ARTICLE IN PRESS

JID: KNOSYS [m5G;November 2, 2017;11:23]

Knowledge-Based Systems 000 (2017) 1–17

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Differential evolution for filter feature selection based on information theory and feature ranking

Emrah Hancer [a,b], Bing Xue [b,*], Mengjie Zhang [b]

[a] Department of Computer Technology and Information Systems, Mehmet Akif Ersoy University, 15039 Burdur, Turkey
[b] Evolutionary Computation Research Group, School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand

ARTICLE INFO

Article history:
Received 19 May 2016
Revised 22 October 2017
Accepted 24 October 2017
Available online xxx

Keywords:
Mutual information
ReliefF
Fisher Score
differential evolution
feature selection

ABSTRACT

Feature selection is an essential step in various tasks, where filter feature selection algorithms are increasingly attractive due to their simplicity and fast speed. A common filter is to use mutual information to estimate the relationships between each feature and the class labels (mutual relevancy), and between each pair of features (mutual redundancy). This strategy has gained popularity resulting a variety of criteria based on mutual information. Other well-known strategies are to order each feature based on the nearest neighbor distance as in ReliefF, and based on the between-class variance and the within-class variance as in Fisher Score. However, each strategy comes with its own advantages and disadvantages. This paper proposes a new filter criterion inspired by the concepts of mutual information, ReliefF and Fisher Score. Instead of using mutual redundancy, the proposed criterion tries to choose the highest ranked features determined by ReliefF and Fisher Score while providing the mutual relevance between features and the class labels. Based on the proposed criterion, two new differential evolution (DE) based filter approaches are developed. While the former uses the proposed criterion as a single objective problem in a weighted manner, the latter considers the proposed criterion in a multi-objective design. Moreover, a well known mutual information feature selection approach (MIFS) based on maximum-relevance and minimum-redundancy is also adopted in single-objective and multi-objective DE algorithms for feature selection. The results show that the proposed criterion outperforms MIFS in both single objective and multi-objective DE frameworks. The results also indicate that considering feature selection as a multi-objective problem can generally provide better performance in terms of the feature subset size and the classification accuracy.

1. Introduction

Classification is typically referred as a supervised learning task in machine learning that infers a relationship between features (characteristics of the dataset) and the class labels. However, the presence of a large number of features often leads to challenges such as overfitting, high computational complexity and low interpretability of the final model [1]. One reason for this is widely known as the curse of dimensionality that arises according to the ratio between the number of features and the number of instances. The most common way to alleviate such problems is to reduce the number of features under consideration using either feature construction or feature selection [1,2].

Feature construction aims to transform the dataset from the high dimensional space to a lower dimensional space by combining the original low-level features to a small number of high-level features, which is better suited for learning process. However, feature construction cannot be easily interpreted since the physical meaning of the original features cannot be retrieved. Feature selection aims to choose a feature subset from the available original features of a dataset, which better contributes to the learning process. In other words, the aim of feature selection is to discard features that are detrimental to the subsequent learning process [3,4]. Feature selection approaches can be categorized into wrappers, embedded and filters based on the evaluation criteria [5]. Wrappers use a learning algorithm (classifier or regression) as a part of evaluation to measure the goodness of the chosen feature subset. Although wrappers are among the most preferred feature selection approaches, there are at least four drawbacks [6]: 1) high computational complexity, 2) the optimal feature subset for a learner may not be optimal for a different learner, 3) determining the user-specified parameters of the learner may be time consuming, and 4) inherent learner limitations (e.g. some learners cannot deal with multi-class classification). Embedded approaches incorporate

* Corresponding author.
E-mail address: bing.xue@ecs.vuw.ac.nz (B. Xue).

https://doi.org/10.1016/j.knosys.2017.10.028

knowledge about the specific structure of the classification algorithm used by a certain learning algorithm. Embedded approaches are computationally less intensive than wrappers. However, they still have high computational complexity and the selected feature subset is dependent on the learning algorithm. Due to these limitations, we specifically focus on filters in this study. Wrapper and embedded approaches are not the focus of this paper and will not be further discussed here. Recent works on wrappers and embedded approaches can be found in [5,7–12].

Filters evaluate feature subsets based on some predefined metrics or information content (e.g. statistical tests) instead of using the learners, i.e., there exists no dependence between the learner (or classifier) and the selected features. Accordingly, filters are more general than wrapper and embedded approaches. In the literature, there have been a wide range of criteria and metrics used for the evaluation of feature subsets such as inconsistency rate, inference correlation, fractal dimension, distance measure and mutual information. Among them, mutual information can be treated as the most preferred and widely investigated for filters due to two main properties [6]: 1) measuring different kinds of relationship between random variables and 2) preserving stability under transformations in the feature space that are invertible and differentiable. Based on mutual information, Battiti [13] proposed the mutual information feature selection (MIFS) method including three fundamental points: 1) features are categorized as relevant and redundant; 2) an heuristic function is used to select features controlling the tradeoff between relevance and redundancy; and 3) a greedy search is applied. Other representative examples of mutual information based approaches are maximum relevance and minimum redundancy (mRmR) [14], uniformly improved MIFS (MIFS-U) [15], and conditional mutual information maximization (CMIM) [16]. Although they are simple to implement and reduce the feature subset size, a selected feature cannot be later removed or changed due to their static greedy search mechanism.

To address these problems, researchers have tried to design mutual information based filter approaches with evolutionary computation (EC) techniques such as particle swarm optimization (PSO) [17], genetic algorithms (GAs) [18], ant colony optimization (ACO) [19] and differential evolution (DE) [20] due to their global search ability. Besides such representative ones, recently developed EC techniques such as artificial bee colony [21], and bacterial colony optimization [22] have also been investigated to obtain better feature subsets for the classification.

However, the potential of EC for feature selection has not been fully investigated. For example, filter based approaches are often computationally cheap, but there is much less work on filters than on wrappers because the fitness functions based on filters are more difficult to design. The most widely used filter measure is mutual information. Although EC with mutual information has achieved better results than classical greedy search, most of such methods just directly adopted existing heuristic/fitness functions as the objective without significant or major improvement, which may limits their performance [5]. Furthermore, although feature selection can be considered as a multi-objective problem, there are only a few works on multi-objective filter feature selection [5,23]. Developing good filter based feature selection methods is still an open issue.

Among EC methods, DE is a relatively recent but highly popular approach. As pointed in [24], DE has been proven to be better than other EC methods in a wide range of problems. Compared to most other EC methods, DE is also much simpler and straightforward to implement, which allows practitioners from other fields, who may not be experts in programming, to implement and tune it to solve the domain-specific problem. Furthermore, DE only has a few parameters to control and the space complexity is low as well. These are particularly important for feature selection since it is a multi-disciplinary area involving researchers from many different fields, but work on DE for feature selection is much less than other EC methods, e.g. GAs and PSO [5]. Furthermore, feature selection is essentially a multi-objective approach, maximizing the classification accuracy and minimizing the number of features [25]. EC methods are particularly good for solving multi-objective problems since their population based mechanism can produce multiple trade-off solutions in a single run [26]. Despite the superior performance of multi-objective DE, there has been almost no work exploring the potential of DE for multi-objective filter feature selection.

1.1. Goals

The overall goal of this paper is to develop filter based feature selection approaches based on information theory, feature ranking and EC techniques to search for a set of non-dominated solutions (feature subsets) yielding a smaller number of features and a similar or even better classification performance on the K-nearest neighbor algorithm than the case that all features are used. To achieve this goal, a novel filter evaluation criterion (named MIRFFS) based on the concepts of mutual relevance, ReliefF [27] and Fisher Score [28] is proposed, and using this proposed criterion, the standard DE and multi-objective DE (MODE) based feature selection approaches are developed. Furthermore, a widely used existing filter based criterion (MIFS) is also redesigned as fitness function for single objective and multi-objective DE to develop filter based approaches. These four developed feature selection approaches will be examined and evaluated on benchmark problems of varying difficulty. Specifically, we will investigate

- the performance of the four algorithms (i.e. single objective and multi-objective DE approaches based on MIRFFS and MIFS) on reducing the number of features and improving the classification performance over using all features,
- the performance of the single objective DE approach based on MIRFFS versus based on MIFS,
- the performance of the multi-objective DE approach based on MIRFFS versus based on MIFS,
- the performance of the multi-objective DE approaches versus the single-objective DE approaches, and
- the performance of all DE filter approaches versus traditional approaches.

1.2. The organization of the paper

The rest of the paper is organized as follows. Section 2 gives an outline of the basic DE algorithm and provides a background on information theory, feature ranking and recent studies related to feature selection, especially filters. Section 3 describes the DE based feature selection approaches using the proposed and existing criteria. Section 4 shows the experimental design and Section 5 presents the experimental results with discussions. Finally, Section 6 concludes the paper and provides an insight into the future trends.

2. Background

This section provides a background concerning the differential evolution, multi-objective optimization, information theory and recent filter approaches.

2.1. Differential evolution

Differential evolution (DE) is a search algorithm proposed by Storn and Price [29] in 1997. DE belongs to the class of evolutionary algorithms in EC techniques that applies biologically inspired

Please cite this article as: E. Hancer et al., Differential evolution for filter feature selection based on information theory and feature ranking, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.10.028