World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:13, No:07, 2019

Evolving Credit Scoring Models using Genetic Programming and Language Integrated Query Expression Trees

Authors : Alexandru-Ion Marinescu

Abstract : There exist a plethora of methods in the scientific literature which tackle the well-established task of credit score evaluation. In its most abstract form, a credit scoring algorithm takes as input several credit applicant properties, such as age, marital status, employment status, loan duration, etc. and must output a binary response variable (i.e. “GOOD” or “BAD”) stating whether the client is susceptible to payment return delays. Data imbalance is a common occurrence among financial institution databases, with the majority being classified as “GOOD” clients (clients that respect the loan return calendar) alongside a small percentage of “BAD” clients. But it is the “BAD” clients we are interested in since accurately predicting their behavior is crucial in preventing unwanted loss for loan providers. We add to this whole context the constraint that the algorithm must yield an actual, tractable mathematical formula, which is friendlier towards financial analysts. To this end, we have turned to genetic algorithms and genetic programming, aiming to evolve actual mathematical expressions using specially tailored mutation and crossover operators. As far as data representation is concerned, we employ a very flexible mechanism – LINQ expression trees, readily available in the C# programming language, enabling us to construct executable pieces of code at runtime. As the title implies, they model trees, with intermediate nodes being operators (addition, subtraction, multiplication, division) or mathematical functions (sin, cos, abs, round, etc.) and leaf nodes storing either constants or variables. There is a one-to-one correspondence between the client properties and the formula variables. The mutation and crossover operators work on a flattened version of the tree, obtained via a pre-order traversal. A consequence of our chosen technique is that we can identify and discard client properties which do not take part in the final score evaluation, effectively acting as a dimensionality reduction scheme. We compare ourselves with state of the art approaches, such as support vector machines, Bayesian networks, and extreme learning machines, to name a few. The data sets we benchmark against amount to a total of 8, of which we mention the well-known Australian credit and German credit data sets, and the performance indicators are the following: percentage correctly classified, area under curve, partial Gini index, H-measure, Brier score and Kolmogorov-Smirnov statistic, respectively. Finally, we obtain encouraging results, which, although placing us in the lower half of the hierarchy, drive us to further refine the algorithm.

Keywords : expression trees, financial credit scoring, genetic algorithm, genetic programming, symbolic evolution
Conference Title : ICCS 2019 : International Conference on Computational Science
Conference Location : Stockholm, Sweden
Conference Dates : July 15-16, 2019