Genet Program Evolvable Mach (2012) 13:69–70
DOI 10.1007/s10710-011-9156-z

EDITORIAL

Evolutionary algorithms for data mining

Pierre Collet · Man Leung Wong

Published online: 14 January 2012

Artificial evolution can be applied to machine learning thanks to genetic programming, for instance, which is a very successful branch of Evolutionary Computing. One would therefore think that by transitivity, because data-mining is one of the main applications of machine learning, artificial evolution could be successful in solving data-mining problems. Whether this is the case or not, it seems that artificial evolution is not much used in data-mining, even though many papers show that EC can provide interesting alternative solutions to standard machine learning approaches.

However, a secret weapon of EC is that it is intrinsically parallel. This very special feature allows EC to exploit multi-core computers as well as the massive parallelism of GPGPU cards, with their hundreds of cores. Then, the island model makes it possible to obtain scalable linear (an even in some cases supra-linear) speedup when several machines are used to solve the same problem, therefore allowing one to efficiently execute EC algorithms on massively parallel machines and super-computers. These architectures were quite rare in the past, but are now widely available as parallelism is the path followed by hardware manufacturers to increase the computing power of their processors.

Therefore, the future of evolutionary computation is quite bright, and any new developments in EC could probably overtake more standard methods if these methods are sequential by nature, or do not parallelize well. This is a reason why it is essential to encourage researchers to tackle problems in new domains where EC may not yet be a major actor.

P. Collet (✉)
Université de Strasbourg, Alsace, France
e-mail: pierre.collet@unistra.fr

M. L. Wong
Lingnan University, Tuen Mun, Hong Kong
e-mail: mlwong@ln.edu.hk

In this special section on data-mining, two important papers were selected that clearly show the important potential of evolutionary computation in data-mining. Doucette et al. describe Symbiotic Bid-Based Genetic Programming (SBB) that employs cooperative and competitive coevolution for discovering knowledge from large databases with many attributes. In order to handle huge databases, the data-mining problem should be decomposed into simpler problems, and suitable attributes as well as training examples should be selected. SBB applies cooperative coevolution to achieve problem decomposition and competitive coevolution to select attributes and training examples. Doucette et al. compare SBB with a kernel method (F+SVM) on databases with many attributes. They observe that the models obtained by SBB and F+SVM have comparable classification performance, while the complexities of the models evolved by SBB are simpler than those of F+SVM.

Veeramachaneni et al. propose an approach for mining knowledge from sparse sensory evaluation data. This data-mining problem is challenging because: (1) the same flavors in the data have been repeatedly measured by different assessors and their responses vary significantly; (2) the flavors are composed of seven ingredients which interact non-linearly; and (3) assessors may use different standards to evaluate the flavors. Veeramachaneni et al. apply ParetoGP to generate a set of models for each assessor. The models can provide collective predictions as well as confidences of the predictions. Based on the sets of models of the assessors, the approach can classify the assessors into three groups which are respectively easy to please, hard to please, and neutral. The approach employs multi-objective Particle Swarm Optimization to identify flavors that maximize liking for individual assessor or a group of assessors.

This research has the potential to impact work in evolutionary computation for data-mining and push forward the frontier of the field.