*(Precision)*(Recall)}{(Precision)+(Recall)}$$

3. Experiments

In this research, we perform classification using the LIBSVM tool. Therefore, we conduct several experiments to compare the classification performance between the features extracted using the Genia Tagger tool and medical experts. In addition, we also compare the performance of classification accuracy for both sets of features that are employing a feature selection method and set of features without employing the feature selection method.

Table 1 and Table 2 illustrate the results of classification experiments using two sets of features extracted from Genia Tagger and medical experts, respectively. Finally, we compare the performance of classification between the result of experiments with Odd Ratio as feature selection and without Odd Ratio.

For the experiments that employ a feature selection method, we use 958 features extracted by medical experts and 1,061 features extracted by Genia Tagger. While, for experiments without feature selection method, we use 30,028 features extracted by medical experts and 25,837 features extracted by Genia Tagger.

Table 1: The performance of classification for experiments with the feature selection method

Experiment	Average Precision (%)	Average Recall (%)	Average F-Measure (%)
Features extracted by medical experts	65.38	42.37	39.57
Features extracted using Genia Tagger	61.14	40.49	35.42

Table 2: The performance of classification for experiments without feature selection

Experiment	Average Precision (%)	Average Recall (%)	Average F-Measure (%)
Features extracted by medical experts	55.99	41.69	39.41
Features extracted using Genia Tagger	67.41	39.00	36.41

4. Results and discussion

In this section, we discuss the performance of classification using a different set of features that are extracted by Genia Tagger and medical experts. From the result of classification experiments, the performance of classification accuracy for a different set of features that employ feature selection method and without employing a feature selection method would be compared. Overall, the results show different performance between the experiments using a different set of features that are extracted by Genia Tagger and medical experts and also the experiments using a different set of features with and without employing feature selection method. In addition, we also compare the performance of our experiment results with other researchers who have published in their work.

Generally, for the experiments with the feature selection method, the results of experiments using a set of features that are extracted by medical experts from PPUKM outperform the results of experiments using a set of features extracted by Genia Tagger. Table 1 and Table 2 show the experimental results produce in our experiments. The results show that the average value for precision, recall, and F-measure are 65.38%, 42.37% and 39.57%, respectively. While, the average value of precision, recall and F-measure for the experiments using a set of features extracted using Genia tagger are 61.14%, 40.49% and 35.42%, respectively. Compared to similar work done by Gong (2018), the researcher using Genia corpus 3.02 version in the experiments

and the experimental results produce the average precision 69.29%, average recall 56.92% and average F-Measure 62.31%, respectively.

From the results obtained in the experiments, we found that the proposed research to extract the relevant and informative features from biomedical literature such as Ohsumed (2005) dataset using Genia Tagger tool and Medical experts such as Cardiologist expert works well within its limitations. Even though, only 958 features are selected using Odd Ratio from 30,028 features that are extracted by medical experts, however, these experiments produce quite good results. This performance might be caused by the use of Odd Ratio as the feature selection method to eliminate most of the general medical terms or features from the original dataset. In addition, most probably the selected features are meaningful and informative features that influence the classification performance.

Subsequently, we compare the performance of classification experiments without employing a feature selection method for both sets of features that are extracted by Genia Tagger and medical experts. From the experiments, we found the performance measure for a set of features extracted from Genia Tagger shows a higher percentage of precision compared to medical experts. However, recall and F-measure values for a set of features extracted by medical experts illustrate a better percentage compared to Genia Tagger. These results indicate that the number of features (30,028 features) that are extracted by medical experts is higher than the number of features (25,837 features) that are extracted by Genia Tagger causes misclassification during the classification process.

5. Conclusion

Due to the excessive amount of biomedical literature in digital form, this causes difficulties in organizing and retrieving relevant information from the web. There are a few solutions that have been proposed to solve this problem, especially in the area of data mining, information retrieval, text mining, text classification, and machine learning techniques. In addition, many researchers are studying on classifying biomedical literature to handle the problem of organizing and navigating the websites and also to improve the accuracy of web searches. However, one of the problems raised in the text classification approach is the high dimensionality problem. One of the effective ways to reduce the high dimensionality of data is by performing feature selection. Employing an effective and efficient feature selection method could improve the performance of classification.

In this paper, we explore the effectiveness of feature selection methods for reducing the high dimensionality of features for text classification. Therefore, we conduct several classification experiments in order to evaluate the effectiveness of the feature selection method for reducing the high dimensionality of features. Generally, we conclude that employing the feature selection method for text classification could reduce the high dimensionality of features in biomedical literature and improve classification accuracy. For future research, we have an interest in increasing the number of Ohsumed (2005) dataset and also perform text classification using different feature selection methods in order to reduce the high dimensionality of features and also increase the performance of text classification, especially in biomedical literature area.

Acknowledgment

This study is supported by the Fundamental Research Grant Scheme (FRGS) under the Vote No. 4F559 that sponsored by the Ministry of Higher Education (MOHE) and Research University Grant Scheme (RUG) under the Vote No. 13J94 and 20H01. The authors are greatly obliged to Universiti Teknologi Malaysia (UTM) and Information Engineering and Behavioral Informatics (INFOBEE) Research Group for support and motivation.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Al-Angari HM, Kanitz G, Tarantino S, and Cipriani C (2016). Distance and mutual information methods for EMG feature and channel subset selection for classification of hand movements. Biomedical Signal Processing and Control, 27: 24-31. https://doi.org/10.1016/j.bspc.2016.01.011

Aljaber B, Martinez D, Stokes N, and Bailey J (2011). Improving MeSH classification of biomedical articles using citation contexts. Journal of Biomedical Informatics, 44(5): 881-896. https://doi.org/10.1016/j.jbi.2011.05.007 **PMid:21683802**

Asdaghi F and Soleimani A (2019). An effective feature selection method for web spam detection. Knowledge-Based Systems, 166: 198-206. https://doi.org/10.1016/j.knosys.2018.12.026

Azam N and Yao J (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications, 39(5): 4760-4768. https://doi.org/10.1016/j.eswa.2011.09.160

Banerjee B and Biswas A (2012). On closeness of the Mantel–Haenszel estimator and the profile likelihood based estimator of the common odds ratio from multiple 2×2 tables. Statistics and Probability Letters, 82(11): 1990-1993. https://doi.org/10.1016/j.spl.2012.06.013

Chandrashekar G and Sahin F (2014). A survey on feature selection methods. Computers and Electrical Engineering, 40(1): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen D, Müller HM, and Sternberg PW (2006). Automatic document classification of biological literature. BMC Bioinformatics, 7: 370. https://doi.org/10.1186/1471-2105-7-370 **PMid:16893465 PMCid:PMC1559726**

Chen YT and Chen MC (2011). Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications, 38(4): 3085-3090. https://doi.org/10.1016/j.eswa.2010.08.100

Cohen AM (2006). An effective general purpose approach for automated biomedical document classification. In the AMIA Annual Symposium Proceedings, American Medical Informatics Association, Washington D.C., USA: 161-165.

Couto FM, Martins B, and Silva MJ (2004). Classifying biological articles using web resources. In the 2004 ACM Symposium on Applied Computing, ACM, Nicosia, Cyprus: 111-115. https://doi.org/10.1145/967900.967925

Dadaneh BZ, Markid HY, and Zakerolhosseini A (2016). Unsupervised probabilistic feature selection using ant colony optimization. Expert Systems with Applications, 53: 27-42. https://doi.org/10.1016/j.eswa.2016.01.021

Ding H, Liang ZY, Guo FB, Huang J, Chen W, and Lin H (2016). Predicting bacteriophage proteins located in host cell with feature selection technique. Computers in Biology and Medicine, 71: 156-161. https://doi.org/10.1016/j.compbiomed.2016.02.012 **PMid:26945463**

El-Alfy ESM and AlHasan AA (2016). Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm. Future Generation Computer Systems, 64: 98-107. https://doi.org/10.1016/j.future.2016.02.018

Fang YC, Huang HC, and Juan HF (2008). MeInfoText: Associated gene methylation and cancer information from text mining. BMC Bioinformatics, 9: 22. https://doi.org/10.1186/1471-2105-9-22 **PMid:18194557 PMCid:PMC2258285**

Fang YC, Lai PT, Dai HJ, and Hsu WL (2011). MeInfoText 2.0: Gene methylation and cancer relation extraction from biomedical literature. BMC Bioinformatics, 12: 471. https://doi.org/10.1186/1471-2105-12-471 **PMid:22168213 PMCid:PMC3266364**

Feng G, Guo J, Jing BY, and Sun T (2015). Feature subset selection using naive Bayes for text classification. Pattern Recognition Letters, 65: 109-115. https://doi.org/10.1016/j.patrec.2015.07.028

Ghareb AS, Bakar AA, and Hamdan AR (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications, 49: 31-47. https://doi.org/10.1016/j.eswa.2015.12.004

Gong L (2018). Application of biomedical text mining. In: Aceves-Fernandez MA (Ed.), Artificial intelligence: Emerging trends and applications: 417-433. Books on Demand, Norderstedt, Germany. https://doi.org/10.5772/intechopen.75924

Gregory M, Ulmer H, Pfeiffer KP, Lang S, and Strasak AM (2008). A set of SAS macros for calculating and displaying adjusted odds ratios (with confidence intervals) for continuous covariates in logistic B-spline regression models. Computer Methods and Programs in Biomedicine, 92(1): 109-114. https://doi.org/10.1016/j.cmpb.2008.05.004 **PMid:18603325**

Hernández-Pereira E, Bolón-Canedo V, Sánchez-Maroño N, Álvarez-Estévez D, Moret-Bonillo V, and Alonso-Betanzos A (2016). A comparison of performance of K-complex classification methods using feature selection. Information Sciences, 328: 1-14. https://doi.org/10.1016/j.ins.2015.08.022

Hliaoutakis A, Zervanou K, and Petrakis EG (2009). The AMTEx approach in the medical document indexing and retrieval application. Data and Knowledge Engineering, 68(3): 380-392. https://doi.org/10.1016/j.datak.2008.11.002

Inbarani HH, Bagyamathi M, and Azar AT (2015). A novel hybrid feature selection method based on rough set and improved harmony search. Neural Computing and Applications, 26(8): 1859-1880. https://doi.org/10.1007/s00521-015-1840-0

Javed K, Babri HA, 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): 465-477. https://doi.org/10.1109/TKDE.2010.263

Joachims T (1997). A probabilistic analysis of the Rocchio algorithm with tf-idf for text categorization. In the 14th International Conference on Machine Learning (ICML'97), Morgan Kaufmann Publishers Inc., Nashville, USA: 143-151.

Kamruzzaman SM, Haider F, and Hasan AR (2005). Text classification using data mining. In the International Conference on Computer and Information Technology, Cyberjaya, Malaysia: 135-139.

Khalid S, Khalil T, and Nasreen S (2014). A survey of feature selection and feature extraction techniques in machine learning. In the Science and Information Conference, IEEE, London, UK: 372-378. https://doi.org/10.1109/SAI.2014.6918213

Lee C and Lee GG (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing and Management, 42(1): 155-165. https://doi.org/10.1016/j.ipm.2004.08.006

Li Y, Lin H, and Yang Z (2007). Two approaches for biomedical text classification. In the 1st International Conference on Bioinformatics and Biomedical Engineering, IEEE, Wuhan, China: 310-313. https://doi.org/10.1109/ICBBE.2007.83

Liu H, Sun J, Liu L, and Zhang H (2009). Feature selection with dynamic mutual information. Pattern Recognition, 42(7): 1330-1339. https://doi.org/10.1016/j.patcog.2008.10.028

Maji P and Paul S (2011). Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning, 52(3): 408-426. https://doi.org/10.1016/j.ijar.2010.09.006

Maldonado S and Weber R (2009). A wrapper method for feature selection using support vector machines. Information Sciences, 179(13): 2208-2217. https://doi.org/10.1016/j.ins.2009.02.014

Mendez JR, Cotos-Yañez TR, and Ruano-Ordás D (2019). A new semantic-based feature selection method for spam filtering. Applied Soft Computing, 76: 89-104. https://doi.org/10.1016/j.asoc.2018.12.008

Mirończuk MM and Protasiewicz J (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106: 36-54. https://doi.org/10.1016/j.eswa.2018.03.058

Ohsumed (2005). DataSet. Available online at: https://bit.ly/2SStkdf

Pinheiro RH, Cavalcanti GD, and Ren TI (2015). Data-driven global-ranking local feature selection methods for text categorization. Expert Systems with Applications, 42(4): 1941-1949. https://doi.org/10.1016/j.eswa.2014.10.011

Raza MS and Qamar U (2016). An incremental dependency calculation technique for feature selection using rough sets. Information Sciences, 343: 41-65. https://doi.org/10.1016/j.ins.2016.01.044

Rehman A, Javed K, Babri HA, and Saeed M (2015). Relative discrimination criterion–A novel feature ranking method for text data. Expert Systems with Applications, 42(7): 3670-3681. https://doi.org/10.1016/j.eswa.2014.12.013

Rizaldy A and Santoso HA (2017). Performance improvement of support vector machine (SVM) with information gain on categorization of Indonesian news documents. In the International Seminar on Application for Technology of Information and Communication, IEEE, Semarang, Indonesia: 227-232. https://doi.org/10.1109/ISEMANTIC.2017.8251874

Sabbah T, Selamat A, Selamat MH, Ibrahim R, and Fujita H (2016). Hybridized term-weighting method for dark web classification. Neurocomputing, 173(Part 3): 1908-1926. https://doi.org/10.1016/j.neucom.2015.09.063

Salton G, Wong A, and Yang CS (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11): 613-620. https://doi.org/10.1145/361219.361220

Sampson M, de Bruijn B, Urquhart C, and Shojania K (2016). Complementary approaches to searching MEDLINE may be sufficient for updating systematic reviews. Journal of Clinical Epidemiology, 78: 108-115. https://doi.org/10.1016/j.jclinepi.2016.03.004 **PMid:26976054**

Saqib SM, Kundi FM, Ahmad S, and Naeem T (2019). Automatic classification of product reviews into interrogative and non-interrogative: Generating real time answer. International Journal of Advanced and Applied Sciences, 6(8): 23-31. https://doi.org/10.21833/ijaas.2019.08.004

Selamat A and Omatu S (2004). Web page feature selection and classification using neural networks. Information Sciences, 158: 69-88. https://doi.org/10.1016/j.ins.2003.03.003

Tang X, Dai Y, and Xiang Y (2019). Feature selection based on feature interactions with application to text categorization. Expert Systems with Applications, 120: 207-216. https://doi.org/10.1016/j.eswa.2018.11.018

Thaoroijam K (2014). A study on document classification using machine learning techniques. International Journal of Computer Science Issues, 11(2): 217-222.

Tutkan M, Ganiz MC, and Akyokuş S (2016). Helmholtz principle based supervised and unsupervised feature selection methods for text mining. Information Processing and Management, 52(5): 885-910. https://doi.org/10.1016/j.ipm.2016.03.007

Uysal AK and Gunal S (2014). The impact of preprocessing on text classification. Information Processing and Management, 50(1): 104-112. https://doi.org/10.1016/j.ipm.2013.08.006

Vinh NX, Zhou S, Chan J, and Bailey J (2016). Can high-order dependencies improve mutual information based feature selection? Pattern Recognition, 53: 46-58. https://doi.org/10.1016/j.patcog.2015.11.007

Wang P, Xu B, Xu J, Tian G, Liu CL, and Hao H (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174: 806-814. https://doi.org/10.1016/j.neucom.2015.09.096

Wei HL and Billings SA (2007). Feature subset selection and ranking for data dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(1): 162-166. https://doi.org/10.1109/TPAMI.2007.250607 **PMid:17108391**

Yang Y and Pedersen JO (1997). A comparative study on feature selection in text categorization. In the 14th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., Nashville, USA: 412-420.

Zhang W, Yoshida T, and Tang X (2008). Text classification based on multi-word with support vector machine. Knowledge-Based Systems, 21(8): 879-886. https://doi.org/10.1016/j.knosys.2008.03.044