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/

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