

Chapter 1

Biologically Inspired Techniques for Data Mining: A Brief Overview of Particle Swarm Optimization for KDD

Shafiq Alam

University of Auckland, New Zealand

Yun Sing Koh

University of Auckland, New Zealand

Gillian Dobbie

University of Auckland, New Zealand

Saeed ur Rehman

Unitec Institute of Technology, New Zealand

ABSTRACT

Knowledge Discovery and Data (KDD) mining helps uncover hidden knowledge in huge amounts of data. However, recently, different researchers have questioned the capability of traditional KDD techniques to tackle the information extraction problem in an efficient way while achieving accurate results when the amount of data grows. One of the ways to overcome this problem is to treat data mining as an optimization problem. Recently, a huge increase in the use of Swarm Intelligence (SI)-based optimization techniques for KDD has been observed due to the flexibility, simplicity, and extendibility of these techniques to be used for different data mining tasks. In this chapter, the authors overview the use of Particle Swarm Optimization (PSO), one of the most cited SI-based techniques in three different application areas of KDD, data clustering, outlier detection, and recommender systems. The chapter shows that there is a tremendous potential in these techniques to revolutionize the process of extracting knowledge from big data using these techniques.

INTRODUCTION

Knowledge Discovery and Data mining (KDD) helps us understand some characteristics of data in large repositories by passing the data through different operations such as data selection, pre-processing, transformation, data mining and post

processing. Data mining, which is the core of KDD, extracts informative patterns such as clusters of relevant data, classification and association rules, sequential patterns and prediction models for different types of data such as textual, audio-visual, and microarray data.

DOI: 10.4018/978-1-4666-6078-6.ch001

Recently a huge increase in the use of Swarm Intelligence (SI) based optimization techniques for KDD has been observed. Particle Swarm Optimization (PSO) is one of the most highly cited SI techniques for KDD because it has the simplicity, and extendibility to be used for different data mining tasks. For instance in data clustering, optimization based techniques have been proposed to address different issues that affect the performance of clustering techniques. These issues include selection of the initial parameters, optimizing the centroids, convergence to a solution, and trapping in unfeasible solutions. When involving optimization in the process, it either uses an optimization technique as a data-clustering algorithm or adds optimization to the existing data clustering approaches. Optimization based clustering techniques treat data clustering as an optimization problem and try to optimize an objective function either to a minima or maxima. In the context of data clustering, a minimization objective function can be the intra-cluster distance and maximization can correspond to the inter-cluster distance.

While adding optimization to the data mining process, the results achieved so far are promising. Optimization has significantly improved accuracy and efficiency while solving some other problems such as global optimization, multi-objective optimization and avoiding being trapped in local optima (Kuo et al., 2011) (Alam et al., 2008) (Das et al., 2009) (Van der Merwe, & Engelbrecht, 2003). The involvement of intelligent optimization techniques has been found effective in enhancing the performance of complex, real time, and costly data mining processing. A number of optimization techniques have been proposed to add to the performance of the clustering process. Swarm Intelligence is one such optimization area where techniques based on SI have been used extensively to either perform clustering independently or add to the existing clustering techniques.

This chapter introduces the use of PSO in three KDD areas; data clustering, outlier detection, and recommender systems. The next section explains the concept of swarm intelligence.

SWARM INTELLIGENCE

Swarm Intelligence, inspired by the biological behavior of animals, birds, and fish, is an innovative intelligent optimization technique (Abraham et al., 2006) (Engelbrecht, 2006). SI techniques are based on the collective behavior of swarms of bees, fish schools, and colonies of insects while searching for food, communicating with each other and socializing in their colonies. The SI models are based on self-organization, decentralization, communication, and cooperation between the individuals within the team. The individual interaction is very simple but emerges as a complex global behavior, which is the core of swarm intelligence (Bonabeau & Meyer, 2001). Although swarm intelligence based techniques have primarily been used and found very efficient in traditional optimization problems, a huge growth in these techniques has been observed in other areas of research. These application areas vary from optimizing the solution for planning, scheduling, resource management, and network optimization problems. Data mining is one of the contemporary areas of application, where these techniques have been found to be efficient for clustering, classification, feature selection and outlier detection. The use of swarm intelligence has been extended from conventional optimization problems to optimization-based data mining.

A number of SI based techniques with many variants have been proposed in the last decade and the number of new techniques is growing. Among different SI techniques, Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) are the two main techniques, which are widely used

8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/biologically-inspired-techniques-for-data-mining/110452

Related Content

Video Stream Mining for On-Road Traffic Density Analytics

Rudra Narayan Hota, Kishore Jonna and P. Radha Krishna (2012). *Pattern Discovery Using Sequence Data Mining: Applications and Studies* (pp. 182-194).

www.irma-international.org/chapter/video-stream-mining-road-traffic/58680/

Toward a Grid-Based Zero-Latency Data Warehousing Implementation for Continuous Data Streams Processing

Tho Manh Nguyen, Peter Brezany, A. Min Tjoa and Edgar Weippl (2005). *International Journal of Data Warehousing and Mining* (pp. 22-55).

www.irma-international.org/article/toward-grid-based-zero-latency/1758/

The Power of Sampling and Stacking for the PaKDD-2007 Cross-Selling Problem

Paulo J.L. Adeodato, Germano C. Vasconcelos, Adrian L. Arnaud, Rodrigo C.L.V. Cunha, Domingos S.M.P. Monteiro and Rosalvo F.O. Neto (2008). *International Journal of Data Warehousing and Mining* (pp. 22-31).

www.irma-international.org/article/power-sampling-stacking-pakdd-2007/1804/

Detecting Pharmaceutical Spam in Microblog Messages

Kathy J. Liszka, Chien-Chung Chan and Chandra Shekar (2012). *Social Network Mining, Analysis, and Research Trends: Techniques and Applications* (pp. 101-115).

www.irma-international.org/chapter/detecting-pharmaceutical-spam-microblog-messages/61514/

Parallel Computing for Mining Association Rules in Distributed P2P Networks

Huiwei Guan (2013). *Data Mining: Concepts, Methodologies, Tools, and Applications* (pp. 107-124).

www.irma-international.org/chapter/parallel-computing-mining-association-rules/73436/