IDEA GROUP PUBLISHING
701 E Chocolate Avenue, Hershey PA 17033, USA
Tel: 717/533-8845; Fax: 717/533-8661; URL: www.idea-group.com
ITB7029

Chapter III

On the Use of Evolutionary Algorithms in Data Mining

Erick Cantú-Paz and Chandrika Kamath
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory, USA

With computers becoming more pervasive, disks becoming cheaper, and sensors becoming ubiquitous, we are collecting data at an ever-increasing pace. However, it is far easier to collect the data than to extract useful information from it. Sophisticated techniques, such as those developed in the multi-disciplinary field of data mining, are increasingly being applied to the analysis of these datasets in commercial and scientific domains. As the problems become larger and more complex, researchers are turning to heuristic techniques to complement existing approaches. This survey chapter examines the role that evolutionary algorithms (EAs) can play in various stages of data mining. We consider data mining as the end-to-end process of finding patterns starting with raw data. The chapter focuses on the topics of feature extraction, feature selection, classification, and clustering, and surveys the state of the art in the application of evolutionary algorithms to these areas. We examine the use of evolutionary algorithms both in isolation and in combination with other algorithms including neural networks, and decision trees. The chapter concludes with a summary of open research problems and opportunities for the future.

INTRODUCTION

Data mining is increasingly being accepted as a viable means of analyzing massive data sets. With commercial and scientific datasets approaching the terabyte

This chapter appears in the book, *Data Mining: A Heuristic Approach*, edited by Hussein Abbass, Ruhul Sarker and Charles Newton.

and even petabyte range, it is no longer possible to manually find useful information in this data. As the semi-automated techniques of data mining are applied in various domains, it is becoming clear that methods from statistics, artificial intelligence, optimization, etc., that comprise data mining, are no longer sufficient to address this problem of data overload. Often, the data is noisy and has a high level of uncertainty. It could also be dynamic, with the patterns in the data evolving in space and time. To address these aspects of data analysis, we need to incorporate heuristic techniques to complement the existing approaches.

In this chapter, we survey the role that one category of heuristic algorithms, namely, evolutionary algorithms (EAs), plays in the various steps of the data mining process. After a brief definition of both the data mining process and evolutionary algorithms, we focus on the many ways in which these algorithms are being used in data mining. This survey is by no means exhaustive. Rather, it is meant to illustrate the diverse ways in which the power of evolutionary algorithms can be used to improve the techniques being applied to the analysis of massive data sets. Following a survey of current work in the use of EAs for data mining tasks such as feature extraction, feature selection, classification, and clustering, we describe some challenges encountered in applying these techniques. We conclude with the exciting opportunities that await future researchers in the field.

AN OVERVIEW OF DATA MINING

Data mining is a process concerned with uncovering patterns, associations, anomalies and statistically significant structures in data (Fayyad et al., 1996). It typically refers to the case where the data is too large or too complex to allow either a manual analysis or analysis by means of simple queries. Data mining consists of two main steps, data pre-processing, during which relevant high-level features or attributes are extracted from the low level data, and pattern recognition, in which a pattern in the data is recognized using these features (see Figure 1). Pre-processing the data is often a time-consuming, yet critical, first step. To ensure the success of the data-mining process, it is important that the features extracted from the data are

Figure 1: Data mining—An iterative and interactive process