Genetic Network Programming with Parallel Processing for Association Rule Mining in Large and Dense Databases

Eloy Gonzales[*]
egonzale@asagi.waseda.jp

Kaoru Shimada
k.shimada@ruri.waseda.jp

Shingo Mabu
mabu@waseda.jp

Kotaro Hirasawa
hirasawa@waseda.jp

Jinglu Hu
jinglu@waseda.jp

Graduate School of Information, Production and Systems
Waseda University
Japan

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Algorithms, Performance

ABSTRACT

Several methods of extracting association rules have been reported. A new evolutionary computation method named Genetic Network Programming (GNP) has also been developed recently and its efectiveness is shown for small datasets. However, it has not been tested for large datasets, particularly in datasets with a large number of attributes. The aim of this paper is to extract association rules from large and dense datasets using GNP considering a real world database with a huge number of attributes. We propose a new method where a large database is divided into many small datasets, then each GNP deals with one dataset having attributes with appropiate size, which was selected randomly from a large dataset and generated genetically. These GNPs are processed in parallel. We then propose some new genetic operations to improve the number of rules extracted and their quality as well. The proposed method improves remarkably on simulations.

Fig. 1 shows the architecture of the proposed method. We use the CLIENT/SERVER model. CLIENT side carries out preprocessing of large database, assignment of files to each server, rule checking, and genetic operations on files. SERVER side carries out processing of each file using conventional GNP based mining method independently. The features and advantages of the proposed method are the following:

Rule extraction is done in parallel. Each file generates its local pool of the rules.

Files or datasets are treated as individuals in order to do new genetic operations over them and improve the rule extraction.

Extracted rules are stored in a global pool. The rules are verified to avoid redundancy among them and it is assured that only new rules are stored.

There are two levels of genetic operations: 1) Crossover, Mutation-1 and Mutation-2 for GNP based mining method on the SERVER side and 2) Crossover (C_a, C_r) and Injection (M_a) for files on the CLIENT side.

Conditions of important association rules are defined flexibly by users for both levels of genetic operations. The definition may include the minimum threshold χ^2 value, points of crossover, number of attributes to be injected and selection of the acquired information to be used. χ^2 value of each rule is measured by the feature of GNP's structure.

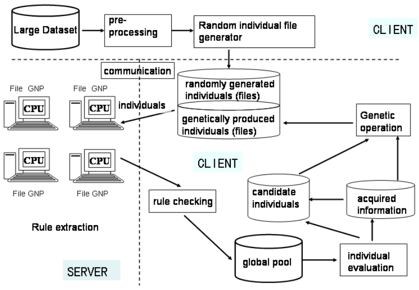

Figure 1: Architecture of the proposed model

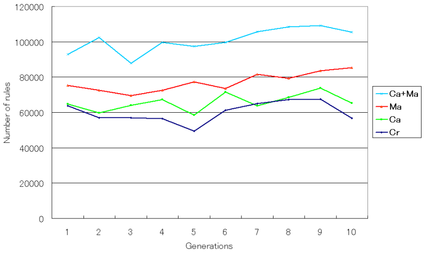

Figure 2: Total number of rules extracted in each generation

Our simulation results show that:

- GNP based parallel data mining method can effectively extract important association rules from large database as shown in Fig. 2, especially when files are regarded as individuals and evolved by genetic operations.
- The general performance of the entire system is improved due to parallel computation.

[*] Hibikino 2-7, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan. Tel/Fax: +81 93 692-5261

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
ACM 978-1-59593-697-4/07/0007.