Chapter 27
Genetic Programming as Supervised Machine Learning Algorithm

Bahrudin Hrnjica
University of Bihac, Bosnia and Herzegovina

Ali Danandeh Mehr
Antalya Bilim University, Turkey

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.

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

DOI: 10.4018/978-1-6684-6291-1.ch027

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.

MACHINE LEARNING

Machine Learning (ML) is a sort of artificial intelligence (AI) that provides learning algorithms, mostly for the computers, with the ability to learn, without being explicitly programmed. The process of ML consists of searching the data to recognize the pattern in the data. The recognizing process can be defined as process of computer learning. Once the patterns are recognized, the computer can make prediction for new or unseen data based on persisted knowledge with more or less accuracy.

ML can be categorized based on the task that is going to be solved:

- Supervised,
- Unsupervised and
- Reinforcement learning.

In supervised ML, the learning process consists of finding the rule that maps inputs (features) to outputs (labels). During the learning process, available data can be divided on the two sets. The training set is used for training and collecting the knowledge from the data. The second set is called validation or testing set which the learning algorithm uses for test against overfitting.

Unsupervised learning is the process of discovering patterns in data without defined output. With unsupervised learning, the correct result cannot be determined because no output variable is defined. Algorithms are left to their capability to discover as much as possible knowledge from the data. Because there is no output variable, there is no need for splitting available data set into training and testing part,