An Enhanced Recursive Firefly Algorithm for Informative Gene Selection

Nassima Dif, EEDIS Laboratory, Djillali Liabes University, Sidi Belabbes, Algeria
Zakaria Elberrichi, EEDIS Laboratory, Djillali Liabes University, Sidi Belabbes, Algeria

ABSTRACT

Feature selection is the process of identifying good performing combinations of significant features among many possibilities. This preprocess improves the classification accuracy and facilitates the learning task. For this optimization problem, the authors have used a metaheuristics approach. Their main objective is to propose an enhanced version of the firefly algorithm as a wrapper approach by adding a recursive behavior to improve the search of the optimal solution. They applied SVM classifier to investigate the proposed method. For the authors experimentations, they have used the benchmark microarray datasets. The results show that the new enhanced recursive FA (RFA) outperforms the standard version with a reduction of dimensionality for all the datasets. As an example, for the leukemia microarray dataset, they have a perfect performance score of 100% with only 18 informative selected genes among the 7,129 of the original dataset. The RFA was competitive compared to other state-of-art approaches and achieved the best results for CNS, Ovarian cancer, MLL, prostate, Leukemia_4c, and lymphoma datasets.

KEYWORDS

Classification, Data Mining, Data Preprocessing, Data Reduction, Feature Selection, Firefly Algorithm, Metaheuristic, Microarray Datasets, Recursion, SVM Algorithm

1. INTRODUCTION

Performance improvement is the challenge in supervised classification. The main objective is to generate a good pattern for the classification task. This can be done, either by improving the learning algorithm, or the quality of the data. Different types of preprocessing, like replacement of noisy and missing values, discretization or feature selection, can improve the quality of data.

Microarrays have been a source of data for a wide range of biomedical investigations. They are useful to distinguish or to diagnostic different types of disease (Saeyes, Inza, & Larrañaga, 2007). A simple classification task consists to separate healthy patients from cancer patients (Bolon-Canedo, Sanchez-Marono, Alonso-Betanzos, Benitez, & Herrera, 2014). These datasets are characterized by a large number of genes associated to a low number of samples. This difference can causes over fitting problems to the classifier, and require a high computational run time. In addition, this type of datasets

DOI: 10.4018/IJSIR.2019040102

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
is noisy and complex (Alshamlan, Badr & Alohali, 2015) which disrupts the classification task and reduce the performance of the classifier. The selection of the most relevant genes and performance improvement in microarray dataset is a challenging task because an important number of genes are irrelevant (Dashtban & Balafar, 2017).

Feature (or gene) selection aims to select a subset of m pertinent attributes were m < N, N is the number of features in the original set. By excluding, the irrelevant, noisy and redundant features (Bolon-Canedo, Sanchez-Maro, & Alonso-Betanzos, 2015), it reduces the dimensionality of the learning dataset, which makes the classification model more appropriate, and reduce the learning time complexity.

Several approaches have been proposed in the literature, to solve the feature selection problem. We distinguish three types of methods: wrapper, filter and hybrid or embedded methods. The filter methods select the most important feature using statistical measures to calculate the pertinence of the selected genes subset. They are characterized by their low computational time (Apolloni, Leguizamón, & Alba, 2016), because they don’t use the classifier. In the other hand, the wrapper approaches connect the learning algorithm with some optimization algorithms such as metaheuristics to select the best subsets (Lv, Peng, Chen, & Sun, 2016). Because they use the performance of the classifier as a value for the fitness function, these methods are better performing compared to the filter ones. The third methods called embedded, combine between the characteristics of the wrapper and the filter methods. The process of selecting the best subsets is done in parallel with the learning process, such as the case of tree algorithms; they are specific to a particular learning algorithm (Zhang, & Deng, 2007). Table1 resume some approaches used in gene selection.

The task of finding the relevant subsets is an NP-hard problem due to the large number of subsets to examine. Therefore, it requires the use of optimization methods like metaheuristics.

Several metaheuristics have been proposed in the literature for the resolution of NP-hard problems. Some of them are old, like genetic algorithm (GA), which was the first proposed evolutionary algorithm (Emary, Ghany, Hassanien, & Pary, 2015). It mimics the natural evolution. Particle swarm optimization (PSO) (Eberhart & Kennedy, 1995) is one of the most used metaheuristics. It is based on the swarm intelligence behavior (birds, fishes ...). Ant colony optimization (ACO) (Dorigo, Maniezzo, & Colorni, 1996) is inspired from the behavior of ants in the search for the shortest path between the colony and

Table 1. Filter’s, wrapper’s and embedded approaches proposed in literature

<table>
<thead>
<tr>
<th>Type</th>
<th>References</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>(Hall &amp; Smith, 1998)</td>
<td>Information gain (IG)</td>
</tr>
<tr>
<td></td>
<td>(Hall, 1999)</td>
<td>Correlation feature selection (CFS)</td>
</tr>
<tr>
<td></td>
<td>(Peng, Long, &amp; Ding, 2005)</td>
<td>minimum Redundancy Maximum Relevance (mRMR)</td>
</tr>
<tr>
<td>Wrapper</td>
<td>(Ruiz, Riquelme, &amp; Aguilar-Ruiz, 2006)</td>
<td>Incremental method (BIRS)</td>
</tr>
<tr>
<td></td>
<td>(Wanderley, Gardeux, Natowicz, &amp; Braga, 2013)</td>
<td>Evolutionary wrapper method (GA-KDE-Bayes)</td>
</tr>
<tr>
<td></td>
<td>(Li, Wang, Chen, Shi, &amp; Qin, 2013)</td>
<td>Ant colony optimization selection method (ACO-S)</td>
</tr>
<tr>
<td></td>
<td>(Prasad &amp; Biswas, 2015)</td>
<td>A progressively refined PSO</td>
</tr>
<tr>
<td></td>
<td>(Vafaei &amp; Mosafer, 2016)</td>
<td>CLACOFS.</td>
</tr>
<tr>
<td></td>
<td>(Anaissi, Kennedy, &amp; Goyal, 2011)</td>
<td>Embedded method based on the random forest algorithm</td>
</tr>
</tbody>
</table>
Related Content

A Differential Evolution Based Multiclass Vehicle Detector and Classifier for Urban Environments
[www.igi-global.com/article/a-differential-evolution-based-multiclass-vehicle-detector-and-classifier-for-urban-environments/181693?camid=4v1a](www.igi-global.com/article/a-differential-evolution-based-multiclass-vehicle-detector-and-classifier-for-urban-environments/181693?camid=4v1a)

The Need for Traffic Based Virtualisation Management for Sustainable Clouds
[www.igi-global.com/article/the-need-for-traffic-based-virtualisation-management-for-sustainable-clouds/137717?camid=4v1a](www.igi-global.com/article/the-need-for-traffic-based-virtualisation-management-for-sustainable-clouds/137717?camid=4v1a)
Constraint Handling in Particle Swarm Optimization
Wen Fung Leong and Gary G. Yen (2012). Innovations and Developments of Swarm Intelligence Applications (pp. 60-80).
www.igi-global.com/chapter/constraint-handling-particle-swarm-optimization/65806?camid=4v1a

iChance: A Web-Based Innovation Support System for Business Intelligence
www.igi-global.com/article/ichance-web-based-innovation-support/60754?camid=4v1a