

You-wei Wang, Li-zhou Feng, 2018. A new feature selection method for handling redundant information in text classification. *Frontiers of Information Technology & Electronic Engineering*, 19(2):221-234. <https://doi.org/10.1631/FITEE.1601761>

# A new feature selection method for handling redundant information in text classification

**Key words:** Feature selection; Dimensionality reduction; Text classification; Redundant features; Support vector machine; Naïve Bayes; Mutual information

Corresponding author: You-wei WANG

E-mail: [ywwang15@126.com](mailto:ywwang15@126.com)



ORCID: <http://orcid.org/0000-0002-3925-3422>

# Motivation

1. Traditional feature selection methods cannot filter the redundant features, showing lower accuracy than that of MI-based methods when the numbers of the selected features are equal.
2. The words those are filtered by the MI-based methods cannot be represented when they occur in a document, possibly missing some helpful category discrimination information.
3. Most MI-based feature selection methods calculate the correlations not only between the candidate features and all categories, but also between the candidate features and the selected features; thus, the time complexities of these methods are high when the numbers of all words and the selected features are both very large.

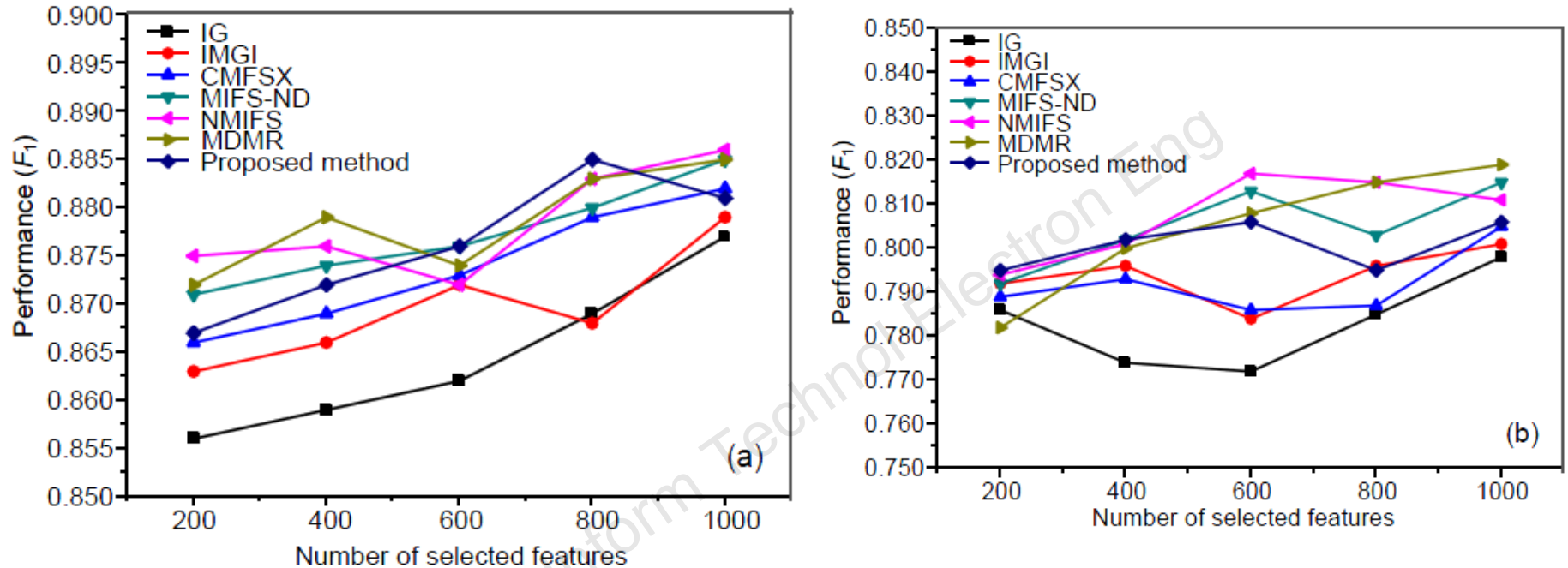
# Main idea

1. A new simple feature selection method, which can efficiently improve the speed of filtering of the redundant features while simultaneously ensuring the classification accuracy is proposed.
2. The definitions of word frequency-based relevance and correlative redundancy are introduced.
3. An optimal feature selection method (OFS) is used to select a feature subset ( $FS_1$ ) and filter the redundant features contained in  $FS_1$  by combining a predetermined threshold.

# Method

1. To calculate the relationship between two words, the definitions of word frequency-based relevance and correlative redundancy are introduced.
2. An optimal feature selection (OFS) method is chosen and used to obtain a feature subset  $FS_1$ .
3. To improve the execution speed, the redundant features in  $FS_1$  are filtered by combining a predetermined threshold, and the filtered features are memorized in the linked lists.

# Major results



**Fig. 5** Performance (values of  $F_1$ ) of different feature selection methods on the WebKB dataset when support vector machine (SVM) (a) and naïve Bayes (NB) (b) are used (References to color refer to the online version of this figure)

When SVM and NB are used, the proposed method outperforms other methods on the aspects of  $F_1$  values in most of the cases.

# Major results (Cont'd)

Tables 6 and 7 show that the proposed method is more advantageous than traditional methods in terms of execution speed and filters the redundant information of the selected features effectively.

**Table 6** Improvements in classification accuracy ( $F_1$ ) of the proposed method over different optimal feature selection methods

| Classifier | Dataset | $F_1$ of the proposed method over different optimal feature selection methods |       |       |       |       |       |       |       |
|------------|---------|-------------------------------------------------------------------------------|-------|-------|-------|-------|-------|-------|-------|
|            |         | DIA                                                                           | IG    | IMGI  | OCFS  | CHI   | MI    | DF    | CMFSX |
| SVM        | WE      | 0.015                                                                         | 0.003 | 0.004 | 0.017 | 0.002 | 0.017 | 0     | 0.015 |
|            | NE      | 0.023                                                                         | 0.013 | 0.021 | 0.012 | 0.011 | 0.017 | 0.008 | 0.005 |
|            | RE      | 0.019                                                                         | 0.017 | 0.003 | 0.009 | 0.024 | 0.003 | 0.003 | 0.012 |
| NB         | WE      | 0.007                                                                         | 0.015 | 0.004 | 0.007 | 0.018 | 0.019 | 0.002 | 0.009 |
|            | NE      | 0.005                                                                         | 0.006 | 0.009 | 0.014 | 0.002 | 0.012 | 0.007 | 0.005 |
|            | RE      | 0.014                                                                         | 0     | 0.008 | 0     | 0.008 | 0.021 | 0.011 | 0.007 |

SVM: support vector machine; NB: naïve Bayes; WE: WebKB; NE: 20-Newsgroups; RE: Reuters-21578

**Table 7** The average of the difference in execution time ( $t_{da}$ ) between the proposed method and other methods

| Dataset | $t_{da}$ between the proposed method and other methods (s) |        |        |         |        |        |
|---------|------------------------------------------------------------|--------|--------|---------|--------|--------|
|         | IG                                                         | IMGI   | CMFSX  | MIFS-ND | NMIFS  | MDMR   |
| WE      | -0.375                                                     | -0.492 | -0.458 | 5.556   | 8.983  | 12.136 |
| NE      | -0.427                                                     | -0.453 | -0.611 | 10.139  | 15.782 | 33.552 |
| RE      | -0.439                                                     | -0.517 | -0.419 | 14.227  | 16.694 | 46.374 |

WE: WebKB; NE: 20-Newsgroups; RE: Reuters-21578

# Conclusions

By comparing the proposed method with several typical feature selection methods on three data sets (WebKB, 20-Newsgroups, and Reuters-21578), we found that the proposed method can effectively enhance the performances in terms of classification accuracy and execution speed when compared with typical traditional feature selection methods and typical MI-based feature selection methods.