Expert Systems With Applications 49 (2016) 31–47

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Hybrid feature selection based on enhanced genetic algorithm for text categorization

Abdullah Saeed Ghareb*, Azuraliza Abu Bakar, Abdul Razak Hamdan

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

ARTICLE INFO

Keywords:
Hybrid feature selection
Enhanced genetic algorithm
Filter feature selection
Text categorization

ABSTRACT

This paper proposes hybrid feature selection approaches based on the Genetic Algorithm (GA). This approach uses a hybrid search technique that combines the advantages of filter feature selection methods with an enhanced GA (EGA) in a wrapper approach to handle the high dimensionality of the feature space and improve categorization performance simultaneously. First, we propose EGA by improving the crossover and mutation operators. The crossover operation is performed based on chromosome (feature subset) partitioning with term and document frequencies of chromosome entries (features), while the mutation is performed based on the classifier performance of the original parents and feature importance. Thus, the crossover and mutation operations are performed based on useful information instead of using probability and random selection. Second, we incorporate six well-known filter feature selection methods with the EGA to create hybrid feature selection approaches. In the hybrid approach, the EGA is applied to several feature subsets of different sizes, which are ranked in decreasing order based on their importance, and dimension reduction is carried out. The EGA operations are applied to the most important features that had the higher ranks. The effectiveness of the proposed approach is evaluated by using naïve Bayes and associative classification on three different collections of Arabic text datasets. The experimental results show the superiority of EGA over GA, comparisons of GA with EGA showed that the latter achieved better results in terms of dimensionality reduction, time and categorization performance. Furthermore, six proposed hybrid FS approaches consisting of a filter method and the EGA are applied to various feature subsets. The results showed that these hybrid approaches are more effective than single filter methods for dimensionality reduction because they were able to produce a higher reduction rate without loss of categorization precision in most situations.

1. Introduction

More than 80% of information is stored as text (Korde & Mahender, 2012); therefore, text categorization is an important task in machine learning and data mining for organizing a massive amount of information (Yun, Jing, Yu, & Huang, 2012). Text categorization is a forecasting process of text document categories that categorizes text documents based on the extracted knowledge from those documents (Manning & Schütze, 1999). In recent years, many categorization methods have been used for the text categorization of many different languages; for example, K nearest neighbor (Abu Tair & Baraka, 2013; Jiang, Pang, Wu, & Kuang, 2012), support vector machine (Joachims, 1998; Mesleh, 2011), Naïve Bayes (NB) (Chen, Huang, Tian, & Qu, 2009; Hattab & Hussein, 2012), decision tree (Harrag, El-Qawasmeh, & Pichappan, 2009; Lewis & Ringuette, 1994), and Associative Classification (AC) (Al-Radaideh, Al-Shawakfa, Ghareb, & Abu-Salem, 2011; Antonie & Zaiane, 2002; Chiang, Keh, Huang, & Chyr, 2008; Ghareb, Hamdan, & Bakar, 2012).

The feature space of textual data can contain a huge number of features, and performing text categorization with such a high dimensional feature space influences the effectiveness of the separation of the different categories of text. Moreover, the existence of noisy and irrelevant features can adversely affect categorization performance and degrade computer resource (Kabir, Shahjahan, & Murase, 2012; Khorsheed & Al-Thubaity, 2013). Therefore, feature selection (FS) has been applied in most of the text categorization methods proposed to date. Many different FS approaches have been developed to reduce text dimensionality and select the informative features for text categorization. Feature selection is defined as "a process that chooses an optimal subset of features according to certain criterion" (Liu & Motoda, 1998). The FS approaches can be grouped into either filter or wrapper approaches based on the evaluation methodology applied to the feature subsets (Blum & Langley, 1997). Filter approaches

* Corresponding authors. Tel.: +60 13 328 5530; +967-715-339998; fax: +60 38 921 6184.

E-mail addresses: aghurieb@yahoo.com, aghurieb@gmail.com (A.S. Ghareb), azuraliza@ukm.edu.my (A.A. Bakar), arh@ukm.edu.my (A.R. Hamdan).

http://dx.doi.org/10.1016/j.eswa.2015.12.004
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

evaluate features independently of categorization techniques, while wrapper approaches employ categorization techniques to evaluate feature subsets (Blum & Langley, 1997; Langley, 1994).

A wide range of effective filtering methods have been proposed and applied for text categorization in the literature, such as term frequency-based and document frequency-based FS methods (i.e. discriminative power measure and Gini index) (Azam, & Yao, 2012), Chi Square (CHI) (Mesleh, 2011; Ogura, Amano, & Kondo, 2009; Yang & Pedersen, 1997), Comprehensively Measure Feature Selection (CMFS) (Yang, Liu, Zhu, Liu, & Zhang, 2012), Odd Ratio (OR) (Mengle & Goharian, 2009; Mladenic & Grobelnik, 1999), compound-features (Figueiredo et al., 2011), distinguishing feature selector (Uysal & Gunal, 2012), Information Gain (IG) (Uysal & Gunal, 2012; Yang & Pedersen, 1997), Improved Gini index (GINI) (Mengle & Goharian, 2009), Poisson distribution (Ogura et al., 2009), binomial hypothesis testing (Yang, Liu, Liu, Zhu, & Zhang, 2011), Class Discriminating Measure (CDM) (Chen et al., 2009), ambiguity measure (Mengle & Goharian, 2009), GSS (Galavotti, Sebastiani, & Simi, 2000), F-measure of training text features (FM) (Forman, 2003; Mesleh, 2011) and many more.

Optimization algorithms (i.e. evolutionary and swarm intelligence algorithms) are considered to fall under the wrapper approach, where categorization techniques are utilized for feature subset evaluation. Several optimization algorithms have been applied successfully for dimensionality reduction and the FS problem in the text categorization field; for instance, Ant Colony Optimization (ACO) (Aghdam, Ghasem-Aghaee, & Basiri, 2009; Janaki Meena, Chandran, Karthik, & Vijay Samuel, 2012; Mesleh & Kanaan, 2008), Particle Swarm Optimization (PSO) (Chantar & Corne, 2011; Zahran & Kanaan, 2009), and the Genetic Algorithm (GA) (Chen & Zou, 2009; Gunal, 2012; Tan, Fu, Zhang, & Bourgeois, 2008; Tsai, Chen, & Ke, 2014; Uğuz, 2011; Uysal & Gunal, 2014).

The GA is an evolutionary algorithm (population-based algorithm) first proposed by Holland (1975). A GA emulates the evolution process found in nature; it is intended to conduct a search for an optimal solution to a given problem by mimicking natural selection. Several research studies have demonstrated the advantages of the GA in solving high dimensionality and FS problems (Gunal, 2012; Uysal & Gunal, 2014; Tsai, Chen & Ke, 2014; Lei, 2012; Uğuz, 2011). However, time consumption, parameter setting and random selection of the initial solution are the main problems associated with the GA. Therefore, enhancement of the GA as a FS strategy is desired to handle these problems and produce more accurate results for text categorization. The hybrid approach attempts to integrate the filter and wrapper approaches in one framework to achieve the best solution rapidly. Recently, many hybrid methods based on the GA have been proposed for text categorization; for example, Uysal and Gunal (2014) proposed a hybrid approach based on filter methods with a GA and latent semantic indexing; Tsai et al. (2014) employed a biological evolution concept to improve the GA; and Uğuz (2011) proposed a hybrid method based on IG, a GA and principal component analysis. Another combination of filter and wrapper approaches was proposed by Gunal (2012) in which the features are first selected by four filtering methods (IG, DF, MI, and CHI) and combined together as an input of the GA in the second stage. Fang, Chen, and Luo (2012) also investigated the performance of a combination of DF, IG, MI, CHI methods with the GA, while Lei (2012) employed the IG with the GA as a FS method for text categorization. These approaches are effective in reducing text dimensionality and improving the performance of text categorization. However, they are parametric-based approaches and this makes it difficult to tune the rate of crossover and mutation operations. Nevertheless, crossover and mutation rates have a real effect on the population diversity and quality of the selected solutions. As suggested by Azam and Yau (2012), the feature frequency can add useful information regarding the important features for text categories; indeed, the improvement of categorization performance is one of main objectives of optimized FS approaches. Therefore, it is desirable that the modification of crossover and mutation is based on useful information about the feature combination instead of the probability rate, which is hard to tune. Furthermore, the randomization in the GA may affect the final solution and the process to generate feature subsets takes a long time, which results in a high computational cost for classifier construction. However, hybridization of filter methods with GA can reduce the adverse impact of randomization, reduce feature dimensionality and speed up the feature subset generation and categorization processes.

In this paper, based on the above arguments, first an enhanced GA (named EGA) is introduced to modify the crossover and mutation operations and overcome their negative effect on the generation process and to increase the diversity of the feature population. In the proposed enhancement of the crossover operation, each of the selected parents was divided into two equivalent parts, and the weight of each part was calculated as the cumulative weight value of features in the part based on the feature frequency and document frequency approach. The features of each category in each chromosome (subset) were ordered based on their weight, so the weight was obtained and the cumulative features weight was computed and then the best two parts from the two parents were concatenated together to form a new feature subset (new child) and the other two parts formed the second subset (second child). Thus, in this approach feature weight information was used to guide the EGA's search for the best subsets. In the modified mutation operation, the source of subset to be mutated was considered and the cumulative accuracy of the original parents was calculated. If it was smaller than a given accuracy threshold, then a specific number of features in the mutated subset that had the least weight were selected and replaced by the most important features (which were not appeared in the mutated subset) from the best found feature subset in the previous generations. In this way a new source was formed for future generations of feature subsets. Thus, we perform crossover and mutation operations based on useful information instead of using probability. Second, we propose six hybrid FS approaches by incorporating six well-known single filter methods with EGA to handle the randomization effect on the selected feature subsets, with the aim of further reducing text dimensionality and the computational cost of classifier construction. In the first stage of the hybrid FS, the importance of the features is evaluated by using one of six filter methods, namely, OR, CDM, GSS, IG, FM and term frequency-inverse term frequency (TF-IDF) (Salton & Buckley, 1988; Zahran & Kanaan, 2009). In the second stage, dimensionality reduction is carried out by applying the EGA to the top ranked features selected by each filter method.

The rest of this paper is organized as follows: Section 2 briefly discusses the related works; Section 3 describes the proposed enhancement of the GA and the hybrid FS approach. Section 4 highlights the used categorization methods. Section 5 discusses the experiments and results and Section 6 concludes the paper.

2. Literature review

The hybrid FS approach attempts to combine the filter and wrapper approaches in one framework, where the features are selected in two stages; in the first stage with the filter and in the second stage using the wrapper approach. Recently, several hybrid FS approaches for text categorization have been proposed, but they have mainly been applied to English text categorization. For instance, Uğuz (2011) proposed a hybrid method based on a combination of an IG filter method, a GA and principal component analysis (PCA). In the first stage, the IG is utilized to assign a rank to each feature in the text datasets and a predefined percentage of features is selected from all features based on their ranks. In the second phase, the GA and PCA are applied separately and work on the selected feature subsets. The KNN and C4.5 (DT) are employed as categorization techniques in the conducted experiments and they are used to evaluate the feature subsets. The fitness function used with the GA to evaluate chromosomes (feature subsets) is based on the average F-measure of KNN and C4.5. The roulette wheel selection method is utilized for the selection of chromosomes to feed the next generation. Their results also showed that the performance of KNN and C4.5 is improved by using a small subset selected by IG-GA and IG-PCA and the results are better than when using IG in isolation. Another combination of filter and wrapper approaches was proposed by Gunal (2012). In his study, the features are first selected by four filtering methods (IG, DF, MI, and CHI) and then combined together as an input of a GA in the second stage. The SVM and DT classifiers are utilized for feature subset evaluation, where the subset fitness depends on the macro- and micro-average F-measure. In other works, Fang et al. (2012) investigated the performance of a combination of DF, IG, MI, CHI filter methods with a GA, Also, Lei (2012) employed IG with GA as an FS method for text categorization. Both studies concluded that the hybrid FS approach can reduce text dimensionality efficiently and significantly improve the performance of text categorization.

Generally, the GA is a good approach to employ to explore the feature space and it can produce many alternative feature subsets through reproduction operations on its way to finding the best subset that includes the most important features. Several approaches have been proposed to try to improve the GA for FS such as the use of subset size control in the fitness function (Tan et al., 2008) and the use of the biological evolution concept with GA (Tsai et al., 2014), while some researchers have focused on the hybridization of filter methods with GA (Gunal, 2012; Lei 2012; Tsai et al. 2014; Uğuz, 2011; Uysal & Gunal, 2014). These approaches are effective in reducing text dimensionality and improve the performance of text categorization. Nevertheless, they are parametric-based approach, so it is difficult to tune the rate of crossover and mutation operations. However, the crossover and mutation rate have a real effect on population diversity and the quality of the selected solutions.

According to Azam and Yau (2012), the feature frequency as well as document frequency can provide useful information regarding the important features for text categories. Also, the improvement of categorization performance is one of the main objectives of optimized or hybrid FS approaches. Therefore, the modification of crossover and mutation needs to be based on useful information about feature combinations instead of a probability rate that is hard to tune. Furthermore, in traditional crossover and mutation, the features to be exchanged or replaced are determined randomly. Also, the initial population is randomly selected; randomization may affect population diversity and create redundant feature subsets and, moreover, some of the less frequent features will have a chance to be chosen. Thus, randomization has a real effect on the final solution. Therefore, based on those arguments, in this research an EGA was developed in which the crossover and mutation operators were enhanced to overcome their negative effect on the generation process and increase the diversity of the feature population. The crossover operation was performed based on chromosome (feature subset) partitioning with term weight of chromosome entries (features), while the mutation was performed based on the classifier performance of the original parents and feature importance. Thus, the crossover and mutation operations were performed based on useful information instead of using probability and random selection. The feature frequency can add useful information regarding the important features for text categories; indeed, the improvement of categorization performance is one of the main objectives of optimized FS approaches. Therefore, it is desirable that the process to produce new feature subsets using crossover and mutation is based on useful information for feature combination instead of the probability rate, which is hard to tune. Furthermore, the process to generate feature subsets takes a long time, which results in a high computational cost for classifier construction. Also, randomization in the GA can adversely affect the final solution. However, the hybridization of filter methods with GA can reduce the impact of randomization, reduce feature dimensionality and accelerate the feature subset generation and categorization processes.

3. Enhanced GA and hybrid FS approaches

The insensitivity of the GA to noise and its need for less knowledge about the problem domain can make it a powerful technique to handle dimensionality reduction and FS for text categorization. However, time consumption, parameter setting and random selection of the initial solution are the main challenges associated with the GA. In addition, the high dimensionality of textual data is a major problem in text categorization due to the existence of noisy and irrelevant features, which adversely affects categorization precision. Therefore, to address these problems, we propose an EGA for FS and then we introduce some hybrid FS approaches that combine a single filter FS method with EGA in one framework.

3.1. Traditional GA for FS

When applying the GA as a FS approach, the search space is the feature dimension space. The GA starts by creating the initial population randomly, which is composed of a set of feature subsets (chromosomes), each subset representing one solution. The GA tries to evolve the best subset by selecting the best subsets (solutions) according to a selection strategy that depends on fitness function. The fitness function evaluates the fittest of each subset, and the fittest subsets survive into the next generation through the crossover and mutation reproduction steps. The crossover operation takes two feature subsets (two parents) and reproduces them to create two new subsets (children), whereas the mutation operation modifies a single subset by changing some of the features' values or replacing them randomly in that subset. Thus, the most important aspects to consider when using a GA as a FS approach are the: (i) definition of the fitness function (evaluation), (ii) definition of the selection strategy, and (iii) definition and implementation of the crossover and mutation operators.

3.2. EGA for FS by modification of crossover and mutation

The GA is a good approach to explore the feature space and it can produce many alternative feature subsets through reproduction operations toward obtaining the best subset that includes the most important features. Several approaches have tried to improve the GA for FS, for instance by the use of subset size control in the fitness function (Tan et al., 2008) and the use of the biological evolution concept with GA (Tsai et al., 2014). However, the GA operators, i.e. crossover and mutation, are the key to achieving a diversity of population and creating a new population for further runs of the algorithm. The adjustment of the probability rate of crossover and mutation is a difficult problem to solve and it is hard to coordinate these operations. In addition, in traditional crossover and mutation, the features to be exchanged or replaced are determined randomly; the initial population is randomly selected, but randomization may adversely affect population diversity and create redundant feature subsets, thus it has a real effect on the final solution. Therefore, we propose a method to modify and enhance the crossover and mutation operators by performing these operations based on other factors, which include chromosome partitioning and determining the weight of each partition based on the importance of the chromosome entries (features). Fig. 1 presents the pseudocode of the suggested modification of the crossover and mutation operations. The following describes the different aspects of the proposed EGA.

3.2.1. Initial population

The initial population is composed of a set of chromosomes that are generated randomly. Each chromosome consists of a set of candidate features (feature subset) and it represents one solution in the

```
Step 1: Generate the initial population
Step 2: Evaluate each subset in the population based on fitness function
Step 3: Create new population by repeating the following operation:
        a)   Select two parents for reproduction
        b)   Apply crossover operation:
           -    Divide each parent into two parts
           -    Compute the cumulative weight of each part based on feature and document frequencies
           -    Concatenate the best two parts that have the highest weight (first child)
           -    Concatenate the other parts that have the lowest weight (second child)
        c)   Apply mutation operation:
 For each subset:
 Get the prior information about original parents (parent accuracy)
 If ([accuracy (parent 1) + accuracy (parent2)] / 2 < 85%) Then:
 - Select features that have the lowest frequencies
 - Replace them by an equivalent number of features from the best found feature subset
 End For.
Update population
Go to step 2
End.
```

Fig. 1. Pseudocode of the EGA.

search space. The features in each chromosome are encoded as *n*-binary vectors, where the bit with value “1” means the corresponding feature is selected and “0” means the feature is not selected.

3.2.2. Fitness function

The fitness function evaluates the performance of the feature subset and plays a major role in the selection of the subsets to be included in the next generation (new population). The better performing subsets will have a higher chance of selection and reproduction to form the new population. In this work we combine the categorization performance of a NB classifier (assessed in terms of macro-average F-measure) with the feature subset size to identify the fitness function and evaluate the feature subset accordingly; the following fitness function is used in our method:

$$Fit - Fun\ (S_i) = Z * C\ (S_i)\ +\ (1\ - - Z)\ (1\ /\ Size\ (S_i)) \tag{1}$$

where (S_i) is the selected feature subset, $C(S_i)$ is the macro-average F-measure of this subset, $Size(S_i)$ is the number of features in this subset, and Z is a parameter between 0 and 1. The fitness of feature subset S_i is increased as the categorization performance increases and it decreases as the size of S_i increases; the higher value of Z means that the categorization performance is more important than the feature subset size. In the present method the Z value is set to 80%.

3.2.3. Selection

The selection process takes place after evaluation of the subsets, where the subsets are selected according to their relative fitness and the adopted selection strategy. Roulette wheel selection (RWS) is the method most frequently used with the GA (e.g. Fang et al. 2012; Gunal, 2012; Tan et al., 2008; Uğuz, 2011). This strategy selects the subset for reproduction based on a probability selecting function that is proportional to subset fitness and inversely proportional to the other subsets' fitness in the explored population. The probability of selecting a subset S_i for reproduction is calculated as:

$$P\ (S_i) = \frac{Fit - Fun\ (S_i)}{\sum_{i=1}^{n} Fit - Fun\ (S_i)} \tag{2}$$

3.2.4. Crossover

Crossover and mutation operators are called reproduction operations and they produce new generations of subsets (children) from the selected subsets (parents). These operations are important to maintain population diversity. The most widely used crossover method involves forming two new subsets from the two parent subsets by swapping their features (Uğuz, 2011). The crossover is performed based on a crossover rate or probability that determines the chance that two parents will swap their features. We use the single point crossover, where the crossover is performed by choosing a random position *i* in the length of the parent subsets and exchanging all the features after that point. Thus, the first *i* features are selected from one subset and the remaining features are chosen from the second subset.

In the proposed enhancement of this crossover operation, each parent of the selected parents is divided into two equivalent parts and the weight of each of those parts is calculated as the cumulative weight value of the features in each part based on the feature frequency and document frequency approach. The weight of the features is already computed in the text preprocessing phase. The features of each category in each chromosome (subset) are ordered based on their weight; we get the weight and compute the cumulative features weight and then the best two parts from the two parents are concatenated together to form a new feature subset (new child) and the other two parts form the second subset (second child). Thus, in this approach we use feature weight information to guide the EGA search toward the best subsets.

3.2.5. Mutation

The mutation operation is applied to a single feature subset; the number of subsets to be mutated is identified according to the mutation rate. A random number of features are selected from the subset to be mutated and their values are changed. In our approach we replace a specific number of features in the mutated subset by features from the best subsets found in the previous generation. Furthermore, in this operation, we look at the source of the feature subset (child) to be mutated and calculate the cumulative accuracy (F-measure) of the original parents; if it is smaller than a given threshold, then we select a specific number of features in that subset that have less weight and replace them with the highest important features that do not appear in the subset from the best found feature subset, and thus we form a new source of future generations.

3.3. Proposed hybrid FS approach based on EGA

Given a subset of features S, which can be selected using any filter FS method, the problem is still the high dimensionality of the features that can be revised with the wrapper approach. The GA is the suggested optimization approach to reduce text dimensionality. However, the GA is impractical for high dimensional text because it takes a long time to locate the relevant feature subset. Furthermore, the GA suffers from randomization in the initial population, which may affect the final result. To avoid these problems and achieve better performance in terms of text categorization, we propose six hybrid

```
Algorithm: Hybrid FS based on EGA
Step 1: Preprocess the text documents
Step 2: Evaluate features based on six filter methods (CDM, IG, OR, FM, GSS and TF-IDF)
Step 3: For each filter method
Begin
Select a predefined number of features for each category based on their scores
Steps 4: Apply GA for the selected subsets
Begin
 Generate the population
 While not reached the termination condition
 For each feature subset (chromosome) in the population
 Calculate the fitness function based on feature size and NB performance
 Create the new population:
 Select parents;
 Apply crossover operation on pairs of parents;
 Apply mutation operation to children;
Store the best subsets (best chromosomes)
End while.
Step 5: Return the best stored feature subsets & combine their features
End For.
Step 6: Go to categorization steps.
 Test with Naïve Bayes
 Test with Associative Classification
```

Fig. 2. Pseudocode of the hybrid FS approach.

approaches, each of which incorporates a different filter method with GA. The general scheme of the proposed hybrid FS approach is presented in the pseudocode shown in Fig. 2. There are two stages in the FS process: the filter stage and the GA stage.

3.3.1. Stage 1: filter methods

This stage is performed for three purposes, first to perform an initial reduction of text dimensionality, second, to further minimize the effect of randomization of the initial population generation in EGA and third to speed up the feature subsets generation process with EGA. In this stage, we use six filter methods that are chosen for their efficiency and diversity. The methods are CDM, FM, GSS, IG, OR and TF-IDF. Details about these methods can be found in the literature (Chen et al., 2009; Forman, 2003; Mesleh, 2011). The following equations (3)–(8) present these methods mathematically, as computed for each feature f in category c_i.

$$TF - IDF = TF(f, c_i) * \log(N/DF(f)) \tag{3}$$

$$OR = (P(f|c_i) * (1 - P(f|c_i^\wedge)))/(P(f\,|c_i^\wedge) * (1 - P(f|c_i))) \tag{4}$$

$$CDM = |\log(P(f|c_i)/P(f|c_i^\wedge))|| \tag{5}$$

$$GSS = \frac{[(P(f,\ c_i) *\ P(f\,^\wedge, c_i^\wedge))\ --\ (P(f, c_i^\wedge)\ *\ P(f\,^\wedge, c_i))]}{N^2} \tag{6}$$

$$\begin{aligned} IG = {} & -\sum_{i=1}^{m} P(c_i)\cdot \log P(c_i) + P(f) + \sum_{i=1}^{m} P(f\,|c_i\,)\cdot \log P(f\,|c_i\,) \\ & + P\,(f^\wedge)\sum_{i=1}^{m} P(f^\wedge\,|c_i\,)\cdot \log P(f^\wedge\,|c_i\,) \end{aligned} \tag{7}$$

$$FM = \frac{2\ (t_p)}{t_p + f_p + t_p + f_n} \tag{8}$$

where N is the total number of training documents in the collection; $P(c_i)$ is the percentage of documents that belongs to category c_i among the total number of documents; $P(c_i^\wedge)$ is the percentage of documents that does not belong to category c_i among the total number of documents; $P(f)$ is the probability of the feature f; $P(f^\wedge)$ is the percentage of documents that does not contains feature f; $P(f, c_i)$ is the joint probability of the category c_i and the occurrence of the feature f; $P(f|c_i)$ is the probability of feature f given category c_i; t_p is true positives (number of positive cases containing a word); f_n is the false negative cases; f_p is the false positives (number of negative cases containing a word); and t_n is the true negative cases.

In this stage, the original feature space is reduced from N to S based on the features' importance, which is calculated by a filter method. In other words, before dimension reduction with EGA, the text features are ranked based on their importance for categorization in decreasing order by using one of six filter methods. Thereby, in the text categorization, the less important features are ignored and the dimension reduction is applied to the most important features. The top features are used in the next stage, which searches for the optimal feature subset in a wrapper way by using EGA.

3.3.2. Stage 2: EGA wrapper/optimization method

In the second stage, the EGA is applied to several subset sizes of features, which are ranked in decreasing order based on their importance, and dimension reduction is carried out. The EGA operations are applied on the most important features, i.e. those which have the higher ranks. Thus, the categorization complexity and computational cost will be reduced. The objective of the EGA is to maximize the reduction of feature dimensionality while minimizing the loss of important information and to improve the classifier accuracy in terms of F-measure. The details of the various aspects of the EGA are provided above in Section 3.2.

4. Text categorization techniques

Categorization techniques are employed to measure the strength of the proposed methods and to show how they affect categorization performance. We investigated two categorization techniques for this purpose; the first is a popular text categorization technique, NB, which is based on the probabilistic information about text features. The second is AC, which is based on the construction of categorization rules that depend on association rule mining algorithms.

4.1. NB

Naïve Bayes is a simple probabilistic text categorization technique, which is based on a Bayesian theorem. We use the multinomial NB model because it is usually better than other variations of NB and it works well for text categorization (McCallum & Nigam, 1998; Chen et al., 2009; Youn & Jeong, 2009). In this paper, NB is utilized for feature evaluation in the training phase and it is tested on other text documents that are not used in the training phase. In text categorization, the aim is to find the correct category of text document. The best category *c* of document *d* in a NB categorization method is the category with the highest probability that document *d* belongs to category *c* and it is calculated as:

$$c = \arg\max \cdot P(c|d) = \arg\max \cdot P(c)\Pi i\, P(f_i|c) \tag{9}$$

$$P(f_i|c) = F_c + cons\ /\ N_c + FV \tag{10}$$

where $P(c|d)$ is the probability that a given document *d* belongs to category *c*; $P(c)$ is the probability of category *c*, which is computed as the ratio of documents that belong to category *c* among the total number of documents; $P(f_i|c)$ is the probability of a set of features f_i given a category *c*; F_c is the number of times the feature *i* occurs in category *c*; N_c is the number of features in category *c*; *FV* is the number of features in the given vocabulary; and *cons* is the positive constant, which is usually set to one to avoid zero probability.

4.2. AC

Associative classification is a branch of rule-based techniques that construct rules based on the concept of association rule mining. Several previous studies that have used an AC method provide directories that show it has the capability to produce classifiers to rival to those learned by other categorization techniques (Thabtah, 2007). In this paper, the main steps of rule generation and categorization using the AC method are summarized as follows (Al-Radaideh et al., 2011; Ghareb et al., 2012; Thabtah, 2007; Yin & Han, 2003):

- Input: a set of documents D_i with their features (feature vectors) for all categories after the FS process; set the minimum support and minimum confidence.
- Output: a set of categorization rules ($f_j \to c_i$) for all categories.
- Step 1: Rule discovery; perform an iterative search through documents D_i and generate the category association rules using a priori-based algorithm.
- Step 2: Rule ordering; order the category association rules according to support, confidence, and length criteria.
- Step 3: Rule pruning; reduce the total number of rules to generate a reasonable number of association rules that compose the associative classifier. This step is based on rule confidence and redundancy.
- Step 4: Save the revised rules in an ordered decision list to be used later for prediction.
- Step 5: Use a multiple rule prediction method to categorize the text document. In this method, all the rules that cover the text document to be categorized are retained. If all the retained rules are associated with only one category, this category will be the text document category. However, when the matched rules are distributed among two or more categories, these rules are divided into n subsets by category, and the text document is assigned to the category which has the largest number of rules.

5. Experiments and discussion

5.1. Datasets and experimental setup

We conducted experiments on three different Arabic text datasets to evaluate the proposed EGA and hybrid FS approaches. The datasets include two relatively small datasets, “Al-Jazeerah (D1)” and “Akhbar Al-Khaleej (D2)” and one large dataset, “Al-waten text datasets (D3)” (Abbas, Smaili, & Berkani, 2011; Chantar & Corne, 2011). Table 1 presents the distribution of documents in the different categories of each dataset. The text in each dataset is processed by using Arabic text preprocessing tools (Al-Radaideh et al., 2011). The function words, punctuation marks and non-Arabic words and letters are removed by a stemming algorithm applied to reduce text sparsity, while the word importance is determined based on a term frequency and document frequency weighting approach. Each text dataset is randomly distributed to training and testing datasets without overlapping. In the case of D1, the training dataset contains 1200 documents (240 documents for each category) and the remainder is used for testing, while in the case of D2 and D3, about 70% is used for training and the rest for testing. A set of documents (10%) is selected randomly from the training datasets for the validation process to estimate the quality of the selected feature subsets (chromosomes) based on NB performance in a wrapper way in the training phase.

The evaluation of the results focuses on categorization performance (macro-average values), execution time and reduction rate of text dimensionality. The categorization effectiveness in terms of precision (P_i), recall (R_i) and F-measure (F_i) for each category is computed as follows:

$$P_i = T/M,\ R_i = T/D,\ \text{and } F_i = (2P_iR_i)/(P_i + R_i)$$

where *T* is the number of documents correctly assigned to category *i*; *M* is the number of documents assigned to category *i*; and *D* is the total documents in category *i*. In the case of multiple categories, the macro-average values are employed to evaluate the performance across multiple categories *C*.

$$P\_macro = \frac{\sum_{i=1}^{C} P_i}{C}, \quad R\_macro = \frac{\sum_{i=1}^{C} R_i}{C},$$

$$F\_macro = \frac{2\left(\sum_{i=1}^{C} P_i * \sum_{i=1}^{C} R_i\right)}{C\left(\sum_{i=1}^{C} P_i + \sum_{i=1}^{C} R_i\right)}$$

The reduction rate (*RR*) can be calculated as $[1 - (N_i/N)]$, where N_i is the number of features selected to construct the categorization model and *N* is the total number of features (Lei, 2012).

In each experiment, we test the quality of the selected features subset by using NB and AC categorization techniques. The NB classifier is utilized in the training phase as an induction algorithm to estimate the fittest among the feature subsets along with the feature size and its tested with another text dataset which does not appears

Table 1
Distribution of “Alwatan”, “Al-jazeera” and “Akhbar-Alkhaleej” text datasets.

Alwatan” text datasets			Al-jazeera text datasets		Akhbar-Alkhaleej text datasets	
Category	# documents	# words	Category	# documents	Category	# documents
Religion	3860	3,144,828	Economy	300	Economy	273
Economy	3468	1,482,009	Politics	300	Sports	429
Sports	4550	1,447,889	Sport	300	Local news	720
Culture	2782	1,411,218	Science	300	Int. news	286
Int. news	2035	865,671	Art	300	Total	1708
Total	16,695	8,351,615	Total	1500	# words	746,307
			# words	389,766		

Table 2
GA and EGA parameter settings.

Description	GA setting	EGA setting
Population size	32	32
Selection technique	RWS	RWS
Crossover type	One point	One point
Crossover rate	0.9	–
Mutation rate	0.001	–
Mutation (F-measure threshold)	–	85%
Number of replaced features	20	20
Iteration number	200	200

in the training phase. The association rule mining algorithm is set to 10% and 50% for minimum support and minimum confidence, respectively. The GA and EGA parameter settings are given in Table 2. The proposed approaches are implemented using C#.net and all experiments are conducted on a PC with an Intel(R) Core^{TM} i3 processor, 2.27 GHz, 4GB of RAM, and Windows operating system.

5.2. Experimental results of the GA and EGA

This section presents a comparison between GA and EGA, which is conducted to see if EGA is better than GA in terms of categorization performance, processing time and reduction rate. Furthermore, it shows the effectiveness of GA and EGA as FS algorithms for reducing text dimensionality and shows their effect on categorization performance.

5.2.1. Analysis of dimensionality reduction and time required with GA and EGA

Dimension reduction is an important aim of any FS approach; therefore, an analysis of the text reduction rate by using GA and EGA is carried out during the performed experiments. In addition, the time required for the whole process with the AC and NB classifiers is assessed. Table 3 shows the dimension reduction rate with GA and EGA when they are tested on three Arabic text datasets and the time required with the AC and NB classifiers. The datasets are abbreviated as D1, D2 and D3 for the Al-Jazeerah datasets, Akhbar Al-Khaleej datasets and Al-waten datasets, respectively.

As shown in Table 3, the highest reduction rate is achieved by EGA with all text datasets; this means that EGA performs better than GA in reducing dimensionality. The reason for this result is that EGA groups the features based on their importance in the crossover operation before evaluation with the induction algorithm, so a large number of features is reduced in the next generation because the fitness values of the feature subsets that contains less important features are low, which prevents their appearance in subsequent generations.

In contrast, in the traditional GA the new feature subsets are generated randomly using crossover and mutation operations. The results also indicate that EGA can speed up the categorization process with the AC and NB classifiers because fewer features are selected by EGA than by GA. When comparing the cumulative time with NB and AC, we can see that the NB is faster than AC with all text datasets. This is because AC also requires a further process to perform the generation of categorization rules, and it requires extra time for these to be discovered.

Table 4
Effect of GA- and EGA-based FS methods on NB performance on three datasets (macro-averages).

Dataset	GA			EGA		
	Precision	Recall	F-measure	Precision	Recall	F-measure
D1	0.8882	0.883	0.8857	0.91	0.9066	0.9083
D2	0.8130	0.869	0.8402	0.8344	0.9007	0.8662
D3	0.8915	0.891	0.8913	0.9067	0.9003	0.9035

Table 5
Effect of GA- and EGA-based FS methods on AC performance for three datasets (macro-averages).

Dataset	GA			EGA		
	Precision	Recall	F-measure	Precision	Recall	F-measure
D1	0.7174	0.6006	0.6529	0.7805	0.6480	0.7065
D2	0.6411	0.6517	0.6463	0.6332	0.6848	0.6575
D3	0.7625	0.6460	0.6975	0.7864	0.6518	0.7095

5.2.2. Analysis of categorization performance with GA and EGA

We also investigate the performance of NB and AC using the feature subsets selected by GA and EGA. The results are presented in Tables 4 and 5 in terms of macro-average precision, macro-average recall and macro-average F-measure. It is clear from both tables that EGA achieves the best performance with both classifiers on all text datasets; however, the difference is slight when the NB is utilized for categorization. For instance, in the case of the D1 datasets NB achieves an 89.13% macro-average F-measure with GA and 90.35% with EGA, which is only about a 1% improvement in performance. The best improvement in NB performance is gained with EGA when tested on D2. In the case of AC, the performance is improved by more than 2% for all text datasets. The results also demonstrate that we cannot achieve the same improvement with different text datasets and different categorization techniques.

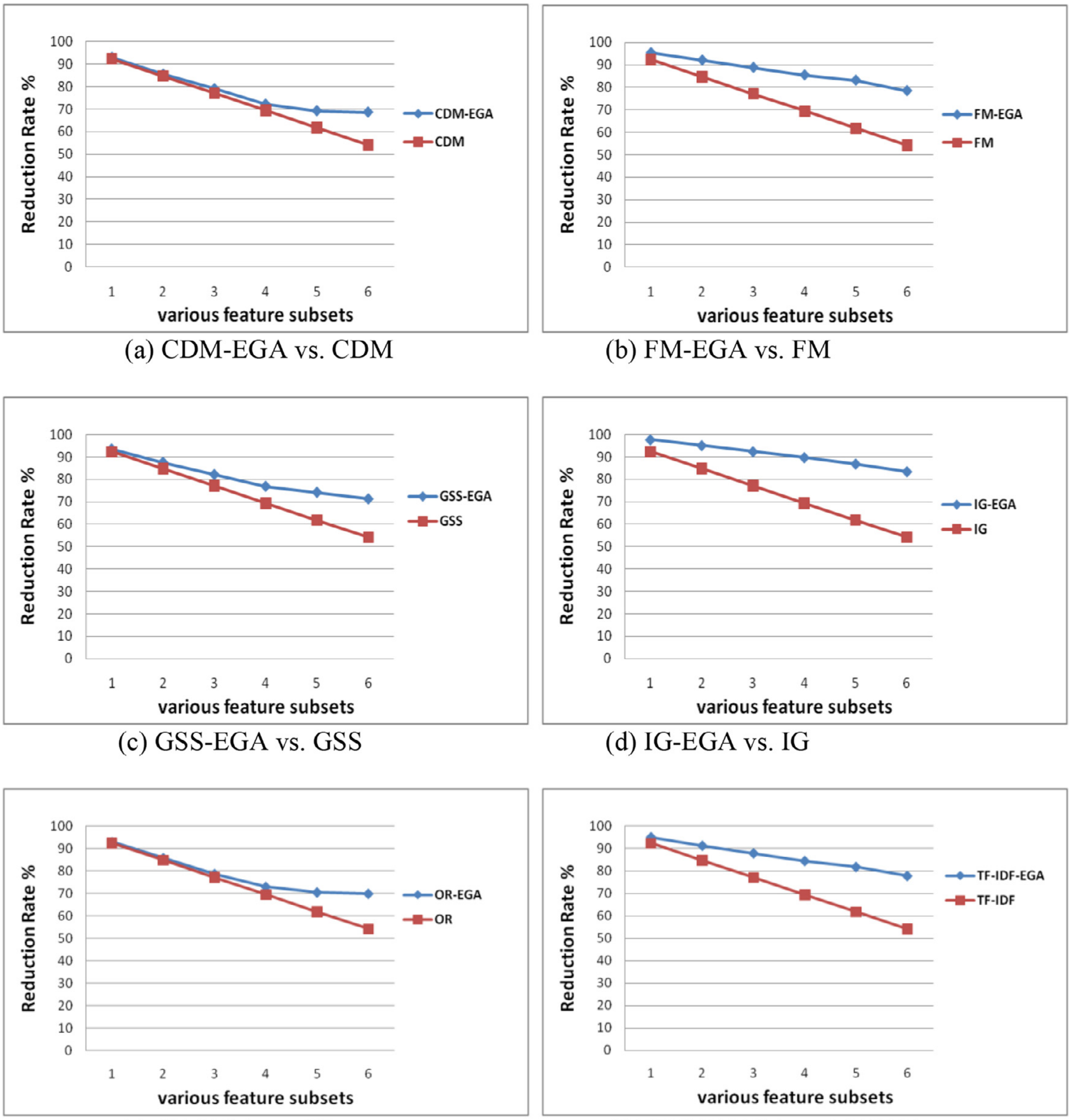
Overall, the results reveal that EGA can produce better results than GA in terms of dimensionality reduction and categorization performance. Therefore, we use EGA for hybridization with six filter methods, as discussed in the next section.

5.3. Experimental results of hybrid FS approaches

A number of experiments are conducted to determine the efficiency of the proposed hybrid FS approaches. The approaches are tested on the three text datasets (D1, D2, and D3) used in the previous section. In the first stage, the most important features in the text documents are identified by using one of the six filter methods (CDM, IG, FM, GSS, OR and TF-IDF). However, the problem of high dimensionality of the selected feature space remains, so the EGA is applied to the top ranked features to achieve an optimal/near-optimal reduction rate of feature dimensions. Then, in the second stage, the EGA is utilized to find the best feature subset from a given subset selected by the filtering methods with the objective of reducing the feature subset and maximizing categorization performance. To assess the effectiveness of the hybrid approaches, we perform an analysis based on three factors: reduction rate, categorization performance and time required for each approach.

Table 3
Reduction rate with GA and EGA and time required with NB and AC classifiers for three datasets.

Dataset	Total # feature	# features		Reduction rate %		Time with NB (min:s)		Time with AC (min:s)	
		GA	EGA	GA	EGA	GA	EGA	GA	EGA
D1	13084	3943	2025	69. 864	84. 523	16:48	12:27	130:00	124:36
D2	13250	3490	2084	73. 660	84.272	17:00	13:15	105:26	77:18
D3	32353	6065	3453	81.254	89.327	149	39:40	520:36	113:20



(a) CDM-EGA vs. CDM

(b) FM-EGA vs. FM

(c) GSS-EGA vs. GSS

(d) IG-EGA vs. IG

(e) OR-EGA vs. OR

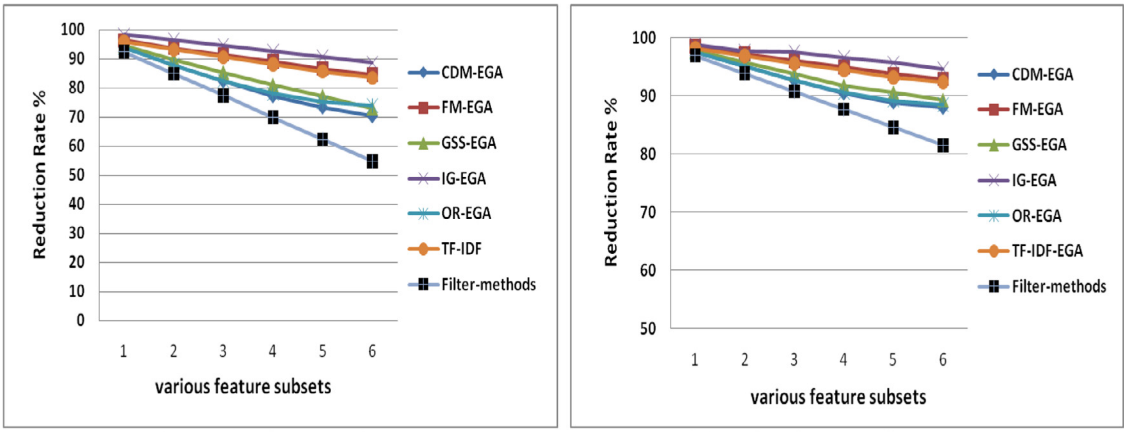
(f) TF-IDF-EGA vs. TF-IDF

Fig. 3. Dimensionality reduction rate with hybrid FS approaches versus single filter methods for D1.

5.3.1. Analysis of dimensionality reduction

For all datasets tested, the dimension of the feature space after text preprocessing is still large, so extra computational cost and space is needed. The size of the space may also affect categorization performance due to the existence of irrelevant features and the random selection of the initial population by EGA. The feature dimension sizes after text preprocessing are 13,084, 13,250 and 32,353 for D1, D2 and D3, respectively. Six different feature subsets (1000, 2000, 3000, 4000, 5000, and 6000) are selected by each filter method from each dataset.

Fig. 3(a)–(f) shows the overall efficiency of the hybrid approaches in reducing the dimensionality of the D1 datasets. It depicts comparisons of the reduction rates achieved by the original filter FS methods alone and those achieved by the hybrid FS approaches for a selection of feature subsets based on the overall number of original features. As shown in Fig. 3(a)–(f), the best reduction rates across different subsets are achieved with the proposed hybrid FS approaches. The reduction rates achieved by most of the hybrid approaches in the case of the smaller feature subsets differ only slightly from those achieved by the original single filter methods because those subsets are already reduced in the first stage by the filter method, so it is difficult for EGA to achieve a high reduction rate with small subsets because they have a negative impact on categorization performance, which is the target objective of the hybrid approaches and which has higher importance (80%) than feature size in the combined fitness function. In contrast, for the larger feature subsets, the reduction rates achieved by the hybrid approaches are significantly better than those gained by the single filter methods alone. This is because those subsets include some of the less important features for categorization and they are reduced by EGA through its search for the best feature subsets based on the features' importance and the reproduction operations, i.e. crossover and mutation. The IG-EGA approach (Fig. 3(d)) is superior to all the other methods, where the reduction rate is notable with all feature subsets. The next best performing hybrid approaches are FM-EGA, TF-IDF-EGA and GSS-EGA, shown in Fig. 3(b), (f) and (c), respectively. Overall, the results indicate that the proposed hybrid FS approaches can achieve a significant improvement in terms of dimensionality reduction rate.



(a) Reduction rate for D2

(b) Reduction rate for D3

Fig. 4. Reduction rate with hybrid FS approaches and filter method for D2 and D3 datasets.

In addition, Fig. 4(a) and (b) shows a summary of the reduction rates achieved by the hybrid FS approaches and the single filter methods based on the ratio of selected features to original features for the D2 and D3 datasets. We can see that the hybrid approaches also reduced the feature dimensions to a greater extent than the single filter methods for these datasets also. This indicates that reduction by filter methods should be revised to achieve the objective function of the hybrid approaches that is intended to reduce the feature dimension and improve categorization performance. The reduction rates achieved by both the hybrid and original methods are much closer in the case of the smallest feature subsets, especially for D3; however, the improvement in the performance of the hybrid approaches increases as the feature size increases and a significant reduction is achieved with the largest feature size. In the case of D3, the reduction rate is high because the feature subsets are much reduced in the filter stage, thus having a reduced subset contributes to improved categorization.

The results demonstrate that a high reduction rate can be achieved with the hybrid FS approaches, so they are advantageous in terms of saving computational resources, which is desired to simplify the machine learning process. The feature dimensions are efficiently reduced by the proposed approaches for all different subsets; however, with these approaches the feature size has less importance (20%) than categorization performance, therefore the results are analyzed with regard to categorization performance, as discussed in the following section.

5.3.2. *Analysis of categorization performance*

In this section, the effect of the proposed hybrid FS approaches on categorization performance is investigated in respect of D1 (Al-Jazeera text datasets). The results of each hybrid approach are compared with those of the filter method without hybridization based on the performance of the NB and AC categorization techniques.

5.3.2.1. Impact of hybrid approaches on NB performance. Table 6(a)–(f) presents the performance of NB with each hybrid approach. Accordingly, dimension reduction is achieved by applying EGA to different sizes of feature subsets (1000–6000) based on their importance, which is calculated by each filter method in the first stage. In addition, the results are compared with the original single filter FS methods in isolation, as shown in Fig. 5(a)–(f).

The results in Table 6 show that the hybrid CDM-EGA approach achieves the best performance with NB with a small feature subset (i.e. 1000), which was reduced by EGA to 922 features. Generally, the performance decreases with this approach when the number of features increases. It is obvious from the results for the first three subsets (1000, 2000, and 3000) that dimensionality is reduced and that this has a positive effect on categorization performance. As shown in Fig. 5(a), the highest performance with the least number of features is achieved by the hybrid CDM-EGA approach compared to CDM in isolation. The hybrid TF-IDF-EGA approach achieves the best performance with 1590 features, which is a reduced subset selected from among 3000 features. This approach is more stable and its result is more consistent with NB. On average, compared to CDM-EGA, TF-IDF-EGA is the best in terms of dimensionality reduction rate and performance of the NB classifier. In addition, a comparison of TF-IDF in isolation and EGA-TF-IDF (Fig. 5(f)) shows that EGA-TF-IDF outperforms TF-IDF alone.

The results of the hybrid OR-EGA approach are lower than those of TF-IDF-EGA in all cases, but the OR-EGA performs better than CDM-EGA on large feature subsets. In terms of dimensionality reduction rate with OR-EGA, it is smaller than TF-IDF-EGA and best or comparable to CDM-EGA. Also, the OR-EGA approach works better than OR alone (Fig. 5(f)) for most feature subsets. In addition, it is clear from the results that the hybrid FM-EGA approach achieves better and comparable results than the other five approaches. The comparison of FM-EGA and single FM is plotted in Fig. 5(b); it is obvious from the results that the FM-EGA approach obtains better performance than FM for most feature subsets.

The hybrid IG-EGA approach has a negative effect on NB performance with small feature subsets; its results with those subsets are the worst among all the approaches. However, this approach has the best reduction rate with all feature subsets and it achieves comparable performance with NB (89.49%) with a significant improvement in dimensionality reduction when the feature subset selected by IG is large (i.e. 5000 and 6000); the IG-EGA reduces the feature dimension to 1715 ($RR = 65.7\%$) and 2165 ($RR = 63.92\%$) features, respectively, which is not achieved by any other approaches for these feature subset sizes. The performance of this approach increases in most cases as the number of features increases. Therefore, it can be suggested that this approach can work well with high dimensional text with a large number of features because it can achieve satisfactory categorization performance with the advantage of a high reduction rate. The comparison of the results of IG and the hybrid IG-EGA approach in Fig. 5(d) show that IG-EGA has a comparable result with IG but with better reduction of feature dimensions; for example, its performance with NB with 1715 features is as good as that of IG with 5000 features.

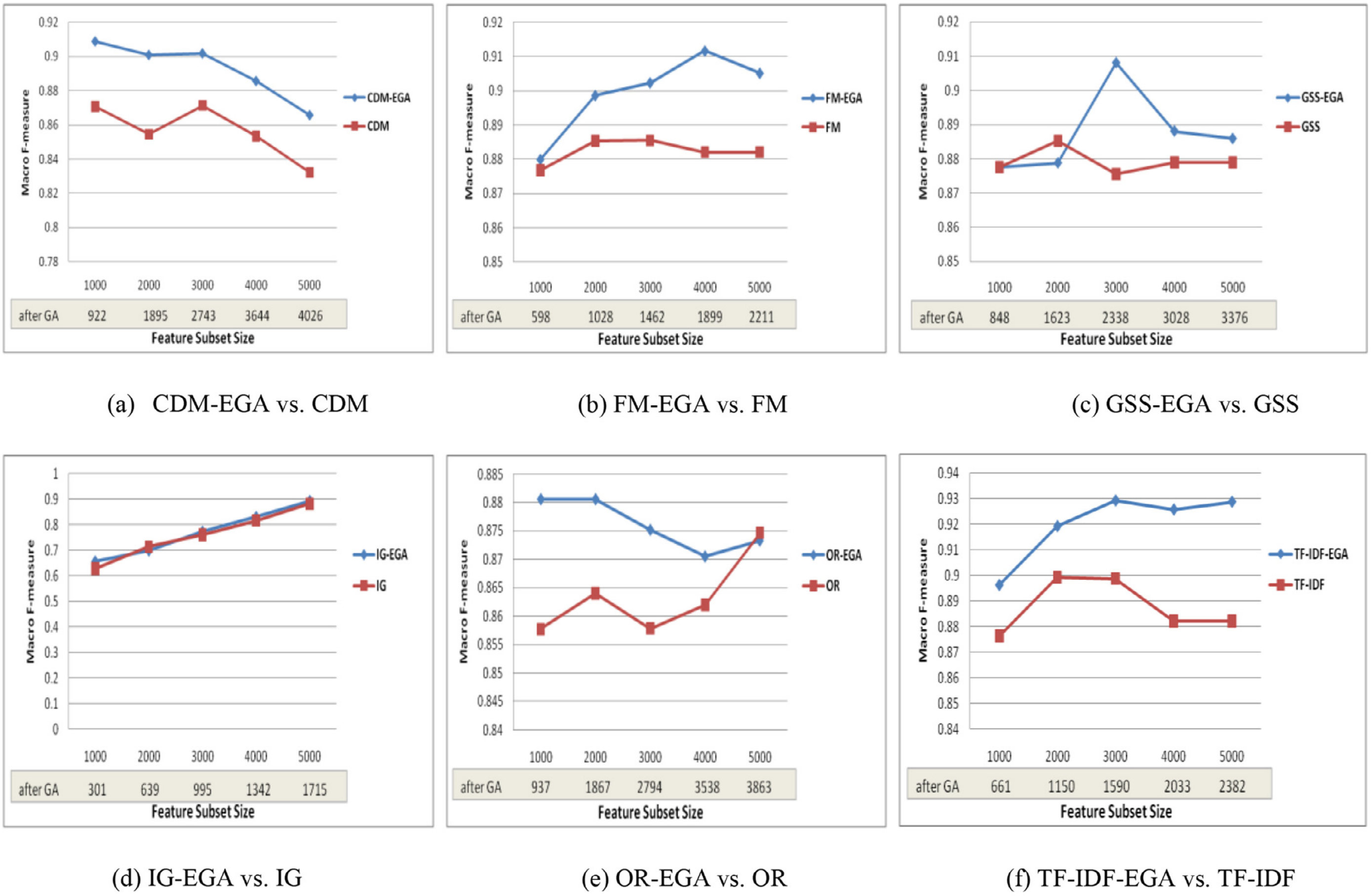
In addition, it can be seen from the results that the hybrid GSS-EGA approach produces comparable results compared to other hybrid approaches and it achieves the highest performance when the number of features is 2338. In addition, the comparison of the GSS-EGA

Table 6
Performance of NB with hybrid FS approaches (macro-averages).

Subset size	# features after EGA	**CDM-EGA**		
		Precision	Recall	F-measure
(a) Result with CDM-EGA				
1000	922	0.9144	0.9033	**0.9088**
2000	1895	0.9019	0.90	0.9009
3000	2743	0.9037	0.90	0.9018
4000	3644	0.8915	0.8799	0.8857
5000	4026	0.8752	0.8566	0.8658
6000	4107	0.8657	0.8533	0.8595
Subset size	# features after EGA	**FM-EGA-**		
		Precision	Recall	F-measure
(b) Result with FM-EGA				
1000	598	0.8834	0.8766	0.880
2000	1028	0.9008	0.8966	0.8987
3000	1462	0.9049	0.9	0.9024
4000	1899	0.9133	0.91	0.9117
5000	2211	0.9071	0.9033	0.9052
6000	2822	0.9043	0.9	0.9021
Subset size	# features after EGA	**GSS-EGA**		
		Precision	Recall	F-measure
(c) Result with GSS-EGA				
1000	848	0.8788	0.8766	0.8777
2000	1623	0.881	0.8766	0.8788
3000	2338	0.9095	0.9066	0.9081
4000	3028	0.8896	0.8866	0.8881
5000	3376	0.8886	0.8833	0.8859
6000	3760	0.8762	0.8733	0.8748
Subset size	# features after EGA	**IG-EGA**		
		Precision	Recall	F-measure
(d) Result with IG-EGA				
1000	301	0.6589	0.6566	0.6578
2000	639	0.7046	0.6966	0.7006
3000	995	0.7748	0.7733	0.7741
4000	1342	0.8342	0.83	0.8321
5000	1715	0.8966	0.8933	0.8949
6000	2165	0.8966	0.8933	0.8949
Subset size	# features after EGA	**OR-EGA**		
		Precision	Recall	F-measure
(e) Result with OR-EGA				
1000	937	0.8846	0.8766	**0.8806**
2000	1867	0.8847	0.8766	0.8806
3000	2794	0.8805	0.87	0.8752
4000	3538	0.8744	0.8666	0.8705
5000	3863	0.876	0.87	0.8733
6000	3943	0.8741	0.87	0.8720
Subset size	# features after EGA	**TF-IDF-EGA**		
		Precision	Recall	F-measure
(f) Result with TF-IDF-EGA				
1000	661	0.8994	0.8933	0.8963
2000	1150	0.9218	0.9166	0.9192
3000	1590	0.9318	0.9266	**0.9292**
4000	2033	0.9281	0.9233	0.9257
5000	2382	0.9307	0.9266	0.9287
6000	2895	0.9147	0.91	0.9123

results with those of GSS (Fig. 5(c)) shows that the hybrid approach is the best in most cases. It can therefore be concluded that, with regard to the dimensionality problem, the proposed hybrid FS approaches are effective in reducing feature dimensions and at the same time they improve categorization performance in most cases. This indicates that the whole features space is not needed to discriminate text categories and that the proposed approaches are good enough to select the optimal or near-optimal feature subsets that contribute to text categorization.

5.3.2.2. Impact of hybrid approaches on AC performance. The experimental results of applying the proposed hybrid FS approaches with AC are presented in Table 7. The aim of this part of the study is to examine the effect of these approaches on AC performance. Fig. 6(a)–(f) shows graphically the comparisons between the hybrid FS approaches and the respective original single filter methods for different feature dimensions. As shown in Table 7, the performance of OR-EGA with AC is more stable than that of the other approaches, but it does not achieve a better performance than CDM-EGA and



(a) CDM-EGA vs. CDM

(b) FM-EGA vs. FM

(c) GSS-EGA vs. GSS

(d) IG-EGA vs. IG

(e) OR-EGA vs. OR

(f) TF-IDF-EGA vs. TF-IDF

Fig. 5. Comparison of hybrid FS methods and original filter methods with NB.

TF-IDF-EGA with some feature subsets. The reason for the stability of OR-EGA is that even though the rules that are discovered with this approach are slightly different from each other for different feature subsets, the features that pass the minimum support requirement and compose the rules are approximately the same features for each feature subset. As shown in Fig. 6(e), the OR-EGA approach also outperforms OR in isolation with most feature subsets.

The hybrid CDM-EGA approach achieves the best results with a small feature subset (i.e. 922), but its efficiency decreases when the number of features increases. As shown in Fig. 6(a), the highest performance of AC (90.2%) with the least number of features (922) is achieved by CDM-EGA; however, CDM alone is more efficient with most other subsets in terms of the macro-average F-measure. As for TF-IDF-EGA, this approach achieves the best performance with 2382 features which is reduced from 5000 features, but it decreases for larger feature subsets. In addition, TF-IDF-EGA outperforms TF-IDF in isolation (Fig. 6(f)). On average, the experimental results show that in most cases, most hybrid approaches perform better than the respective single filter methods. The results also demonstrate the efficiency of the proposed approaches in reducing dimensionality, which is the main problem of text categorization. In the proposed approaches, the features dimensions are reduced in multiple stages, which produce more relevant features that able to discriminate the different categories of Arabic texts in most cases. However, in some cases, such as with CDM-EGA, we have the advantage of dimensionality reduction with only a small effect on categorization precision. Overall, TF-IDF-EGA, OR-EGA and CDM-EGA are the best in terms of AC performance.

The hybrid FM-EGA approach does not work well with AC and its results are unsatisfactory. The reason is that with this method a large set of categorization rules (more than 3000 rules) are used for categorization, which adversely affects the decisions made about the correct categories of texts. However, this method is among the best performing method with NB, which highlights that each FS approach has inconsistent results with different categorization techniques. However, FM-EGA obtained higher performance than FM with most feature subsets. The performance of AC with the hybrid IG-EGA approach increases in most cases as the number of features increases. The comparison of the results for IG and IG-EGA (Fig. 6(d)) shows that IG-EGA obtains better results, particularly with large feature subsets. The performance of IG-EGA with AC is better than that of FM-EGA, but it is poorer than that of the other hybrid approaches in some cases.

5.3.3. Analysis of processing time

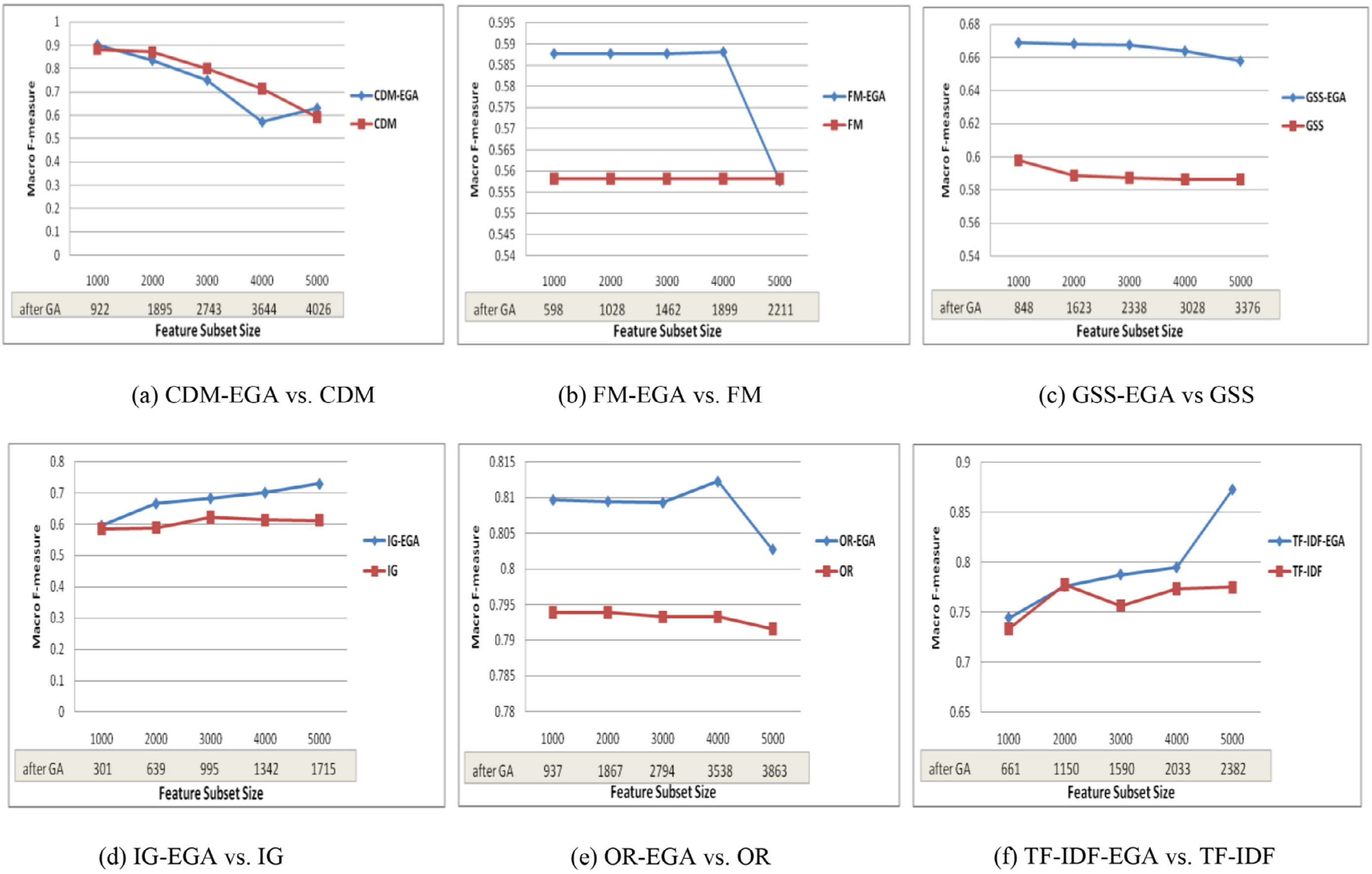
One of the objectives of the hybrid approaches is to minimize the time taken to perform the whole categorization process, where a subset from the original feature space is selected and revised by EGA and then utilized for categorization. Fig. 7 presents the time required to do this using each hybrid approach with AC and with NB when tested on D1. The results indicate that OR-EGA is the fastest among all the approaches with both categorization techniques (AC and NB), followed by CDM-EGA with NB. It is obvious from these results that NB has optimal time complexity compared to AC because with AC further operations (i.e. generation of categorization rules, ordering and pruning) needs to be performed before the final categorization, therefore AC requires more time than NB, particularly in the training phase. The long period of time required when using FM-EGA with AC is because a large set of frequent features has to be generated, which requires a lot of time for rule discovery, so this produces a large set of categorization rules and adversely affects the performance of AC as discussed earlier. However, if we compare the time required by the hybrid

Table 7
Performance of AC with hybrid FS approaches (macro-averages).

Subset size	# features after EGA	**CDM-EGA**		
		Precision	Recall	F-measure
(a) Result with CDM-EGA				
1000	922	0.9252	0.8813	0.902
2000	1895	0.9084	0.7722	0.8348
3000	2743	0.8627	0.6635	0.7501
4000	3644	0.8230	0.4372	0.5711
5000	4026	0.7915	0.5238	0.6304
6000	4107	0.7745	0.4290	0.5522
Subset size	# features after EGA	**FM-EGA**		
		Precision	Recall	F-measure
(b) Result with FM-EGA				
1000	598	0.6516	0.5356	0.5877
2000	1028	0.6516	0.5356	0.5877
3000	1462	0.6516	0.5356	0.5877
4000	1899	0.6513	0.5365	0.5881
5000	2211	0.6216	0.5056	0.5577
6000	2822	0.6216	0.5056	0.5576
Subset size	# features after EGA	**GSS-EGA**		
		Precision	Recall	F-measure
(c) Result with GSS-EGA				
1000	848	0.7433	0.610	0.6689
2000	1623	0.7420	0.6096	0.6682
3000	2338	0.7412	0.6092	0.6676
4000	3028	0.7381	0.6053	0.6639
5000	3376	0.737	0.597	0.658
6000	3760	0.7154	0.6200	0.6637
Subset size	# features after EGA	**IG-EGA**		
		Precision	Recall	F-measure
(d) Result with IG-EGA				
1000	301	0.6277	0.5683	0.5958
2000	639	0.7104	0.6294	0.6664
3000	995	0.7310	0.6442	0.6837
4000	1342	0.7451	0.6645	0.7016
5000	1715	0.7954	0.6763	0.7292
6000	2165	0.7516	0.6339	0.6864
Subset size	# features after EGA	**OR-EGA**		
		Precision	Recall	F-measure
(e) Result with OR-EGA				
1000	937	0.8480	0.7748	0.8097
2000	1867	0.847	0.7745	0.8095
3000	2794	0.8479	0.7740	0.8093
4000	3538	0.8616	0.7683	**0.8123**
5000	3863	0.8469	0.7628	0.8027
6000	3943	0.8571	0.7351	0.7914
Subset size	# features after EGA	**TF-IDF-EGA**		
		Precision	Recall	F-measure
(f) Result with TF-IDF-EGA				
1000	661	0.7485	0.7400	0.7442
2000	1150	0.7832	0.7692	0.7762
3000	1590	0.7937	0.7815	0.7875
4000	2033	0.8066	0.7829	0.7945
5000	2382	0.8686	0.8767	**0.8726**
6000	2895	0.8830	0.8512	0.8668

approaches with that required by EGA for all features, it could be argued that most of the hybrid approaches can speed up the categorization process because they require less time in most cases to select the important features for categorization.

From the experimental results on the performance of the proposed hybrid FS approaches with NB and with AC, we can see that these hybrid FS approaches can efficiently reduce text dimensionality; the dimension of each selected subset is further reduced using the EGA, which searches for the best feature subset. They also minimize the randomization effect of the GA because the EGA is applied to subsets of important features that are ordered based on their importance, which is computed by filter methods and also enhanced through crossover and mutation reproduction operations. Furthermore, categorization performance is enhanced in most cases with the hybrid approaches when compared to the respective single filter method, and the categorization process is faster with a subset of features than with the whole feature space.



(a) CDM-EGA vs. CDM

(b) FM-EGA vs. FM

(c) GSS-EGA vs GSS

(d) IG-EGA vs. IG

(e) OR-EGA vs. OR

(f) TF-IDF-EGA vs. TF-IDF

Fig. 6. Comparison of hybrid FS methods and single filter methods with AC.

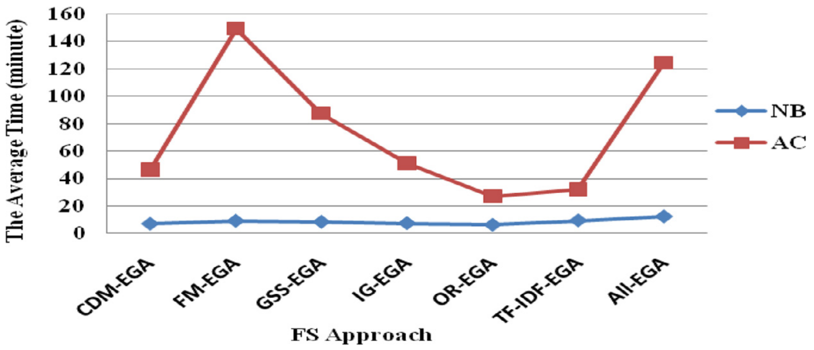
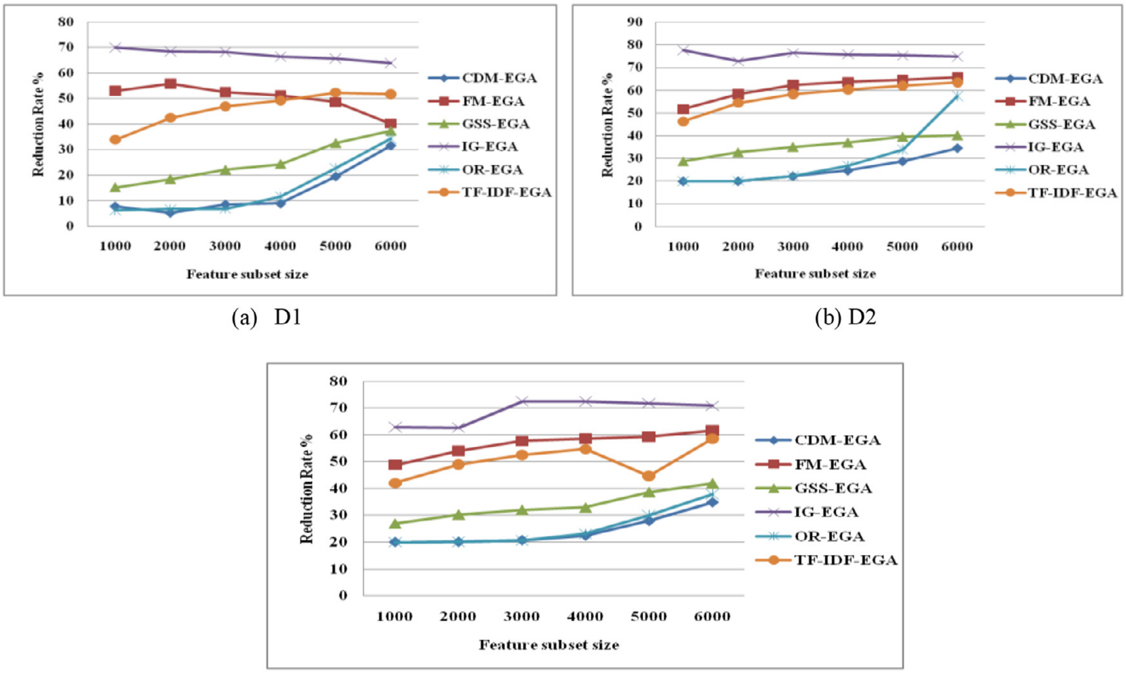


Fig. 7. Total processing time using hybrid FS approaches with NB and with AC.

To sum up, the results demonstrate that we can reduce text dimensionality and perform categorization based on a subset of features by using the hybrid approaches without degrading categorization performance in most cases with NB and in some cases with AC. It is hard to identify the most effective hybrid FS approach for all situations; however, based on the conducted experiments, TF-IDF-EGA, CDM-EGA and OR-EGA are among the best hybrid FS approaches for Arabic text categorization with AC, while TF-IDF-EGA, FM-EGA and CDM-EGA are the most effective with NB. In addition, the different hybrid FS approaches can select different feature subset sizes from the same feature dimension, which is evident from the number of features in each subset that are selected by each approach. In terms of categorization performance with most of the reduction approaches, NB seems to perform better than AC. In general, most FS approaches have different results with different feature dimensions and different categorization techniques. This highlights that the usage of a dimensionality reduction approach is affected by the characteristics of the dataset, feature dimensions and behavior of the categorization algorithms.

5.4. Comparison of the proposed hybrid approaches for different text datasets

From the above results (Section 5.3.2), it is clear that NB is superior to AC in most situations with all FS methods, therefore in this section, we focus on NB only and compare the efficiency of the proposed hybrid approaches based on the reduction rate achieved with each hybrid approach and its effect on NB performance in



(a) D1

(b) D2

(c) D3

Fig. 8. Reduction rate with hybrid FS approaches on each dataset with different feature subsets.

terms of macro-average F-measure for three text datasets (D1, D2, and D3).

5.4.1. *Dimensionality reduction comparisons*

Different numbers of features are used in this experiment; the feature dimensions selected by each filter method range from 1000 to 6000 features for each subset based on the highest ranked features. The dimension reduction of the selected feature subset is carried out using EGA. Fig. 8(a)–(c) shows the reduction rate for the three text datasets after applying EGA based on the selected feature subset. The results show that the hybrid approaches can add the advantage of dimensionality reduction. A different reduction rate is obtained by EGA with different filter methods. The reduction rate with small feature subsets (i.e. 1000 and 2000) is small with most approaches because these subsets are already reduced in the first stage. The hybrid IG-EGA approach maintains its superiority, achieving the best reduction rate. For most feature subsets, the highest reduction rates for D1, D2 and D3 are achieved by IG-EGA, followed by FM-EGA and TF-IDF-EGA with a slight difference in reduction rate. The reduction rate with GSS-EGA is not high like IG-EGA and FM-EGA and not low like CDM-EGA and OR-EGA. However, if we analyze the reduction results from another perspective, the lowest reduction rates are achieved by CDM-EGA and OR-EGA, which may be because the CDM and OR are designed to establish the candidate features in the first stage and the EGA is designed to reduce the selected dimensions of features and improve categorization performance at the same time. However, the reduction rates with CDM-EGA and OR-EGA are not high like other approaches; in spite of the reduction rate, this indicates that the features selected in the first stage are more associated with text categories and they are the best option to represent texts, which prevents EGA improving the reduction rate further because any more reduction will affect categorization performance with these methods. In general, the reduction rate increases as the feature size increases, and this indicates that the filter methods are good at selecting the candidate features and removing a wide range of noisy and irrelevant features in the first stage, and that EGA is a good approach to further reduce dimensionality and to explore the best solution in terms of feature subsets.

5.4.2. *Performance comparison*

The above results (Section 5.4.1) indicate that the use of the proposed hybrid FS approaches can result in a significant improvement in dimensionality reduction for all three of the text datasets examined. Nevertheless, this observation needs more investigation regarding categorization performance with reduction rate. Tables 8 and 9 summarize the reduction rate and categorization performance results in terms of macro-average F-measure. The reduction rates in bold denote the best reduction with regard to categorization performance (macro average F-measure %) with each hybrid approach. The annotation #F means the feature size after EGA, RR% is the reduction rate and F-m% is the macro-average F-measure.

It can be seen from the results (Tables 8 and 9) that in the case of D1 and D2 the best performance is achieved with TF-IDF-EGA when the number of features is 1590 and 1899, respectively, and the best result in the case of D3 is achieved by FM-EGA with 919 features, which is reduced from 2000 features. The results show that the proposed hybrid FS methods can reduce text dimensionality without affecting categorization performance in most situations with most methods. The best method in terms of reduction rate is IG-EGA, where the number of features does not exceed 2200 features; however, unfortunately, this method achieves the lowest performance, particularly with small feature subsets. Nevertheless, its efficiency improves as the number of features increases and it produces a comparable performance with the best reduction rate when the number of selected features in the first stage is 5000 and 6000. The reason for this is that this method in the first stage (IG) selects features in a global way, so the pattern of different text categories is not clear with small feature subsets, which means the correlation between the features and their category is not strong with small subsets. Overall, in terms of both reduction rate and categorization performance the best methods are TF-IDF-EGA and FM-EGA because they achieve both a high reduction rate and high categorization performance. The hybrid IG-EGA approach is the most effective with respect to dimensionality reduction, achieving the highest reduction rate with a comparable performance with large feature dimensions.

Table 8
Reduction rate (%) and categorization performance (macro-average F-measure %) of NB for D1 and D2.

Hybrid method	Measure	D1						D2					
		1000	2000	3000	4000	5000	6000	1000	2000	3000	4000	5000	6000
CDM-EGA	#F	922	1895	2743	3644	4026	4107	800	1600	2334	3013	3526	3925
	RR%	**7. 8**	5. 25	8. 56	8. 9	19.0	31.55	20.00	**20.00**	22.20	24.67	28.76	34.58
	F-m%	**90.88**	90.09	90.18	88.57	86.58	85.95	81.11	**83.07**	83.72	83.02	80.01	78.34
FM-EGA	#F	598	1028	1462	1899	2211	2822	483	832	1136	1455	1769	2051
	RR%	40.20	48.6	51.26	**52.52**	55.78	52.96	51.7	58.4	62.33	63.62	64.62	**65.82**
	F-m%	88.01	89.87	90.24	**91.17**	90.52	90.21	83.52	87.13	86.81	87.24	87.59	**87.60**
GSS-EGA	#F	848	1623	2338	3028	3376	3760	713	1347	1948	2521	3025	3589
	RR%	15.2	18.4	**22.06**	24.3	32.48	37.33	28.7	32.65	35.06	36.97	39.5	**40.18**
	F-m%	87.77	87.88	**90.81**	88.81	88.59	87.48	82.64	83.22	83.06	83.41	82.84	**83.44**
IG-EGA	#F	301	639	995	1342	1715	2165	223	454	707	968	1234	1510
	RR%	69.9	68.35	68.16	66.45	65.7	**63.92**	77.70	72.75	76.43	75.80	75.32	**74.83**
	F-m%	65.78	70.06	77.41	83.21	89.49	**89.49**	64.06	72.21	78.33	81.62	81.73	**82.26**
OR-EGA	#F	937	1867	2794	3538	3863	3943	800	1598	2331	2921	3302	3453
	RR%	6.3	**6.65**	6.86	11.55	22.74	34.28	20.00	**20.10**	22.30	26.79	33.96	57.55
	F-m%	88.05	**88.06**	87.52	87.05	87.33	87.20	85.56	**86.02**	85.32	84.31	84.575	80.96
TF-IDF-EGA	#F	661	1150	1590	2033	2382	2895	537	910	1251	1588	1899	2192
	RR%	33.9	42.5	**47.0**	49.175	52.36	51.75	46.3	54.5	58.3	60.30	**62.02**	63.46
	F-m%	89.63	91.92	**92.92**	92.57	92.87	91.23	84.04	86.05	87.49	87.54	**87.81**	87.54

Table 9
Reduction rate (%) and categorization performance (macro-average F-measure %) of NB for D3.

Hybrid method	Measure	D3					
		1000	**2000**	**3000**	**4000**	**5000**	**6000**
CDM-EGA	#F	800	1599	2378	3105	3606	3907
	RR%	20.00	20.05	**20.73**	22.37	27.88	34.88
	F-m%	85.25	85.57	**89.04**	86.27	81.71	83.50
FM-EGA	#F	512	919	1269	1658	2032	2305
	RR%	48.8	**54.05**	57.7	58.55	59.36	61.58
	F-m%	89.11	**92.09**	91.22	90.96	88.82	88.79
GSS-EGA	#F	730	1393	2037	2683	3066	3483
	RR%	27.00	**30.35**	32.10	32.92	38.68	41.95
	F-m%	88.41	**89.39**	88.73	87.69	87.92	86.62
IG-EGA	#F	372	746	823	1104	1410	1749
	RR%	62.8	62.7	72.56	72.4	71.8	**70.85**
	F-m%	67.11	75.42	79.37	81.83	85.49	**86.38**
OR-EGA	# F	800	1597	2383	3077	3503	3735
	RR%	20.00	**20.15**	20.56	23.07	29.94	37.75
	F-m%	91.13	**91.42**	89.53	89.66	88.68	85.90
TF-IDF-EGA	#F	580	1022	1425	1814	2230	2490
	RR%	42.00	**48.90**	52.5	54.65	44.6	58.50
	F-m%	88.03	**91.10**	90.66	90.96	89.45	88.82

The objective of the hybrid FS approach is to minimize the feature size (reduce dimensionality) and maximize categorization accuracy, and, as mentioned earlier, in our experiments we gave performance higher importance than feature size. With respect to this objective, we can see that the best performing methods are TF-IDF-EGA, FM-EGA and MCDM-EGA. The main reasons for the superiority of TF-IDF-EGA is that TF-IDF is used twice; in the first stage to select the candidate features and secondly to guide EGA when the crossover operation is performed. Furthermore, this method ranks features based on their importance inside categories and their importance in the whole collection, so that the features selected by this method are revised more, which produces a notable performance in some cases. The good result achieved by FM-EGA is due to the nature and usage of this approach where the FM combines the precision and recall of features inside the training dataset to select the most important feature in the first stage and then the EGA searches for the best feature size along with best performance in terms of the macro-average F-measure. The hybrid CDM-EGA approach works well with small feature subsets, but its efficiency declines when the feature size is large. This method, in the first stage, depends on the feature importance inside the category, which is based on document frequency; however, when a large subset is selected, the features may be correlated with other categories, which confuses the EGA in its search for the best subsets and affects the categorization decision. The other hybrid approaches also show promising results with regard to reduction rate and categorization performance.

5.4.3. *Time comparisons*

Fig. 9 presents the time averages of the hybrid FS approaches across different feature subsets for D1, D2, and D3. The best time for D1 was achieved with OR-EGA, while for D2 and D3 the best time was recorded with CDM-EGA and IG-EGA, respectively. The efficiency of the hybrid approaches in terms of time were especially notable for D3 compared to EGA alone because much-reduced feature subsets are used for feature generation and categorization compared to the original features dimension. On average, the hybrid approaches can reduce time complexity and high dimensionality, which is highly desired for text categorization, particularly when a few computational resources are available.

Generally, the results demonstrate the potential enhancement of feature selection and categorization that can be achieved with the proposed hybrid FS approaches. The effectiveness of using hybrid FS based on EGA for dimensionality reduction has been proved for Arabic text categorization. Indeed, it is unnecessary to use the whole feature space for categorization due high computational cost. As the experimental results indicate, we can achieve a higher reduction rate

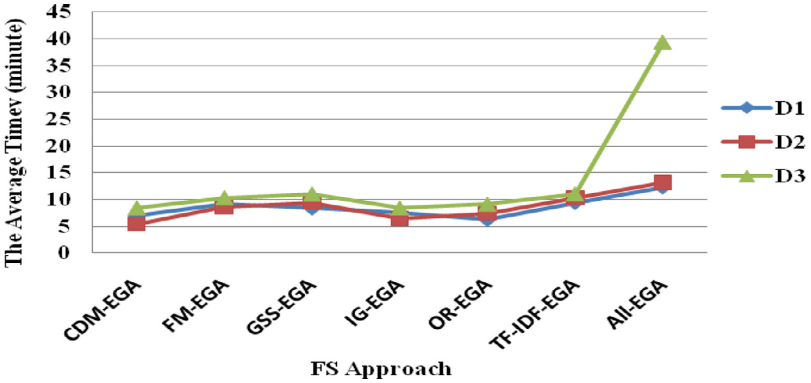


Fig. 9. Efficiency of hybrid FS approaches in terms of average time for each dataset.

with hybrid approaches and still maintain satisfactory categorization performance in most cases.

Overall, the proposed methods in this research have advantages for preserving the important features in Arabic text datasets and for simplifying and improving text categorization. The research presented some contributions in the area of expert and intelligent systems; first, an improved version of GA is presented which named EGA. Secondly, six hybrid FS approaches based on the EGA were introduced to handle the high dimensionality of the feature space and enhance text categorization. From the experimental results, it can be seen that the proposed EGA and hybrid FS approaches based on the EGA are efficient at reducing the complexity of text categorization with a high dimensional feature space. The results demonstrate that text dimensionality can be reduced and categorization can be based on a subset of features by using the hybrid approaches without degrading categorization performance in most cases. In sum, promising results were achieved with the proposed hybrid approaches compared to the original filter methods and EGA individually. The proposed approaches gave better performance; they significantly reduce the high dimensionality of feature space with less time across the selected feature subsets and achieve higher categorization performance. The high dimensionality of feature space is a major problem in text categorization because it degrades computational resources and categorization performance. The proposed hybrid FS approaches treat this problem by combining filter methods with an enhanced version of GA named the EGA. This approach is desired to preserve the useful knowledge in text datasets and simplify the categorization process.

On average, the hybrid approaches can reduce time complexity and high dimensionality, which is highly desired for text categorization particularly when limited computational resources are available. In the hybrid approach, the feature space is reduced in the first stage based on the filter method. It also minimizes the randomization effect of the GA where the EGA is applied to subsets with the most important features that are ordered based on their importance, which is computed by the filter method. Furthermore, in the EGA, the crossover and mutation operations are performed based on further analysis of feature importance besides the original feature subset's (parent subsets) fitness. The dimension of the selected feature space is reduced in the second stage by using the EGA, which searches for the best feature subset by employing a multi-objectives function that is designed to reduce feature dimensionality and improve categorization performance simultaneously. Hence, a reduced subset that contains the best features is generated with the hybrid approach through direct interaction with the categorization algorithm and this contributes to final categorization. Thus, categorization performance is enhanced in most situations with the various hybrid approaches compared to each individual filter method and the EGA for the full feature space.

6. Conclusion

This paper presented enhanced version of GA named EGA in which the GA operators (crossover and mutation) were modified to reduce the adverse effect of randomization and to guide search for the best feature subsets and create population diversity with the useful knowledge. The EGA superior the GA in terms of dimensionality reduction, time and categorization performance. In addition, some hybrid FS methods based on the EGA was also proposed in this paper. The hybrid FS methods were introduced to reduce the high dimensionality of text in an efficient way and produce revised feature subsets that would have the ability to produce accurate categorization. At this stage, the effectiveness of the hybrid FS based on EGA was explored, six hybrid FS approaches were proposed that incorporated six filtering methods with the EGA. The results of the hybrid FS approaches showed that the hybrid FS approaches were more effective in reducing dimensionality and they could produce a higher reduction rate and higher categorization precision in most situations compared to single filter method and GA individually.

As the experimental results indicate, the potential of the proposed EGA was proven individually and when it was utilized in the second stage of the hybrid FS approaches. However, the usage of the EGA is limited in the construction stage in terms of its effect on improving FS and simplifying the categorization process that is performed by another categorization algorithm. Thus, the EGA as categorization algorithm can be used to create a rule-based text classifier that combines several advantages instead of using it as preprocessing tool for another algorithm. In addition, the enhancement of GA was applied to two operators of GA (crossover and mutation); however the GA could be enhanced by working on other operations such as fitness function and selection schemes.

In the future; one possible solution that could be applied to enhance the GA is to adopt dynamic probability for cross validation and mutation besides the mentioned modification. Moreover, the fitness function is a major step in the GA that needs to be investigated and enhanced in the future; it is therefore suggested that a combined fitness function be used that is composed from many factors that are related to text characteristics and categorization performance to identify the fittest of the candidate feature subsets. Another suggestion for future work is to investigate the performance of the proposed FS approaches with other categorization techniques such as the Support Vector Machine (SVM), which has generated promising results in some researches for text categorization. Another suggestion for future is to improve EGA to generate a complete set of useful rules for all text portions for text categorization based only on rules. Moreover, the performance of this approach could be investigated in terms of application to healthcare datasets in which the features are more dependent and correlated.

Acknowledgment

This research project was supported in part by the Universiti Kebangsaan Malaysia (UKM) under Grant (DIP-2012-15).

References

Abbas, M., Smaili, K., & Berkani, D. (2011). Evaluation of topic identification methods on Arabic corpora. *Journal of Digital Information Management, 9*(5), 185–192.

Abu Tair, M. M., & Baraka, R. S. (2013). Design and evaluation of a parallel classifier for large-scale Arabic text. *International Journal of Computer Applications, 75*(3), 13–20.

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert systems with applications, 36*(3), 6843–6853.

Al-Radaideh, Q. A., Al-Shawakfa, E., Ghareb, A. S., & Abu-Salem, H. (2011). An approach for Arabic text categorization using association rule mining. *International Journal of Computer Processing Of Languages, 23*(01), 81–106.

Antonie, M. L., & Zaiane, O. R. (2002). Text document categorization by term association. In Proceedings of IEEE international conference on data mining, *(ICDM'02)* (pp. 19–26).

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications, 39*, 4760–4768.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence, 97*(1), 245–271.

Chantar, H. K., & Corne, D. W. (2011). Feature subset selection for Arabic document categorization using BPSO-KNN. In *Proceedings of third world congress on nature and biologically inspired computing (NaBIC)* (pp. 546–551).

Chen, H., & Zou, B. (2009). Optimal feature selection algorithm based on quantum-inspired clone genetic strategy in text categorization. In *Proceedings of the first ACM/SIGEVO summit on genetic and evolutionary computation* (pp. 799–802).

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36*(3), 5432–5435.

Chiang, D. A., Keh, H.-C., Huang, H.-H., & Chyr, D. (2008). The Chinese text categorization system with association rule and category priority. *Expert Systems with Applications, 35*(1), 102–110.

Fang, Y., Chen, K., & Luo, C. (2012). The algorithm research of genetic algorithm combining with text feature selection method. *Journal of Computational Science and Engineering, 1*(1), 9–13.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira, W. (2011). Word co-occurrence features for text classification. *Information Systems, 36*(5), 843–858.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research, 3*, 1289–1305.

Galavotti, L., Sebastiani, F. &, & Simi, M. (2000). Experiments on the use of feature selection and negative evidence in automated text categorization. *Research and advanced technology for digital libraries* (pp. 59–68). Springer.

Ghareb, A. S., Hamdan, A. R., & Bakar, A. A. (2012). Text associative classification approach for mining Arabic data set. In *Proceedings of the 4th international conference on data mining and optimization (DMO)* (pp. 114–120). IEEE.

Gunal, S. (2012). Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering & Computer Sciences, 20*(2), 1296–1311.

Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009). Improving Arabic text categorization using decision trees. In *Proceedings of the first international conference on networked digital technologies, NDT'09* (pp. 110–115). IEEE.

Hattab, A. M., & Hussein, A. K. (2012). Arabic content classification system using statistical Bayes classifier with words detection and correction. *World of Computer Science & Information Technology Journal, 2*(6), 193–196.

Holland, J. H. (1975). *Adaption in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

Janaki Meena, M., Chandran, K., Karthik, A., & Vijay Samuel, A. (2012). An enhanced ACO algorithm to select features for text categorization and its parallelization. *Expert Systems with Applications, 39*(5), 5861–5871.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K nearest-neighbor algorithm for text categorization. *Expert Systems with Applications, 39*(1), 1503–1509.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European conference on machine learning (ECML)* (pp. 137–142).

Kabir, M., Shahjahan, M., & Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. *Expert Systems with Applications, 39*, 3747–3763.

Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language resources and evaluation, 47*(2), 513–538.

Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: a survey. *International Journal of Artificial Intelligence & Applications, 3*(2), 85–99.

Langley, P. (1994). Selection of relevant features in machine learning. In *Proceedings of the AAAI fall symposium on relevance* (pp. 1–5).

Lei, S. (2012). A feature selection method based on information gain and genetic algorithm. In *Proceedings of international conference on computer science and electronics engineering (ICCSEE)* (pp. 355–358). IEEE.

Lewis, D. D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of third annual symposium on document analysis and information retrieval: 33* (pp. 81–93). ISRI; University of Nevada.

Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Boston: Kluwer Academic.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. *MIT Press*.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization, 752*, 41–48.

Mengle, S. S., & Goharian, N. (2009). Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology, 60*(5), 1037–1050.

Mesleh, A. (2011). Feature sub-set selection metrics for Arabic text classification. *Pattern Recognition Letters, 32*(14), 1922–1929.

Mesleh, A., & Kanaan, G. (2008). Support vector machine text classification system: using ant colony optimization based feature subset selection. In *Proceeding of international conference on computer engineering & systems, ICCES* (pp. 143–148).

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. In *Proceeding of the 16th international conference on machine learning (ICML)* (pp. 258–267).

Ogura, H., Amano, H., & Kondo, M. (2009). Feature selection with a measure of deviations from Poisson in text categorization. *Expert Systems with Applications, 36*(3), 6826–6832.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection. *Soft Computing, 12*(2), 111–120.

Tsai, C-F., Chen, Z-Y., & Ke, S-W. (2014). Evolutionary instance selection for text classification. *Journal of Systems and Software, 90*, 104–113.

Thabtah, F. (2007). A review of associative classification mining. *The Knowledge Engineering Review, 22*(01), 37–65.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024–1032.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems, 36*, 226–235.

Uysal, A. K., & Gunal, S. (2014). Text classification using genetic algorithm oriented latent semantic features. *Expert Systems with Applications, 41*(13), 5938–5947.

Yang, J., Liu, Y., Liu, Z., Zhu, X., & Zhang, X. (2011). A new feature selection algorithm based on binomial hypothesis testing for spam filtering. *Knowledge-Based Systems, 24*(6), 904–914.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management, 48*(4), 741–754.

Yang, Yiming, & Pedersen, Jan O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning ICML* (pp. 412–420).

Yin, X., & Han, J. (2003). CPAR: classification based on predictive association rules. In *Proceedings of the SIAM international conference on data mining* (pp. 369–376). SIAM Press.

Youn, E., & Jeong, K. (2009). Class dependent feature scaling method using naive Bayes classifier for text datamining. *Pattern Recognition Letters, 30*(5), 477–485.

Yun, J., Jing, L., Yu, J., & Huang, H. (2012). A multi-layer text classification framework based on two-level representation model. *Expert Systems with Applications, 39*(2), 2035–2046.

Zahran, B. M., & Kanaan, G. (2009). Text feature selection using particle swarm optimization algorithm. *World Applied Sciences Journal (Special Issue of Computer & IT), 7*, 69–74.