Available online at www.sciencedirect.com

ScienceDirect

European Journal of Operational Research 180 (2007) 723–737

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Computing, Artificial Intelligence and Information Management

Framework for efficient feature selection in genetic algorithm based data mining

Riyaz Sikora [a,*], Selwyn Piramuthu [b]

[a] *Department of Information Systems, University of Texas at Arlington, P.O. Box 19437, Arlington, TX 76019, United States*
[b] *Department of Decision and Information Sciences, University of Florida, Gainesville, FL 32611-7169, United States*

Received 9 November 2003; accepted 9 February 2006
Available online 26 May 2006

Abstract

We present the design of more effective and efficient genetic algorithm based data mining techniques that use the concepts of feature selection. Explicit feature selection is traditionally done as a wrapper approach where every candidate feature subset is evaluated by executing the data mining algorithm on that subset. In this article we present a GA for doing both the tasks of mining and feature selection simultaneously by evolving a binary code along side the chromosome structure used for evolving the rules. We then present a wrapper approach to feature selection based on Hausdorff distance measure. Results from applying the above techniques to a real world data mining problem show that combining both the feature selection methods provides the best performance in terms of prediction accuracy and computational efficiency.

Keywords: Genetic algorithms; Rule learning; Knowledge discover; Data mining; Evolutionary algorithms

1. Introduction

The ubiquity of databases in almost every area of human endeavor has resulted in the rapid increase in the amount of data collected and stored in industry, government, and scientific organizations. Data in these databases vary in format and content ranging from trillions of point-of-sale transactions and credit card purchases to pixel images of distant galaxies. It is not uncommon to come across databases that are measured in terabytes of data. For example, Wal-Mart, the chain of over 2000 retail stores, uploads over 20 million point-of-sale transactions to an AT&T massively parallel system with more than 1000 processors running a centralized database every day [4]. Wal-Mart operates a data warehouse with over 583 terabytes of sales and inventory data. Although Moore's law does not cover the growth curve of databases, the amount of data stored by businesses nearly doubles every 12–18 months. These collections of massive amounts of data have created opportunities to monitor, analyze and predict processes of interest. In today's fierce competitive business environment, firms need to rapidly turn these terabytes of data into significant insights into their

[*] Corresponding author. Tel.: +1 817 272 5397; fax: +1 817 272 5801.
E-mail address: rsikora@uta.edu (R. Sikora).

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.02.040

customers, markets, products and processes to guide their marketing, investment, and management strategies to gain competitive advantage.

Although these data contain buried valuable information that can be beneficially utilized, raw data by itself does not provide much information. Raw data must first be processed to extract patterns or useful knowledge. To this end, the development of effective and efficient methods for deriving knowledge from these data is becoming increasingly important [16]. Recent efforts under the general rubric of data mining represent a strategic response to this growing need for analytic tools that deliver in situations where traditional analytical methods fail. This is especially evident when dealing with unstructured data that are noisy and incomplete, and where scalability of traditional methods is an issue. By discovering hidden, implicit, and previously unknown and potentially useful patterns and relationships in the data, data mining enables users to extract greater value from their data than simple query and analyses approaches.

Some of the commonly used data mining algorithms fall under the following categories: decision trees and rules [26] nonlinear regression and classification methods [14], example-based methods [18], probabilistic graphical dependency models [30], and relational learning models [12]. Over the years genetic algorithms have been successfully applied in learning tasks in different domains, like chemical process control [27], financial classification [28], manufacturing scheduling [22], among others.

In this paper we propose frameworks for data mining using genetic algorithms, implement these, and evaluate their performance using examples. We chose genetic algorithm due to its simplicity and its capability as a powerful search mechanism. We present the design of more effective and efficient genetic algorithm based data mining techniques that use the concepts of self-adaptive feature selection together with a wrapper feature selection method based on Hausdorff distance measure. A genetic algorithm [17] uses a population of individual solution structures called *chromosomes*. The *fitness* of an individual solution is its performance measure. This measure is used to favor selection of successful *parents* for new *offspring*, such that the whole *population* of solutions incrementally evolves towards greater fitness. Offspring solutions are produced from parent solutions by the application of *crossover* and *mutation* operators. Theory shows that the knowledge about desirable solutions is advantageously stored in the population itself, implicitly contained in the surviving chromosomes. We take advantage of this principle in developing modified frameworks essentially using the genetic algorithm at its core.

The rest of this paper is organized as follows. We present a brief discussion on feature selection in the next section. In Section 3, we discuss genetic algorithm and its basic dynamics. We also include the first framework for the basic genetic algorithm case and show how we implement and evaluate this methodology. This is followed by the presentation of a modified filter-based genetic algorithm with embedded feature selection in Section 4. In Section 5, we provide a brief discussion of the Hausdorff distance measure, and then present the proposed framework of genetic algorithm with Hausdorff distance as wrapper-based feature selection. We discuss the effectiveness of feature selection in Section 6, including experimental results using a real world data set. Section 7 concludes this paper.

2. Feature selection

A typical real world data set consists of as many features that are deemed necessary. This is constrained by (1) knowledge of the domain of interest, and in turn knowledge of essential features that capture knowledge in this domain, (2) availability of these essential features, (3) resources available to gather these available essential features, (4) resources available to store, maintain, and retrieve these collected features. Given these constraints, it is clear that not all features that end up in the resulting data set are necessary or sufficient to learn the concept of interest. Assuming all necessary features are present in this data, feature selection is the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept [19]. A goal of feature selection is to avoid selecting too many or too few features than is necessary. If too few features are selected, there is a good chance that the information content in this set of features is low. On the other hand, if too many (irrelevant) features are selected, the effects due to noise present in (most real-world) data may overshadow the information present. Hence, this tradeoff between noise and useful information is addressed by feature selection methods to alleviate effects due to noise while accentuating effects due to useful information in the data.