Enabling End-User Datawarehouse Mining
Contract No.: IST-1999-11993
Deliverable No. D17.0

Mining data with the MiningMart system – Evaluation Report

Marco Richeldi and Alessandro Perrucci

Telecom Italia Lab
Via G. Reiss Romoli 274 – 10148 Torino, Italy
{marco.richeldi, alessandro.perrucci}@tilab.com

February 26, 2003

Abstract

This document summarizes the experiences of TILAB with the MiningMart system and compares knowledge discovery with and without MiningMart.

TILAB applied MiningMart to perform a real-world discovery task, namely, predicting the likelihood of customers of a mobile TLC operator to become churners. This churn analysis system is based upon data mining technology and sifts through the customer database to model customer behavior and predict customer turnaround.

TILAB's evaluation of the MiningMart is positive. The MiningMart system meets the success criteria that were defined at the beginning of the project, i.e., efficient support of pre-processing tasks, pre-processing speed up, user friendliness and scalability. The MiningMart system makes it possible to strongly improve the quality of data mining output.

1 Introduction

The goal of the MiningMart is to develop an environment for the support of knowledge discovery from databases (KDDSE) that provides decision-makers with advanced knowledge extraction from large distributed data sets.

Within the project framework, new techniques for selecting and constructing features on the basis of given data have been developed that are supposed to ease knowledge discovery where currently most time is spent in pre-processing. MiningMart let domain knowledge be exploited during the data mining task, in order to enhance the quality of data mining results. In addition, a case-base of discovery tasks together with the required pre-processing techniques offers an adaptive interface to the KDDSE. This should speed-up similar applications of knowledge discovery and make the KDDSE self-improving.

Telecom Italia Lab (TILAB) carried out extensive experiments to verify if MiningMart project fulfilled its objectives. Project results were applied to perform a real-world discovery task that was accomplished by TILAB in the past by using a number of pre-processing and predictive modeling technologies. The case study concerns developing a Churn Analysis system based upon data mining technology to analyze the customer database of a telecommunication company and predict customer turnaround.

Evaluation results are definitely positive and are presented in the next chapters of this documents. Indeed,

- Speed up for most of the pre-processing tasks increased by at least 50% by using MiningMart.
- Power users will find MiningMart system as much easy to use as the leading commercial data mining platforms (e.g., SAS Enterprise Miner or SPSS Clementine).
- MiningMart makes it possible to build libraries of predefined data mining applications that can be easily reused and customized to fulfill specific needs.
- MiningMart features the highest scalability, since it builds upon and fully exploits state-of-the-art database technology.
- Quality of data mining output increased as domain knowledge may be exploited during the discovery process and the number of preprocessing trials decreased in number.

The rest of the document is organized as follows. Chapter 2 and Chapter 3 present the metrics and the case study used to assess MiningMart, respectively. Chapter 4 describes the results of the evaluation task.

2 Evaluation criterias

Goal of the MiningMart project was that knowledge discovery becomes a powerful, but not difficult, query facility for very large databases. End users should ideally be able to run application-driven queries against large and heterogeneous data sets in a simplified language.

In order to make knowledge discovery a powerful and easy query facility for very large databases, the current tools needed to be enhanced in the following ways:

- Supporting advanced, partially automated pre-processing of data.

- Supporting the view of the end-user by a case base of pre-processing and analysis tasks for re-use.
- Reducing the number and complexity of trial and error pre-processing and analysis cycles for each discovery task.
- Minimizing the amount of data that is kept within the data mining operators.
- Improving the quality of data mining results by improving the quality of data

The success criterion that was set for the overall project was that by the end of the project, some discovery tasks (for which entries in the case-base exist) could be solved with only 20% of the time for pre-processing, where the time for the data mining step remained the same as before the project.

Further, the measures listed below were considered to enable in-depth evaluation of MiningMart systems:

- Measure to creating user-friendly access to data mining for non-expert users.

  Is MiningMart easy to use, transparent, and supportive? Is the flow of control natural to the user? Can the user make good use of the results? How do the user assess the results? Is the user faster when using the system? Can he/she do more when using the system?
- Speed-up the discovery process.

  The time for finding the appropriate data transformations with and without MiningMart has to be measured.
- Minimize the amount of data kept within data mining operators

  The aim is to perform extensive processes directly within the datawarehouse. A clear and operational measure of success is whether a huge datawarehouse can be handled by the MiningMart.
- Improve the quality of mining by improving the quality of data.

  The main measure for the quality of data mining results is their accuracy on test data. Mining results with and without MiningMart have to be compared.

3 The Case Study

Evaluation of MiningMart was performed by carrying out a difficult knowledge and data discovery task concerning understanding the behavior of customers of a telecommunication company that unsuscribe and move their business to a competitor and predicting the likelihood of this event to occur. Customers who turnaround are called churners and the process of exploring and modeling their behavior is named Churn Analysis.

The Churn Analysis case study was selected as the massive volume and complexity of data collected by telcos about their customers and the difficulty of modeling this data make churn analysis a very good test bed for evaluating MiningMart features. In addition, TILAB developed several Churn Analysis systems for different companies of the Telecom Italia Group. The knowledge and data discovery process required to build effective churn prediction models has been widely explored along the years, and it reveals quite easy to assess how MiningMart supports it.

4 Evaluation

As mentioned in the chapter “Evaluation Criterias”, the evaluation task focussed on the following topics:

- Usability.
- Mining process speed-up.
- Scalability.
- Mining process quality.
- Integrability (into the business processes).

4.1 Usability

MiningMart’s Human Computer Interface builds upon a state-of-the-art visual programming approach. Preprocessing chains may be created just dragging operators icon into the application framework and adding links to chain them. This is a very effective manner to perform discovering tasks. (Preprocessing chains are sequence of preprocessing operators that depend on each other. They are of paramount importance since while preparing raw data several transformation steps are necessary, where the output of the previous operator becomes the input of the one which follows in the chain).

The interface look-and-feel is quite good and it reveals user-friendly. Few steps are required to implement any data mining process, and meta data helps the user in selecting the right ones. The flow of control is rather clear and well documented.

Interface quality compares to the one of leading commercial tools (e.g., SPSS Clementine, SAS Enterprise Miner).

With MiningMart, best practices can be easily pre-packaged. Indeed, the tool allows each user to store entire chains of pre-processing and analysis steps for later re-use in a case-base (for example, a case of pre-processing for churn analysis, or a case of pre-processing for monthly business reports). This way, libraries of data mining applications may be developed and customized to satisfy new business requirements.

Further enhancements

Should MiningMart be enhanced, we suggest putting some effort to further simplify the Definition of Concepts stage, for example by making it possible to define new attributes by directly editing table column names.

4.2 Mining process speed-up

Less trial and error pre-processing and analysis cycles are required to develop the data mining solution by using MiningMart. Indeed:

- Meta data and operator constraints drive unskilled users to build correct and effective analytical applications;
- The case-base of pre-processing and analysis tasks assists in the exploitation of experienced guidance from past successful applications. More, when analysis tasks are repeated regularly, the case base of stored discovery tasks can be used and the same steps need not to be repeated.

  In the churn analysis case study, the preprocessing stage took 2.5 days by using MiningMart (2 data miners working full time). The same tasks was carried out in 12 days by using standard sql-based preprocessing tools.

A saving of almost 80% time was obtained.

- Users achieve a better understanding of data underlying structure by browsing source and processed data while computing descriptive statistics and transforming data.
- Processing chains can be quickly tested during chain set-up.
- Processing chains can be saved an restored, allowing versioning.
- Multistep and loopable operators enable users to define parallel mining tasks consistently and effectively.

Further enhancements

Should MiningMart be enhanced, we suggest developing more powerful graphical investigation features and improving workgroup enabling features, e.g., multiple users capabilities, definition of user roles and access rights.

4.3 **Scalability**KDD systems that have to load all data for data mining into main memory cannot handle large data volumes. MiningMart provides a new work-share between datawarehouse and KDD operators so that it can minimize the amount of data kept within KDDSE and perform extensive processes directly within the datawarehouse.

This claim was verified by running the pre-processing task described in section **Fehler! Verweisquelle konnte nicht gefunden werden.** on datasets of increasing size.

MiningMart server (compiler and data base) was installed on a SUN E450, 4 400MHz CPUs and 4Gb RAM. Raw data tables were stored in a Oracle 9.1 data base with table partitioning. Data sets of 8000, 800,000, and 2,000,000 records with 20 attributes were created. Notice that 40 more new attributes are created during the construction stage aimed at increasing the information power of data.

The time required to perform the pre-processing task is shown in the table below. As it can seen, MiningMart scales linearly. This is due to its fully leveraging on Oracle scaling capability. Overhead due to parsing of operators is negligible, unless for very small datasets.

data set size (# of tuples)	Pre-processing time (mins)
8,000	60
800,000	600
2,000,000	1200

Conversely, modeling operators currently implemented in MiningMart have not been optimized, so they scale exponentially.

4.4 **Mining process quality**Transforming raw data into a high-value basis for discovery is a time-consuming task, but it is also the most important step in the knowledge discovery cycle. and a particular challenge in real world applications. •

MiningMart makes available a set of transformation tools/operators to ease this task. In particular, a number of tools are aimed at assisting the user in selecting appropriate samples, aggregating data to change its level of detail, discretizing numeric data, and reducing the dimensionality of data. and changing the level of detail of the data by means of aggregation operators, according to the task and/or the algorithm used. By automating these time-consuming tasks, the likelihood of doing processing mistakes decreases.

Similarly, MiningMart framework avoids potential conceptual mistakes by enforcing chain consistence and correctness.

In addition, quality of mining results improves as MiningMart makes it possible to build libraries of data mining applications that can be tailored to meet new business requirements.

However, the main measure for the quality of data mining results is their accuracy on test data. Consequently, a comparison was done in the framework of the case study between the predictive performance of MiningMart with the one obtained in the past using alternative mining technologies. The predictive accuracy of the model generated by MiningMart is slightly less than the one of the original model (79% vs. 82%). This is a very good result since MiningMart implements just a basic version of a source-free decision tree algorithm.

4.5 Integrability into the business process

The MiningMart system can be integrated into an Analytical CRM platform as the analytical extension of either the enterprise data warehouse or the business-oriented data marts.

5 Conclusions

Using a data warehouse for decision support or applying tools for knowledge discovery are difficult and time-consuming tasks. Pairing of data with algorithms and effectively pre-processing of the data are still a matter of trial and error.

We believe that applying the MiningMart system we can reduce by one third the effort spent on the pre-processing task of the knowledge discovery process. As pre-processing usually takes the 80% of the time required by the overall discovery project, the MiningMart system may reduce the project time by almost the 25%.

This estimate has been proved correct by the experimental study that we carried out. In summary, we built again a churn prediction system that we fielded for a mobile operator in the past.

MiningMart allowed us to save 50% of the effort spent in pre-processing the data (i.e., finding an appropriate transformation of the given data, finding appropriate sampling of the data, etc.) and to achieve the same predictive performance.

Indeed, the knowledge discovery task was greatly streamlined, as the number and complexity of trial and error pre-processing and analysis cycles was dramatically reduced. Moreover, the re-use of pre-defined building blocks made it possible to speed up the entire analysis task and to minimize the number of conceptual and development mistakes.