

Enabling End-User Datawarehouse Mining
Contract No. IST-1999-11993
Deliverable No. D17.3B

Evaluation report by NIT

Janusz Granat, Wieslaw Traczyk, and Cezary Chudzian

National Institute of Telecommunications, Szachowa Str. 1, 04-894 Warsaw
{J.Granat, W.Traczyk, C.Chudzian}@it1.waw.pl
<http://www.it1.waw.pl>

February 27, 2003

Abstract

This document summarizes our experiences with *MiningMart* system. It is complementary with deliverable 17.2B presenting a case study developed as the National Institute of Telecommunications contribution to IST Project *MiningMart*. Our case study concerns a problem of selection of subgroup of customers that are believed to be potential subscribers of a service offered them by a telecom company via a call center. Success criteria achievement is discussed here along with some performance results and additional remarks.

Chapter 1

Evaluation

One of the most important objectives was supporting end-user view of a knowledge discovery task (the subtitle of the project is *Enabling End-User Datawarehouse Mining*). This goal has been achieved since *MiningMart* is:

- end-user oriented terminology of steps, concepts, features, etc.
- base of best-practise cases
- graphical interface providing the user the ability to model, edit and reuse cases
- almost self-documenting

With *MiningMart* the user is not forced to deal with the database relations and SQL queries. His task is just to formulate a business problem and to express it with well-defined semantics of *MiningMart*, that is closer to his way of thinking about the problem.

The good illustration of how the *MiningMart* supports end-user view is our talk with colleagues from the lab, not involved in the project, who wanted to know what the *MiningMart* and the case we are working on are. After brief introduction to *MiningMart* fundamentals and presentation of some of the pre-processing operators, we tried to explain our case model. With the *MiningMart* terminology it was quite easy to show them what is going on in call center case and we felt well-understood. Even in spite of the fact we have not made use of the system itself, we just explained it in words.

It is enough to compare figures below to have good imagination on understandability of *MiningMart* models. The figures represent approximately the same phase of pre-processing. In the figure 1.1 one can see *SAS 4GL* code responsible for this stage. Figures 1.2, 1.3 are how the *MiningMart* user sees the same thing. Difference is significant.

```
data mm.decattrib;
  merge mm.list mm.ccxam;
  by caller;

  format dec $2.;

  if one ne . then dec=one;
  else dec=decision;
run;

data mm.miningtable;
  merge mm.decattrib mm.allstatistics;
  by caller;

  if dec eq . then dec = 0;
  if NumOfIntCalls = 0 then AnyIntCalls=0;
  else AnyIntCalls=1;
  if NumOfIntCallsPH = 0 then AnyIntCallsPH=0;
  else AnyIntCallsPH=1;
  if NumOfIntCallsNPH = 0 then AnyIntCallsNPH=0;
  else AnyIntCallsNPH=1;
  if NumOf0700Calls = 0 then Any0700Calls=0;
  else Any0700Calls=1;
  if NumOf0700CallsPH = 0 then Any0700CallsPH=0;
  else Any0700CallsPH=1;
  if NumOf0700CallsNPH=0 then Any0700CallsNPH=0;
  else Any0700CallsNPH=1;
  if NumOf080Calls=0 then Any080Calls=0;
  else Any080Calls=1;
  if NumOf080CallsPH=0 then Any080CallsPH=0;
  else Any080CallsPH=1;
  if NumOf080CallsNPH=0 then Any080CallsNPH=0;
  else Any080CallsNPH=1;
run;
```

Figure 1.1: Pre-processing with *SAS 4GL*

The *MiningMart* pre-processing operators are well-defined, so the chain of them is almost self-documenting. Additionally each case, step, concept, feature may be described in free-text form. Business layer documentation

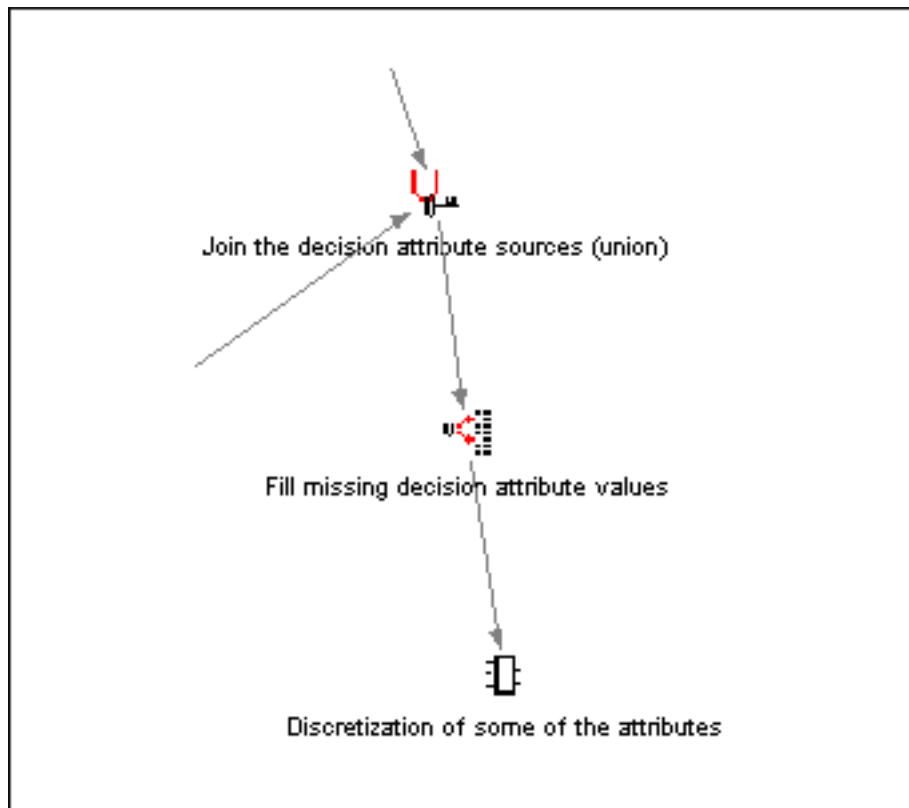


Figure 1.2: MiningMart steps

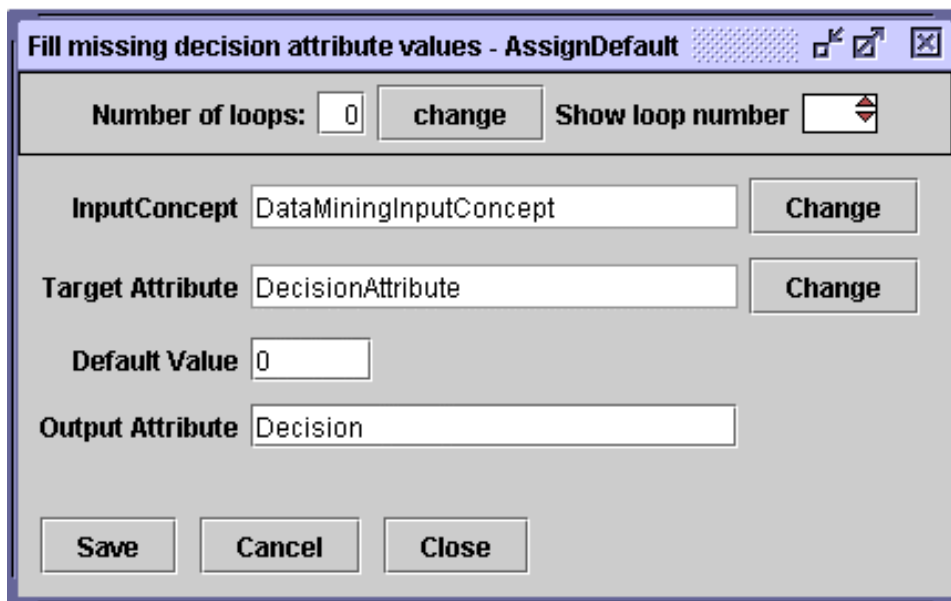
connected with the case and accessible via web interface is also valuable source of information and moreover a good start point for the user. It places *MiningMart* in contrast to other tools used for pre-processing which often suffer from the lack of any documentation.

Structure of the operators, associated conditions, constraints and assertions are helpful while tracking proper steps construction and execution. Each step may be validated at any time with respect to conditions, constraints or assertions. Thanks to it user can avoid mistakes often arising while coding manually.

Chains stored in database, validated and correct, may be used again in further modeling.

Automatization of sampling, discretization, feature selection or aggregation, along with multistepability and loopability are the key to shorten the time of modeling, since they are typical transformations needed in order to solve a number of problems.

The factors mentioned above have an influence on both the speed-up of the KD process and its quality. It is obvious that quality of data mining results depends strongly on quality of the data itself.



Fill missing decision attribute values - AssignDefault

Number of loops: Show loop number

InputConcept

Target Attribute

Default Value

Output Attribute

Figure 1.3: Step definition

One of the primary objectives of *MiningMart* was integration of operators into the database. They work directly inside the datawarehouse as SQL queries and stored procedures (PL/SQL and Java). It allows to minimize the amount of data kept in KDDSE and to take advantage of DBMS capabilities.

The call center case has been developed in parallel with the system. We can tell about it as of an iterative process. After modeling phase we had given our feedback to system developers and they introduced changes or additions we needed to the *MiningMart* software. As the consequence, we cannot say that we managed to prepare the case study in some days or weeks. Comparison with the previous *SAS 4GL* modeling would be difficult in this context.

Chapter 2

Performance issues

Here we would like to present results of the tests we made in order to assess performance of *MiningMart* system. Whole the pre-processing chain of call center case was fired with business data sets differing in number of customers being characterized with the data and overall number of records (CDRs). Experiments were held with all the software needed to run *MiningMart* (including Oracle9i and JBoss servers), installed on Toshiba Satellite 5100-503 notebook with Pentium IV 1.8 GHz processor and 512 MB of RAM memory.

num of CDRs/num of cust.	pre-processing time
3/455	00:01:07
19/2556	00:04:29
35/4845	00:09:13
50/6878	02:17:22
75/9626	04:34:20
101/12742	07:24:18

Figure 2.1: Performance tests results

We suppose the main factor that decided on the time of pre-processing was number of considered customers, since it influences directly the amount of database objects created during the chain compilation.

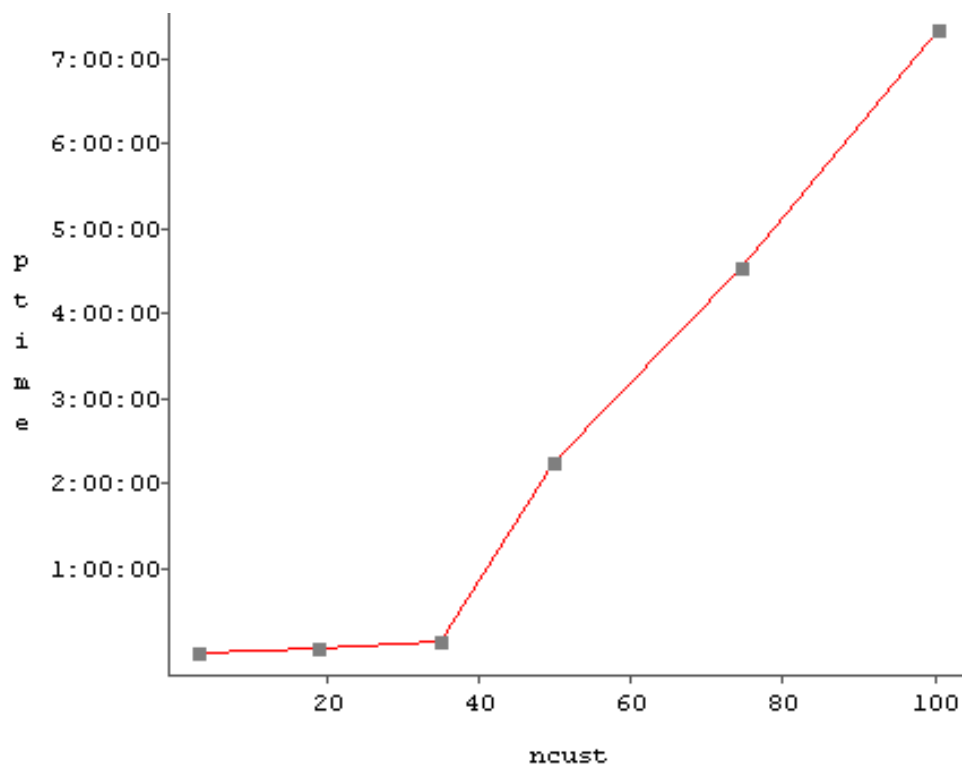


Figure 2.2: Performance tests visualization

Chapter 3

Summary and further enhancements

The main value of MiningMart system for case developers:

1. Conceptual modeling improves:
 - the understanding of the data preparation process;
 - maintenance of the data preparation process;
 - knowledge transfer to other people.
2. There is no need to use programming language.

We have some suggestions, coming from our experiences with *MiningMart*, concerning further enhancements of the system.

We believe that optimization of pre-processing operators and chains should considerably increase efficiency of the compiler.

Definition of interfaces for interaction with external tools (data mining, visualization) could improve functionality of *MiningMart*. It could be also possible to use external tools to perform some parts (subchains) of pre-processing task, especially in the case of operations that are not covered by *MiningMart* set of operators.