Association Rule Mining for Continuous Attributes using Genetic Network Programming

Karla Taboada [*]
k.taboada@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

Most association rule mining algorithms make use of discretization algorithms for handling continuous attributes. However, by means of methods of discretization, it is difficult to get highest attribute interdependency and at the same time to get lowest number of intervals. We propose a method using a new graph-based evolutionary algorithm named "Genetic Network Programming (GNP)" that can deal with continues values directly, that is, without using any discretization method as a preprocessing step. GNP is one of the evolutionary optimization techniques, which uses directed graph structures as solutions and is composed of three kinds of nodes: start node, judgment node and processing node. Once GNP is booted up, firstly the execution starts from the start node, secondly the next node to be executed is determined according to the judgment and connection from the current activated node. The features of GNP are described as follows. First, it is possible to reuse nodes; because of this, the structure is compact. Second, GNP can find solutions of problems without bloat, which can be sometimes found in Genetic Programming (GP), because of the fixed number of nodes in GNP. Third, nodes that are not used at the current program executions will be used for future evolution. Fourth, GNP is able to cope with partially observable Markov processes. In this paper, we propose a method that can deal with continuous attributes, where attributes in databases correspond to judgment nodes in GNP and each continuous attribute is checked whether its value is greater than a threshold value and the association rules are represented as the connections of the judgment nodes. Threshold a_i is firstly determined by calculating the mean μ_i and standard deviation σ_i of all attribute values of A_i. Then, initial threshold a_i is selected randomly between the interval [μ_i - $\alpha_i\sigma_i$, μ_i + $\alpha_i\sigma_i$] where α_i is a parameter to determine the range of the interval. Once the threshold a_i is selected for all attributes, each value of the attribute A_i is checked if it is greater than the threshold a_i in the judgment nodes of the proposed method. In addition to that, the threshold a_i is also evolved by mutation between [μ_i - $\alpha_i\sigma_i$, μ_i + $\alpha_i\sigma_i$] in every generation in order to obtain as many association rules as possible. The features of the proposed method are as follows compared with other methods: 1) Extracts rules without identifying frequent itemsets used in Apriori-like mining methods. 2) Stores extracted important association rules in a pool all together through generations. 3) Measures the significance of associations via the chi-squared test. 4) Extracts important rules sufficient enough for user's purpose in a short time. 5) The pool is updated in every generation and only important association rules with higher chi-squared value are stored when the identical rules are stored. We have evaluated the proposed method by doing two simulations. Simulation 1 uses fixed threshold values; that is, they remain fixed at initial thresholds during evolution. In simulation 2, thresholds are evolved by mutation in every generation. Fig. 1 shows the number of rules extracted in the pool in simulation 2. It is found that the number of rules extracted has been increased, which means simulation 2 outperforms simulation 1.

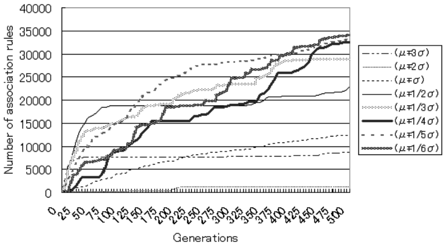

Figure 1: Number of rules extracted in the pool in Simulation 2.

The proposed system can evolve GNP and threshold parameters by an evolutionary method and measures the significance of associations via the chi-squared value. Extracted association rules are stored in a pool all together through generations in order to find new important rules. The results showed that the proposed method extracts the important association rules in the database effectively. We are currently studying a data mining method for continuous attributes using Genetic Network Programming and fuzzy membership functions.

[*] 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.