Chapter 2
Genetic Programming as Supervised Machine Learning Algorithm

ABSTRACT

This chapter presents the theory and procedures behind supervised machine learning and how genetic programming can be applied to be an effective machine learning algorithm. Due to simple and powerful concept of computer programs, genetic programming can solve many supervised machine learning problems, especially regression and classifications. The chapter starts with theory of supervised machine learning by describing the three main groups of modelling: regression, binary, and multiclass classification. Through those kinds of modelling, the most important performance parameters and skill scores are introduced. The chapter also describes procedures of the model evaluation and construction of confusion matrix for binary and multiclass classification. The second part describes in detail how to use genetic programming in order to build high performance GP models for regression and classifications. It also describes the procedure of generating computer programs for binary and multiclass calcification problems by introducing the concept of predefined root node.

DOI: 10.4018/978-1-5225-6005-0.ch002

INTRODUCTION

As an evolutionary computation technique, Genetic Programming (GP) is one of the most popular and widely used. It is inspired by biological evolution, where each individual in the population represents possible solution of the problem. Individuals in the population are breeding, by exchanging their genetic materials and producing offspring. Individuals in the population have fitness values which play a role in the process of selecting and creating the new population. The offspring and its parents are potential members of the new population. Who will be the 'inhabitant' of the new population depends of the fitness value and selection method. The new population is created by selecting individuals with one of defined selection method by taking into account the fitness value of the members. Selection method takes the fitness of each individual from that population and decide if this individual is good to be the member of the new population. Different selection methods treat each individual differently. Once the new population is created, members are tested how to solve the problem. The best individual in the current population represents solution of the problem for current evolution (iteration). At the end of the evolution process the best individual, which is the best result of all previous evolutions, represents the best solution for the problem. Individuals in GP are called computer programs.

The structure of the computer program is the reason why the GP is so popular and widely used. There are several ways how to represent the computer programs in GP. The tree structure is the most popular representation, where nodes of computer programs are initialized from the function and terminal sets. They can be in different shape which allow diversity of the population and better condition of the mating and breeding.

As possible solution of machine learning problem, GP computer programs are generated on different way for each type of the problem. In case of regression problems, GP programs are usually generated from algebraic, logical and conditional functions and terminals. Due to the fact that the solution for regression problems can take any continuous numeric value, computer structure does not have any constrains in its construction except protected operations. For classification problems construction of computer programs should be carefully planned, including that the root node must be pre-defined. This chapter will cover all aspects of supervised machine learning and how each category of the ML can be applied in GP.