Contribution ID: **58** Type: **Oral**

An investigation of the effect of clustering-based initialization on Learning Classifier Systems' effectiveness: leveraging the Grid infrastructure

Wednesday 14 April 2010 11:40 (20 minutes)

Strength-based Learning Classifier Systems (LCS) are machine learning systems designed to tackle both sequential and single-step decision tasks by coupling a gradually evolving population of rules with a reinforcement component. ZCS-DM, a Zeroth-level Classifier System for Data Mining, is a novel algorithm in this field, recently shown to be very effective in several benchmark classification problems. In this paper, we evaluate the effect of clustering-based initialization on the algorithm's performance, utilizing the EGEE infrastructure as a robust framework for an efficient parameter sweep.

Detailed analysis

Clustering-based initialization is based on the idea that starting from a non-random set of rules may help the evolutionary process focus on the search-space optima (the optimal ruleset for the given classification task in our case) more effectively and quickly. Intuitively, this non-random set of rules should be based on the given dataset and provide an effective summary of the knowledge available in it. Our solution tries to leverage the potential of clustering algorithms to provide a representative set of centroids for a given dataset, that we then try to transform into rules suitable for the initialization of ZCS-DM. The ultimate goal is to boost the algorithm's performance, both in terms of predictive accuracy and in terms of training times, through the reduction of the evolutionary process'execution time. In our current investigation, after detailing the proposed initialization process, we report the results of deploying the algorithm on the Grid infrastructure by means of a DAG workflow process. The conducted series of experiments evaluates alternative initialization parameter sets, aiming towards the optimization of the algorithm in terms of both efficiency and accuracy.

Conclusions and Future Work

Our studies so far have proven ZCS-DM to be a robust and accurate data mining tool, which can outperform its rival algorithms in most of the benchmark datasets used and to achieve a prediction accuracy well above the baseline on all of them. However, given the evolutionary nature of the algorithm, further optimization in terms of time efficiency is necessary. In this direction, we have employed a clustering-based initialization phase and evaluated its effect on algorithm performance through an extensive set of experiments conducted by leveraging the Grid infrastructure.

Impact

Among the various methods used to tackle classification problems, rule-based (or tree-structured) classifiers are particularly popular, because they combine: i) an intuitive representation that allows for easy interpretation of the resulting classification model; ii) a nonparametric nature that is especially suited for exploring datasets where there is no prior knowledge of the attributes'probability distributions; iii) fast, computationally inexpensive construction methods that produce models storable in a compact form; and iv) fast classification of new observations, once the model has been constructed. Inspecting the above list, one can easily conclude

that LCS share most of the advantages of these methods, with the exception of the third point, as genetic algorithm-based search is an arguably slow and computationally expensive search method. Towards this end, the optimization of the ZCS-DM algorithm using Grid resources may provide researchers with an invaluable tool for performing data-mining tasks, and end-users with an efficient application for enhancing decision making tasks.

Keywords

Classification, Learning Classifier Systems, Parameter Sweep, Algorithm Optimization

URL for further information

http://issel.ee.auth.gr/

Authors: Ms TZIMA, Fani (Aristotle University of Thessaloniki); Mr PSOMOPOULOS, Fotis (Aristotle University of Thessaloniki)

Co-author: Prof. MITKAS, Pericles (Aristotle University of Thessaloniki)

Presenters: Ms TZIMA, Fani (Aristotle University of Thessaloniki); Mr PSOMOPOULOS, Fotis (Aristotle University of Thessaloniki)

Session Classification: Computer Science

Track Classification: Scientific results obtained using distributed computing technologies