A Review of Feature Selection on Text Classification

Nur Syafiqah Mohd Nafis
Soft Computing and Artificial Intelligence Research Group
Faculty of Computer Systems & Software Engineering. University Malaysia Pahang, UMP
Lebuhraya Tun Razak 26300, Kuantan, Pahang, Malaysia
nsyafiqahmnafis@gmail.com

Suryanti Awang
Soft Computing and Artificial Intelligence Research Group
Faculty of Computer Systems & Software Engineering. University Malaysia Pahang, UMP
Lebuhraya Tun Razak 26300, Kuantan, Pahang, Malaysia
suryanti@ump.edu.my

Abstract— Textual data is a high-dimensional data. In high-dimensional data, the number of features exceeds the number of samples. Hence, it equally increased the amount of noise, and irrelevant features. At this point, dimensionality reduction is necessary. Feature selection is an example of dimensionality reduction techniques. Moreover, it had been an indispensable component in classification. Thus, in this paper, we presented three feature selection approaches; filter, wrapper and embedded. Their aims, advantages and disadvantages are also briefly explained. Besides, this study reviews several significant studies for each feature selection approach for text classification. Based on the studies, we found that wrapper approach is less used by researchers since it is prone to over-fit and exposed local-optima for text classification while filter and embedded achieved an amount of research. However, in filter approach, the classification accuracies cannot be guaranteed because it does not incorporate with any learning algorithm. Therefore, it concludes that embedded feature selection can offer a promising classification performance regarding classification accuracy and computational time.

Keywords— feature selection; text classification; high-dimensional

1. INTRODUCTION

With the rapid increment of technologies nowadays, the electronic or digital text is growing and evolving fast. Due to this phenomenon, textual data are developing from low-dimensional data to high-dimensional data. This high-dimensional text data are saturated with features. They are beneficial for text classification. However, not all features are relevant for text classification. Because of that, the researcher's had a lot of attention in reducing data dimensionality to improve text classification performance. Feature selection and feature transformation are the examples of dimensionality reduction techniques. Feature selection had an advantage of preserving the original semantics of the text data while feature transformation will transform feature into new independence features. Also, feature selection will help in avoiding the curse of dimensionality, imbalanced class data distribution, and decreasing over-fitting issue as well as to improve text classification accuracy. Hence, in this review, we focus on the application of various feature selection approaches for text classification.

In standard, there are three types of feature selection approaches; filter, wrapper, and embedded or hybrid. Filter approach is known as a classifier-independent feature selection since it does not interact with any classification algorithm. It measures feature based on feature's importance and relevance. Meanwhile, wrapper approach is a classifier-dependent approach since the outcome of the learning algorithm guides it. The third approach is embedded approach. It is also a classifier-dependent approach but can works together with filter or wrapper approaches to improve the classification accuracy. Among these three approaches, the embedded approach is claimed to be a good trade-off between performance and computational time of filter

and wrapper approaches [1]. Later in this paper, we will explain in detail on these feature selection approaches. Hence, we organized this paper as follows; in Section 2, it introduces feature selection as one of the dimensionality reduction techniques. Section 3 reviews on previous studies for three approaches of feature selection. Lastly, in Section 4, the conclusion will be made.

2. FEATURE SELECTION APPROACHES

Textual data are the most common data types available in all over the internet. They are unstructured since it does not have fix format [2]. Textual data is also claimed as high-dimensional data since the number of features always exceed the number of samples. Therefore, it required a dimensionality reduction technique. One of the techniques in dimensionality reduction is feature selection. Feature selection technique had advantages over feature transformation in dimensionality reduction. Feature selection had been an indispensable process in text classification as well as other Knowledge Discovery for Database (KDD) studies. Feature selection approach comprises of three categories which are a filter, wrapper, and embedded approaches as shown in Figure 1.

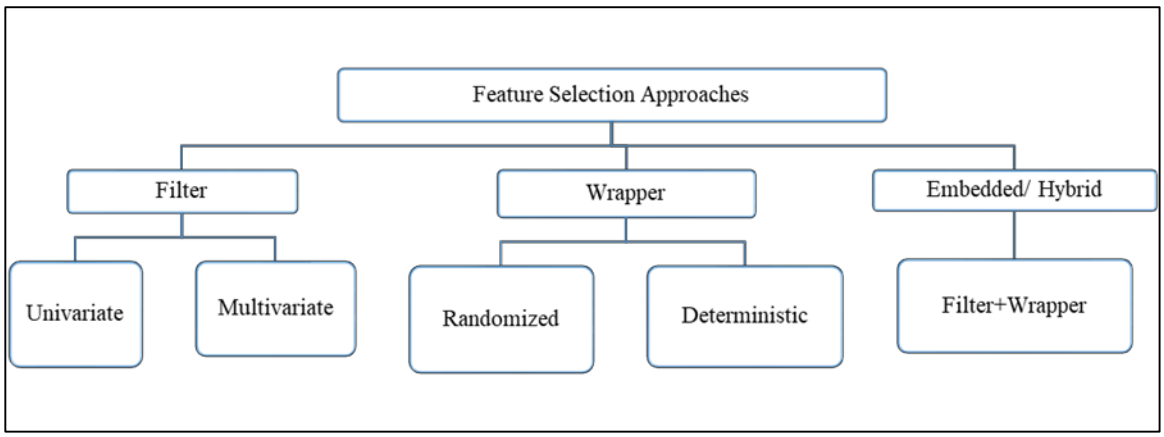

Figure 1: Feature selection approaches.

Filter approach will estimate the relevance index of the feature to evaluate how relevant a feature to the target. Then, this approach rank features based on relevance indices and search is performed according to rank or based on a statistical criterion [3]. However, filter approach has its drawback. The calculation of relevance index is on one single feature without considering the values on other features. Hence, it neglects the dependencies and interactions between features. Then, it may cause best pair features to be omitted. Even so, due to the simple approach in the filter, it proved to be more efficient and more robust to overfitting. Initially, filter approach only considers on univariate feature selection. However, multivariate is introduced to resolve some problem encounters on univariate filter feature selection approach. Multivariate filter approach models feature dependencies [4]. But, it is slower than univariate filter approach since it takes time to measure the feature dependencies and interactions. Mutual info, information gain and odd ratio can be categorized as filter feature selection techniques.

Unlike a filter approach, wrapper approach selects features by utilizing prediction performance of a classifier on a given subset. Thus, wrapper approach is claimed to be a classifier dependent approach. The wrapper approach can be applied to search through all possible feature subset and considering if there is also potential mutual information among features. Since it is classifier dependent, the choice of classifier became crucial. For text classification, classifier selection must be able to handle high-dimensional feature as well as significant data size; ability to deal with single-label or multi-label dataset and ability to manage noises because the wrapper approach measures the prediction performance by cross-validation or theoretical performance bound. Thus, careful classifier selection is essentials to reduce over-fitting, achieve high accuracy as well as to minimize computational time. Since the wrapper approach depends on search strategies [3], it is prone to over-fit problems, especially when dealing with high- dimensional data such as text classification. Moreover, wrapper approach needs higher computational resources and are often troublesome for large-scale issues [5]. Wrapper feature selection can be categorized into two-types; deterministic and randomized. The traditional way of wrapper approach is called deterministic technique. The deterministic approach commonly used greedy search in the learning process. Thus, it is exposed to get stuck in local-optima. Sequential forward selection, backward sequential elimination and beam search are the examples of deterministic wrapper approaches. Meanwhile, the randomized approach used Genetic Algorithm as a search strategy. Therefore, it has higher chances of over-fitting compared to deterministic approach.

Lastly, the embedded feature selection approach. It works a similarly with wrapper approach except embedded approach embeds either filter or wrapper feature selection approach and classifier into a single algorithm. In short, the embedded approach applied wrapper approach on features that already been selected by filter approach [6]. That is a reason why wrapper

approach reports better result compared to single filter and wrapper feature selection approaches. Moreover, it is also claimed to reduce computational time compared to wrapper approach. Support Vector Machine with Recursive Feature Elimination [7, 8], Weighted Naïve Bayes [9] and Odd Ratios + SVM-RFE [10] are the examples of embedded feature selection techniques.

As a whole, all feature selection approaches had their advantages and disadvantages and it is summarized in Table 1. Nevertheless, embedded or hybrid feature selection approach is claimed to be a good trade-off among the approaches. Furthermore, embedded feature selection approach is introduced to overcome the limitations exist in both traditional filter and wrapper approaches.

Table 1: Feature selection approach advantages and disadvantages.

Approach	Advantages	Disadvantages	Example
Filter	Simple, fast, classifier independent	Ignore feature dependencies and interactions	Mutual info, information gain, odd ratio (feature ranking and feature weighting)
Wrapper	Incorporate with classifier	Computationally intensive and risk for over-fitting	SFS, SBE, Genetic algorithm, randomized hill climbing
Embedded/hybrid	Interact with classifier but better computational complexity compared to wrapper	Classifier dependent selection but less risk for over-fitting	Weighted Naïve Bayes, Weight Vector of SVM, SVM-RFE

3. RELATED WORKS

In this section, we discussed related works in the features selection approaches.

A. Filter Feature Selection Approach for Text Classification

Some studies had been done to investigate how stable filter selection approach in classifying textual data. Abdul-Rahman, Mutalib, Khanafi, & Ali compare two filter techniques which are Information Gain (IG) and Gain Ratio (GR) [11]. IG is the favorite filter feature selection techniques [12]. It identified the presence and absence of a feature in the document. Then, IG calculates a score for each feature depending on how much information is gained concerning the class. Meanwhile, GR technique is an enhancement technique from IG. GR was introduced to solve bias issue in IG. In IG technique, it only selects feature with a higher score value. Thus, it contributes to a biased measurement. Besides, GR will maximize the feature's information gain score value and minimize the number of its value simultaneously. It maximizes the feature's information gain by using a split information value. The split information value measures the potential information generated by splitting the training data set into some partitions. The experiments were conducted to measure the performance of both filter selection techniques using Support Vector Machine (SVM). Evaluation methods used are the hold-out and k-fold in cross-validation method. K-fold in cross-validation strategy was not affected by any feature selection techniques. And no significant differences in modeling time are shown for both methods. Hence, authors concluded that SVM is competent in handling noise despite high dimensional data tested. Moreover, SVM can deal with complex problems such as text classification which has a significant amount of feature space and large samples. However, feature selection technique tested does not consider feature's importance to represent the whole text document.

Rehman, Javed and Babri in their study proposed a new filter selection approach [13]. The suggested technique is known as normalized difference measure (NDM), a feature ranking-based technique. NDM trying to overcome problem arose in balanced accuracy measure (ACC2) technique by [14]. In ACC2, it measures a feature by taking the difference of its document frequency in the positive class (also known as true positives) and its document frequency in the negative class (also known as false positives). However, these approach resulting in assigning equal ranks to features having a similar difference, neglecting their relative document frequencies in the classes. Therefore, in NDM corresponding document frequencies are considered. It works as a regularizer to the balanced accuracy of ACC2 by minimizing document frequency. Odds ratio (OR), chi-squared (CHI), information gains (IG), distinguishing feature selector (DFS), Gini index (GINI), balanced accuracy measure (ACC2) and the Poisson ratio (POIS) are the seven well-known feature ranking metrics investigated to compared with NDM. Seven datasets which are bACE (WAP, K1a, and K1b), Reuters (RE0, RE1), spam email dataset and 20 newsgroups are used to apply to all the techniques. Then, they chose the multinomial naive Bayes (MNB) and supports vector machines (SVM) as classifiers. The experimental trails on those datasets show that NDM metric outperforms the seven metrics in 66% cases regarding macro-F1 measure and 51% cases regarding the micro F-1 measure. This study successfully solved incorrect assessment of two terms having same ACC2 value by Forman [14]. However, feature selected from this proposed technique might not represent the whole document collection since feature's importance is not considered.

A study by Lioa and Pan obtains better performances when combining two filter selection approach classifying Chinese textual dataset [15]. They had proposed new averaged interaction gain (AIG) based filter with weight adjustment (WA) technique. This new technique is known as AIG-WA. This technique aims to overcome drawback in term of frequency-inverse document frequency (TF-IDF) technique which does not consider inner-class distribution and intra-class distribution of features. AIG algorithm is used to select features based on averaged interaction gain. Then, WA is assigned to the selected feature to take an inner-class distribution and intra-class distribution into consideration. They evaluate the proposed technique performance by comparing with other technique, for instance, AIG, mutual information (MI) and term frequency and mutual information (TF+MI). AIG-WA successfully helps to quicken the classification to achieve best performances. Even so, measuring average interaction gain between two features is still computationally expensive.

The study by Uysal & Gunal presented the new filter based probabilistic feature selection method, namely distinguishing feature selector (DFS) [16]. The authors concern about the computational time and accuracy when dealing with high-dimensional text classification. Hence, they proposed this technique. DFS selects distinctive features while removing uninformative ones considering specific requirements in term of characteristics. One example of the requirement is; a term or feature which frequently appears in a single class and does not appear in the other classes is a distinctive feature; therefore, it a high score is assigned. The proposed technique is compared with other well-known filter-based technique, for instances chi-square, information gain, Gini index and deviation from Poisson distribution. Besides, this comparative study experimenting for different classification algorithms, datasets (Reuters, Newsgroups, SMS and Enron1 dataset) and success measures with distinct characteristics, for example, feature similarity, accuracy, dimension reduction and processing time. Thus, in different circumstances, the effectiveness of DFS can be observed. From the experimental results, it shows that DFS obtained a considerably successful performance regarding accuracy, dimension reduction rate and processing time. This research has advantages, instead of calculating contributions of features to the class discrimination in a probabilistic approach; it also assigns specific importance scores to the features which strengthen the distinguishing quality of a feature. However, since this study implementing filter approach, the classification accuracy can be disputed because any learning algorithm is not use to measured classification performance.

Table 2: Previous studies on filter selection approach.

Study	**Feature Selection Method**	**Strength**	**Limitation**
Abdul-Rahman, Mutalib, Khanafi, & Ali (2013) [11]	Information Gain and Odd Ratio	No significant difference when tested using Support Vector Machine	Feature selection technique tested does not consider feature's importance to represent the whole text document.
Rehman, Javed, & Babri (2017) [13]	Normalized difference measure (NDM)	NDM metric outperforms the other seven feature ranking metrics	Feature selection technique tested does not consider feature's importance to represent the whole text document.
Lioa & Pan (2009) [15]	Chi-square (CHI) + GA + ant colony optimization (ACO)	AIG-WA successfully helps to quicken the classification to achieve best performances	Measuring average interaction gain between two features is still computationally expensive.
Uysal & Gunal (2012) [16]	Distinguishing feature selector (DFS)	Calculating contributions of features to the class discrimination in a probabilistic approach and assigns specific importance scores to the features	Classification accuracy can be disputed since it is a filter approach

Table 2 list the previous works on filter selection approach. From the studies on filter feature selection approach, it can be summarized that filter feature selection approaches offer multiple ways to calculate the relevance of a feature for text classification. Even so, it cannot guarantee the classification accuracies since it does not cooperate with any learning algorithms or classifiers.

B. Wrapper Feature Selection Approach for Text Classification

In text classification, the implementation of wrapper approach seems to be unsuitable since it incorporates with the classifier and textual data are high in dimensional [17]. Hence, it is prone to over-fitting.

However, Aghdam *et al*. [18] able to conduct a study using a wrapper feature selection approach. The motivation of this research is as no perfect heuristic can guarantee optimality even for medium size datasets while stochastic approaches offer a promising feature selection mechanism compared to heuristic search. They proposed a modified ant-colony optimization-based (ACO) feature selection. The classification performance and the length of selected feature subset are the heuristic information for ACO. Hence, no prior knowledge is required. The new modified ACO is implementing to text features of a bag of words. Then, they compared the performance of new modified ACO to the performance of Genetic Algorithm (GA), chi-squared and Information Gain (IG) by using Reuter's dataset. The suggested method presents some superiority on the Reuter's dataset compared to CHI and IG since ACO can converge quickly. It also has a robust search capability in the feature space and can efficiently find the least number of feature subset. Nevertheless, for a large dataset, the parallel algorithm may be required to speed the computation of reducts in ACO.

C. Embedded Feature Selection Approach for Text Classification

Previous studies had highlighted several embedded feature selection techniques for text classification. For embedded feature selection, Forman researched to classify text data by integrating Odd Ratio (OR) into the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithm [10] . OR filter selection approach is utilized to roughly and rapidly select a feature subset. Besides, the OR acts as one-sided feature selection because it gives the highest preference to the unusual features occurring in only the positive class. It calculates the probability of a feature occurring in the positive class normalized by the possibility of the feature appearing in the negative class alone [19]. Moreover, OR technique is developed for the binary dataset. Then, SVM-RFE is implemented to the selected feature (from OR) to acquire a smaller but significant feature subset. This integration technique is known as OR+SVM-RFE. Although OR+SVM-RFE produces lesser feature subset, the experimental results prove that it provides a good classification performance. Moreover, it did calculate the importance of the feature to represent the whole text collections. However, the OR is one-sided feature selection technique which leads to bias evaluation. Hence, it decreased the classification accuracy.

Next, Uğuz conducted a study to implement the two-phase embedded feature selection approach for text categorization [20]. These two-phases feature selection utilized information gain (IG) at the filter phase as the first stage. IG ranked each feature within the document regarding on the feature importance for classification. Next, in the second phase, GA and principal component analysis (PCA) are used independently for feature selection and feature extraction. The feature is sorted in decreasing order of importance. Then, features with high importance score or principal components are selected for feature selection using GA and feature extraction using PCA. Therefore, it decreased the computational time and complexity when a fewer number of features undergo feature selection and feature extraction. Reuters-21,578 and Classic3 datasets collection are used to assess the usefulness of the proposed technique. The results acquired show that suggested technique (IG-GA and IG-PCA) able to give high classification accuracy. The IG feature selection technique successfully selects features according to the importance of classification, but it still required for feature selection and feature extraction. In this study, it reveals that the success of text classification using the C4.5 decision tree and KNN algorithms with fewer features selected via IG-PCA and IG- GA is better than the success obtained using features chosen via single IG. Hence, it shows that this two-stage feature selection approach can enhance the classification performance. However, since this is the two-stage feature selection approach, it required a longer runtime.

Meanwhile, Taylor et.al. demonstrated on how an embedded approach feature selection works by integrating Chi-square (CHI) (filter feature selection) with the positive feedback mechanism of ant colony optimization (ACO) (wrapper feature selection) [6]. Besides, the fast global search ability of GA is also applied. At the first phase, CHI technique will select the most distinguished features. Hence, it will lessen the processing time for next wrapper feature selection phase. In the next phase, these selected features will be the input features. In the second phase, ACO and GA are cooperating with each other. ACO and GA are the randomized search method. They work efficiently in a way that both of them use the better search results of the other on the selected features generated from the first phase. The results from experimenting series of Reuters-21578 corpus reveal that proposed technique acquired better performance than other techniques. The suggested technique successfully improves the classification accuracy and efficiency. However, Chi-square technique does not calculate the feature importance. It estimates independence between features and category only. Feature importance is crucial to select the most useful features of the text. Furthermore, GA requires an extra run-time since is a stochastic search method.

Jiang *et al*. [9] had performed a study on embedding feature selection approach. In this technique, it incorporates deep feature weighting (DFW) for Naïve Bayes. At the first place, they purposed of this technique is to reduce the assumption of Naïve Bayes feature independency since it had less attention from researchers. They indict that previous feature weighting approaches only incorporate the trained feature weights into the classification of the formula of Naive Bayes but do not incorporate the learned feature weights into its conditional probability estimates at all. Therefore, they suggested DFW for Naïve Bayes. In this technique, it measures the conditional probabilities of Naive Bayes by deeply calculating the weighted feature frequencies from training data. Feature weight acts as a predictive weight on the features based on the degree to which

they rely on the values of other features. They had implemented proposed technique on 36 benchmark dataset from UCI Repository. The experimental result proved that proposed method rarely reduce the quality of the model compared to standard Naive Bayes but it further enhanced it dramatically. Then, they applied proposed DFW technique on other three Naive Bayes-based classifiers; multinomial naive Bayes (MNB), complement naive Bayes (CNB) and the one-versus-all-but-one model (OVA) to access the effectiveness of the proposed method on the state-of-art Naïve Bayes text classifiers. The results also reveal that some significant improvement on MNB, CNB, and OVA technique. However, all these Naïve Bayes-based techniques still correlated with Naive Bayes assumption. Furthermore, DFW with the standard Naïve Bayes is lacking in estimating the feature importance.

Lastly, Guo *et al*. [1] had conducted an outstanding research for clinical text classification. They introduced a new technique called as an ensemble embedded feature selection (EEFS). In this research, it deals with multi-label feature selection. EEFS train a classification model by selecting partial training example randomly which is an ensemble method. Then, it used the evaluation measure and averaged training examples for each column to test the trained models repeatedly to determine the final feature importance ranking. However, EEFS implicitly find out the correlations among labels, but it can sufficiently utilize the label correlations by multi-label classifiers and the evaluation measures. As a result, EEFS achieved an outstanding improvement. EEFS use the prediction risk and forward search strategy to calculate the importance of features to generate the feature subset. Hence, the classifier performance is better. Even so, this proposed technique is only appropriate for multi-label textual data. Implementing this technique on single-label dataset will lead to worse the classification performance.

As a conclusion, from the review of the previous studies, embedded feature selection successfully able to overcome weakness and limitation on filter and wrapper feature selection approaches. Table 3 summarized the strength and limitation for all reviewed papers. Overall, techniques proposed in these papers do not measure or calculate the feature importance or feature relevance for text classification. Feature importance or relevance is crucial in text classification since feature will represents the whole text documents for text classification.

Table 3: Previous studies on embedded feature selection method for text classification.

Study	**Feature Selection Method**	**Strength**	**Limitation**
Forman [10]	Odd Ratio + SVM-RFE	Roughly and quickly select a feature subset	OR is one-sided feature selection, it prone to bias.
Uğuz [20]	Information gain (IG)+Genetic Algorithm (GA) + principal component analysis (PCA)	Technique is applied to features with high importance order	GA technique requires a longer runtime.
Taylor *et al*. [6]	Chi-square (CHI) + GA + ant colony optimization (ACO)	Use the better search results of the other on the selected features produced from the first stage	Chi-square - measures independence between features and category but not the importance of the feature
Jiang *et al*. [9]	Deep feature weighting (DFW) for Naïve Bayes	Incorporate the learned feature weights into its Naïve Bayes	Lacking in calculating the feature importance.
Guo *et al*. [1]	Ensemble embedded feature selection	Achieved a significant classification improvement for multi-label clinical data	Only applicable for multi-label textual data

4. CONCLUSION

In this paper, we had review feature selection approaches and how it contributes to dimensionality reduction for high-dimensional data classification, especially for text classification. Efforts had been made in past few years to develop multiple feature selection techniques for text classification. From our broad investigation, filter approach had been emphasizing as the most frequently used feature selection for text classification because filter approach offers fast and straightforward computation time. Hence, it is acceptable to deal with high-dimensional text data. Even so, filter approach is often neglecting feature's interaction and dependency. Moreover, it is not classifier friendly approach. Meanwhile, wrapper approach had the least number of studies in text classification domain since wrapper approach is a classifier dependent approach with high-risk for over-fitting and highly exposed to local-optima. Nevertheless, the embedded approach is the most acceptable one to

handle high-dimensional textual data since embedded approach can guarantee the classification performance which also a classifier-dependent approach with less risk to over-fitting and without losing the essential features compared to wrapper approach. In this paper, it also presents several significant studies on each feature selection approach. As a result, embedded feature selection approach has a good trending in this current years.

As a conclusion, the most crucial step is to select the most appropriate feature selection approach for text classification. Dash and Liu [21] had summarized three characteristics to choose the right feature selection approaches. They are data types, data size, and noise. Features and class label are the two things to be considered when choosing right feature selection. Feature value can be continuous, discrete, nominal and Boolean. Meanwhile, the class label can be a single class label or multiple class labels. Next characteristic is data size. Feature selection approach can be separated based on their ability to handle with small training size or large training size. However, a method which can deal larger data size is preferable for text classification since textual data are high in dimensional. Last but not least is the ability to handle noises.

ACKNOWLEDGMENT

We would like to show our gratitude to Universiti Malaysia Pahang (RDU vote number RDU180380) for supporting this study.

REFERENCES

[1] Y. Guo, F. Chung, and G. Li, "An ensemble embedded feature selection method for multi-label clinical text classification," in: Proc. - 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016, pp. 823–826, 2017.
[2] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," Inf. Fusion, vol. 28, pp. 45–59, 2016.
[3] Y. Kuang, "A Comparative Study on Feature Selection Methods and Their Applications in Causal Inference," Computer, 2009.
[4] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.
[5] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Mach. Learn. Proc. 1994, pp. 121–129, 1994.
[6] P. Taylor, M. B. Imani, M. R. Keyvanpour, and R. Azmi, "A Novel Embedded Feature Selection Method : A Comparative Study in The Application of Text Categorization," pp. 37–41, 2013.
[7] X. Li, S. Peng, J. Chen, B. Lü, H. Zhang, and M. Lai, "SVM-T-RFE: A novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles," Biochem. Biophys. Res. Commun., vol. 419, no. 2, pp. 148–153, 2012.
[8] S. Mishra and D. Mishra, "SVM-BT-RFE : An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm," Karbala Int. J. Mod. Sci., vol. 1, no. 2, pp. 86– 96, 2015.
[9] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," Eng. Appl. Artif. Intell., vol. 52, pp. 26–39, 2016.
[10] G. Forman, "Feature Selection for Text Classification using OR+SVM-RFE," Comput. Methods Featur. Sel., vol. 16, pp. 257–274, 2010.
[11] S. Abdul-Rahman, S. Mutalib, N. A. Khanafi, and A. M. Ali, "Exploring Feature Selection and Support Vector Machine in Text Categorization," 2013 IEEE 16th Int. Conf. Comput. Sci. Eng., pp. 1101–1104, 2013.
[12] G. Forman, H. Labs, P. M. Road, and P. Alto, "A Pitfall and Solution in Multi-Class Feature Selection for Text Classification," 2004.
[13] Rehman, K. Javed, and H. A. Babri, "Feature selection based on a normalized difference measure for text classification," Inf. Process. Manag., vol. 53, no. 2, pp. 473–489, 2017.
[14] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Mach. Learn.Res., vol. 3, pp. 1289–1305, 2003.
[15] Y. Lioa and X. Pan, "A Novel Feature Selection Approach and Feature Weight Adjustment Technique in Text Classification," Trial, pp. 6–9, 2009.
[16] K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Syst., vol. 36, pp. 226–235, 2012.
[17] W. Zong, F. Wu, L. K. Chu, and D. Sculli, "A discriminative and semantic feature selection method for text categorization," Int. J. Prod. Econ., vol. 165, pp. 215–222, 2015.
[18] M. H. Aghdam, N. Ghasem-aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," Expert Syst. Appl., vol. 36, no. 3, pp. 6843–6853, 2009.
[19] M. Alibeigi, S. Hashemi, and A. Hamzeh, "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," Data Knowl. Eng., vol. 81–82, pp. 67–103, 2012.
[20] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," Knowledge-Based Syst., vol. 24, no. 7, pp. 1024–1032, 2011.
[21] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, no. 1–4, pp. 131–156, 19