Available online at www.sciencedirect.com
ScienceDirect
BioSystems 88 (2007) 76–91
BioSystems
www.elsevier.com/locate/biosystems

Evolving fuzzy rules to model gene expression

Ricardo Linden [a,*], Amit Bhaya [b]

[a] *FSMA-RJ, R Monte Elíseo S/N°, CEP 27943-180, Macaé, RJ, Brazil*
[b] *Department of Electrical Engineering, PEE/COPPE/UFRJ, P.O. Box 68504, CEP 21945-970, Rio de Janeiro, RJ, Brazil*

Received 5 October 2004; received in revised form 29 March 2006; accepted 21 April 2006

Abstract

This paper develops an algorithm that extracts explanatory rules from microarray data, which we treat as time series, using genetic programming (GP) and fuzzy logic. Reverse polish notation is used (RPN) to describe the rules and to facilitate the GP approach. The algorithm also allows for the insertion of prior knowledge, making it possible to find sets of rules that include the relationships between genes already known. The algorithm proposed is applied to problems arising in the construction of gene regulatory networks, using two different sets of real data from biological experiments on the *Arabidopsis thaliana* cold response and the rat central nervous system, respectively. The results show that the proposed technique can fit data to a pre-defined precision even in situations where the data set has thousands of features but only a limited number of points in time are available, a situation in which traditional statistical alternatives encounter difficulties, due to the scarcity of time points.
© 2006 Elsevier Ireland Ltd. All rights reserved.

Keywords: Genetic programming; Fuzzy logic; Microarrays; Reverse engineering; Gene regulatory network

1. Introduction

Fuzzy logic is based on fuzzy set theory and especially on the concept of a fuzzy set. Informally, a fuzzy set is a set with imprecise boundaries, in which the transition from membership to non-membership is gradual rather than abrupt. A fuzzy set F in a universe of discourse U is characterized by a membership function μ_F, which associates each element $u \in U$ with a grade of membership $\mu_F(u) \in [0,1]$ in the fuzzy set F. Note that a classical set A in U is a special case of a fuzzy set with all membership values $\mu_A(u) \in \{0,1\}$ (Hiirsalmi et al., 2000).

A fuzzy implication is viewed as describing a fuzzy relation between the fuzzy sets forming the implication. A fuzzy rule, such as "if x is A then y is B" is implemented by a fuzzy implication (fuzzy relation) which has a membership function $\mu_{A\to B}(x, y) \in [0,1]$. Note that $\mu_{A\to B}(x, y)$ measures the degree of truth of the implication relation between x and y. A set of related fuzzy rules forms a fuzzy rule base that can be used to infer fuzzy results in the form of fuzzy sets.

Fuzzy logic offers an appealing method for describing phenomena by a set of rules and data sets. These data sets relate directly to concepts used on a daily basis, such as "fast", "strong" or "high", while the rules express knowledge approximately the same way a human expert would. An example of a fuzzy rule would be "if the car is fast, then the force applied to the brakes is strong".

Given these characteristics, fuzzy rules are easy to understand, verify and extend, since they are very simi-

* Corresponding author. Tel.: +55 21 2526 2344; fax: +55 21 2260 6211.
E-mail addresses: rlinden@pobox.com (R. Linden), amit@nacad.ufrj.br (A. Bhaya).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.04.006

lar to the way a person might express knowledge. This also makes them attractive for use in domains where experts are available and can seed the systems with a number of effective rules from the outset (Bentley, 1999).

The goal of this paper is to propose an algorithm that finds a set of fuzzy rules that could represent the actual regulation of gene expression performed in the cell. This problem is analogous to fuzzy control, because there is an unknown control process determining how each gene will be expressed. In fact, a major challenge in current fuzzy control research is learning good controllers for large-scale, non-linear systems with many input and output variables where no training data are available from an expert (Carse et al., 1996).

The optimization abilities of evolutionary algorithms (EA) could be used to develop a good set of rules to be used by a fuzzy inference engine and to optimize the choice of membership functions. This has been done in other situations, starting as far back as (De Jong and Spears, 1991) and, more recently, in Bentley (1999), Dounias et al. (2002) or Yang et al. (2003) and in the review work in Freitas (2003) for example. EAs are used in this paper as a way to find a fuzzy rule base that might explain phenomena at hand.

Evolutionary algorithms are inspired by Nature. The idea is to mimic the natural evolution of the species in order to create a new kind of search technique that is robust and intelligently seeks solutions in a possibly infinite search space (Mitchell, 1996). Some of the techniques that are part of this branch of computer science are genetic algorithms (GA), genetic programming (GP) and evolutionary programming (EP).

All evolutionary algorithms use a population of competing solutions subjected to random variation and selection for a specific purpose (Fogel and Corne, 2003) which is to evolve the population to one that contains a higher proportion of superior ("fitter") individuals. The fitness of each individual in the population (its quality) is a measure of how well that individual achieves the desired goal. The variation and selection are usually based on two operators, the crossover operator which combines two different individuals into a new one and the mutation operator, which randomly changes parts of one individual in order to increase diversity.

An evolutionary algorithm could be described by the following pseudo-code

```
Create Initial Population
While termination criteria not met
    Select from current population parents which will generate
    offspring
    Apply genetic operators to the selected parents and generate
    offspring
    Select next population from current individuals and
    generated offspring
End While
Present best solution(s)
```

In this algorithm, the termination criteria are usually time based (a number of generations has elapsed), quality based (a certain performance has been achieved) or stagnation based (the best individuals has not improved for a certain number of generations), while parent selection is usually based on a roulette approach, where the parents with highest evaluation ("fittest") correspond to a bigger fraction of the roulette wheel.

Therefore, in order to define an EA one must define the coding scheme (how each individual will be represented in the computer), the operators (both mutation and crossover and any other specific one that will be used), the evaluation or fitness function (i.e., a measure of the quality of the current solutions to the problem at hand).

In genetic algorithms (GA), a form of EA, each solution is encoded by a binary string or another simple structure called a chromosome. Genetic programming (GP), which is another form of EA, can be used to evolve programs to perform certain tasks (Koza, 1992). In GP, the simple chromosome structure is replaced by a tree structure, in which a solution is either an algebraic equation or a program based on the input variables (Yang et al., 2003).

Many problems of current interest in bioinformatics have high dimensionality without a corresponding number of examples (data points) that would allow the application of well-known statistical techniques. One such area is the reverse engineering of genetic networks, further described below. The data set available for this reverse engineering generally consists of hundreds to thousands of signals, usually measured for no more than twenty time-steps. The paucity of data renders network models inferred from this data statistically insignificant (Van Someren et al., 2000). Since GP methods are able to generate a broad spectrum of solutions, even from incomplete or insufficient data sets, they are adequate for application in this area, generating testable hypotheses for the biological laboratory. GP methods might even succeed when the data scarcity might make statistical methods unable to generate solutions.

The ongoing revolution in the field of genomics can be attributed mainly to the development of new tools that have flooded scientists with huge amounts of data. These new tools have made it clear that the current

Download English Version:

https://daneshyari.com/en/article/2076938

Download Persian Version:

https://daneshyari.com/article/2076938

Daneshyari.com