Using Penguins Search Optimization Algorithm for Best Features Selection for Biomedical Data Classification

Noria Bidi, Djillali Liabès University, Sidi Bel Abbès, Algeria
Zakaria Elberrichi, Djillali Liabès University, Sidi Bel Abbès, Algeria

ABSTRACT

Feature selection is essential to improve the classification effectiveness. This paper presents a new adaptive algorithm called FS-PeSOA (feature selection penguins search optimization algorithm) which is a meta-heuristic feature selection method based on “Penguins Search Optimization Algorithm” (PeSOA), it will be combined with different classifiers to find the best subset features, which achieve the highest accuracy in classification. In order to explore the feature subset candidates, the bio-inspired approach PeSOA generates during the process a trial feature subset and estimates its fitness value by using three classifiers for each case: Naïve Bayes (NB), Nearest Neighbors (KNN) and Support Vector Machines (SVMs). Our proposed approach has been experimented on six well known benchmark datasets (Wisconsin Breast Cancer, Pima Diabetes, Mammographic Mass, Dermatology, Colon Tumor and Prostate Cancer data sets). Experimental results prove that the classification accuracy of FS-PeSOA is the highest and very powerful for different datasets.

KEYWORDS
Classification, Feature Selection, Meta-Heuristic, Penguins Search Optimization Algorithm, Support Vectors Machines, Wisconsin Diagnostic Breast Cancer

INTRODUCTION

In biomedical data mining, classification is one of the most important techniques. The massive volume of data makes feature selection an essential data pre-processing step to obtain the best classification models. Because there are several statements of the feature selection problem several methods were proposed, which make it difficult to choose the most suitable feature one for a given application. Some methods are specialized to particular cases and other methods can be used only for small numbers of features or only for large numbers of features. Many meta-heuristics strategies are among the used methods. Meta-heuristics have emerged in the 1980s to solve difficult optimization problems (Dreo, Petrowski, Siarry, & Taillard, 2003). These methods have been successfully proposed to improve the resolution of the most complex problems; they are generally designed initially for discrete problems, but can be adapted to other types of problems. There are several ways to classify meta-heuristics; memory usage or memory less methods, nature inspired or non-nature inspired, dynamic or static objective function, one or various neighborhood structures, and their operating principle during the search of the solution. We chose the latter classification where we can distinguish two categories:

DOI: 10.4018/IJOCI.2017100103
The trajectory-based approaches also called iterative methods in a single solution. They are all based on a neighborhood search algorithm, beginning with an initial solution and then getting better by choosing a new solution in its neighborhood (Bachelet, 1999). These algorithms start from an initial solution (obtained accurately, or by random selection) and gradually recede to realize a trajectory. In this category, we find: descent method, simulated annealing, tabu search method and a variable neighborhood. The population-based approaches use a set of solutions and in each iteration, the algorithms explore the solution space through a predefined set of operators. In this second category, we find: genetic algorithms, ant colony algorithms, etc. A novel population based meta-heuristic approach called penguins search optimization algorithm (PeSOA) which was developed by Youcef Gheraibia and Abdelouahab Moussaoui in 2013 (Gheraibia & Moussaoui, 2013). PeSOA is inspired by the penguin’s hunting behavior. Like all nature-inspired meta-heuristic algorithms, the two very important components are intensification and diversification. The diversification strategy of PeSOA allows penguins to explore more effectively the space of solutions, whereas the intensification strategy allows penguins to converge rapidly to an optimal solution. In the present work, a new adaptive algorithm called FS-PeSOA (feature selection-penguins search optimization algorithm) is used to obtain the optimal subset of features. We propose our algorithm in order to achieve either higher accuracy with smaller size of the feature subsets selected. The organization of the rest of this paper is as follows: Brief review of the related work in using machine learning techniques is discussed in section 2. Section 3 presents feature selection methods. The PeSOA and FS-PeSOA methods are presented in section 4 and section 5. The selected classifier SVM is presented in section 6. In section 7, our experimental results for the implementation of FS-PeSOA on different datasets are reported. Finally, we conclude this work.

RELATED WORK

Sensitive data such as patients’ records and body images such as tumor and surgery related information, should not be in public domains. All these data should only be within the hospital and not in any public clouds. Hence, the design and implementation of private clouds is essential for biomedical scientists to generate, process, update, archive and store their data. (Chang & Wills, 2016). Six benchmark datasets are used in this paper, where Wisconsin Breast Cancer, Pima Diabetes, Mammographic Mass, and Dermatology datasets were obtained from the UCI machine learning repository (UCI), the colon cancer and the prostate cancer datasets were taken from Kent Ridge Biomedical Data Repository. The main characteristics of these datasets are depicted in Table 1.

Firstly, the Wisconsin Diagnostics Breast Cancer (WDBC) dataset, it was created by William H. Wolberg at University of Wisconsin. This dataset includes a total of 569 observations with benign and malign cases being 357 and 212 observations respectively. For each observation, there are 30 featured variables. These features are computed from digital images of Fine Needle Aspirates (FNA)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Features</th>
<th>Instances</th>
<th>Class</th>
<th>Missing Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wisconsin Breast cancer</td>
<td>32</td>
<td>569</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Pima Diabetes</td>
<td>8</td>
<td>768</td>
<td>2</td>
<td>Yes</td>
</tr>
<tr>
<td>Mammographic Mass</td>
<td>5</td>
<td>961</td>
<td>2</td>
<td>Yes</td>
</tr>
<tr>
<td>Dermatology</td>
<td>33</td>
<td>366</td>
<td>6</td>
<td>Yes</td>
</tr>
<tr>
<td>Colon tumor</td>
<td>2000</td>
<td>62</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>12600</td>
<td>21</td>
<td>2</td>
<td>No</td>
</tr>
</tbody>
</table>
Enhancing E-Service Collaboration with Enforcement and Relationship Management: A Methodology from Requirements to Event Driven Realization
www.igi-global.com/chapter/enhancing-service-collaboration-enforcement-relationship/65785?camid=4v1a

A SOA-Based Module for the Production of Geo-Summaries
Elena Roglia and Rosa Meo (2014). International Journal of Organizational and Collective Intelligence (pp. 64-76).
www.igi-global.com/article/a-soa-based-module-for-the-production-of-geo-summaries/105089?camid=4v1a