Chapter 5
Advances of Applying Metaheuristics to Data Mining Techniques

Simon Fong  
University of Macau, Macau SAR

Jinyan Li  
University of Macau, Macau SAR

Xueyuan Gong  
University of Macau, Macau SAR

Athanasios V. Vasilakos  
Kuwait University, Kuwait

ABSTRACT

Metaheuristics have lately gained popularity among researchers. Their underlying designs are inspired by biological entities and their behaviors, e.g. schools of fish, colonies of insects, and other land animals etc. They have been used successfully in optimization applications ranging from financial modeling, image processing, resource allocations, job scheduling to bioinformatics. In particular, metaheuristics have been proven in many combinatorial optimization problems. So that it is not necessary to attempt all possible candidate solutions to a problem via exhaustive enumeration and evaluation which is computationally intractable. The aim of this paper is to highlight some recent research related to metaheuristics and to discuss how they can enhance the efficacy of data mining algorithms. An upmost challenge in Data Mining is combinatorial optimization that, often lead to performance degradation and scalability issues. Two case studies are presented, where metaheuristics improve the accuracy of classification and clustering by avoiding local optima.

DOI: 10.4018/978-1-4666-8513-0.ch005
1. INTRODUCTION

With the ever increase in data growth and ease of data collection, data mining has always been a popular computing discipline for business and many other industry domains. Data mining offers insights from the data by analyzing, modeling and detecting outliers or anomalies over the underlying patterns hidden among the data. It is a process of knowledge discovery founded on well-established statistical principles which are deterministic in nature (Fayyad et al., 1996). In general there are two types of learning approaches for uncovering and understanding the representative patterns of the data – one is called supervised learning and the other is unsupervised learning. Supervised learning as the name suggests requires observing through the dataset in order to induce a mathematical model that essentially maps the relations between the variables (attributes) of the data which characterize the whole dataset and the predefined class labels as targets. Unsupervised learning often just let the data to reveal their characteristic patterns, by for examples clustering into groups, visualization, or generating their frequent itemsets or associations. In both types learning, though they are very popular among modern data mining techniques, the process is deterministic and sequential – one can trace from the steps of the process as well as the data that have been so far processed. This typical learning process is by greedy search or divide-and-conquer where the model updates itself whenever new data becomes available.

One drawback about deterministic learning is that often a local optimum is obtained and the learning process ceased, without considering all the combinations of possibilities. The other reason is that it may simply take too long the time to evaluate through all the possible solutions because of the combinatorial limitation such as problems of NP-hard in nature. The limitation introduced by combinatorial property in numerical analysis hinders the efficacy of data mining in producing an intermediate solution of local optimum, or else taking forever to find a globally best solution. Such multidimensional combinatorial problems that constitute to the “curse-of-dimensionality” are often found in data mining problems. The dataset from which a data mining model has to learn from is usually characterized by a very high dimensionality depending on the amount of features or attributes that characterize the data. Measuring the distance between a pair of data points in clustering algorithms, or finding a best fit regression curve in prediction model among a series of data point in a high-dimensional hyper-space is a norm in data mining. Generally the higher the dimension space, the greater the combinatorial explosion on the search space results, making the corresponding model induction extremely difficult and time-consuming. The difficulty lies in the search space of candidate solutions that arise in many combinations expands faster than exponentially, when the size of the problem grows. As a result the increasingly huge search space makes exhaustive search for the best solution infeasible, which cripples the efficacy of the analytical methods far from the globally best.

Having known of this obstacle stemmed mainly from the combinatorial problems, researchers resorted to combinatorial optimization in the hope of breaking through this limitation and improving the data mining performance. Metaheuristics are designed for combinatorial optimization in which an optimal solution is explored over a large search-space. In the context of theoretical optimization and computer algorithms, a metaheuristic is an abstract procedure programmed to manage or control a lower-level search mechanism called heuristic that looks for a sufficiently good solution to an optimization problem. The lower-level search works by searching partially of the whole space but it retains information gained from previous iteration heuristically. This way, without the need of going through the whole search space which is either impossible or too time consuming, metaheuristics is able to work on incomplete information with a known upper bound on the required computation capacity. Metaheuristics are also
www.igi-global.com/e-resources/library-recommendation/?id=79

Related Content

Data Field for Hierarchical Clustering
Shuliang Wang, Wenyan Gan, Deyi Li and Deren Li (2013). Developments in Data Extraction, Management, and Analysis (pp. 303-324).
www.igi-global.com/chapter/data-field-hierarchical-clustering/70803?camid=4v1a

Web Semantics for Personalized Information Retrieval
www.igi-global.com/chapter/web-semantics-for-personalized-information-retrieval/178373?camid=4v1a

Combining User Co-Ratings and Social Trust for Collaborative Recommendation: A Data Analytics Approach
www.igi-global.com/chapter/combining-user-co-ratings-and-social-trust-for-collaborative-recommendation/159504?camid=4v1a

A Workload Assignment Strategy for Efficient ROLAP Data Cube Computation in Distributed Systems
Ilhyun Suh and Yon Dohn Chung (2016). International Journal of Data Warehousing and Mining (pp. 51-71).
www.igi-global.com/article/a-workload-assignment-strategy-for-efficient-rolap-data-cube-computation-in-distributed-systems/168486?camid=4v1a