Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 32 (2014) 596 – 603

Procedia Computer Science

The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014)

Espresso for Rule Mining

Elhoussini F. Ashmouni[a], Rabie A. Ramadan[b], Ali. A. Rashed[c] *

[a,c] *Systems and Computers Department , Alazhar University, Cairo, Egypt*
[b] *Computer Engineering Departement , Cairo Univerisity , Cairo , Egypt*

Abstract

The Rule-based systems generate many of the redundant rules. Such rules are expensive especially in online systems. Currently, there are many of the available rule minimization techniques; however, they still suffer from many challenges in exploiting parallelism, load balancing, efficient memory usage, minimization of communication cost, efficient data, task decomposition and others. This paper introduces a new approach for minimizing association rules based on the adaptation of Espresso algorithm, used in reducing Boolean expressions. We believe that our proposed method is a simple and efficient method that supports a large number of input and output variables. The proposed method starts by the usage of binary encoding followed by the minimization. In the last step, data decoding is utilized generating the final rules. Such rule minimization could be used in many applications including the Wireless Sensor Networks collected data. For testing purposes, a car data set has been used and the results seem promising compared to the original rules

Keywords: rule association mining ; Car evaluation dataset; WSNs

1. Introduction and Overview

There is a dramatic exponential increase in the amount of information or data being manipulated, in which mining for data becomes essential to support decision making. Data mining is a key step in the knowledge discovery process, where the main tasks are divided into Predictive or Descriptive. Predictive in terms of predicting the value of a particular attribute based on the values of other attributes or extracting previously unknown and useful

* Corresponding author. Tel.: +2-012-276-19419; +2-010-676-43227; fax: +2-02-353-70273.
E-mail address: rabie@rabieramadan.org ; housainy@gmail.com

doi:10.1016/j.procs.2014.05.465

information such as patterns, associations, changes, anomalies and significant structures, from large datasets. There are several techniques satisfying these objectives of data mining such as clustering, classification, association rule mining, sequential pattern discovery and analysis. this paper focuses on Association Rule Mining for knowledge discovery where a large number of current systems are based on association rules including the WSNs. Rules are usually generated from different set of items and can be expanded dramatically where the increase of the item set frequency increases the number of the association rules in a proportional or exponential relationship.

The mining for association rules between items in large datasets is an important area of data mining research[20]. Rules can be manipulated to discover unknown relationships and produce the results that provide a basis for forecasting and decision making. A number of association rule mining algorithms and techniques have been developed in the last few years[1, 17, 13, 15, 5, 19]. Such algorithms can be classified into two classes: 1) candidate generation/test approach such as Apriori[2] and 2) pattern growth approach[5, 19]. However, both approaches suffer from the following drawbacks:

1) Inefficiency in reading large datasets (where it was designed in forms of several passes so that the whole dataset needs to be read several times); consequently the performance will be affected.
2) A "try and error" approach may be used to get suitable number of rules where there is no clear guide for the suitable constraints' settings like the support and the confidence. This is very time consuming and inefficient.
3) The algorithms generate a large number of association rules. Most of the time it is not easy to understand or validate such large number of complex association rules. Many of these rules are redundant and/or irrelevant and thereby limiting the usefulness of the data mining results. Again, this is very time consuming and inefficient.
4) Some of the algorithms are based on constructing a decision tree in which it is not efficient to be constructed for large data.

Therefore, one of the main challenges in mining association rules is to reduce such number of rules. This paper introduces a new approach for association rule minimization based on the basics of Boolean algebra and the digital logic minimization. The approach starts by using encoding to compress the data; then it adapts the Espresso heuristic logic minimizer simplification method for the rule minimization. The paper proposes an efficient binary encoding method for data/database compression. Thus, the data size is reduced and its scanning time is minimized. Moreover, our proposal handles multiple output states with high performance.

Due to the limited space, the following subsections contain an overview on the most related association rules techniques to our proposal including Quine-McCluskey and Espresso as well as some other heuristics[11, 18, 3, 16].

1.1. Quine–McCluskey

Quine-McCluskey is a tabular method to find all prime implicants systematically. The method starts with the truth table for a set of logic functions. By combining the minterms for which the functions are active (the ON-cover) or for which the function value is irrelevant (the Don't-Care-cover or DC-cover), a set of prime implicants is composed. Finally a systematic procedure is followed to find the smallest set of prime implicants that the output functions can be realized with. However, Quine-McCluskey is considered as an exact method for logic minimization which takes a very long time when large number of input and outputs are used. Certainly, such computation time will not be suitable for real time machines that use such method. On the other hand, the method calculates all prime implicants to derive the optimal solution(s) and derives all covers to determine minimum cover set(s). Unfortunately the number of prime implicants grows quickly in which the solution space becomes huge. Finding the minimum cover set in a class is NP-Hard problem where determining optimal solution is difficult. For majority of application, there is a need for moving to heuristics. In[11], Quine-McCluskey method is introduced for rule minimization. New encoding technique is used allowing rule reduction. However, due to the method restrictions, it is not suitable to be used in large data. In order to avoid the restriction to the number of variables, output functions and product terms of a combinational function, Espresso algorithm could be a suitable heuristic to do so which is our proposal in this paper.