# Investigation of Feature Selection Techniques on Performance of Automatic Text Categorization

**Dilip Singh Sisodia and Ankit Shukla**

## 1 Introduction

Nowadays, digital documentation is increasing at a very fast pace, and it is very important to maintain the classification of digital documents. The main aim of digital document classification is to categorize the documents into predefined classes. It is an active research area for the information retrieval [1] and machine learning from the digital text documentation. There are many supervised algorithms which are employed on the digital text documents for the classification such as support vector machine [2], Naïve Bayes [3], decision tree [4], and nearest neighbors [5].

There are two phases of text categorization [6] of digital documents: One is the training phase, and the second is classification testing phase. Earlier, subject indexing and feature extraction method [7] were used for text categorization. However, these methods are not very much successful for the classification. Text categorization methods are based on the term frequency and inverted term frequency and count the frequencies of the term but not consider the position of the term. Therefore, these methods were not efficient in articulating the class for the text data. In each data, the position of the term is very relevant for the identification of the documents.

The remaining paper is organized as follows: Sect. 2 discusses the related work. In Sect. 3, material and methodology used for this work are discussed. Section 4 describes the experimental results and discussions. Lastly, Sect. 5 concludes this study.

D. S. Sisodia (✉)
National Institute of Technology, Raipur, India
e-mail: dssisodia.cs@nitrr.ac.in

A. Shukla
Jaypee University of Engineering & Technology, Guna, India
e-mail: ankeetshk@gmail.com

R. K. Shukla et al. (eds.), *Data, Engineering and Applications*,
https://doi.org/10.1007/978-981-13-6347-4_7

## 2 Related Work

Earlier, the text classification was done manually, but those classifications were not at all efficient. After that, many classification schemes came to existence such as subject indexing [8], term frequency [9], Gini index [10], mutual information, and information gain [11]. Till now, a significant amount of research has been done in automatic text categorization (ATC). Term frequency and subject indexing also used for classification, but these techniques were using the phenomenon of term redundancy [12] and subject index but missing the relevancy of the term. Gini index is also a global feature selection method for text classification. It is an improved version attribute selection algorithm. Currently, the weighted feature selection [13] algorithms are used for automatic text categorization since it is based on the mutual information [14, 15] of the term of the dataset. Mutual information and maximum entropy classification [16] are the basic techniques which are used by the researcher for machine learning and information retrieval from the text document.

## 3 Material and Methodology

### *3.1 Data Source*

Four datasets have been taken from the Knowledge Extraction based on Evolutionary Learning (KEEL) repository text classification datasets. It contains preprocessed data of text document of Ohsumed test collection which is a subset of the MEDLINE database. The MEDLINE database is a collection of the bibliographic database of important, peer-reviewed medical literature maintained by the National Library of Medicine. Ohsumed test classification [17] is the collection of each dataset which contains 100 attributes which are enough to test various feature selection algorithms. The brief description of the used dataset is given in Table 1.

### *3.2 Methodology*

Before doing any classification, we need to do preprocessing of dataset. Since dataset is very large and has an enormous number of the attribute, we have to reduce the number of the attribute in the dataset using preprocessing step known as feature selection. There are various feature selection algorithms available, but we will use only those feature selection algorithms which are available in the feature selection toolbox developed at UTIA of the Czech Academy of Sciences. The methodology works as shown in Fig. 1.

**Table 1** A brief description of used data sets

| Dataset name | Number of instances | Class labels | Number of features |
|---|---|---|---|
| 6 Ketoprostaglandin F1 $\alpha$ | 1003 | negative 6-ketoprostaglandin F1 α,<br>positive 6-ketoprostaglandin F1 α, | 100 |
| Brain chemistry data | 1003 | negative brain chemistry<br>positive brain chemistry | 100 |
| Heart valve data | 1003 | negative heart valve<br>positive heart valve | 100 |
| Uric acid data | 1003 | negative uric acid<br>positive uric acid | 100 |

**Feature Selection Algorithms** *CMIM*. The conditional mutual information maximization (CMIM) [18] algorithm selects a subset of a feature from dataset to minimize the number of features. The selected features carry more relevant information of data according to mutual information and save computational time.

*mRMR*. The minimum redundancy maximum relevance (mRMR) [18] select features are having less redundant data to minimize feature redundancy and high correlation to maximize feature relevance. The two usually used objective functions in mRMR are mutual information difference criterion (MID) and mutual information quotient criterion (MIQ).

*JMI*. The joint mutual information (JMI) [19] uses information theory to calculate the mutual information and entropy between any random variables together for feature selection. The representation is shown in Eq. (1).

$$I(x,\ y) = H(x) - H\ (x|y) \tag{1}$$

where $I$ is mutual information and $H$ is entropy.

*Condred*. In the condition redundancy [20] feature selection method, the race condition is overcome. The race condition is occurred due to the redundancy of term which is not related to the classification of a text document to predefined classes and statistical property.

*MIFS*. Mutual information feature selection (MIFS) [21] algorithm is entirely based on the mutual information computed for each term of the dataset. Mutual information of dataset is more reliable data as compared to the frequency of data for the classification of a text document. MIFS gives a more precise result, but it is a little bit slower since it calculates the weight of each data and then weight frequency.

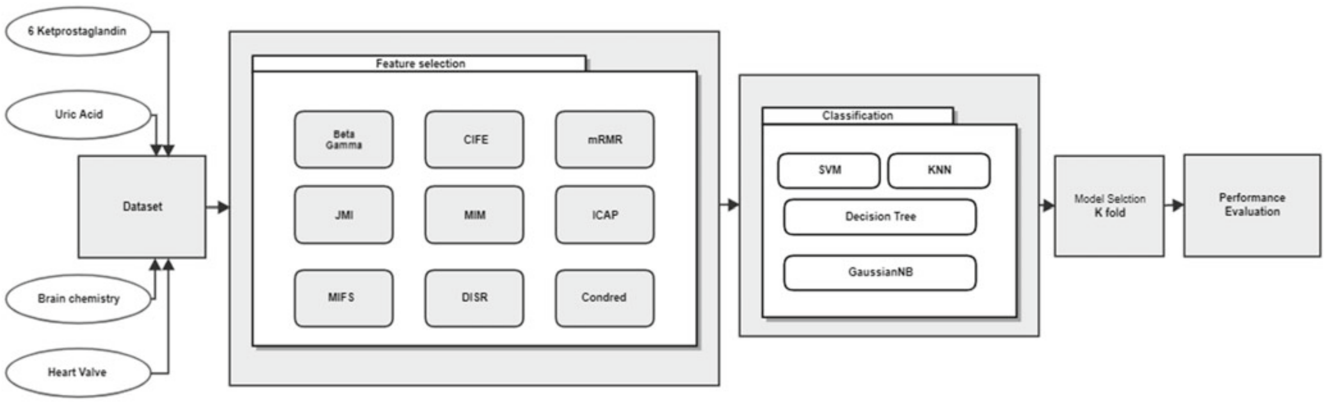

**Fig. 1** Methodology of text classification via feature selection

*ICAP*. In Interaction capping (ICAP) [22] feature selection algorithm features are sorted using the interaction of their term with other term using the information capping.

*DISR*. The double-input symmetrical relevance (DISR) [23] feature selection algorithm combines two main properties of variable complementarily, and the collection of the feature gives a different result. The most promising set is $d - 1$ if there is no information about the relation of the variable in datasets.

*CIFE*. The conditional infomax feature extraction (CIFE) [24] algorithm is based on information theory. In this feature selection, the systematical study of the structure of the document is done. It improves the performance of joint-class relevant detail by reducing class redundancy of dataset [8].

*BetaGamma*. The BetaGamma [25] is conditional mutual information-based feature selection algorithm. In this algorithm, beta and gamma are two values that maintain the weight of a feature by their relevance. Normally, the value of $\beta$ (beta) and $\gamma$ (gamma) is zero.

**Classification Algorithm** *Support Vector Machine*. SVM is a supervised learning technique and segregates classes using hyperplane for the classification using feature values of $N$ instances.

*Decision Tree*. A decision tree is a predictive model used for the classification based on the tree model. In this supervised algorithm, the datasets are broken down into a subset and create an association with that subset in a form decision tree having the node as decision node, intermediate node, and leave node.

*K-nearest neighbors*. In this method, the prediction function depends on the approximated locality, and Euclidean distance [25], Chebyshev norm, or Mahalanobis distance is used for distance computation.

*Gaussian Naïve Bayes*. This algorithm is based on the probabilistic classifier and relies on the well-known theorem—Bayes' theorem. It is also very popular for text categorization method. In this algorithm, the following formula in Eq. (2) are implied

$$P(c/X) = \sum_{n=1}^{\infty} P(xi/c) \tag{2}$$

where $P(c/X)$ is the posterior probability.

## 4 Experimental Results and Discussions

This section summarizes the simulation result performed on four text categorical data. We consider four classifiers and nine feature selection technique for the sake of performance evaluation. Results are annotated for each classifier and feature selection technique pair. The accuracy values are recorded and listed in Tables 2, 3, 4, and 5. The feature selection techniques referred in this study are from the filter-based approach, which requires the number of the feature as an input parameter. Due to

**Table 2** Impact of feature selection algorithm on 6-ketoprostaglandin F1 $\alpha$

| Classifiers | Feature selection technique | Number of features | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| SVM | BetaGamma | 0.996 | 0.980 | 0.988 | 0.972 | 0.980 |
| | CIFE | 0.976 | 0.980 | 0.988 | 0.984 | 0.984 |
| | CMIM | 0.992 | 1 | 0.988 | 0.992 | 0.992 |
| | Condred | 0.988 | 1 | 0.984 | 0.992 | 0.988 |
| | DISR | 1 | 0.984 | 0.984 | 0.996 | 0.996 |
| | ICAP | 0.988 | 0.972 | 0.984 | 0.980 | 0.988 |
| | JMI | 0.976 | 0.992 | 0.996 | 0.992 | 0.988 |
| | MIFS | 0.984 | 0.992 | 0.984 | 0.996 | 0.992 |
| | MRMR | 0.984 | 1 | 0.996 | 0.996 | 0.996 |
| KNN | BetaGamma | 0.964 | 0.976 | 0.972 | 0.988 | 0.964 |
| | CIFE | 0.988 | 0.976 | 0.968 | 0.972 | 0.980 |
| | CMIM | 0.996 | 0.984 | 0.972 | 0.964 | 0.964 |
| | Condred | 0.976 | 0.980 | 0.988 | 0.980 | 0.980 |
| | DISR | 0.980 | 0.992 | 0.972 | 0.988 | 0.980 |
| | ICAP | 0.980 | 0.976 | 0.968 | 0.996 | 0.976 |
| | JMI | 0.996 | 0.980 | 0.984 | 0.972 | 0.972 |
| | MIFS | 0.996 | 0.976 | 0.988 | 0.976 | 0.960 |
| | MRMR | 0.984 | 0.988 | 0.996 | 0.964 | 0.976 |
| DT | BetaGamma | 0.976 | 0.980 | 0.972 | 0.956 | 0.972 |
| | CIFE | 0.984 | 0.992 | 0.972 | 0.972 | 0.940 |
| | CMIM | 0.988 | 0.976 | 0.984 | 0.984 | 0.984 |
| | Condred | 0.984 | 0.984 | 0.988 | 0.972 | 0.980 |
| | DISR | 0.980 | 0.984 | 0.984 | 0.984 | 0.988 |
| | ICAP | 0.984 | 0.992 | 0.968 | 0.980 | 0.972 |
| | JMI | 0.988 | 0.984 | 0.976 | 0.980 | 0.968 |
| | MIFS | 0.992 | 0.984 | 0.988 | 0.972 | 0.968 |
| | MRMR | 0.976 | 0.972 | 0.984 | 0.972 | 0.964 |
| GaussianNB | BetaGamma | 0.984 | 0.984 | 0.964 | 0.972 | 0.960 |
| | CIFE | 1 | 0.992 | 0.984 | 0.960 | 0.972 |
| | CMIM | 0.996 | 0.980 | 0.992 | 0.976 | 0.984 |
| | Condred | 0.964 | 0.908 | 0.940 | 0.928 | 0.920 |
| | DISR | 0.988 | 1 | 0.984 | 0.984 | 0.980 |
| | ICAP | 0.976 | 0.996 | 0.984 | 0.984 | 0.992 |
| | JMI | 0.984 | 0.980 | 0.952 | 0.964 | 0.952 |
| | MIFS | 0.992 | 0.988 | 0.984 | 0.964 | 0.964 |
| | MRMR | 0.992 | 0.984 | 0.972 | 0.988 | 0.976 |

**Table 3** Impact of feature selection algorithm on uric acid data

| Classifiers | Feature selection technique | Number of features | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| SVM | BetaGamma | 0.996 | 0.992 | 0.984 | 0.984 | 0.984 |
| | CIFE | 0.996 | 0.980 | 0.992 | 0.992 | 0.988 |
| | CMIM | 0.980 | 0.996 | 0.996 | 0.988 | 1 |
| | Condred | 0.972 | 0.968 | 0.984 | 0.980 | 0.988 |
| | DISR | 0.992 | 0.992 | 0.984 | 0.992 | 0.988 |
| | ICAP | 0.984 | 0.988 | 0.988 | 0.992 | 0.968 |
| | JMI | 0.992 | 0.988 | 0.972 | 0.984 | 0.988 |
| | MIFS | 0.980 | 0.988 | 0.980 | 0.980 | 0.984 |
| | MRMR | 0.984 | 0.992 | 0.996 | 0.992 | 0.988 |
| KNN | BetaGamma | 0.976 | 0.984 | 0.964 | 0.964 | 0.964 |
| | CIFE | 0.992 | 0.968 | 0.952 | 0.960 | 0.956 |
| | CMIM | 0.984 | 0.944 | 0.948 | 0.948 | 0.956 |
| | Condred | 0.992 | 0.960 | 0.960 | 0.960 | 0.984 |
| | DISR | 0.964 | 0.968 | 0.936 | 0.948 | 0.960 |
| | ICAP | 0.980 | 0.980 | 0.972 | 0.976 | 0.940 |
| | JMI | 0.984 | 0.976 | 0.984 | 0.976 | 0.952 |
| | MIFS | 0.988 | 0.960 | 0.964 | 0.964 | 0.964 |
| | MRMR | 0.976 | 0.972 | 0.964 | 0.960 | 0.964 |
| DT | BetaGamma | 0.960 | 0.944 | 0.976 | 0.944 | 0.968 |
| | CIFE | 0.984 | 0.960 | 0.968 | 0.960 | 0.932 |
| | CMIM | 0.996 | 0.976 | 0.976 | 0.964 | 0.960 |
| | Condred | 0.980 | 0.980 | 0.964 | 0.952 | 0.956 |
| | DISR | 0.968 | 0.972 | 0.968 | 0.988 | 0.964 |
| | ICAP | 0.988 | 0.964 | 0.972 | 0.968 | 0.932 |
| | JMI | 0.980 | 0.976 | 0.964 | 0.968 | 0.956 |
| | MIFS | 0.980 | 0.960 | 0.980 | 0.960 | 0.956 |
| | MRMR | 0.988 | 0.980 | 0.968 | 0.968 | 0.972 |
| GaussianNB | BetaGamma | 0.992 | 0.952 | 0.976 | 0.936 | 0.912 |
| | CIFE | **1** | 0.972 | 0.952 | 0.944 | 0.928 |
| | CMIM | 0.992 | 0.988 | 0.992 | 0.972 | 0.980 |
| | Condred | 0.972 | 0.972 | 0.972 | 0.960 | 0.972 |
| | DISR | 0.984 | 0.988 | 0.980 | 0.980 | 0.960 |
| | ICAP | 0.996 | 0.980 | 0.964 | 0.956 | 0.928 |
| | JMI | 0.980 | 0.972 | 0.980 | 0.988 | 0.960 |
| | MIFS | 0.972 | 0.992 | 0.984 | 0.988 | 0.984 |
| | MRMR | 0.980 | 0.976 | 0.996 | 0.980 | 0.984 |

**Table 4** Impact of feature selection algorithm on heart valve data

| Classifiers | Feature selection technique | Number of features | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| SVM | BetaGamma | 0.984 | 0.964 | 0.988 | 0.976 | 0.992 |
| | CIFE | 0.972 | 0.988 | 0.976 | 0.984 | 0.976 |
| | CMIM | 0.976 | 0.976 | 0.980 | 0.968 | 0.976 |
| | Condred | 0.960 | 0.944 | 0.980 | 0.988 | 0.980 |
| | DISR | 0.968 | 0.972 | 0.984 | 0.968 | 0.980 |
| | ICAP | 0.976 | 0.968 | 0.976 | 0.992 | 0.961 |
| | JMI | 0.956 | 0.976 | 0.980 | 0.980 | 0.988 |
| | MIFS | 0.980 | 0.992 | 0.980 | 0.968 | 0.968 |
| | MRMR | 0.980 | 0.996 | 0.984 | 0.976 | 0.988 |
| KNN | BetaGamma | 0.964 | 0.992 | 0.940 | 0.960 | 0.944 |
| | CIFE | 0.968 | 0.968 | 0.956 | 0.952 | 0.936 |
| | CMIM | 0.972 | 0.976 | 0.952 | 0.952 | 0.940 |
| | Condred | 0.952 | 0.976 | 0.960 | 0.968 | 0.960 |
| | DISR | 0.968 | 0.972 | 0.968 | 0.928 | 0.968 |
| | ICAP | 0.972 | 0.968 | 0.984 | 0.960 | 0.960 |
| | JMI | 0.980 | 0.976 | 0.972 | 0.948 | 0.964 |
| | MIFS | 0.980 | 0.960 | 0.972 | 0.956 | 0.976 |
| | MRMR | 0.980 | 0.968 | 0.960 | 0.976 | 0.948 |
| DT | BetaGamma | 0.968 | 0.992 | 0.952 | 0.964 | 0.952 |
| | CIFE | 0.968 | 0.960 | 0.980 | 0.968 | 0.960 |
| | CMIM | 0.980 | 0.964 | 0.980 | 0.964 | 0.940 |
| | Condred | 0.940 | 0.952 | 0.952 | 0.952 | 0.972 |
| | DISR | 0.956 | 0.992 | 0.968 | 0.944 | 0.940 |
| | ICAP | 0.984 | 0.980 | 0.960 | 0.976 | 0.936 |
| | JMI | 0.960 | 0.952 | 0.960 | 0.940 | 0.928 |
| | MIFS | **0.992** | 0.964 | 0.956 | 0.964 | 0.952 |
| | MRMR | 0.968 | 0.968 | 0.968 | 0.968 | 0.964 |
| GaussianNB | BetaGamma | 0.956 | 0.940 | 0.936 | 0.948 | 0.948 |
| | CIFE | 0.972 | 0.944 | 0.956 | 0.936 | 0.920 |
| | CMIM | 0.940 | 0.976 | 0.976 | 0.976 | 0.956 |
| | Condred | 0.948 | 0.928 | 0.920 | 0.940 | 0.944 |
| | DISR | 0.948 | 0.972 | 0.968 | 0.976 | 0.972 |
| | ICAP | 0.976 | 0.968 | 0.968 | 0.960 | 0.944 |
| | JMI | 0.972 | 0.972 | 0.972 | 0.964 | 0.960 |
| | MIFS | 0.972 | 0.972 | 0.956 | 0.956 | 0.964 |
| | MRMR | 0.960 | 0.968 | 0.956 | 0.964 | 0.956 |

**Table 5** Impact of feature selection algorithm on brain chemistry data

| Classifiers | Feature selection technique | Number of features | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| SVM | BetaGamma | 0.992 | 0.992 | 0.988 | 0.980 | 0.988 |
| | CIFE | 0.992 | 0.988 | 0.988 | 0.984 | 0.992 |
| | CMIM | 0.992 | 0.992 | 0.996 | 0.980 | 0.988 |
| | Condred | 0.984 | 0.980 | 0.980 | 0.968 | 0.968 |
| | DISR | 0.992 | 0.984 | 0.976 | 0.964 | 0.992 |
| | ICAP | 0.984 | 0.976 | 0.952 | 0.976 | 0.948 |
| | JMI | 0.976 | 0.984 | 0.960 | 0.976 | 0.992 |
| | MIFS | 0.996 | 0.992 | 0.988 | 0.958 | 0.980 |
| | MRMR | 0.980 | 0.980 | 0.992 | 0.944 | 1 |
| KNN | BetaGamma | 0.976 | 0.952 | 0.972 | 0.964 | 0.964 |
| | CIFE | 0.964 | 0.968 | 0.944 | 0.932 | 0.952 |
| | CMIM | 0.960 | 0.956 | 0.964 | 0.948 | 0.956 |
| | Condred | 0.988 | 0.980 | 0.968 | 0.976 | 0.968 |
| | DISR | 0.972 | 0.980 | 0.984 | 0.984 | 0.964 |
| | ICAP | 0.972 | 0.964 | 0.968 | 0.976 | 0.948 |
| | JMI | 0.992 | 0.972 | 0.980 | 0.956 | 0.944 |
| | MIFS | 0.980 | 0.980 | 0.980 | 1 | 0.936 |
| | MRMR | 0.988 | 0.976 | 0.992 | 0958 | 0.936 |
| DT | BetaGamma | 0.984 | 0.968 | 0.956 | 0.956 | 0.924 |
| | CIFE | 0.988 | 0.980 | 0.952 | 0.952 | 0.944 |
| | CMIM | 0.992 | 0.992 | 0.952 | 0.956 | 0.964 |
| | Condred | 0.992 | 0.968 | 0.964 | 0.956 | 0.988 |
| | DISR | 0.984 | 0.956 | 0.976 | 0.984 | 0.980 |
| | ICAP | 0.992 | 0.992 | 0.992 | 0.924 | 0.924 |
| | JMI | 0.980 | 0.948 | 0.972 | 0.980 | 0.956 |
| | MIFS | 0.976 | 0.980 | 0.972 | 0.988 | 0.980 |
| | MRMR | 0.980 | 0.960 | 0.980 | 0.980 | 0.972 |
| GaussianNB | BetaGamma | 0.996 | 0.988 | 0.972 | 0.932 | 0.908 |
| | CIFE | 0.992 | 0.980 | 0.964 | 0.932 | 0.924 |
| | CMIM | **1** | 0.992 | 0.992 | 0.992 | 0.992 |
| | Condred | 0.964 | 0.948 | 0.928 | 0.968 | 0.968 |
| | DISR | 0.984 | 0.968 | 0.980 | 0.956 | 0.988 |
| | ICAP | 0.968 | 0.984 | 0.992 | 0.992 | 0.984 |
| | JMI | 0.992 | 0.984 | 0.992 | 0.980 | 0.984 |
| | MIFS | 0.972 | 0.996 | 0.980 | 0.980 | 0.972 |
| | MRMR | 0.984 | 0.936 | 0.988 | 0.976 | 0.968 |

uncertainty in the selection of optimal features, we performed multiple experiments by initializing the ten features with an interval of 10. In total, we capture the five instances in multiple of ten features.

Table 2 lists the experimental result performed on ketoprostaglandin dataset. The accuracy values in the table indicate the highest value when support vector machine is aligned with DISR on ten features. When selected features are 20, then the three feature selection techniques, CMIM, conditional reduced (Condred), and MRMR, produce 100% accuracy.

The performance on uric acid data is noted in Table 3. The table reveals that the CIFE feature selection techniques with Naïve Bayesian classifiers achieved 100% accuracy with ten features only.

Decision tree performance in heart valve data is delivering the maximum accuracy. MIFS feature selection technique with ten features achieves 99.2% of accuracy. Experimental results are listed in Table 4.

In brain chemistry data, the combination of Naïve Bayes algorithm along with CMIM feature selection technique has the highest classification rate. This pair of classifier feature selection achieves 100% accuracy when the number of the feature is selected as 10. Table 5 reports the experimental outcome.

The objectives of these experiments were to extract out the best feature selection and classifier combination so that the choice of making an efficient model could be effortless. However, there is no single pair that can be identified, but still, the use of Naïve Bayes classifier along with CMIM feature selection technique could be an optimal choice for text categorization model.

## 5 Conclusion

In this paper, nine weighted feature selection algorithms are used in the four-text classification preprocessed dataset from KEEL repository. The feature selection is performed with the different number of features ranging from 10 to 50 at the interval of 10 features. This experiment shows improvement in the performance of classification for text documentation categorization on using weighted feature selection. The experimental results concluded that mutual information based feature selection algorithm improves the result of text classification significantly. The weighted feature selection methods are also work well because of the relevance of position of the term used in the text document, and it also reduces factor of redundancy while classification of documents.

## References

1. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning ECML ’98, pp. 137–142 (1998)
2. Markowetz, F.: Classification by support vector machines. In: Discrete Methods in Epidemiology, pp. 1–9 (2000)
3. Leung, K.M.: Naive Bayesian Classifier (2007)
4. Friedl, M.A., Brodley, C.E.: Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. **61**, 399–409 (1997)
5. Cai, Y., Ji, D., Cai, D.: A KNN research paper classification method based on shared nearest neighbor. In: Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp. 336–340 (2010)
6. Ladha, L., Deepa, T.: Feature selection methods and algorithms. Int. J. Comput. Sci. Eng. **3**, 1787–1797 (2011)
7. Brown, G., Pocock, A., Zhao, M.-J., Lujan, M.: Conditional likelihood maximisation: a unifying framework for mutual information feature selection. J. Mach. Learn. Res. **13**, 27–66 (2012)
8. Albrechtsen, H.: Subject analysis and indexing. From automated indexing to domain analysis. Indexer **18**, 219–224 (1993)
9. Amati, G., van Rijsbergen, C.J.: Term frequency normalization via Pareto distributions. Adv. Inf. Retr. **2291**, 183–192 (2002)
10. Gini coefficient
11. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E-Stat. Nonlinear Soft Matter Phys. **69** (2004)
12. Boulis, C., Ostendorf, M.: Text classification by augmenting the bag-of-words representation with redundancy compensated bigrams. In: Workshop on Feature Selection in Data Mining, pp. 9–16 (2005)
13. Agre, G., Dzhondzhorov, A.: A weighted feature selection method for instance-based classification. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 14–25 (2016)
14. Pluim, J.P.W., Maintz, J.B.A.A., Viergever, M.A.: Mutual-Information-Based Registration of Medical Images: A survey (2003)
15. Li, W.: Mutual information functions versus correlation functions. J. Stat. Phys. **60**, 823–837 (1990)
16. Nigam, K., Lafferty, J., Mccallum, A.: Using maximum entropy for text classification. In: IJCAI-99 Workshop on Machine Learning for Information Filtering, vol. 1, pp. 61–67 (1999)
17. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Mult. Valued Log. Soft Comput. **17**, 255–287 (2011)
18. Fleuret, F.: Fast binary feature selection with conditional mutual information. J. Mach. Learn. Res. **5**, 1531–1555 (2004)
19. Bennasar, M., Hicks, Y., Setchi, R.: Feature selection using joint mutual information maximisation. Expert Syst. Appl. **42**, 8520–8532 (2015)
20. Long, W.C., Swiney, K.M., Harris, C., Page, H.N., Foy, R.J.: Effects of ocean acidification on juvenile red king crab (*Paralithodes camtschaticus*) and tanner crab (*Chionoecetes bairdi*) growth, condition, calcification, and survival. PLoS ONE **8** (2013)
21. Rades, M., Ewins, D.: Mifs and macs in modal analysis. In: Modal Analysis Conference (IMAC-20), pp. 771–778 (2002)
22. Jakulin, A.: Machine learning based on attribute interactions. PhD thesis, pp. 1–252 (2005)
23. Bar-Nun, A., Dimitrov, V., Tomasko, M.: Titan’s aerosols: comparison between our model and DISR findings. Planet. Space Sci. **56**, 708–714 (2008)

24. Fischer, M., Stone, M., Liston, K., Kunz, J., Singhal, V.: Multi-stakeholder collaboration : the CIFE iRoom. In: International Council for Research and Innovation in Building and Construction. CIB W78 Conference, pp. 12–14 (2002)
25. Lewis, D.: Feature selection and feature extract ion for text categorization. In: Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, 23–26 Feb 1992