An Efficient Ant Colony Instance Selection Algorithm for KNN Classification

Amal Miloud-Aouidate, University of Science and Technology, Houari Boumediene, Bab Ezzouar Algiers, Algeria
Ahmed Riadh Baba-Ali, University of Science and Technology, Houari Boumediene, Bab Ezzouar Algiers, Algeria

ABSTRACT

The extraordinary progress in the computer sciences field has made Nearest Neighbor techniques, once considered impractical from a standpoint of computation (Dasarathy et al., 2003), became feasible for real-world applications. In order to build an efficient nearest neighbor classifier two principal objectives have to be reached: 1) achieve a high accuracy rate; and 2) minimize the set of instances to make the classifier scalable even with large datasets. These objectives are not independent. This work addresses the issue of minimizing the computational resource requirements of the KNN technique, while preserving high classification accuracy. This paper investigates a new Instance Selection method based on Ant Colonies Optimization principles, called Ant Instance Selection (Ant-IS) algorithm. The authors have proposed in a previous work (Miloud-Aouidate & Baba-Ali, 2012a) to use Ant Colony Optimization for preprocessing data for Instance Selection. However to the best of the authors’ knowledge, Ant Metaheuristic has not been used in the past for directly addressing Instance Selection problem. The results of the conducted experiments on several well known data sets are presented and compared to those obtained using a number of well known algorithms, and most known classification techniques. The results provide evidence that: (1) Ant-IS is competitive with the well-known kNN algorithms; (2) The condensed sets computed by Ant-IS offers also better classification accuracy then those obtained by the compared algorithms.

Keywords: Accuracy Rate, Ant Colonies Optimization, Evolutionary Algorithms, Instance Selection (IS), K-Nearest Neighbor Classification (kNN)

1. INTRODUCTION

The emergence of the computer revolution, in terms of inexpensive memory and high processing speed, has created a renewed interest in nearest neighbor (NN) techniques. This has positioned the NN rule among popular nonparametric classification techniques. The K-Nearest Neighbor classification rule (KNN) is a powerful classification method allowing the classification of an unknown instance using a set of classified training instances.

DOI: 10.4018/ijamc.2013070104
Compared to other well known classifiers, neighborhood techniques remain very attractive thanks to their easy use. One of the limitations of the k-Nearest Neighbor rule (Suguna & Thanushkodi, 2010; Yu & Zhengguo, 2007) is the size of the training set. If the number of training instances is too small, the accuracy of the KNN classifier is not acceptable. But if the training set is too large, many KNN classifiers need excessive running time. This problem can be solved in 3 ways (Shi et al., 2007): by reducing the dimensions of the feature space; by using smaller data sets; or by using an improved algorithm which can accelerate the calculation time.

The calculation of a consistent training subset with a minimal cardinality for the KNN rule turns to be an NP-hard problem (Cover & Hart, 1967; Wilfong, 1992). Researchers have proposed instances selection techniques to address this problem. Like many other combinatorial problems, the instance selection (IS) requires an exhaustive search to obtain an optimal solution. Among the existing methods, metaheuristics and especially genetic algorithms and tabu search have proven their efficiency in dealing with the said problem (Gil-Pita & Yao, 2007; Acı et al., 2010; Suguna & Thanushkodi, 2010; Ceveron & Ferri, 2010; Wu et al., 2011).

Recently, many Ant based algorithms have been successful in solving different types of combinatorial optimization problems, such as: symmetric and asymmetric traveling salesman problems (Gambardella & Dorigo, 1995; Gambardella & Dorigo, 1996; Dorigo & Gambardella, 1997); the sequential ordering problem (Gambardella & Dorigo, 1997; Gambardella & Dorigo, 1999); the quadratic assignment problem (Gambardella & Dorigo, 1995); and the vehicles routing problem (Gambardella & Taillard, 1999).

In this paper we propose a new approach, to reduce the number of instances in the training set of the 1NN classifier involving the concepts of ant colony optimization.

We note that the proposal described in this paper is an extended algorithm to the one described in our previous work (Miloud-Aouidate & Baba-Ali, 2012b). Our previous version presented some weaknesses related principally to the stopping criterion. In this paper we present an improvement for the previous algorithm introducing a new stopping criterion and a new selection equation which allows summarizing the selection calculation, prescrip in two stages before. In addition, this paper investigates the reduction aspect of the proposed algorithm, which constitutes one of two bases of the instance selection.

Furthermore, more data sets and appropriate statistical tools have been used to justify the conclusions achieved, also more investigations and an extensive comparison to key algorithms in instance selection field are presented in this paper.

This paper is organized as follow: Section 2 describes the k-Nearest Neighbor classifier. Section 3 explains the instance selection problem. Section 4 surveys some of the most known instance selection algorithms. Section 5 introduces the proposed Ant-IS. Section 6 investigates the experiments and comparisons. Finally, section 7 concludes the work.

2. K-NEAREST NEIGHBOR CLASSIFIER

The k-Nearest Neighbor rule is a supervised learning algorithm. This algorithm allows classifying a new instance according to the most frequent classe of its k-nearest neighbors’ classes.

The purpose of this algorithm is to classify a new instance according to the instances in the training set. Each instance \( I_i \) of the training set is composed of inputs and an output, where \( I_i = <\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ij}, c_i> \). The inputs are the features of the instance \( (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ij}) \) and the output is its respective class \( c_i \).

To make a prediction for a test instance, the algorithm first computes its distance to every training instance. Then, it keeps the \( k \) closest training instances, where \( k \geq 1 \) is a fixed integer. The algorithm looks for the most common class among these instances’ classes. This class is the predicted class for the test instance.
Investigating the Efficiency of GRASP for the SDST HFS with Controllable Processing Times and Assignable Due Dates
www.igi-global.com/chapter/investigating-the-efficiency-of-grasp-for-the-sdst-hfs-with-controllable-processing-times-and-assignable-due-dates/82704?camid=4v1a

Online Adaptive Neuro-Fuzzy Based Full Car Suspension Control Strategy
www.igi-global.com/chapter/online-adaptive-neuro-fuzzy-based-full-car-suspension-control-strategy/82707?camid=4v1a