MINING MART

Enabling End-User Datawarehouse Mining
Contract No. IST-1999-11993
Deliverable No. D17.2B

Call Center Case

Cezary Chudzian, Janusz Granat, and Wieslaw Traczyk

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

February 21, 2003

Abstract

This document presents **the Call Center Case** developed by National Institute of Telecommunications. This case has been developed for Mining-Mart system and is focusing on selecting the group of prospect clients for targeting marketing campaign. The call center is used as a communication channel with the clients. We are presenting a business problem, the data sources and preprocessing steps.

Chapter 1

Introduction

Dynamic growth of the telecommunications market forces telecommunications operators to take up necessary actions, leading to improve their competitiveness in the market. The business departments are supported by various information systems. Customer Relationship Management (CRM) systems become one of the most important tools of the modern telecommunications company. The data mining module is one of the essential component of such system.

The CRM process consists of four main phases:

- Acquisition and analysis of clients data
- Marketing planning
- Communication with client
- Analysis of the marketing action

In the above process the correct data is a solid base for efficiency and effectiveness of the CRM process. The data preparation is one of the most time consuming task and needs dedicated tools. MiningMart project goal was to provide software tools for advanced pre-processing of data. This paper presents a case that has been prepared for verification of MiningMart system as well as one of the examples of using MiningMart, stored in the case base of MininMart system.

We restrict our case to one communication channel (the call center) with the clients and we focus on selling services by this channel. The process of selling services by call center is as follows:

- a group of potential service subscribers is selected manually by the telecommunications specialists (*primary selection* - figure 1.1); the selection is based e.g. on invoice data but other rules are possible;

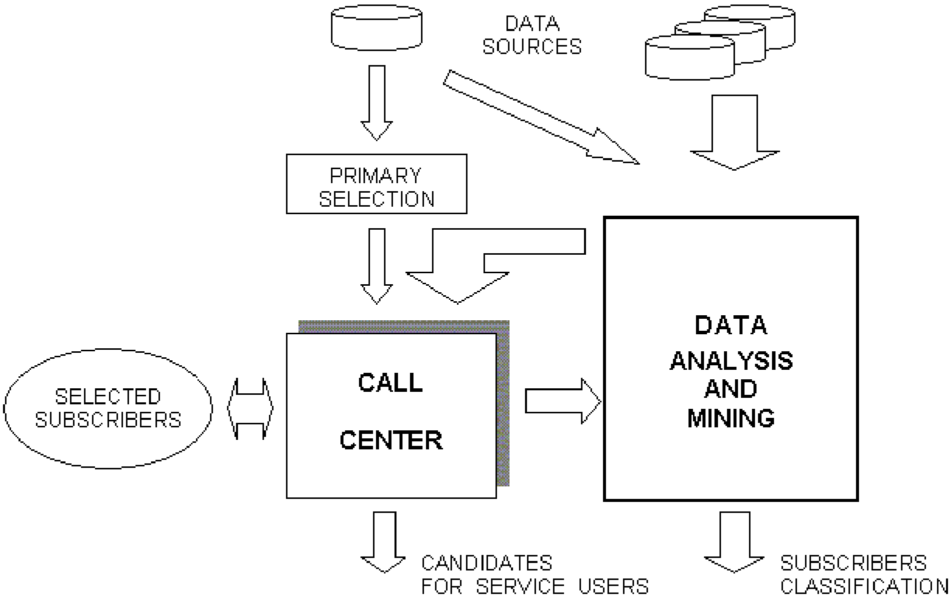

Figure 1.1: Data flow in the classification system

- these candidates are contacted by call center agents; the answers of the potential clients are stored in the data base;
- data stored in the call center, the list of real service users and CDR (Call Detail Records) are the data sources that are used for building data mining table; this table is an input for data mining algorithms;
- results of mining - the list of potential candidates for buying the service - is sent to the call center;
- the clients from the new list are contacted by an agents and a new responses data are stored
- the new data are used once again for building a data mining table and modeling; and this steps can be repeated as long as the list of candidates is not empty.

Simplified model of the call center is shown on the figure 1.2. We can distinguish there the input list of a clients that will be called. The call center agent calls potential client and asks him whether he is interested in a service subscription. The answer is coded as an integer number and kept in the database for further analysis:

- client is not interested in service subscription (code -1)
- client is already service subscriber (code 1)

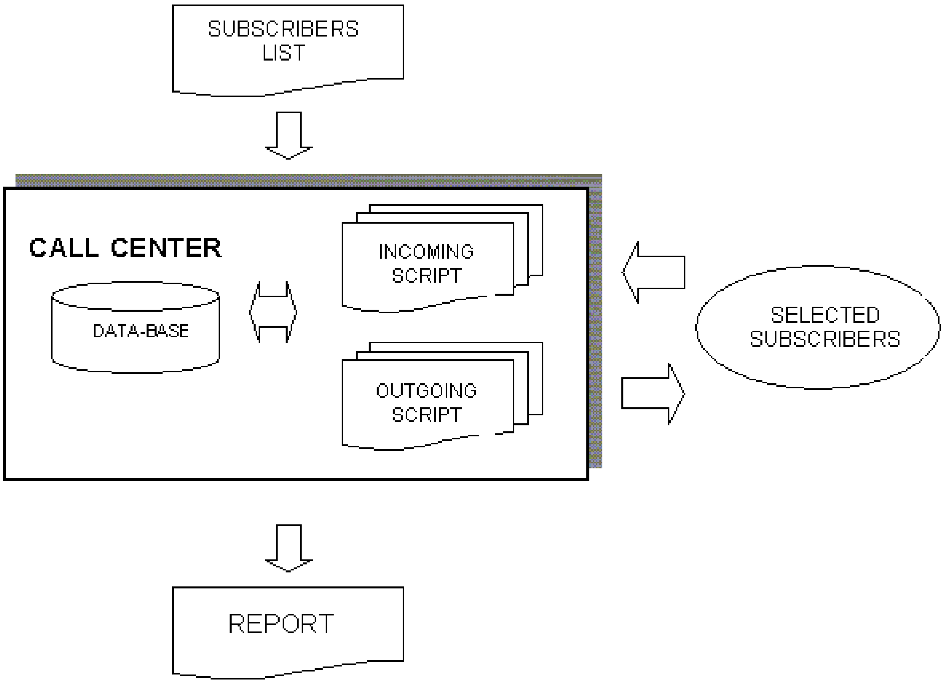

Figure 1.2: Simplified model of call center

- client agrees on service activation (code 2)

Additionally the company is storing the list of people being subscribers of the service. This source is even more reliable than the results of the call center review. Sometimes it happens that questioned people do not deliver real information. Call center questioning results will serve as description of the telecom company's clients preferences on offered service. In the case of the person which cannot be found in any of these data sets, we assign code 0, that means ,,unknown preference".

Each client's profile is created on the basis of data acquired from company's billing system. It is very typical problem of transformation clients activities in past (here we have Call Detail Records (CDR)) into their behavioral description. There are many CDR forms, but all of them contain such information on single call as:

- caller number
- called number
- timestamp
- length of the call

- tariff units per call

We use also class attribute, indicating type of call (e.g. local, long distance, etc.) and disconnection state (may be one of: properly disconnected, improperly disconnected, not answered).
Taking into consideration different tariffications between 8 a.m. and 18 p.m. and between 18 p.m. and 8 a.m. we distinguish calls performed within peak hours and within non-peak hours.

In our model, a client is characterized by the following set of attributes:

- number of calls
- number of calls within peak hours
- number of calls within non-peak hours
- number of different called parties
- number of specific type connections
  - local
  - long distance
  - abroad
  - cellular
  - special numbers prefixed with 0700 (party lines, etc.)
  - special numbers prefixed with 080 (info lines)
  - internet provider (dial-up modem connections)
- number of specific type connections (as above) within peak hours
- number of specific type connections (as above) within non-peak hours
- total length of calls
- total length of calls within peak hours
- total length of calls within non-peak hours
- sum of tariff units
- sum of tariff units for peak hours calls
- sum of tariff units for non-peak hours calls
- number of properly disconnected calls
- number of improperly disconnected calls
- number of not answered calls

Chapter 2

The data model

Three concepts are introduced at the beginning. *CallData* corresponds to data set containing CDRs. *ActualClients* has underlying database table holding the list of actual service subscribers. The results of questioning the clients by call center are present in the model as *CCExamination* concept.

At the beginning we present the class diagrams presenting data associations. Notation is based on that proposed in [1] (there are some extensions to standard UML class diagrams). We will use:

- ⟨⟨base⟩⟩ stereotype for concept having direct correspondence at the database level,
- ⟨⟨aggregate⟩⟩ stereotype for concepts holding summaries of groups in the data (like *sum*, *count*, etc.),
- ⟨⟨summarize⟩⟩ stereotype to illustrate the dependency between concept holding summarized data and the concepts the data that is to be summarized come from,
- ⟨⟨construct⟩⟩ stereotype for visualizing associations linking concepts used to construct new feature and concept holding that feature

Concepts corresponding to data sources are depicted in figure 2.1. There is one class for each concept.

Person and *Client* do not have their column sets, but we introduce them to show how the initial concepts are associated. The attribute that is used to identify particular client of the company is his phone number (*ClientNumber* or *PhoneNumber* or *Caller* in fig. 2.1).

Attribute *CallClass* of *CallData* allows us to distinguish between seven main types of calls. In the figure 2.2 one can find class connected with each type of call (*LocalCallDetails*, *LDstCallDetails*, etc.). There are also three classes for properly, improperly and not answered calls. Attributes of class *Statistics* are summaries of features of all these classes. *Caller* is *group by*

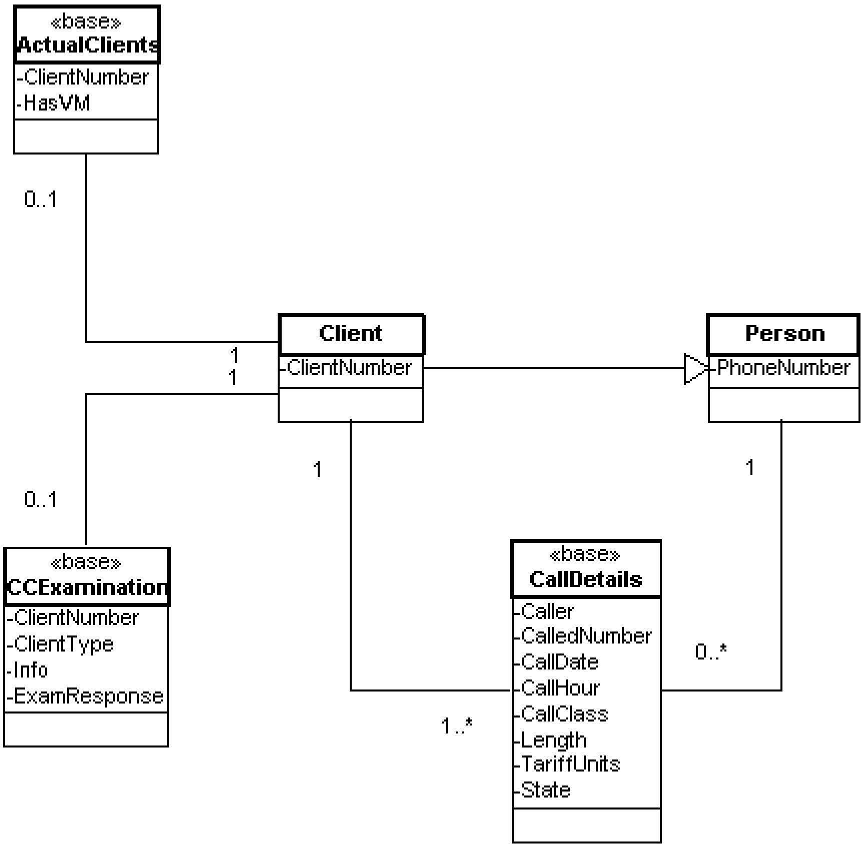

Figure 2.1: Initial concepts

attribute of *Statistics* class and as the consequence appropriate statistics are computed separately for subsets connected with clients.

The price of the tariffication unit differs in time intervals 8-18 (denoted as peak hours) and 18-8 (non-peak hours). *CallDetailsPeak* and *CallDetailsNonPeak* are derived from *CallDetails.* Additional client statistics - features are computed by the summarizations (figures 2.3, 2.4). *Statistics* states complete clients' profiles description.

Decision attribute, that is client preference on the offered service, is constructed basing on the list of current service subscribers (*ActualClients*) and an outcome of the call center review (*CCExamination*) - figure 2.5. The list of subscribers is more reliable source of an information. In the case the data is available in both of them, the value from this list must be taken.

All the statistics - clients profiles merged together with the decision attribute values - form data mining data set.

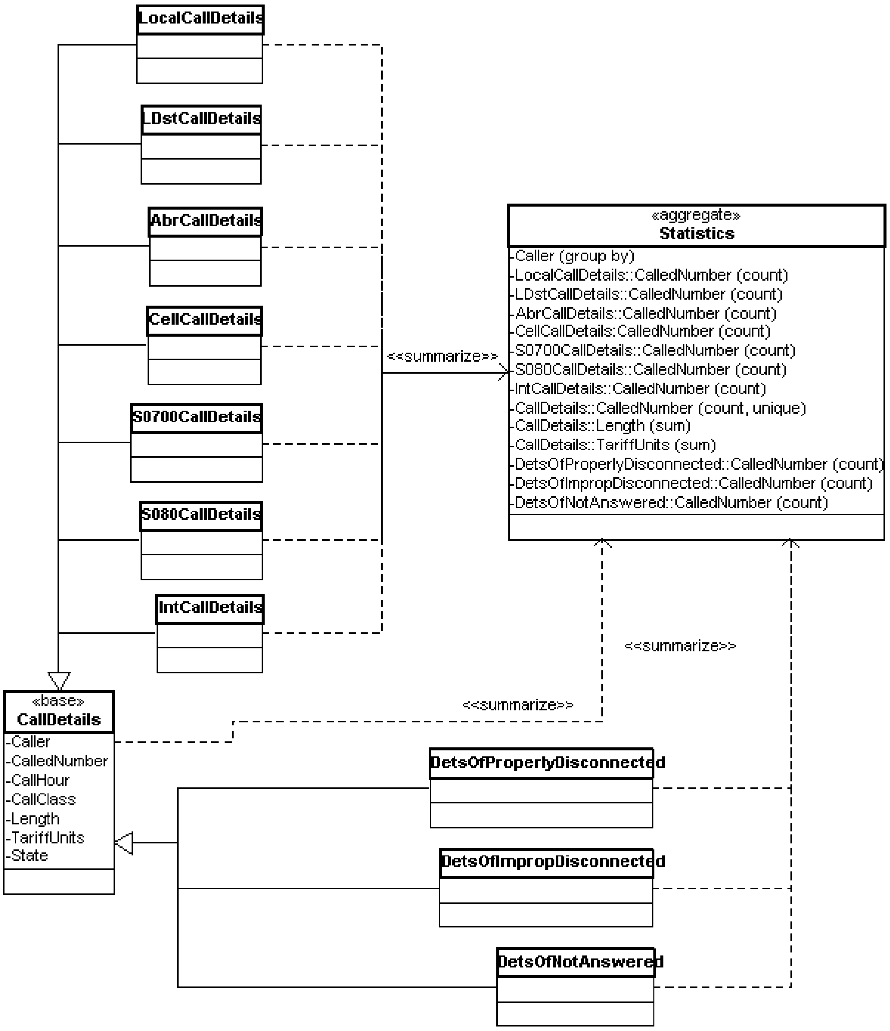
LocalCallDetails
LDstCallDetails
AbrCallDetails
CellCallDetails
S0700CallDetails
S080CallDetails
IntCallDetails
«aggregate»
Statistics
-Caller (group by)
-LocalCallDetails::CalledNumber (count)
-LDstCallDetails::CalledNumber (count)
-AbrCallDetails::CalledNumber (count)
-CellCallDetails:CalledNumber (count)
-S0700CallDetails::CalledNumber (count)
-S080CallDetails::CalledNumber (count)
-IntCallDetails::CalledNumber (count)
-CallDetails::CalledNumber (count, unique)
-CallDetails::Length (sum)
-CallDetails::TariffUnits (sum)
-DetsOfProperlyDisconnected::CalledNumber (count)
-DetsOfImpropDisconnected::CalledNumber (count)
-DetsOfNotAnswered::CalledNumber (count)
<<summarize>>
<<summarize>>
<<summarize>>
«base»
CallDetails
-Caller
-CalledNumber
-CallHour
-CallClass
-Length
-TariffUnits
-State
DetsOfProperlyDisconnected
DetsOfImpropDisconnected
DetsOfNotAnswered

Figure 2.2: Summarization of features of *CallData* (1)

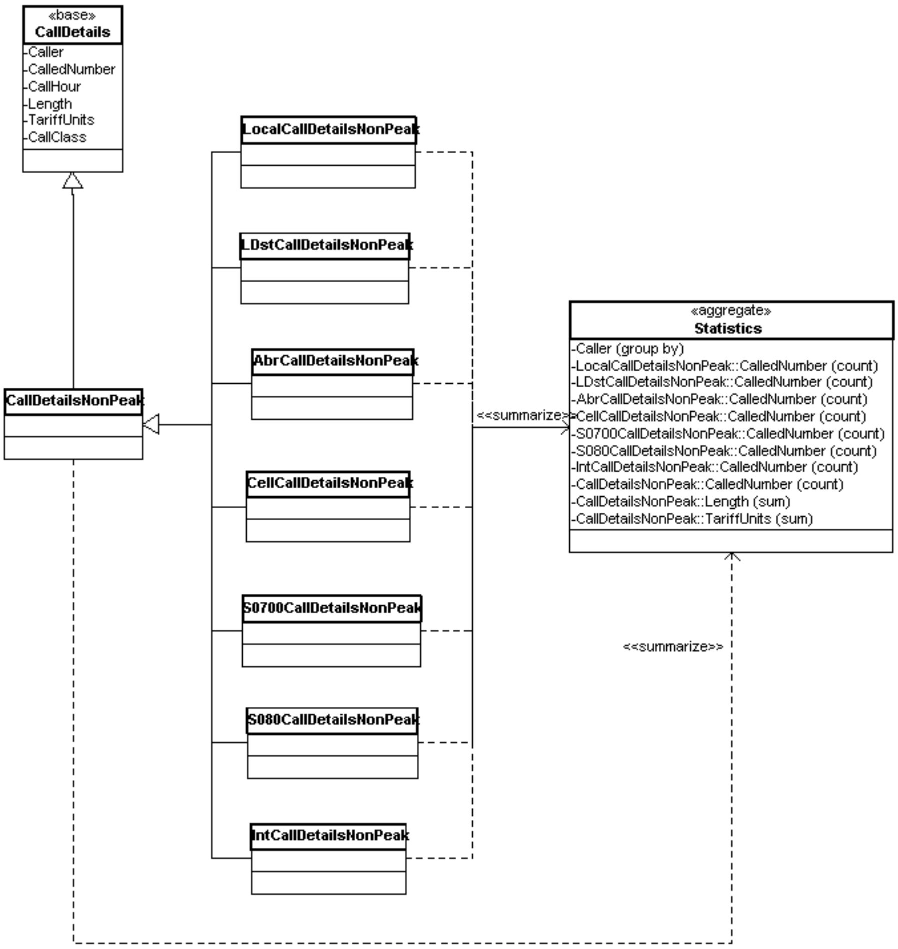
«base»
CallDetails
-Caller
-CalledNumber
-CallHour
-Length
-TariffUnits
-CallClass
CallDetailsNonPeak
LocalCallDetailsNonPeak
LDstCallDetailsNonPeak
AbrCallDetailsNonPeak
CellCallDetailsNonPeak
S0700CallDetailsNonPeak
S080CallDetailsNonPeak
IntCallDetailsNonPeak
<<summarize>>
«aggregate»
Statistics
-Caller (group by)
-LocalCallDetailsNonPeak::CalledNumber (count)
-LDstCallDetailsNonPeak::CalledNumber (count)
-AbrCallDetailsNonPeak::CalledNumber (count)
-CellCallDetailsNonPeak::CalledNumber (count)
-S0700CallDetailsNonPeak::CalledNumber (count)
-S080CallDetailsNonPeak::CalledNumber (count)
-IntCallDetailsNonPeak::CalledNumber (count)
-CallDetailsNonPeak::CalledNumber (count)
-CallDetailsNonPeak::Length (sum)
-CallDetailsNonPeak::TariffUnits (sum)
<<summarize>>

Figure 2.3: Summarization of features of *CallData* (2)

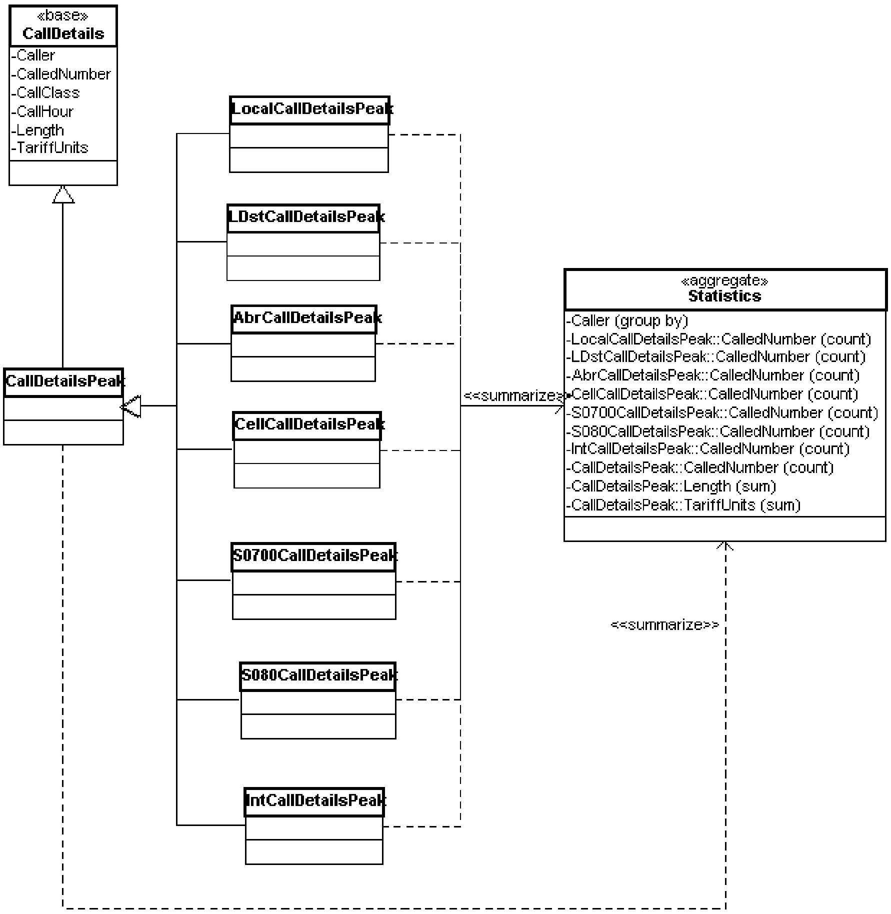
«base»
CallDetails
-Caller
-CalledNumber
-CallClass
-CallHour
-Length
-TariffUnits
CallDetailsPeak
LocalCallDetailsPeak
LDstCallDetailsPeak
AbrCallDetailsPeak
CellCallDetailsPeak
S0700CallDetailsPeak
S080CallDetailsPeak
IntCallDetailsPeak
<<summarize>>
«aggregate»
Statistics
-Caller (group by)
-LocalCallDetailsPeak::CalledNumber (count)
-LDstCallDetailsPeak::CalledNumber (count)
-AbrCallDetailsPeak::CalledNumber (count)
-CellCallDetailsPeak::CalledNumber (count)
-S0700CallDetailsPeak::CalledNumber (count)
-S080CallDetailsPeak::CalledNumber (count)
-IntCallDetailsPeak::CalledNumber (count)
-CallDetailsPeak::CalledNumber (count)
-CallDetailsPeak::Length (sum)
-CallDetailsPeak::TariffUnits (sum)
<<summarize>>

Figure 2.4: Summarization of features of *CallData* (3)

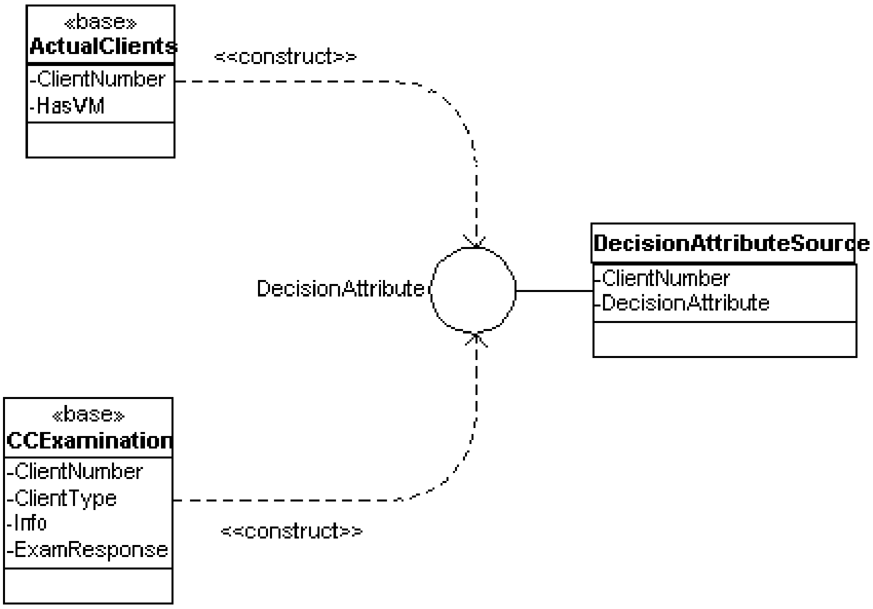

Figure 2.5: Decision attribute sources

Chapter 3

Case modeling by MiningMart

In this chapter we will present a process of building the pre-processing steps with the help of Human Computer Interface. We are expressing our data transformations in terminology of steps, concepts and features, that is specific to MiningMart modeling.

At the conceptual level we focus on just single client for calculating clients attributes. To deal with each client (or *Caller*) separately, *SegmentationStratified* operator is applied at the beginning (figure 3.1).

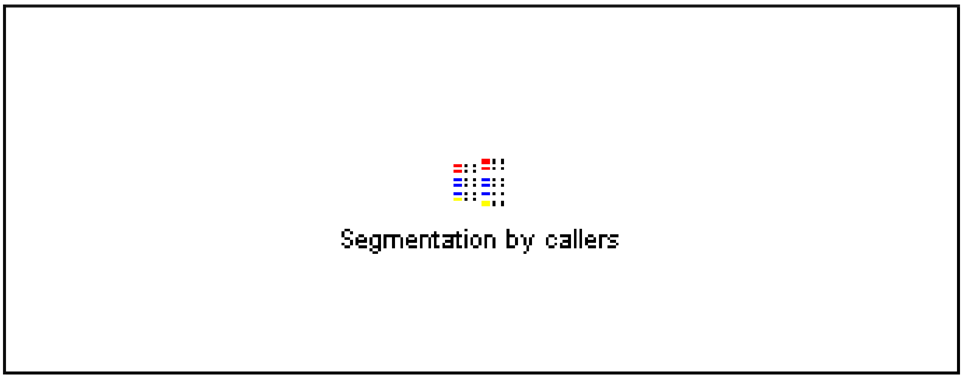

Figure 3.1: Segmentation by *Caller*s

After that the new binary attribute is constructed (according to *CallHour*) that have value ,,*peak*” for peak hours calls and ,,*nonpeak*” otherwise. We use here *TimeIntervalManualDiscretization* operator. Subsets of peak hours and non-peak hours calls are selected (figure 3.2) by *RowSelectionByQuery* operator.

Three possible values of the *state* attribute indicate three states of the call disconnection, that may be:

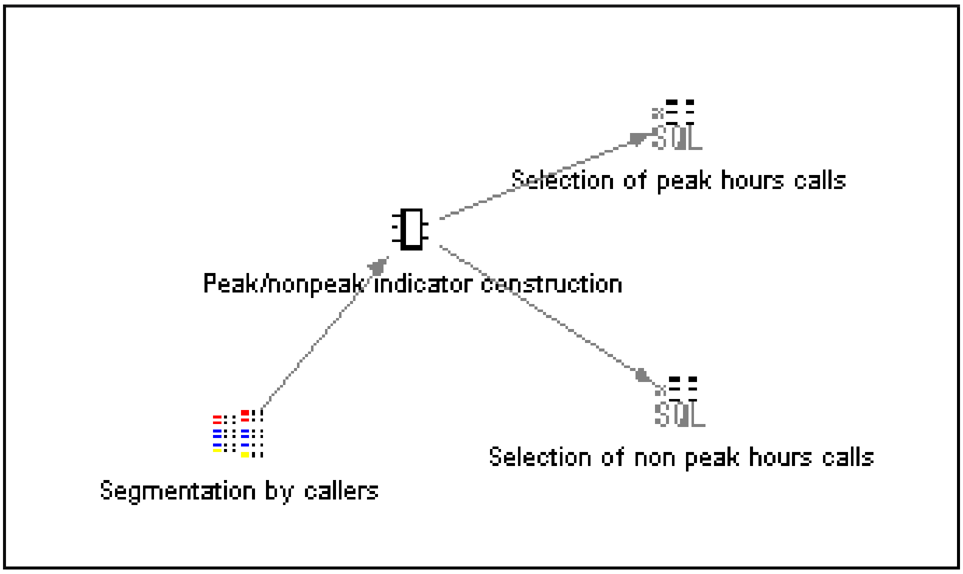
Selection of peak hours calls
Peak/nonpeak indicator construction
Segmentation by callers
Selection of non peak hours calls

Figure 3.2: Selection of peak/non-peak hours calls

- disconnected properly,
- improperly disconnected - broken because of e.g. technical reasons
- not answered by the called party.

Three subsets of calls differing in disconnection state are selected.

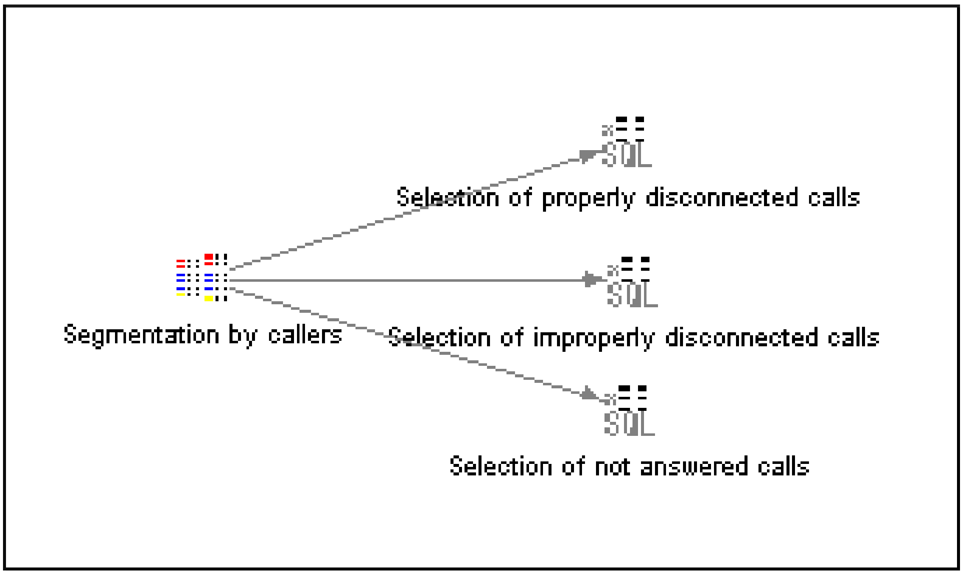
Selection of properly disconnected calls
Segmentation by callers
Selection of improperly disconnected calls
Selection of not answered calls

Figure 3.3: Subsets of calls selected according to disconnection state

Peak and non-peak hours calls are again divided into 7 subsets according to classes of calls (local, long-distance, abroad, cellular, 0700, 080, internet provider), see figure 3.4.

The steps introduced above, effects in creation of some new concepts, that enhance available information on client. We have concepts with details of client's calls within peak hours and non peak hours, specific type calls details (like local calls details), details of calls disconnected properly, improperly, not answered and so on. But still we have no client profile that we need. Interesting client's features may be now calculated with the help of *SpecifiedStatistics* operator. It may count calls, sum up durations of calls, compute total amount of tariff units, count number of different called numbers. In the figures 3.4, 3.5 one can see an application of *SpecifiedStatistics*. As the input concept we take information, mentioned above. The output concept of the step embedding the operator holds interesting, from our point of view, features of the client.

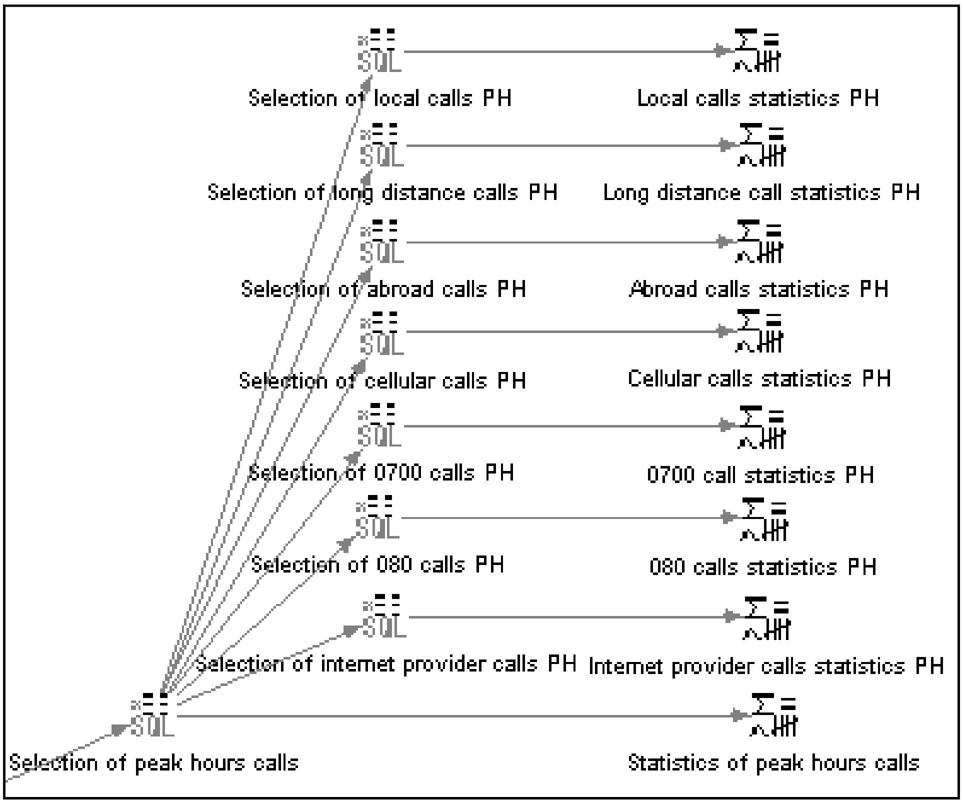
Selection of local calls PH
Local calls statistics PH
Selection of long distance calls PH
Long distance call statistics PH
Selection of abroad calls PH
Abroad calls statistics PH
Selection of cellular calls PH
Cellular calls statistics PH
Selection of 0700 calls PH
0700 call statistics PH
Selection of 080 calls PH
080 calls statistics PH
Selection of internet provider calls PH
Internet provider calls statistics PH
Selection of peak hours calls
Statistics of peak hours calls

Figure 3.4: Statistics of peak hours calls

The first step we introduced while modeling the case was *Segmentation-Stratified* operator. From that point, data set underlying to the input concept *CallDetails*, was split into subsets, one for each client. Then identical

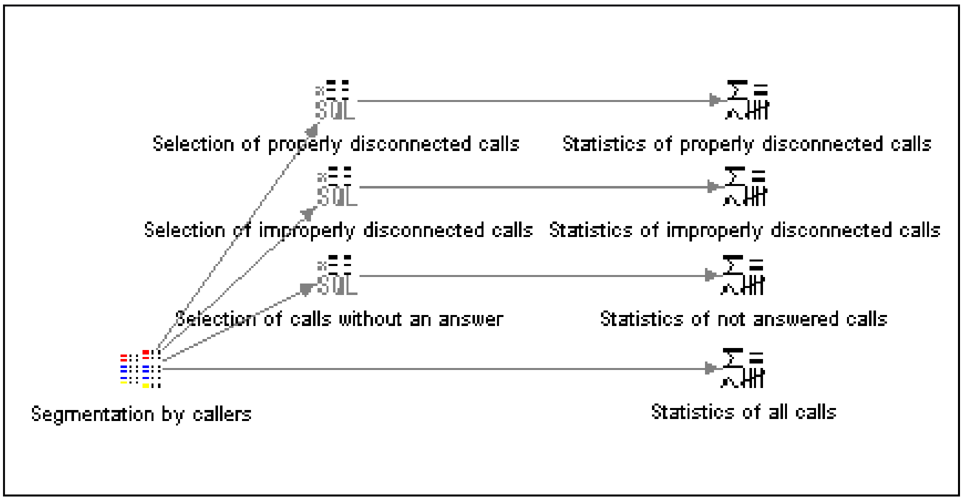
Selection of properly disconnected calls
Statistics of properly disconnected calls
Selection of improperly disconnected calls
Statistics of improperly disconnected calls
Selection of calls without an answer
Statistics of not answered calls
Segmentation by callers
Statistics of all calls

Figure 3.5: More statistics

operations in subsequent steps were performed on these subsets independently and we could simply focus on single client information processing. Now we would like to have the concept corresponding to features of all the clients. To realize this task we use *Unsegment* operator, that is inverse to any operator of type *Segmentation* (figure 3.6). Our focus changes back from one client to all the clients found in initial data set.

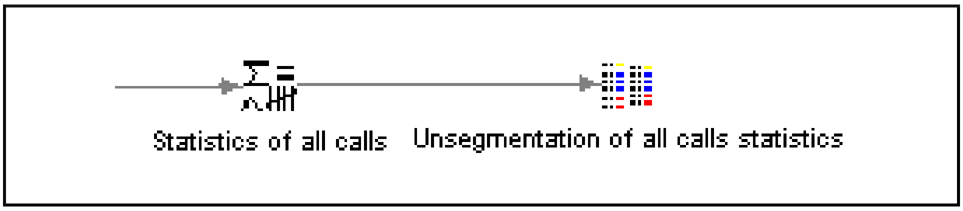

Figure 3.6: Unsegment - inverse of the segmentation

The *Unsegment* steps give as the output the total number of 20 concepts. Features of these concepts, assembled together, constitute almost complete profiles of the clients we want to achieve. *JoinByKey* operator is the best way to create a new concept with all currently available clients' features (see figure 3.7).

In order to build full profiles we need additionally attributes describing e.g. number of local calls or number of all calls, without differentiation of peak hours and non-peak hours calls. The goal is achieved here by summing up these numbers computed by *SpecifiedStatistics* operator for peak and non-peak hours calls. *GenericFeatureConstruction* is applied for this purpose (figure 3.8).

We have prepared clients' profiles. Now we can construct the decision

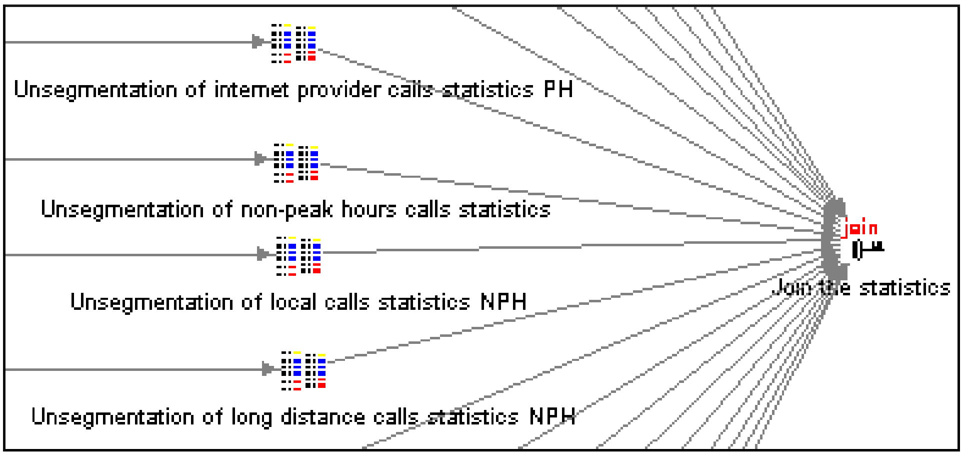
Unsegmentation of internet provider calls statistics PH
Unsegmentation of non-peak hours calls statistics
Unsegmentation of local calls statistics NPH
Unsegmentation of long distance calls statistics NPH
join
Join the statistics

Figure 3.7: All the statistics gathered

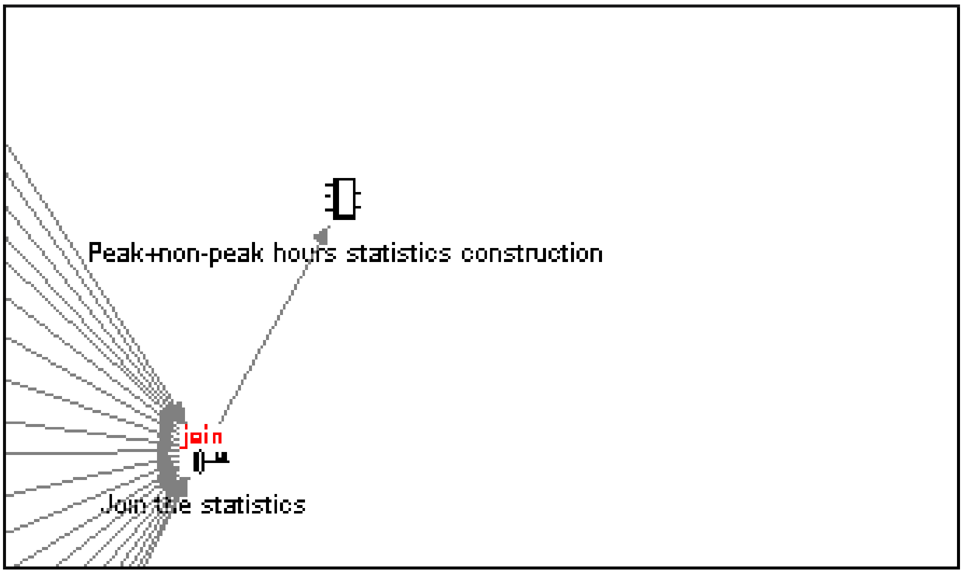

Figure 3.8: Construction of new features

attribute (figure 3.9). *ActualClients* and *CCExamination* are the available sources for the decision attribute. With the use of *UnionByKey* operator we join both the sources. The operator *JoinByKey* joins data underlying to its concepts according to key attribute values. The output column set of the operator contains only records marked with common for all its input column sets subset of key attribute values (intersection). Here we would like to have a tuple in the output columnset for every key attribute value found in any input columnset (union of keys rather than intersection).

Two attributes of the output concept *UnifiedDecAttrSources* of this *Union-*

ByKey step are now considered, namely *ExamResponse* (from *CCExamination*) and *HasVM* (from *ActualClients*). As mentioned before, the value present in the clients list delivered by the company is more reliable source than outcome of the call center review.
Having this on mind, order of actions to take looks as follows:

For every key attribute value found in *UnifiedDecAttrSources* concept:

1. Look for *NOT NULL* value of *HasVM* attribute. If found there, assume it to be the newly constructed decision attribute value.

2. If the value of *HasVM* for the particular key is *NULL*, take *ExamResponse* value.

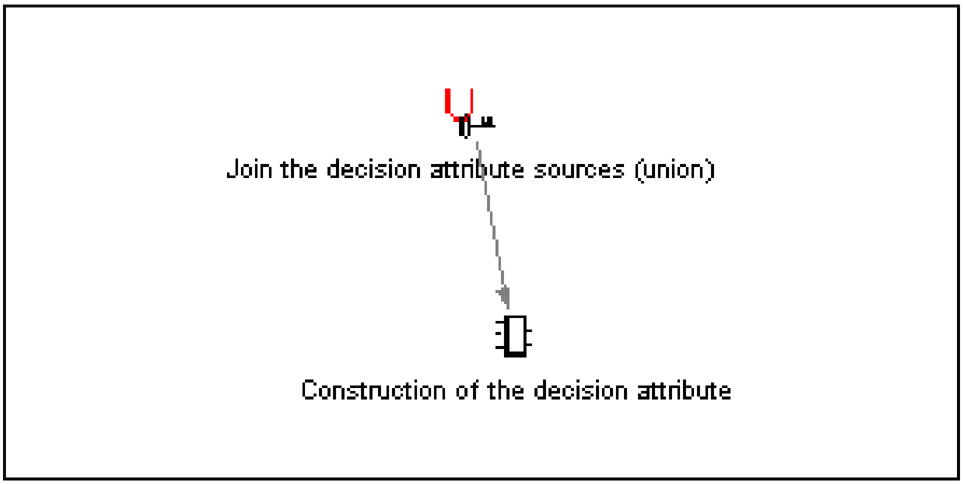
Join the decision attribute sources (union)
Construction of the decision attribute

Figure 3.9: Decision attribute construction

The last use of *UnionByKey* attaches the decision attribute to profiles created before (see figure 3.10).

There are many of clients for whom no decision attribute have been found, neither in actual clients list, nor in the response given to the call center review. We fill (figure 3.11) these missing values of the decision attribute with the default value of 0 (which means ,,clients relation to offered service is unknown").

To improve the efficiency of mining algorithms, some of the attributes are discretized manually. During profile data analysis, we noticed that for majority of callers some of the attributes, like for example number of calls to party-lines, take value 0. We decided to create a binary attribute, reflecting the fact that the client did or did not make such a type of call (figure 3.11).

Figure 3.12 gives an overall look at all the preprocessing steps.

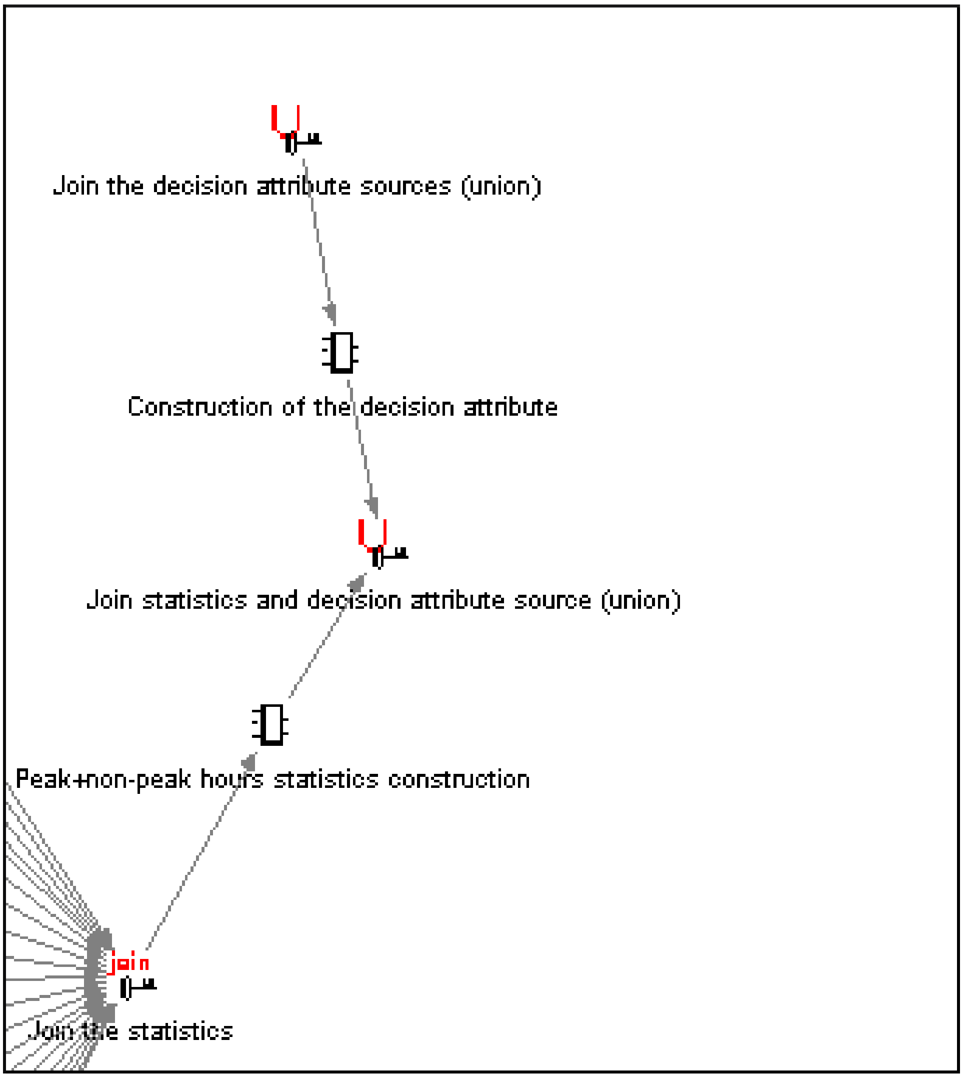
Join the decision attribute sources (union)
Construction of the decision attribute
Join statistics and decision attribute source (union)
Peak+non-peak hours statistics construction
join
Join the statistics

Figure 3.10: Adding decision attribute to clients' profiles

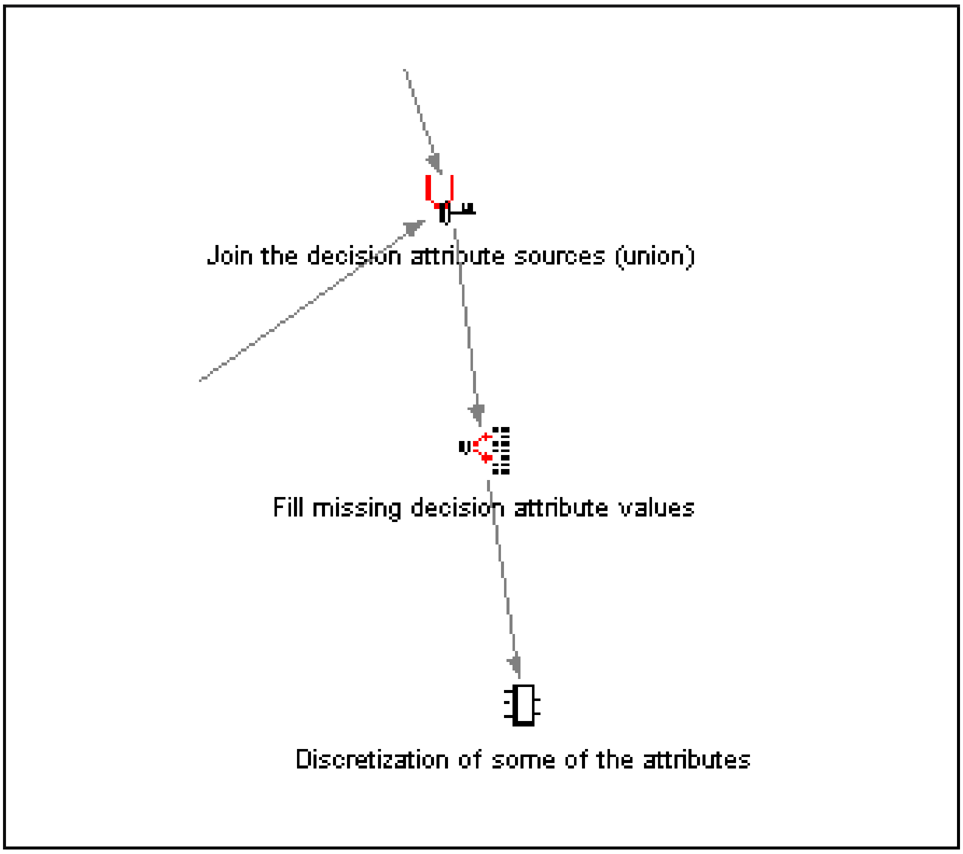
Join the decision attribute sources (union)
Fill missing decision attribute values
Discretization of some of the attributes

Figure 3.11: Final steps

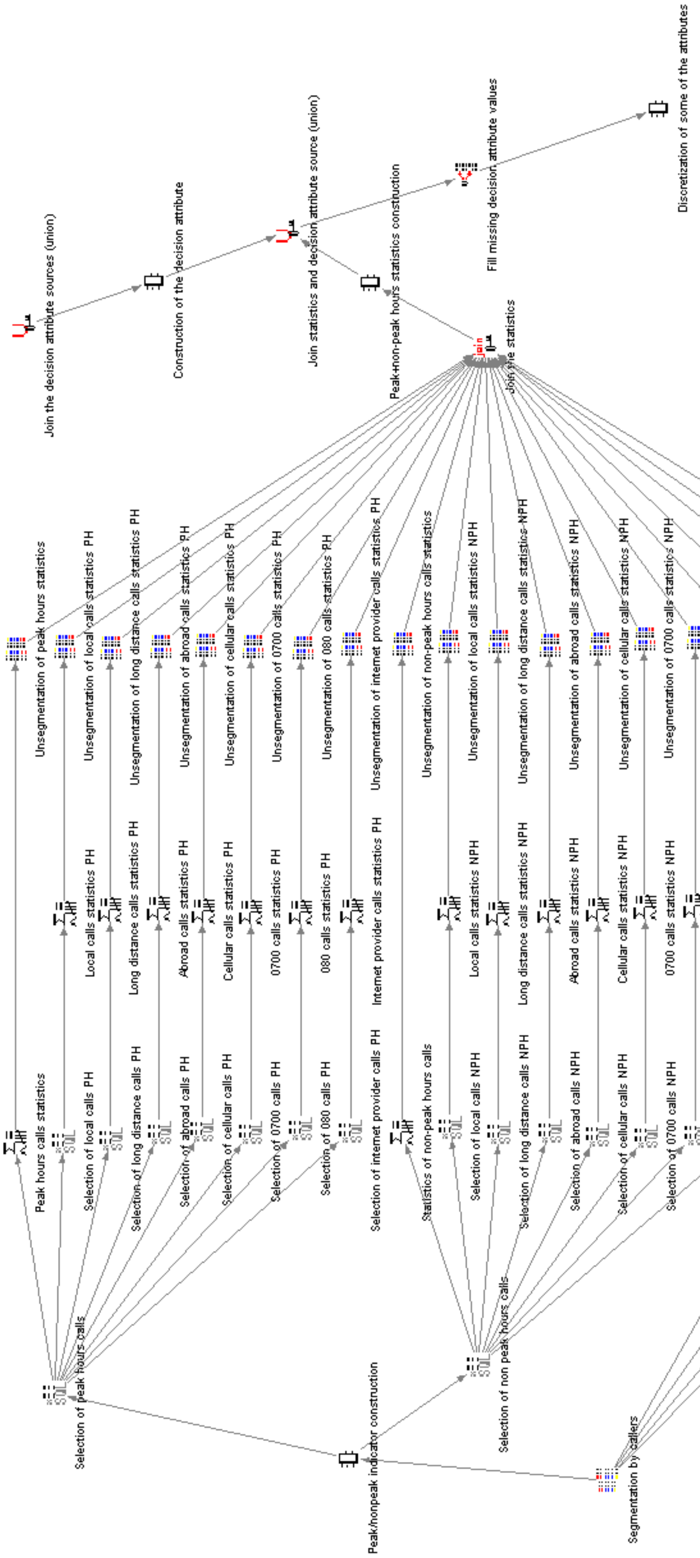
Join the decision attribute sources (union)
Construction of the decision attribute
Join statistics and decision attribute source (union)
Peak+non-peak hours statistics construction
Fill missing decision attribute values
Discretization of some of the attributes
Join the statistics
Unsegmentation of peak hours statistics
Unsegmentation of local calls statistics PH
Unsegmentation of long distance calls statistics PH
Unsegmentation of abroad calls statistics PH
Unsegmentation of cellular calls statistics PH
Unsegmentation of 0700 calls statistics PH
Unsegmentation of 080 calls statistics PH
Unsegmentation of internet provider calls statistics PH
Unsegmentation of non-peak hours calls statistics
Unsegmentation of local calls statistics NPH
Unsegmentation of long distance calls statistics NPH
Unsegmentation of abroad calls statistics NPH
Unsegmentation of cellular calls statistics NPH
Unsegmentation of 0700 calls statistics NPH
Unsegmentation of 080 calls statistics NPH
Unsegmentation of internet provider calls statistics NPH
Unsegmentation of properly disconnected calls statistics
Unsegmentation of improperly disconnected calls statistics
Unsegmentation of statistics of calls without an answer
Unsegmentation of all calls statistics
Local calls statistics PH
Long distance calls statistics PH
Abroad calls statistics PH
Cellular calls statistics PH
0700 calls statistics PH
080 calls statistics PH
Internet provider calls statistics PH
Local calls statistics NPH
Long distance calls statistics NPH
Abroad calls statistics NPH
Cellular calls statistics NPH
0700 calls statistics NPH
080 calls statistics NPH
Internet provider calls statistics NPH
Statistics of properly disconnected calls
Statistics of improperly disconnected calls
Statistics of not answered calls
Peak hours calls statistics
Selection of local calls PH
Selection of long distance calls PH
Selection of abroad calls PH
Selection of cellular calls PH
Selection of 0700 calls PH
Selection of 080 calls PH
Selection of internet provider calls PH
Statistics of non-peak hours calls
Selection of local calls NPH
Selection of long distance calls NPH
Selection of abroad calls NPH
Selection of cellular calls NPH
Selection of 0700 calls NPH
Selection of 080 calls NPH
Selection of internet provider calls NPH
Selection of properly disconnected calls
Selection of calls improperly disconnected
Selection of calls without an answer
Statistics of all calls
Selection of peak hours calls
Selection of non peak hours calls
Peak/nonpeak indicator construction
Segmentation by callers

Figure 3.12: Whole preprocessing schema in HCI

Bibliography

[1] R. Bathoorn N. Brandt M. de Haas O. Rem. Problem modeling. MiningMart Deliverable D19 IST research project, 2001.