Computers and Electrical Engineering 45 (2015) 1–9

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Investigating the effect of fixing the subset length on the performance of ant colony optimization for feature selection for supervised learning ☆,☆☆

Nadia Abd-Alsabour

Cairo University, Cairo, Egypt

ARTICLE INFO

Article history:
Received 28 April 2013
Received in revised form 1 May 2015
Accepted 1 May 2015
Available online 15 May 2015

Keywords:
Ant colony optimization
Feature selection
Subset problems
Supervised learning

ABSTRACT

This paper studies the effect of fixing the length of the selected feature subsets on the performance of ant colony optimization (ACO) for feature selection (FS) for supervised learning. It addresses this concern by investigating: (1) determining the optimal feature subset from datamining perspective, (2) demonstrating the solution convergence in case of fixing the length of the selected feature subsets, (3) determining the subset length in ACO for subset selection problems, and (4) different stopping criteria when solving FS by ACO. Besides, two types of experiments on ACO algorithms for FS for classification and regression problems using artificial and real world datasets in two cases fixing and not fixing the length of the selected feature subsets with the use of a support vector machine. The obtained results showed that not fixing the length of the selected feature subsets is better than fixing the length of the selected feature subsets.

1. Introduction

Ants are able to find the shortest path between their nest and food sources because of the chemical substance, called *pheromone*, which they deposit on their way. The pheromone evaporates over time so the shortest paths will contain more pheromone and will subsequently attract a greater number of ants. Ant colony optimization algorithms simulate the foraging behavior of some ant species [1]. ACO algorithms have been used successfully for solving many optimization problems. Recently, they have been adopted to solve feature selection problems.

The main goal of feature selection is to find a subset of features with predictive performance comparable to the full set of features [2] i.e., the results of feature selection should be data having potential for good solutions. In feature selection, we try to avoid selecting too many or too few features than necessary. If insufficient features are selected, the information content to keep the concept of data is degraded. If too many features (including redundant or irrelevant features) are selected, the prediction accuracy may be lower. This is due to the interference of irrelevant information. The removal of irrelevant features is significant for discovering hidden relationships between features and between features and class targets of patterns [3].

In many applications, we do not know which features (and subsequently how many features) will lead to higher prediction accuracy. The main target of feature selection algorithms is to select the most informative features regardless of their

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Fangyang Shen.
☆☆ A short version of this paper was published in the proceedings of PDCAT 2012.

E-mail address: nadia.abdalsabour@ieee.org

http://dx.doi.org/10.1016/j.compeleceng.2015.05.003

number. For example, if the most informative features in a dataset is ten features but we restrict the number of the selected features in advance to be eight features only, this means that we exclude important features that will affect the prediction accuracy.

As a subset problem, solutions for feature selection problems do not have fixed length i.e., different ants may have solutions of different lengths. However, some literature (such as in Nemati et al. [4], Sivagaminathan and Ramakrishnan [5], and Gunturiet et al. [6]) solved feature selection problems using ant algorithms by fixing the length of the selected feature subsets. This contradicts with the main concepts of ant algorithms as heuristic algorithms. In this paper, we study the effect of fixing the length of the selected feature subsets on the performance of ant colony optimization algorithms for feature selection for supervised learning. This is achieved by investigating that from datamining perspective, operations research perspective, and ant algorithms for subset problems perspective. Moreover, two types of experiments with and without fixing the length of the selected feature subsets using two ant algorithms for feature selection for classification problems and regression problems [7,8] were performed.

The rest of this paper is organized as follows. Section 2 briefly explains ant colony optimization. Section 3 introduces the basics of supervised learning. The fourth section explains solution convergence in heuristic search. Section 5 describes the process of optimal feature selection from data mining perspective. The sixth section explains building solutions for subset problems by ACO algorithms. Section 7 presents different stopping criteria when solving feature selection by ACO algorithms. Section 8 details the experiments carried out and presents the obtained results. The discussion of the results is presented in Section 9. Section 10 concludes this paper and highlights future work in this area.

2. Ant colony optimization

Ant colony optimization is a strand of swarm intelligence (collective intelligence) that was inspired from the foraging behavior of some ants' species. The novelty of this foraging behavior of ants rises from the fact that the collective behavior of some unintelligent decentralized small entities results in intelligent outputs. Ant colony optimization was introduced by means of the proof-of-concepts to the traveling salesman problem (TSP). Since then, ACO algorithms have been applied to many optimization problems. First, classical problems (other than the TSP) were tackled such as routing, assignment, scheduling, and subset problems. More recent applications include for example bioinformatics, and dynamic problems. Moreover, it has shown a great potential to cope with multi-objective optimization problems. Ant Colony Optimization has reached state-of-the-art results for several important problem classes, such as quadratic assignment, scheduling, and routing problems.

The first ant algorithm is known as ant system (AS) and it was proposed in the early nineties [9]. Ant system could not compete with state of-the-art algorithms for the traveling salesman problem. Nevertheless, it had the important role of stimulating further research both on ant algorithmic variants (that obtain much better computational performance), and on applications to a large variety of different optimization problems. Since its appearance, it became the basis for many successive ACO algorithms and it is known as the progenitor of all ACO algorithms.

Ant colony system (ACS) is considered one of the most successful ACO algorithms [10]. Since its appearance, it has been used to solve a variety of optimization problems. It addresses the main disadvantages of ant system that are the lack of diversification and the lack of intensification i.e., it intensively explores areas of the search space with high quality solutions and moves to unexplored areas of the search space [1,9–11].

3. Supervised learning

In supervised learning, the goal is to *predict* for a given record the value of the output feature based on the values of the input features [12,13]. The relationship between the target feature and the input features is *learned* from the training data in which the target feature is already known [13]. The learning process is *supervised* i.e., it is told to which class each sample belongs [14].

If the labels are *categorical* (such as employed, unemployed, or retired) that is are chosen from some fixed set of possibilities such as predicting whether or not a prospective borrower should be given a mortgage, then the prediction is called *classification*. If the labels are real-valued (rankings, ratings, counts, or quantities) such as predicting the size of mortgage a prospective borrower should be allowed, or predicting price changes, then the prediction is called *regression* [12,13,15,16].

4. Solution convergence in heuristic search

Exhaustive search (that searches over all the possible feature subsets of a feature set) is usually time-consuming. Can one make some smart choices based on the minimum information available but without looking at the whole picture? This is all what heuristic search is about. The rationale of this non-optimal strategy is threefold:

- it is quick to find a solution (a subset of features),
- it usually can find near optimal solution if not optimal, and
- the trade-off of optimality with the speed is often worthwhile because of a gain in speed with little loss of optimality [17].