Chapter XIX
Ant Colonies and
Data Mining

Miroslav Bursa
Czech Technical University in Prague, Czech Republic

Lenka Lhotska
Czech Technical University in Prague, Czech Republic

ABSTRACT

The chapter concentrates on the use of swarm intelligence in data mining. It focuses on the problem of medical data clustering. Clustering is a constantly growing area of current research. Medicine, market, trade, and meteorology belong to the numerous fields that benefit of its techniques. First an introduction into data mining and cluster validation techniques is presented, followed by a review of ant-inspired concepts and applications. The chapter provides a reasonably deep insight into the most successful ant colony and swarm intelligence concepts, their paradigms and application. The authors present discussion, evaluation and comparison of these techniques. Important applications and results recently achieved are provided. Finally, new and prospective future directions in this area are outlined and discussed.

INTRODUCTION

This chapter concentrates on the use of swarm intelligence in data mining. It focuses on the problem of data clustering in biomedical data processing. Clustering is a constantly growing area of current research. Medicine, market, trade, and meteorology are some of the numerous fields that benefit of its techniques.

The objective of this chapter is to introduce the methods for clustering together with the methods for evaluation of different clusterings. It presents the fundamentals of ant inspired methods, followed by a compact review of the basic ant clustering models together with the most successful variations and modifications. In the second part, it presents the application of ant-colony clustering in biomedical data processing.

In the last two decades, many advances in computer sciences have been based on the observation and emulation of processes of the natural world.

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.
The coordination of an ant colony is of local nature, composed mainly of indirect communication through pheromone (also known as stigmergy; the term has been introduced by Grassé et al. (1959)), although direct interaction communication from ant to ant (in the form of antennation) and direct communication have also been observed (Trianni, Labella, & Dorigo, 2004). In studying these paradigms, we have high chance to discover inspiring concepts for many successful metaheuristics. More information on the ant colony metaphors can be found in the section Ant Colony Optimization.

The author himself specializes on the use of such kind of methods in the area of biomedical data processing. The application is described in the section Applications.

The chapter is organized as follows: First, an introduction to data mining and clustering is presented together with a brief survey on ant colony inspired methods in clustering. Then, a natural background of applied methods is presented. It summarizes the most important properties of ant colonies that served as an inspiration source for many algorithms that are described in the following part. The next section describes the most successful methods in data clustering: first, the pioneering ant-inspired clustering algorithms are described followed by the evolution of further ant-inspired algorithms for clustering. Finally, applications of the algorithms and paradigms published by the author are presented, followed by conclusion and future directions. At the end, relevant literature has been carefully selected to provide the reader with additional resources containing the state-of-the-art information in the area.

BACKGROUND

About Data Mining

This section provides more thorough introduction into data mining and reasons the use of methods that provide only approximate results (with an acceptable error) in much more reasonable time.

In many industrial, business, healthcare and scientific areas we can see still growing use of computers and computational appliances. Together with the boom of high-speed networks and increasing storage capacity of database clusters and warehouses, a huge amount of various data can be collected. Such data are often heterogeneous and seldom come without any errors (or noise). Data mining is not only an important scientific discipline, but also an important tool in industry, business and healthcare, that is concerned with discovering (hidden) information in the data.

Many data-mining algorithms with growing number of modifications exist nowadays. Such modifications aim at speeding up the data mining process, increase its robustness and stability. But even with rapidly increasing computational power of modern computers, the analysis of huge databases becomes very expensive (in terms of computer time and/or memory – and therefore also financially). This is why scientists instantly create, develop and evaluate novel approaches to analyze and process these data. In contrast to classical methods, nature inspired methods offer many techniques that can increase speed and robustness of classical methods. Clustering techniques inspired by nature already exist – self-organizing maps, neural networks, evolutionary algorithms, etc. This chapter concentrates on techniques, inspired by ant colonies, where various branches can be distinguished: (1) methods inspired solely by ant behavior, and (2) hybrid methods (combining ant-colony approach with traditional methods).

Clustering belongs to the most useful data-mining techniques. It significantly reduces the amount of data to be processed. Data clustering can significantly help in the case of electrocardiogram (ECG) processing. Clustering of long-term ECG record should reduce expert’s (cardiologist’s) work and furthermore reduce his load and fatigue. Some methods can produce structures, which can reveal the inherent structure of the data.

Clustering

As the application section is targeted towards the task of data clustering, this part provides a general introduction into the clustering problem with an overview of cluster validation techniques, which can be used in the case the correct classification is not known.