Turk J Elec Eng & Comp Sci, Vol.20, No.Sup.2, 2012, © TÜBİTAK
doi:10.3906/elk-1101-1064

Hybrid feature selection for text classification

Serkan GÜNAL
Department of Computer Engineering, Anadolu University, Eskişehir-TURKEY
e-mail: serkangunal@anadolu.edu.tr

Received: 30.01.2011

Abstract

Feature selection is vital in the field of pattern classification due to accuracy and processing time considerations. The selection of proper features is of greater importance when the initial feature set is considerably large. Text classification is a typical example of this situation, where the size of the initial feature set may reach to hundreds or even thousands. There are numerous research studies in the literature offering different feature selection strategies for text classification, mostly focused on filters. In spite of the extensive number of these studies, there is no significant work investigating the efficacy of a combination of features, which are selected by different selection methods, under different conditions. In this study, a hybrid feature selection strategy, which consists of both filter and wrapper feature selection steps, is proposed to comprehensively analyze the redundancy or relevancy of the text features selected by different methods in the case of different feature set sizes, dataset characteristics, classifiers, and success measures. The results of the experimental study reveal that a combination of the features selected by various methods is more effective than the features selected by the single selection method. The profile of the combination is, however, influenced by characteristics of the dataset, choice of the classification algorithm, and the success measure.

Key Words: *Feature extraction, feature selection, pattern recognition, text classification*

1. Introduction

The aim of text classification, or categorization, is simply to classify texts of interest into appropriate classes or categories. A typical text classification system mainly consists of a feature extraction mechanism that computes numerical information from a raw text document, and a classifier that executes a classification process using prior knowledge of the labeled data.

Though there exist lexical, semantic, and syntactic approaches for document representation [1,2], the majority of text classification studies make use of a bag-of-words approach [3] to represent documents, where the exact ordering of words, or terms, in the documents is ignored but the number of occurrences of each term is considered. Each distinct term in a document collection therefore constitutes an individual feature. Terms are individually assigned certain weights that represent the importance of the terms in a given document [4]. Widely used weighting schemes are term frequency (TF), which corresponds to the number of occurrences of a term in a document, and term frequency-inverse document frequency (TF-IDF), which scales down the TF weight by considering the number of documents in the collection containing the regarding term [5]. Hence, a

document is represented by a multidimensional feature vector, where each dimension in the vector corresponds to the weighted value for a distinct word within the document collection. Representation of documents as features or identifier vectors in such a way is known as the vector space model [6].

The size of the feature vector may reach to considerable values, even for moderate numbers of documents. Thus, the processing time of text classification increases drastically. Moreover, classification accuracy might even be degraded due to the phenomenon known as the "curse of dimensionality" [7,8]. Therefore, feature dimension should be reduced in such a way that the features that are irrelevant or that have low discriminatory power are eliminated. Dimension reduction can be achieved using either feature transformation or feature selection. In feature transformation, the original feature space is projected into a lower dimensional subspace that carries more relevant or discriminative information. Popular feature transformation techniques in text classification studies are latent semantic analysis [9,10], principal component analysis [11,12], and linear discriminant analysis [13,14]. On the other hand, feature selection operates on the original feature space and aims to eliminate irrelevant dimensions, whereas relevant dimensions are kept with their original values. A comprehensive overview of various aspects of feature selection can be found in [15–17].

Though a vast amount of methodologies for feature selection are available in the literature, all of the feature selection methods broadly fall into 3 categories: filters, wrappers, and embedded methods. Filters assess feature relevancies using various scoring schemes independently from a learning model or classifier [17]. Filter techniques are easily scalable to high-dimensional datasets, computationally simple and fast. On the other hand, wrappers evaluate features using a specific learning model and search algorithm, where the search algorithm is wrapped around the model to examine the feature space [18]. Wrapper techniques consider feature dependencies and provide interaction between feature subset search and choice of a learning model, but are computationally expensive with respect to filters. In embedded methods, an optimal feature subset search is built into the classifier setup. In other words, feature selection is integrated in the classifier training process; therefore, embedded methods are specific to the utilized learning model, just like wrappers. However, they are less computationally intensive than wrappers [15,17]. While the individual employment of filters or wrappers is more common in the literature, there are also a number of studies utilizing filters and wrappers together within the feature selection scheme [19,20]. The evaluation of features in a feature selection scheme can be carried out with either a univariate or multivariate approach. A univariate approach examines the features independently, provides individual discriminatory powers of the features, and is fast, but it ignores possible correlation among the features [21]. On the contrary, a multivariate approach takes feature dependencies into consideration during evaluation of the features but is relatively slow [15]. Among all of the feature selection strategies, only the exhaustive search and branch-and-bound algorithms [22] may yield optimal results. Their computational complexity is, however, significantly high for even moderate numbers of features. This leads researchers to consider suboptimal selection methods most of the time.

In text classification studies, widely used feature selection methods are univariate filter approaches due to the mass amount of features that require significant processing time. Once the individual discriminatory powers of the features are obtained, the best N features are selected while the others are eliminated. Hence, a compact subset of features is attained, although feature dependencies are ignored. The most popular examples of these approaches in the literature are term strength [23], odds ratio [24], document frequency [25], mutual information [26], chi-square [27], and information gain [28]. A number of comparative studies on different feature selection metrics can be found in [25,29], as well.

In spite of the extensive number of feature selection studies on text classification, there is no significant

work investigating the efficacy of a combination of features, which are selected by a variety of selection methods, under different conditions. Therefore, a hybrid feature selection strategy, which consists of both filter and wrapper feature selection steps, is proposed in this study. This hybrid selection process is repeated under particular conditions, including different feature set sizes, dataset characteristics, classifiers, and success measures. The results of the study enable us to discover which features or feature combinations are better identifiers for text classification, whether there is a distinct relation among the useful features, the desired feature size, the utilized classification method, and the success measure.

The rest of the paper is organized as follows: feature selection methods utilized in the study are briefly described in Section 2; Section 3 introduces the hybrid feature selection scheme; Section 4 presents the experimental work and related results; and, finally, the conclusion of the paper is given in Section 5.

2. Feature selection

The mathematical background and detailed explanation of the filter and wrapper methods that are utilized within the proposed hybrid feature selection scheme are given in the following subsections.

2.1. Filter

The filter methods, or actually the scoring schemes, utilized in this study are document frequency, mutual information, chi-square, and information gain. The rationale of these methods is as follows.

2.1.1. Document frequency

Document frequency (DF) is one of the simplest approaches to assess feature relevance in text classification problems. The DF of a specific term simply corresponds to the number of documents in a class containing that term [5,25,29]. Hence, the DF of each term constitutes the relevancy score of the term.

2.1.2. Mutual information

The mutual information (MI) of 2 random variables indicates the mutual dependence of the variables. Therefore, the MI related to term t and class c describes the amount of information the presence of that term carries about the relevant class [25]. Therefore, MI can be formulated as:

$$MI(t,c) = \log \frac{P(t|c)}{P(t)}, \tag{1}$$

where $P(t)$ is the probability of term t and $P(t|\,c)$ is the probability of term t given class c.

2.1.3. Chi-square

Another popular selection approach is chi-square (CHI2). In statistics, the CHI2 test is applied to examine the independence of 2 events. The events, X and Y, are assumed to be independent if:

$$p(XY) = p(X)p(Y). \tag{2}$$

In text feature selection, these 2 events correspond to the occurrence of a particular term and class, respectively. CHI2 information can be computed using:

$$CHI2(t,c) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} \frac{(N_{t,c} - E_{t,c})^2}{E_{t,c}}, \tag{3}$$

where N is the observed frequency and E is the expected frequency for each state of term t and class c [5]. CHI2 is a measure of how much the expected counts E and observed counts N deviate from each other. A high value of CHI2 indicates that the hypothesis of independence is not correct. If the 2 events are dependent, then the occurrence of the term makes the occurrence of the class more likely. Consequently, the term in question is relevant as a feature.

2.1.4. Information gain

Information gain (IG) measures how much information the presence or absence of a term contributes to making the correct classification decision for any class [25,29]. IG reaches its maximum value if a term is an ideal indicator for class association, that is, if the term is present in a document, if and only if the document belongs to the respective class. IG for term t with respect to class c can be obtained using:

$$IG(t,c) = -\sum_{i=1}^{M} P(c)\log P(c) + P(t)\sum_{i=1}^{M} P(c|t)\log P(c|t) + P(\overline{t})\sum_{i=1}^{M} P(c|\overline{t})\log P(c|\overline{t}), \tag{4}$$

where M is the number of classes, $P(c)$ is the probability of class c, $P(t)$ and $P(\overline{t})$ are the probability of the presence and absence of term t, and $P(c|\,t)$ and $P(c|\,\overline{t})$ are the probability of class c given the presence and absence of term t, respectively.

2.2. Wrapper

The wrapper method employed in this study is a genetic algorithm (GA)-based selection (GS), which has been proven to be relatively capable and fast among many suboptimal search algorithms such as sequential forward and backward selections [30]. The GA is a probabilistic search method inspired by the biological evolution process [31]. The principle of the GA is the survival of the fittest solutions among a population of potential solutions for a given problem. Thus, new generations produced by the surviving solutions are expected to provide better approximations to the optimum solution. The solutions correspond to chromosomes that are encoded with an appropriate alphabet. The fitness value of each chromosome is determined by a fitness function. New generations are obtained using genetic operators, namely crossover and mutation, with certain probabilities on the fittest members of the population. The initial population can be randomly or manually defined. Population size, number of generations, probability of crossover, and mutation are defined empirically.

As a simple but useful GS approach, chromosome length is equal to the dimension of a full feature set. The chromosomes are then encoded with a {0, 1} binary alphabet. In a chromosome, the indices represented with "1" indicate the selected features, whereas "0" indicates the unselected ones. For example, a chromosome defined as

$$\{1\ \ 0\ \ 1\ \ 0\ \ 1\ \ 1\ \ 0\ \ 0\ \ 0\ \ 1\} \tag{5}$$

specifies that the 1st, 3rd, 5th, 6th, and 10th features are selected while the others are eliminated. The fitness value corresponding to a chromosome is determined by a particular success measure that is obtained with the selected features. Some examples of genetic feature selection studies are given in the literature in [30,32–34].

3. Hybrid feature selection scheme

Hybrid feature selection strategy, which is proposed in this study, consists of both filter and wrapper selection stages. In the first stage, the features are selected using DF-, MI-, CHI2-, and IG-based filter methods. Next, the features selected by the filters are combined together and fed into the GS in the second stage. This 2-stage hybrid feature selection process is repeated under particular conditions, including different feature set sizes, dataset characteristics, classifiers, and success measures. Hence, the redundancy or relevancy of a combination of text features selected by different methods can be extensively analyzed for each condition.

3.1. Feature set size

Feature set size is the first varying parameter, such that the experiments are carried out with various numbers of features ranging from 1 feature per class to 10 features per class.

3.2. Dataset characteristics

The second altering item is the characteristic of the datasets utilized. Two celebrated text datasets, namely Reuters-21578 and 20 Newsgroups [35], are employed in this experimental study. The top 10 classes of the Reuters-21578 ModApte split and 10 classes of the 20 Newsgroups dataset are considered during the experiments. All of the information regarding these 2 datasets, including class descriptions and the number of training and testing samples, are listed in Tables 1 and 2. Reuters is a skewed dataset; that is, the numbers of documents in each class are quite different. On the contrary, the Newsgroups dataset has a uniform distribution with an equal number of documents per class. Hence, the effectiveness of the features can be observed in 2 separate datasets with different characteristics.

Table 1. Reuters dataset.

No.	Class label	Number of training samples	Number of testing samples
1	earn	2877	1087
2	acq	1650	719
3	money-fx	538	179
4	grain	433	149
5	crude	389	189
6	trade	369	117
7	interest	347	131
8	ship	197	89
9	wheat	212	71
10	corn	181	56

Table 2. Newsgroups dataset.

No.	Class label	Number of training samples	Number of testing samples
1	alt.atheism	500	500
2	comp.graphics	500	500
3	comp.os.ms-windows.misc	500	500
4	comp.sys.ibm.pc.hardware	500	500
5	comp.sys.mac.hardware	500	500
6	comp.windows.x	500	500
7	misc.forsale	500	500
8	rec.autos	500	500
9	rec.motorcycles	500	500
10	rec.sport.baseball	500	500

3.3. Classifiers

Another altering parameter is the classification algorithm. The first classifier utilized in the study is the decision tree (DT), which is a nonlinear classifier [36]. The other classifier is a linear form of support vector machine (SVM) [37]. Thus, 2 different classification algorithms are tested within the wrapper stage of the hybrid feature selection to assess feature subsets, which enables us to observe variations of the selected features in each case. Both classifiers have previously been widely employed for text categorization research [38–40].

3.3.1. Decision tree

Decision, or classification, trees are multistage decision systems in which classes are consecutively rejected until an accepted class is reached [36]. For this purpose, the feature space is split into unique regions corresponding to the classes. The most commonly used type of DT is the binary classification tree, which splits the feature space into 2 parts sequentially by comparing the feature values with a specific threshold. Thus, an unknown feature vector is assigned to a class via a sequence of Yes/No decisions along a path of nodes of a DT. One has to consider the following items in the design of a classification tree:

i. Splitting criterion,

ii. Stop-splitting rule,

iii. Class assignment rule.

The fundamental aim of the splitting feature space is to generate subsets that are more class-homogeneous compared to former subsets. In other words, the splitting criterion at any node is to obtain the split providing the highest decrease in node impurity. Entropy is one of the most widely used types of information to define impurity and can be computed as:

$$I(t) = -\sum_{i=1}^{M} P(c_i|t) \log_2 P(c_i|t), \tag{6}$$

where $\mathrm{P}(\mathrm{c}_i|\ \mathrm{t})$ denotes the probability that a vector in subset X_t, associated with node t, belongs to class c_i, $\mathrm{i} = 1, 2, \ldots, \mathrm{M}$. Assume now that performing a split, N_{tY} points are sent into the "Yes" node (X_{tY}) and N_{tN}

into the "No" node (X_{tN}). The decrease in node impurity is then defined as:

$$\Delta I(t) = I(t) - \frac{N_{tY}}{N_t} I(t_{YES}) - \frac{N_{tN}}{N_t} I(t_{NO}), \tag{7}$$

where $I(t_Y)$, $I(t_N)$ are the impurities of the t_{YES} and t_{NO} nodes, respectively. If the highest decrease in node impurity is less than a certain threshold or a single class is obtained following a split, then the splitting process is stopped. Once a node is declared to be terminal or leaf, then a class assignment is made. A commonly used assignment method is the majority rule, which assigns a leaf to a class to which the majority of the vectors in the corresponding subset belong.

3.3.2. SVM

The SVM is one of the most effective classification algorithms in the literature. The SVM algorithm has both linear and nonlinear versions. In this study, the linear version of the SVM is employed. The crucial point of the SVM classifier is the notion of the margin [36,37]. Classifiers utilize hyperplanes to separate classes. Every hyperplane is characterized by its direction (w) and its exact position in space (w_0). Thus, a linear classifier can be simply defined as:

$$w^T x + w_0 = 0. \tag{8}$$

Figure 1 illustrates a 2-class classification problem. Obviously, the position of the hyperplane should be defined in such a way that maximizes the separation between classes.

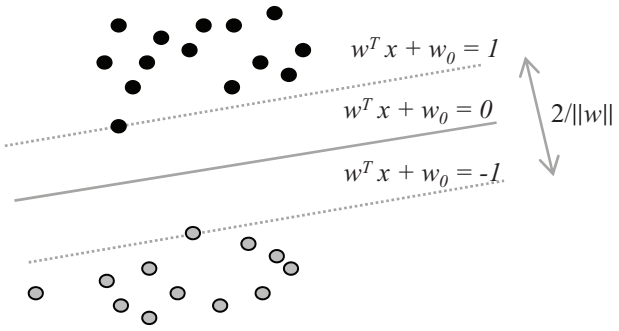


Figure 1. A linear classifier and the associated margin lines for a 2-class classification problem.

Here, the region between 2 hyperplanes,

$$w^T x + w_0 = 1, \quad w^T x + w_0 = -1, \tag{9}$$

is called the margin. The width of the margin is equal to $2/\|w\|$. Achieving the maximum possible margin is the underlying idea of the SVM algorithm. Maximization of the margin requires the minimization of:

$$J(w, w_0, \varepsilon) = \frac{1}{2}\|w\|^2 + K\sum_{i=1}^{N} \varepsilon_i, \tag{10}$$

which is subject to:

$$\begin{aligned} &w^T x_i + w_0 \geq 1 - \varepsilon_i, \quad \text{if} \quad x_i \in c_1 \\ &w^T x_i + w_0 \leq -1 + \varepsilon_i, \quad \text{if } x_i \in c_2 \\ &\varepsilon_i \geq 0. \end{aligned} \tag{11}$$

In Eq. (10), K is a user-defined constant and ε is the margin error. A margin error occurs if data belonging to one class are on the wrong side of the hyperplane. Minimizing the cost is therefore a trade-off issue between a large margin and a small number of margin errors. The solution of this optimization problem is obtained as:

$$w = \sum_{i=1}^{N} \lambda_i y_i x_i, \tag{12}$$

which is the weighted average of the training features. Here, λ_i is a Lagrange multiplier of the optimization task and y_i is a class label. The values of the λs are nonzero for all of the points lying inside the margin and on the correct side of the classifier. These points are known as support vectors, and the resulting classifier is known as the SVM.

In the case of multiclass classification problems, 1 of 2 common approaches, namely one-against-all and one-against-one, can be chosen to adopt 2-class classification to a multiclass case [41].

3.4. Success measures

The final varying parameter during the experiments is the success measure that is used within the classification process. The success measures employed in this study are well-known F1 measures, namely Macro-F1 and Micro-F1.

3.4.1. Macro-F1

In macroaveraging, the F-measure is computed for each class within the dataset and then the average over all of the classes is obtained. Hence, equal weight is assigned to each class regardless of the class frequency [5]. Computation of Macro-F1 can be formulated as:

$$MacroF1 = \frac{\sum_{k=1}^{C} F_k}{C}, \quad F_k = \frac{2 \times p_k \times r_k}{p_k + r_k}, \tag{13}$$

where pair (p_k, r_k) corresponds to precision and recall values of class k, respectively. These values are calculated by:

$$p_k = \frac{TP_k}{TP_k + FP_k} \tag{14}$$

and

$$r_k = \frac{TP_k}{TP_k + FN_k}, \tag{15}$$

where TP_k (named as the true positive) is the number of documents classified correctly to class k, FP_k (named as the false positive) is the number of documents that do not belong to class k but are classified to this class incorrectly, and FN_k (named as the false negative) corresponds to the number of documents that actually belong to class k but are not classified to the respective class.

3.4.2. Micro-F1

In microaveraging, the F-measure is computed globally without class discrimination. Hence, all classification decisions in the entire dataset are considered [5]. In the case that the classes in a collection are biased, large

classes would dominate small ones in microaveraging. Computation of Micro-F1 can be formulated as:

$$MicroF1 = \frac{2 \times p \times r}{p + r}, \tag{16}$$

where pair (p, r) corresponds to precision and recall values, respectively, over all of the classification decisions within the entire dataset. These values are calculated similar to Eqs. (14) and (15), but by considering the entire dataset, not individual classes.

4. Experimental study

This section describes the experimental study and the related results. First, the preprocessing steps are briefly explained, and then the outcomes of the analysis in each stage of the hybrid selection scheme are provided. The software used to carry out all of the experiments was developed by the author on a MATLAB platform.

4.1. Preprocessing

Before proceeding with the hybrid feature selection scheme, stop-word removal [6] and a stemming process [42] are carried out during the feature extraction from the text documents. Moreover, the TF-IDF approach is preferred for the term weighting process.

4.2. Stage 1: Filter

Using DF, MI, CHI2, and IG, relevancy scores of all of the terms in each class are obtained first. The terms are then sorted in descending order based on these scores. Finally, top N terms are selected from the sorted list of each class and combined together. However, the value of N may not be the same for each method because the same terms might have the same rank for more than one class. For instance, the most informative terms in the Newsgroups dataset determined by IG are listed in Table 3. One can easily note that the term "rec" is common to both class 2 and class 4. Thus, the selection of 1 term per class would yield 9 unique terms out of 10. In other words, the uniqueness rate of this selection is 90%. Similarly, the uniqueness rates for varying numbers of terms per class are provided in Figure 2 for both the Reuters and Newsgroups datasets. These rates can be used as an indicator of how different the features selected by the utilized selection method are for each class. An increase in the number of common terms among classes decreases the uniqueness rate. One can observe from Figure 2 that DF has the lowest uniqueness rate in both datasets. On the other hand, MI always provides unique terms for each class, whereas CHI2 and IG find themselves at a level between DF and MI.

Table 3. The most informative terms determined by IG in the Newsgroups dataset.

Class	1	2	3	4	5	6	7	8	9	10
Term	atheist	rec	apr	rec	appl	window	sale	car	dod	team

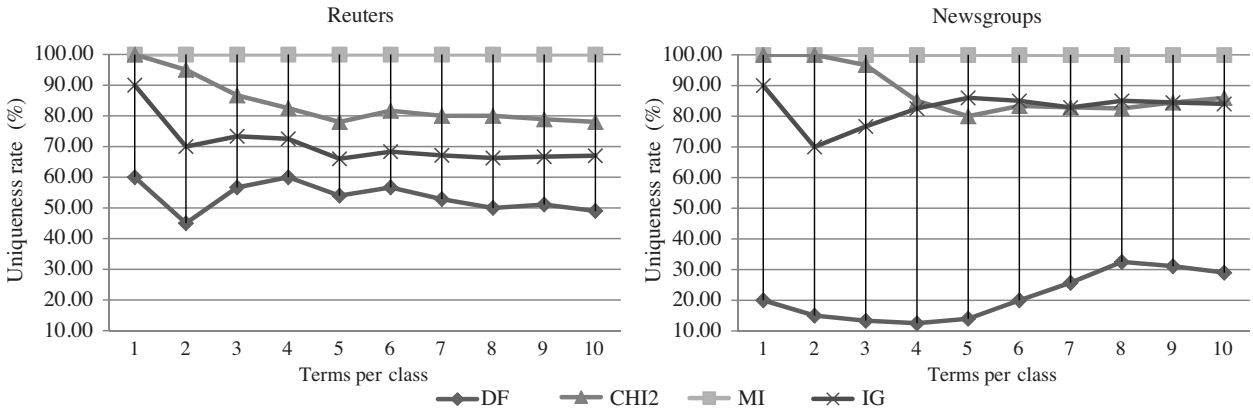


Figure 2. Uniqueness rates of the terms selected by the filter methods.

The features selected by each filter method are then combined together to be able to analyze the effectiveness of the feature combinations. The combining process is carried out by removing duplicates of common terms that are selected by distinct methods. The resulting feature set sizes for different numbers of terms per class are presented as a part of Figures 3 and 4. For instance, in the Reuters dataset, selecting 1 term per class for 4 different filter methods and combining those terms yields just 24 features instead of 40, due to several duplicates. The features selected by the abovementioned filter methods are now ready to go through the second stage of the hybrid feature selection.

4.3. Stage 2: Wrapper

The combined feature set, which is obtained during the first stage of the hybrid selection scheme, is fed into the GS in the second stage. Optimal GS parameters are empirically obtained as follows: the population size is 50, the number of generations is 30, the probability of crossover is 0.8, and the probability of mutation is 0.08. The fitness values are Micro-F1 and Macro-F1 measures obtained by 2 different classification algorithms. Figures 3 and 4 illustrate the outcome of the GS for each combination of the utilized success measures, classifiers, datasets, and number of terms per class. In each case, the size of the feature set and the value of the success measures before and after hybrid feature selection are provided.

Since the Reuters dataset is skewed, one can easily observe the significant difference between the values of Micro-F1 and Macro-F1 measures obtained by the classification algorithms. On the other hand, the values of the 2 success measures are almost the same in the Newsgroups dataset, which is uniformly distributed. In the Reuters dataset, the highest Micro-F1 score (85.83%) is achieved by the SVM, whereas the highest Macro-F1 score (66.19%) is attained using DT. However, in the Newsgroups dataset, both the highest Micro-F1 (98.48%) and Macro-F1 (98.44%) scores are obtained by the DT classifier.

Furthermore, it is obvious from Figures 3 and 4 that there is a considerable reduction in the size of the feature sets after the selection in all of the cases. While the maximum dimension reduction rate reaches as high as 56% in the Reuters dataset, this value goes up to approximately 54% in the Newsgroups dataset. In spite of this amount of reduction in the feature set size, both the Micro-F1 and Macro-F1 values after hybrid selection are even higher with respect to their corresponding values before the selection.

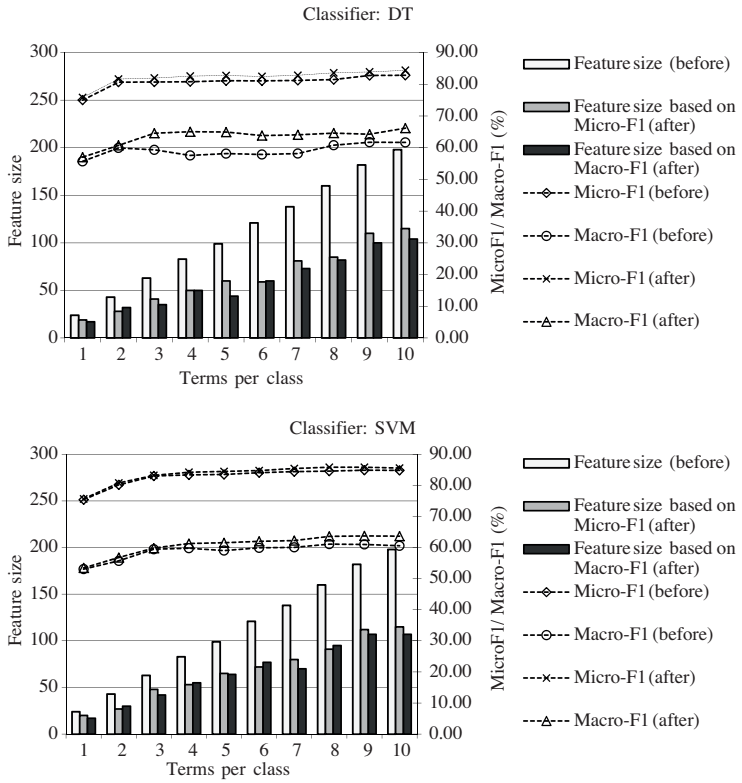


Figure 3. Success measures and feature set sizes before and after hybrid feature selection in the Reuters dataset.

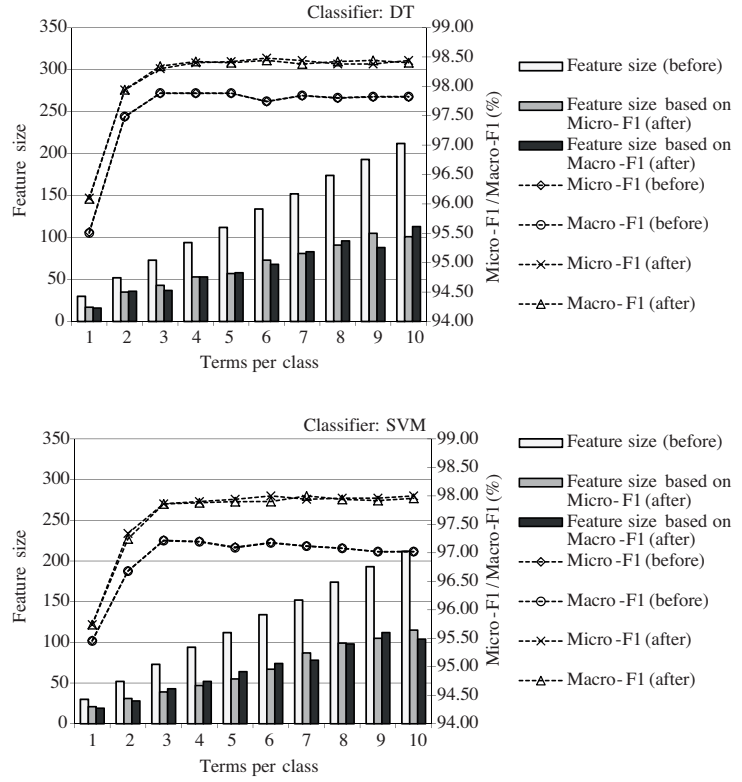
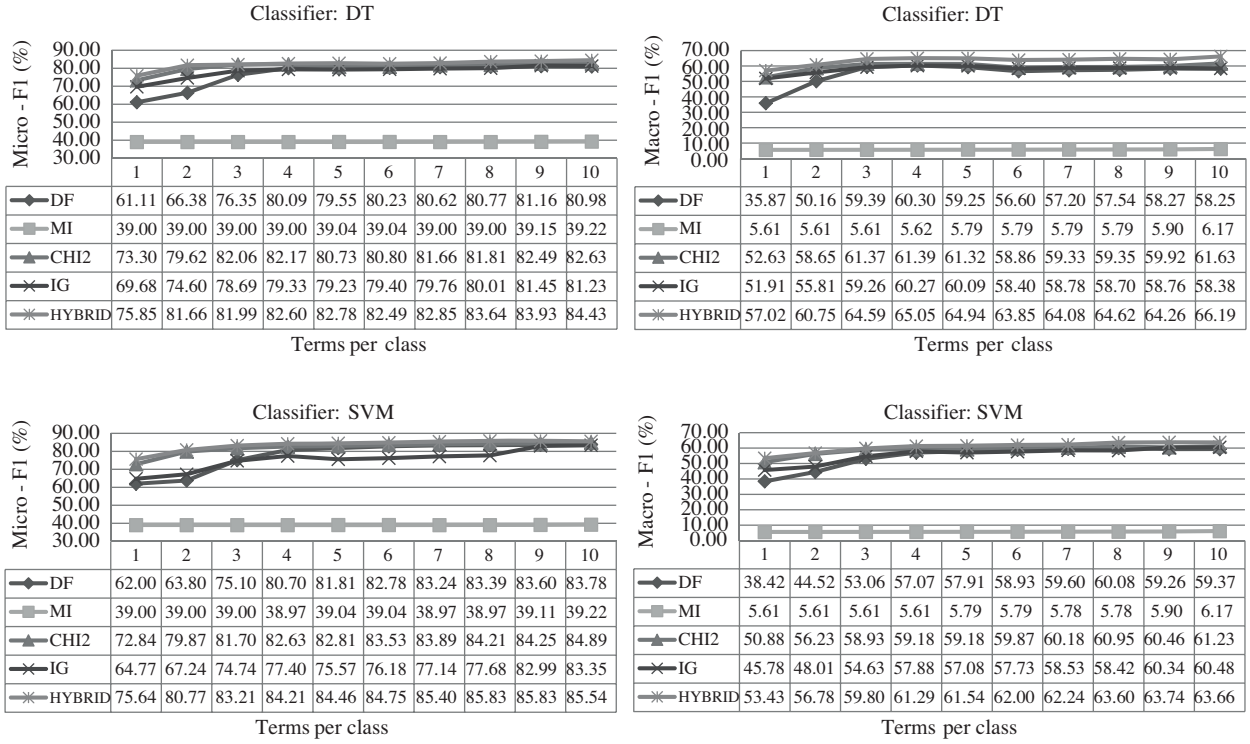


Figure 4. Success measures and feature set sizes before and after hybrid feature selection in the Newsgroups dataset.

As the next step, the performance of the hybrid feature selection scheme is compared with every single filter method. The comparison is provided in Figures 5 and 6 for the Reuters and Newsgroups datasets, respectively. When the individual performances of the filters are analyzed, it is clear that MI offers poor scores, although its uniqueness rate is higher than the other filter methods, as presented in Figure 2. This situation reveals that MI-based features are not individually discriminative, although they are unique for each class. On the contrary, DF, IG, and CHI2 deliver similar performances as the feature set size increases. These results are in agreement with those presented in [25,29]. When the filters are compared with the proposed feature selection scheme, it is apparent that the hybrid selection offers improved performance with respect to individual filter methods for almost every combination of the utilized success measures, classifiers, datasets, and number of terms per class. The amount of improvement is much more noticeable, especially in the Reuters dataset, which has overlapping classes that make it difficult to classify correctly. For instance, the highest Micro-F1 score obtained by the filters (84.89%) is improved to 85.83% whereas the highest Macro-F1 score obtained by the filters (61.63%) is increased to 66.19% when the hybrid selection is employed in the Reuters dataset. Similarly, in the Newsgroups dataset, the highest Micro-F1 and Macro-F1 scores (98.00% and 98.01%) obtained by the filters are improved to 98.48% and 98.44%.



Micro-F1 (%)	1	2	3	4	5	6	7	8	9	10
DF	61.11	66.38	76.35	80.09	79.55	80.23	80.62	80.77	81.16	80.98
MI	39.00	39.00	39.00	39.00	39.04	39.04	39.00	39.00	39.15	39.22
CHI2	73.30	79.62	82.06	82.17	80.73	80.80	81.66	81.81	82.49	82.63
IG	69.68	74.60	78.69	79.33	79.23	79.40	79.76	80.01	81.45	81.23
HYBRID	75.85	81.66	81.99	82.60	82.78	82.49	82.85	83.64	83.93	84.43

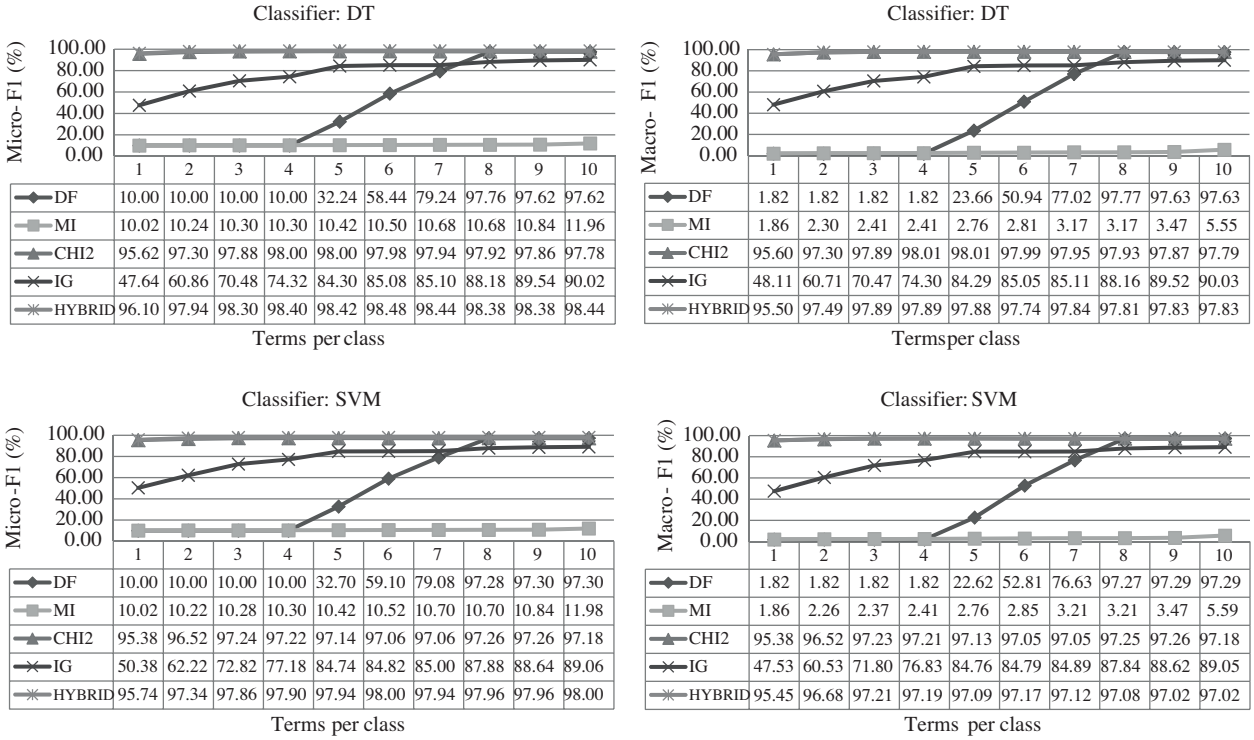
Macro-F1 (%)	1	2	3	4	5	6	7	8	9	10
DF	35.87	50.16	59.39	60.30	59.25	56.60	57.20	57.54	58.27	58.25
MI	5.61	5.61	5.61	5.62	5.79	5.79	5.79	5.79	5.90	6.17
CHI2	52.63	58.65	61.37	61.39	61.32	58.86	59.33	59.35	59.92	61.63
IG	51.91	55.81	59.26	60.27	60.09	58.40	58.78	58.70	58.76	58.38
HYBRID	57.02	60.75	64.59	65.05	64.94	63.85	64.08	64.62	64.26	66.19

Micro-F1 (%)	1	2	3	4	5	6	7	8	9	10
DF	62.00	63.80	75.10	80.70	81.81	82.78	83.24	83.39	83.60	83.78
MI	39.00	39.00	39.00	38.97	39.04	39.04	38.97	38.97	39.11	39.22
CHI2	72.84	79.87	81.70	82.63	82.81	83.53	83.89	84.21	84.25	84.89
IG	64.77	67.24	74.74	77.40	75.57	76.18	77.14	77.68	82.99	83.35
HYBRID	75.64	80.77	83.21	84.21	84.46	84.75	85.40	85.83	85.83	85.54

Macro-F1 (%)	1	2	3	4	5	6	7	8	9	10
DF	38.42	44.52	53.06	57.07	57.91	58.93	59.60	60.08	59.26	59.37
MI	5.61	5.61	5.61	5.61	5.79	5.79	5.78	5.78	5.90	6.17
CHI2	50.88	56.23	58.93	59.18	59.18	59.87	60.18	60.95	60.46	61.23
IG	45.78	48.01	54.63	57.88	57.08	57.73	58.53	58.42	60.34	60.48
HYBRID	53.43	56.78	59.80	61.29	61.54	62.00	62.24	63.60	63.74	63.66

Figure 5. Comparison of hybrid selection and the filters for the Reuters dataset.

Following the hybrid feature selection process, the distribution profile of the features within the selected feature subset enables us to assess the efficacy of different features. For this purpose, the coverage rates of the DF-, MI-, CHI2-, and IG-based features are analyzed within the selected subset. The coverage rate indicates the ratio of the number of unique features selected by a particular filter method to the total number of features in the entire feature subset. This analysis is carried out considering the cases in which the highest success measures are attained in each dataset. The results of the analysis are presented in Figure 7. It is apparent



	1	2	3	4	5	6	7	8	9	10
DF	10.00	10.00	10.00	10.00	32.24	58.44	79.24	97.76	97.62	97.62
MI	10.02	10.24	10.30	10.30	10.42	10.50	10.68	10.68	10.84	11.96
CHI2	95.62	97.30	97.88	98.00	98.00	97.98	97.94	97.92	97.86	97.78
IG	47.64	60.86	70.48	74.32	84.30	85.08	85.10	88.18	89.54	90.02
HYBRID	96.10	97.94	98.30	98.40	98.42	98.48	98.44	98.38	98.38	98.44

	1	2	3	4	5	6	7	8	9	10
DF	1.82	1.82	1.82	1.82	23.66	50.94	77.02	97.77	97.63	97.63
MI	1.86	2.30	2.41	2.41	2.76	2.81	3.17	3.17	3.47	5.55
CHI2	95.60	97.30	97.89	98.01	98.01	97.99	97.95	97.93	97.87	97.79
IG	48.11	60.71	70.47	74.30	84.29	85.05	85.11	88.16	89.52	90.03
HYBRID	95.50	97.49	97.89	97.89	97.88	97.74	97.84	97.81	97.83	97.83

	1	2	3	4	5	6	7	8	9	10
DF	10.00	10.00	10.00	10.00	32.70	59.10	79.08	97.28	97.30	97.30
MI	10.02	10.22	10.28	10.30	10.42	10.52	10.70	10.70	10.84	11.98
CHI2	95.38	96.52	97.24	97.22	97.14	97.06	97.06	97.26	97.26	97.18
IG	50.38	62.22	72.82	77.18	84.74	84.82	85.00	87.88	88.64	89.06
HYBRID	95.74	97.34	97.86	97.90	97.94	98.00	97.94	97.96	97.96	98.00

	1	2	3	4	5	6	7	8	9	10
DF	1.82	1.82	1.82	1.82	22.62	52.81	76.63	97.27	97.29	97.29
MI	1.86	2.26	2.37	2.41	2.76	2.85	3.21	3.21	3.47	5.59
CHI2	95.38	96.52	97.23	97.21	97.13	97.05	97.05	97.25	97.26	97.18
IG	47.53	60.53	71.80	76.83	84.76	84.79	84.89	87.84	88.62	89.05
HYBRID	95.45	96.68	97.21	97.19	97.09	97.17	97.12	97.08	97.02	97.02

Figure 6. Comparison of hybrid selection and the filters for the Newsgroups dataset.

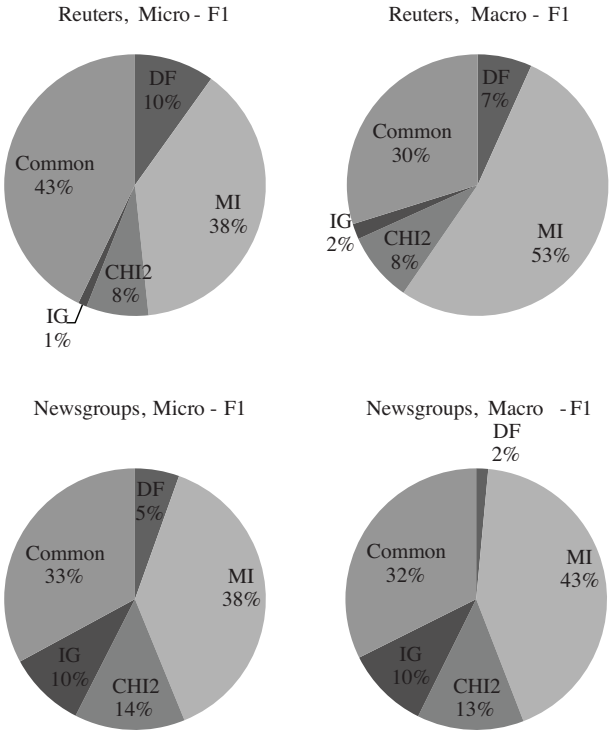


Figure 7. Distribution of the features after hybrid feature selection.

that the coverage rate of the MI-based features is significantly high with respect to other features, although the MI-based features offer a poor performance when used alone (see Figures 5 and 6). For example, in the Reuters dataset, 38% of the selected feature subset providing the highest Micro-F1 score is constituted by only MI-based features. On the other hand, the coverage rates of the DF-, CHI2-, and IG-based features are only 10%, 8%, and 1%, respectively, within the same distribution. The remaining part with 43% coverage, namely common, is constituted by the features that are common to at least 2 different filter methods. The dominance of MI-based features is valid for the Newsgroups dataset as well, while the coverage rates of the DF-, CHI2-, and IG-based unique features vary between 2% and 14%.

The bottom line is that a combination of the features selected by various methods is more effective than the features selected by the single selection method. However, the profile of the combination may vary based on the utilized dataset, classifier, and success measure. A feature that is not capable of providing satisfactory discrimination individually may unexpectedly have a positive effect when used together with some other features. These observations show that when new features are adopted for a text classification problem, one should consider all of the varying parameters to decide whether the new features are actually relevant or not.

5. Conclusions

In this study, a hybrid feature selection scheme, which is composed of filter and wrapper selection stages, is proposed for text classification problems so that a combination of various types of features can be assessed. The selection process is carried out under different conditions: varying feature set sizes, dataset characteristics, classification algorithms, and the success measures. This approach enables us to discover which features or feature combinations are better identifiers for text classification and whether there is a correlation among the useful features, the desired feature size, the utilized classification method, and the success measure. During the assessment, information such as the uniqueness and coverage rate of the features are utilized. The results of the experimental study reveal that a combination of the features selected by various methods is more effective than the features selected by the single selection method. The profile of the combination, however, depends on the characteristics of the dataset and the choice of the classification algorithm and success measure. The incorporation of other features and feature weighting schemes is a possible extension of this research and remains as an interest for future study.

References

[1] Z. Li, Z. Xiong, Y. Zhang, C. Liu, K. Li, "Fast text categorization using concise semantic analysis", Pattern Recognition Letters, Vol. 32, pp. 441–448, 2011.

[2] İ. Pehlivan, Z. Orhan, "Automatic knowledge extraction for filling in biography forms from Turkish texts", Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 19, pp. 59–71, 2011.

[3] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", Proceedings of the 14th International Conference on Machine Learning, pp. 143–151, 1997.

[4] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, Vol. 24, pp. 513–523, 1988.

[5] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge, Cambridge University Press, 2008.

[6] G. Salton, A. Wong, C. Yang, "A vector space model for automatic indexing", Communications of the ACM, Vol. 18, pp. 613–620, 1975.

[7] R.E. Bellman, Adaptive Control Processes, Princeton, New Jersey, Princeton University Press, 1961.

[8] A.K. Jain, R.P. Duin, J. Mao, "Statistical pattern recognition: a review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, pp. 4–37, 2000.

[9] B. Yu, Z. Xu, C. Li, "Latent semantic analysis for text categorization using neural network", Knowledge-Based Systems, Vol. 21, pp. 900–904, 2008.

[10] C.H. Lee, H.C. Yang, "Construction of supervised and unsupervised learning systems for multilingual text categorization", Expert Systems with Applications, Vol. 36, pp. 2400–2410, 2009.

[11] A. Selamat, S. Omatu, "Web page feature selection and classification using neural networks", Information Sciences, Vol. 158, pp. 69–88, 2004.

[12] J. Ye, Q. Li, "A two-stage linear discriminant analysis via QR-decomposition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, pp. 929–941, 2005.

[13] K. Torkkola, "Discriminative features for text document classification", Pattern Analysis and Applications, Vol. 6, pp. 301–308, 2003.

[14] C.H. Park, M. Lee, "On applying linear discriminant analysis for multi-labeled problems", Pattern Recognition Letters, Vol. 9, pp. 878–887, 2008.

[15] Y. Saeys, I. Inza, P. Larrañaga, "A review of feature selection techniques in bioinformatics", Bioinformatics, Vol. 23, pp. 2507–2517, 2007.

[16] H. Liu, L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, pp. 491–502, 2005.

[17] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection", Journal of Machine Learning Research, Vol. 3, pp. 1157–1182, 2003.

[18] G. Kohavi, R. John, "Wrappers for feature subset selection", Artificial Intelligence, Vol. 97, pp. 273–324, 1997.

[19] Y. Zhang, C. Ding, T. Li, "Gene selection algorithm by combining reliefF and mRMR", BMC Genomics, Vol. 9, doi 10.1186/1471-2164-9-S2-S27, 2008.

[20] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection", Proceedings of the 18th International Conference on Machine Learning, pp. 74–81, 2001.

[21] S. Gunal, R. Edizkan, "Subspace based feature selection for pattern recognition", Information Sciences, Vol. 178, pp. 3716–3726, 2008.

[22] P.M. Narendra, K. Fukunaga, "A branch and bound algorithm for feature subset selection", IEEE Transactions on Computers, Vol. C-26, pp. 917–922, 1977.

[23] Y. Yang, "Noise reduction in a statistical approach to text categorization", Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 256–263, 1995.

[24] D. Mladeni, M. Grobelnik, "Feature selection on hierarchy of web documents", Decision Support Systems, Vol. 35, pp. 45–87, 2003.

[25] Y. Yang, J.O. Pedersen, "A comparative study on feature selection in text categorization", Proceedings of the 14th International Conference on Machine Learning, pp. 412–420, 1997.

[26] H. Liu, J. Sun, L. Liu, H. Zhang, "Feature selection with dynamic mutual information", Pattern Recognition, Vol. 42, pp. 1330–1339, 2009.

[27] Y.T. Chen, M.C. Chen, "Using chi-square statistics to measure similarities for text categorization", Expert Systems with Applications, Vol. 38, pp. 3085–3090, 2011.

[28] C. Lee, G.G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization", Information Processing and Management, Vol. 42, pp. 155–165, 2006.

[29] G. Forman, "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research, Vol. 3, pp. 1289–1305, 2003.

[30] S. Gunal, O.N. Gerek, D.G. Ece, R. Edizkan, "The search for optimal feature set in power quality event classification", Expert Systems with Applications, Vol. 36, pp. 10266–10273, 2009.

[31] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Reading, Massachusetts, Addison-Wesley, 1989.

[32] C.L. Huang, C.J. Wang, "A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications, Vol. 31, pp. 231–240, 2006.

[33] W. Siedlecki, J. Sklansky, "A note on genetic algorithms for large-scale feature selection", Pattern Recognition Letters, Vol. 10, pp. 335–347, 1989.

[34] J. Yang, V. Honavar, "Feature subset selection using a genetic algorithm", IEEE Intelligent Systems, Vol. 13, pp. 44–49, 1998.

[35] A. Asuncion, D.J. Newman, UCI Machine Learning Repository, Irvine, California, University of California School of Information and Computer Science, 2007.

[36] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Waltham, Massachusetts, Academic Press, 2009.

[37] T. Joachims, Advances in Kernel Methods – Support Vector Learning, Cambridge, MIT Press, 1999.

[38] D.E. Johnson, F.J. Oles, T. Zhang, T. Goetz, "A decision-tree-based symbolic rule induction system for text categorization", IBM Systems Journal, Vol. 41, pp. 428–437, 2002.

[39] M.A. Kumar, M. Gopal, "A comparison study on multiple binary-class SVM methods for unilabel text categorization", Pattern Recognition Letters, Vol. 31, pp. 1437–1444, 2010.

[40] H. Drucker, D. Wu, V.N. Vapnik, "Support vector machines for spam categorization", IEEE Transactions on Neural Networks, Vol. 10, pp. 1048–1054, 1999.

[41] C.W. Hsu, C.J. Lin, "A comparison of methods for multiclass support vector machines", IEEE Transactions on Neural Networks, Vol. 13, pp. 415–425, 2002.

[42] M.F. Porter, "An algorithm for suffix stripping", Program: Electronic Library and Information Systems, Vol. 14, pp. 130–137, 1980.