Pattern Recognition 76 (2018) 416–433

Contents lists available at ScienceDirect

Pattern Recognition

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

A two-dimensional (2-D) learning framework for Particle Swarm based feature selection

Faizal Hafiz[a,*], Akshya Swain[a], Nitish Patel[a], Chirag Naik[b]

[a] Department of Electrical & Computer Engineering, The University of Auckland, Auckland, New Zealand
[b] Sarvajanik College of Engineering & Technology, Surat, India

ARTICLE INFO

Article history:
Received 4 November 2016
Revised 1 October 2017
Accepted 21 November 2017
Available online 22 November 2017

Keywords:
Classification
Dimensionality reduction
Feature selection
Particle Swarm Optimization
Machine learning

ABSTRACT

This paper proposes a new generalized two dimensional learning approach for particle swarm based feature selection. The core idea of the proposed approach is to include the information about the subset cardinality into the learning framework by extending the dimension of the velocity. The 2D-learning framework retains all the key features of the original PSO, despite the extra learning dimension. Most of the popular variants of PSO can easily be adapted into this 2D learning framework for feature selection problems. The efficacy of the proposed learning approach has been evaluated considering several benchmark data and two induction algorithms: *Naive–Bayes* and *k-Nearest Neighbor*. The results of the comparative investigation including the time-complexity analysis with GA, ACO and five other PSO variants illustrate that the proposed 2D learning approach gives feature subset with relatively smaller cardinality and better classification performance with shorter run times.

1. Introduction

In recent years, the feature selection problem has been a major concern amongst the researchers in the machine learning field due to availability of huge volume of data from diverse fields such as finance, bio-medical, physical sciences and consumer electronics, etc. The feature selection problem arises from the fundamental question of the machine learning: *How many input features are required to sufficiently capture the characteristics of a data pattern/model?* In the absence of this information large number of input features are used to represent a pattern which often leads to inclusion of many redundant, irrelevant or noisy features. The determination of effective feature subset has been a fundamental problem in machine learning and a topic of active research since last few decades [1].

Recognition of a pattern by a machine learning method involves induction of hypotheses by an induction algorithm which maps the input features to the output class/label. It is highly desirable to induce a classifier using fewer input features as possible for various reasons. For example, in most cases, only limited samples are available for training, whereas training samples required to achieve required accuracy rise exponentially with the number of input features [2]. Further, earlier studies [2,3] suggest that good generalization capabilities can be achieved by a smaller feature subset following the *principle of parsimony* or *Occam's razor*. Moreover, there may be constraints imposed on number of input features due to limited resources, such as number and/or cost of measurements associated with features, storage requirements.

To understand the feature selection problem, consider a dataset with 'n' input features ($U = \{u_1 \dots u_n\}$) and 'm' output labels ($V = \{v_1 \dots v_m\}$). The task of the induction algorithm is to induce hypotheses in the classifier using the learning data pairs $\{U, V\}$ which can later be used to determine appropriate label, $v_k \in V$, corresponding to any future input pattern. The objective is to find a subset of input features,'X' ($X \subset U$, $|X| < n$), through which this task can be accomplished with same or improved accuracy. This can further be represented as follows,

$$J(X) = \max_{Y \subset U,\ |Y| < n} J(Y) \tag{1}$$

where, '$J(\cdot)$' is a criterion function which estimates the '*goodness*' of the given subset. An exhaustive search of all possible feature subsets to solve (1) requires the examination of large number of feature subsets ($\sum_{k=1}^{n-1} \binom{n}{k} \approx 2^n$), which often becomes intractable even for moderate size problems. The feature selection problem is NP-hard [2,4,5] and its optimal solution is not guaranteed unless all possible 2^n feature subsets are examined [4]. The optimal solution of this problem requires the selection of both the subset size (*cardinality*) and the features themselves.

* Corresponding author.
E-mail addresses: faizalhafiz@ieee.org, faizalhafiz@icloud.com (F. Hafiz), a.swain@auckland.ac.nz (A. Swain), nd.patel@auckland.ac.nz (N. Patel), naik.chirag@gmail.com (C. Naik).

https://doi.org/10.1016/j.patcog.2017.11.027
0031-3203/

Most of the methods which have been proposed to address the feature selection problem can be broadly be classified into two categories: *deterministic* and *meta-heuristic*. Majority of deterministic search methods such as *sequential search* [1,6–8], *branch and bound* [9,10] require *monotonic* criterion function, $J(\cdot)$, and/or neglects correlation among features. Therefore, some of the recent search methods uses alternate search approach based on meta-heuristics [11]. The meta-heuristic search methods have proven to be very effective on various discrete and combinatorial problems. If properly adapted, they can provide optimal solution to the feature selection problem. The pioneer effort in this direction was the application of Genetic Algorithm (GA) for the feature selection problem in [12]. Apart from GA, various other meta-heuristics search such as Tabu Search (TS), Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) have been applied to the feature selection task [11,13–16]. Note that all of the meta-heuristic search paradigms have to be adapted for the feature selection task. Unlike PSO, the other search paradigms like GA, ACO can be adapted comparatively easily without any major changes in their learning strategies. For example, GA can be applied with a binary string representation and graph representation can be used for ACO. On the other hand, the similar task is quite challenging with canonical PSO due to *Euclidean distance* based learning at its core[17]. Nevertheless, in the recent research, PSO is preferred due to its simplicity, its ability to avoid local minima and it does not require any heuristic information other than the criterion function, $J(\cdot)$ [11].

Unlike other evolutionary search paradigms, PSO has a single learning mechanism; known as *velocity update*. In canonical PSO, the velocity update involves evaluation of the *Euclidean distance* of the particle from its learning exemplars, referred to as *learning* [17,18]. The next move of the particle on the search landscape is evaluated based on this learning. Since the Euclidean distance does not convey useful information in the discrete domain, a binary version of PSO (BPSO) was proposed in [19], where the velocity is represented as the *selection probability*. However, BPSO has several limitations which severely affect its search performance (discussed at length in Section 2.3). Moreover, BPSO is intended to be a general search paradigm for any discrete problem and for this reason does not contain any specific search mechanism to cope with the feature selection problem. Most of the recent research on PSO based feature selection are either application of BPSO or some extension of it. The new search strategies dedicated to feature selection problem is still an open issue [11].

The objective of this study is to bridge this gap by proposing a new learning framework for the PSO and its variants. The proposed learning framework is designed as *generic learning framework* which can be used to adapt any PSO variant for the feature selection problems. Since its introduction, PSO has attracted many researchers and over the years many PSO variants have been proposed to improve the performance of the original algorithm albeit the research has been mostly restricted to the '*continuous domain*', *i.e.*, $x \in \mathbb{R}$. The introduction of a generalized learning framework can help in transferring most of the existing PSO related research from *continuous domain* ($x \in \mathbb{R}$) to the feature selection problem ($x \in \mathbb{N}$). This is the main motivation of this research.

The search for the optimal feature subset involves two aspects; selection of both *cardinality* and *features*. However, most of the search methods focus only on the significance of feature/feature subset and to the best of our knowledge, none of the search methods directly exploit the information on subset cardinality to guide the search process. In the proposed learning framework, information about both the cardinality and features are jointly exploited to effectively guide the search. Contrary to the earlier practice of storing only selection probabilities of *features* in n-dimensional *velocity* vector, in this work, the velocity records the *selection likelihood* of both *cardinality* and *features* in a two-dimensional matrix. Due to this distinctive quality, the proposed learning framework is named "*2D learning approach*". Moreover, a simple method is proposed to update the cardinality and feature selection likelihoods and generate a new feature subset based on this comprehensive information.

The efficacy of the proposed approach is evaluated on wide variety of real-life datasets obtained from UCI Machine Learning repository [20]. Note that the search landscape of the feature selection problem is jointly defined by the dataset and the induction algorithm used to induce the classifier. For this reason, two widely used induction hypotheses based on *Naive–Bayes* and *k-Nearest Neighbor* (k-NN) [21] have been used to induce the classifier. Further, the well known PSO variant, *Unified Particle Swarm Optimization* (UPSO) [22], is adapted for the feature selection problem using the 2D learning approach. The performance of the adapted UPSO (2D-UPSO) is compared with GA and five other PSO based feature selection methods.

The rest of the manuscript is organized as follows: the brief overview of the feature selection methods is provided in Section 2, followed by the detailed description of the proposed 2-D learning approach in Section 3. The application of 2-D learning approach to adapt PSO variants is illustrated in Section 4. The experimental setup, compared algorithms and results are discussed in Sections 5 and 6, respectively. Finally, the conclusions of this study are discussed in the Section 7.

2. Brief review of feature selection methods

Since, in this study a new feature selection method has been proposed in a 2D learning framework, it is appropriate to briefly discuss the search behavior of the existing feature selection methods for the sake of completeness. Most of the existing feature selection methods can be classified into two distinctive categories on the basis of the nature of search: *deterministic vs. meta-heuristic*. Another important distinction arises from the different approaches to evaluate the criterion function, $J(\cdot)$ (*filters vs. wrappers*).

2.1. *Deterministic* vs. *meta-heuristic search*

Over the years several deterministic search methods have been proposed such as sequential search [1,6–8], *branch and bound* (BAB) [9,10]. These methods are *deterministic*, since for a given dataset, each independent run of these methods will provide same solution. The core idea of the sequential search is to operate on a single feature, e.g., forward search (SFS) [6] starts with an empty set and a single feature is included in each step where as in backward search (SBS) [1], the search begins with all the features and a single feature is discarded in each step. The drawback of this approach is the *nesting effect*, i.e., once the feature is included/excluded it cannot be discarded/included. To overcome this problem, "*plus-l-minus-r*" and *floating* search [7,8] methods were suggested. The common drawback of the sequential search is the emphasis on an isolated feature which completely ignores the correlation among the features. Due to correlation, an isolated insignificant feature may become very effective when considered with others [23]. Further, most of deterministic search methods operate on strict assumption of *monotonic* criterion function, i.e., adding a feature will always lead to improvement. This assumption is impractical as in many cases, due to limited training samples larger input feature subset often leads to over-fitting and deteriorates the classifier's performance. Moreover, *a priori* selection of subset size (d) is required for most of the deterministic methods, consequently the search space reduces to only $\binom{n}{d}$ subsets instead of all possible 2^n subsets. Hence, it is highly possible that the optimal feature subset is not even included in the search space.

The meta-heuristic search methods such as Genetic Algorithm (GA), Tabu Search (TS), Particle Swarm Optimization (PSO), Ant