Chapter 81
Data Mining using Advanced Ant Colony Optimization Algorithm and Application to Bankruptcy Prediction

Vishal Arora
Indian Institute of Technology Delhi, India

Vadlamani Ravi
Institute for Development and Research in Banking Technology, India

ABSTRACT

Ant Colony Optimization (ACO) is gaining popularity as a data mining technique in the domain of Swarm Intelligence for its simple, accurate and comprehensive nature of classification. In this paper, the authors propose a novel advanced version of the original ant colony based miner (Ant-Miner) in order to extract classification rules from data. They call this Advanced ACO-Miner (ADACOM). The main goal of ADACOM is to explore the flexibility of using a different knowledge extraction heuristic approach viz. Gini’s Index to increase the predictive accuracy and the simplicity of the rules extracted. Further, the authors increase the information and the prediction level of the set of rules extracted by dynamically changing specific parameters. Simulations are performed with ADACOM on a few benchmark datasets Wine, WBC (Wisconsin Breast Cancer) and Iris from UCI (University of California at Irvine) data repository and compared with Ant-Miner (Parpinelli, Lopes, & Freitas, 2002), Ant-Miner2 (Liu, Abbass, & McKay, 2002), Ant-Miner3 (Liu, Abbass, & McKay, 2003), Ant-Miner+ (Martens, De Backer, Haesen, Vanthienen, Snoeck, & Baesens, 2007) and C4.5 (Quinlan, 1993). The results show that ADACOM outperforms the above mentioned algorithms in terms of predictive accuracy, simplicity of rules, sensitivity, specificity and AUC values (area under ROC curve). In addition, the ADACOM is also employed to extract rules from bank datasets (UK, US, Spanish and Turkish) for bankruptcy prediction and the results are compared with that obtained by Ant-Miner. Again ADACOM yielded better results and is proven to be the better choice for solving bankruptcy prediction problems in banks.
INTRODUCTION

The exponential advancement in digital and hardware technology over the last few decades has led to tremendous increase in the amount of data received and stored in various data warehouses and data repositories around the world. The data collection and handling is not of much concern but the useful and relevant information retrieval from the so called raw data is one area that has clearly seen loads of efforts coming into it recently.

Data Mining (DM) is one commonly used name to address the efforts put in to derive a set of high level knowledge, from data, which is said to be both comprehensible and accurate (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). In other words we can say that DM is the process of automatically seeking, constructing, analyzing and validating the structural patterns in data and in turn using them, for prediction, on unseen data (Witten, Frank, 1999; Lu, Setiono, & Liu, 1995).

Decision trees (Quinlan, 1986), neural networks (Odom & Sharda, 1990) and genetic algorithms (Goldberg, 1989) are some of the data mining techniques employed to extract classification rules. Currently many classification algorithms are used to extract some kind of relevant pattern amongst the data in the form of a model which can be used in prediction phase for the classification of unseen data. These include decision tree (Quinlan, 1986), neural networks (Odom & Sharda, 1990), support vector machine (Vapnik, 1995). It is an undeniable fact that neural networks, support vector machine and statistical algorithms are strong when it comes to predictive accuracy but fall short in terms of simplicity and comprehensibility. It is nearly impossible to extract useful high level knowledge, from the model developed, which can be easily understood and comprehended by the user. Thus they are called ‘Black Box’. This black box nature is overcome by the new algorithms which generate a set of easily interpreted rules as the model itself.

Recently efforts have been put into research, with success, related to nature-based approaches for finding the solution to classification problems. These include Genetic Algorithms (GA) (Mahfoud & Mani, 1996), Ant Colony Optimization (ACO) (Dorigo & Maniezzo, 1996., Dorigo & Stutzle, 2004) and Particle Swarm Optimization (PSO) (Sousa, Silva, & Neves, 2003) etc. which fall under Swarm intelligence (SI).

Since we developed a new variant of ACO-Miner we discuss more on ACO in the following. The basic ACO technique consists of a population of ants (artificial) which searches in the solution space of the given problem, guided by the pheromone trails and a heuristic approach. The pheromone update and evaporation take place depending upon the past exposure and experience of the ants (candidate solutions). This eventually increases the probability of finding the optimum solution to the problem in the next iteration (Dorigo & Gambardella, 1997). Its significance and efficiency in various fields has been successfully demonstrated by successful applications to hard combinatorial optimization problems (Dorigo & Stutzle, 2004). In this paper we developed an advanced ant colony optimization algorithm (ADACOM) which makes use of a different but efficient heuristic approach. The new algorithm makes use of Gini’s Index, instead of Entropy, as the measure of information level in a particular term. In addition to this the ADACOM tries to capture the overlooked information resulting from the parameter constraints. This is accomplished by dynamically varying the parameters so as to retrieve more knowledge from the data by compensating the simplicity of the rules extracted but significantly improving upon the predictive accuracy., Ripper, C4.5 (Quinlan, 1993), INN, logit (Ohlson, 1980) and SVM (Quinlan, 1986) are compared with ADACOM for the benchmark datasets.