Paper

Data mining and complex telecommunications problems modeling

Janusz Granat

Abstract — The telecommunications operators have to manage one of the most complex systems developed by human beings. Moreover, the new technological developments, the convergence of voice and data networks and the broad range of services still increase this complexity. Such complex object as telecommunication network requires advanced software tools for their planning and management. Telecommunications operators collect large volumes of the data in various databases. They realize that the knowledge in these huge databases might significantly improve various organizational strategic and operational decisions. However, this knowledge is not given explicitly, it is hidden in data. Advanced methods and algorithms are being developed for knowledge extracting. In this paper we will focus on using data mining for solving selected problems in telecommunication industry. We will provide a systematic overview of various telecommunications applications.

Keywords* — *decision support systems, telecommunications, dynamic information system, temporal data mining.

1. Introduction

The problems that are specified in the domain terms might be classified into three main levels of analysis (Fig. 1):

- *Business level* (e.g. better understanding and prediction of customer behavior, identification of customer needs, customer-oriented supply of new services, improvement of business processes). On this level we use a client oriented data.
- *Product or service level* (e.g. web mining). On this level we use service oriented data.
- *Network and information infrastructure analysis level* (e.g. fault detection, supporting network management, resource planning). On this level we use a network oriented data.

We can distinguish three main steps of describing data mining problems:

1. Problem formulation in the domain terms. This is usually textual description of the business requirements that have to be fulfilled by data mining.
2. The transformation of business requirements into a class of data mining problems like classification, prediction, associations etc. It is a bridge between business description and detailed model specification.
3. The detailed model specification. This is a model specification that is used by data mining modeler for a specific software tools.

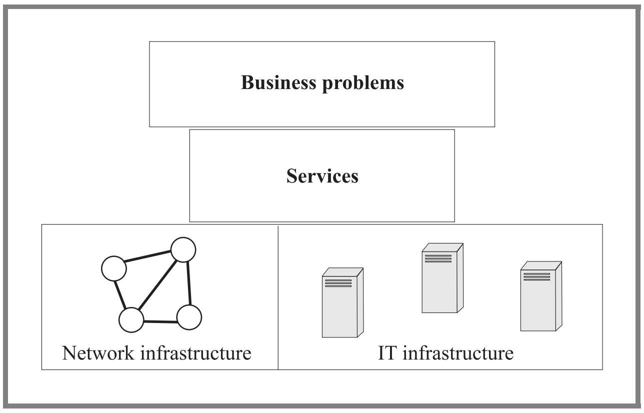

Fig. 1. Levels of problem analysis.

An overview of data mining problems in the context of business problems in telecommunication is given in [1, 5]. It can be observed that one of the main areas of applications of data mining on business level is a support for various task of the marketing departments. The data mining becomes a key part of analytical subsystem of customer relationship management systems. On business level of analysis there are many similarities to other industries. The applications of data mining for marketing can be found in [11]. The following main problems for marketing and sales departments of telecommunication operators can be distinguished:

- customer segmentation and profiling,
- churn prediction,
- cross selling and up-selling,
- live-time value,
- fraud detection,
- identifying the trends in customer behavior.

On product or service level there is a focus on analysis of incomes, quality of the service, grade of the service and others. There are formal agreements called service level agreements (SLA) [4] between providers of the service and the customers. Service level management (SLM) are becoming the prevailing business model for delivering a products and services. Such approaches need advanced computerized tools.

On the level of infrastructure and network analysis we can distinguish the following problems:

- network planning,
- IT resources planning,
- fault detection, location and identification.

2. The formal description of a data mining process

The typical data mining process consist of the following steps:

- problem formulation,
- data preparation,
- model building,
- interpretation and evaluation of the results.

In the industry environment these steps as follows (in the brackets there is information about responsible persons):

- problem formulation (business users),
- developing programs for preprocessing the data (data mining analyst),
- building the model (data mining analyst, business users),
- prepare the processes of the use of the data mining models in the business (business users),
- repetitive running of the model (data mining analyst),
- running programs for loading and transformation of the data,
- running the data mining models – scoring,
- export the scoring results to the operational systems.

There are a lot of publications related to data mining but these publications are focusing on algorithms, description of problems etc. but there is no common formal description of data mining process in the context of enterprise application. In this section we will provide such a formal description of a data mining process. We will start with source data description by information system, then preprocessing of data in order to prepare input for data mining algorithms, and finally the results of the algorithms.

2.1. Source information systems

As the input for a data mining process there are various tables of the databases, text files etc. These source data might be described formally by *the information systems*. We define, following [8, 9] or [3], an *information system* as a 4-tuple:

$$S = (X, A, V, \rho), \tag{1}$$

where:

X – is the finite and nonempty set of objects or observations,

A – is finite and nonempty set of attributes,

$V = \bigcup_{a \in A} V_a$, V_a is a set of values of attribute $a \in A$, called the domain of a,

ρ – is an information function: $\rho : A \times X \to V$.

Information system S define a relation $R_s \subset V_{a_1} \times V_{a_2} \times \times \ldots \times V_{a_k}$, so that $R_s(v_{i_1}, v_{i_2}, \ldots, v_{i_k}) \Leftrightarrow (a_1, v_{i_1}), (a_2, v_{i_2}), \ldots, (a_k, v_{i_k})$ is nonempty information in S. The relational approach is often used in data processing, but in data mining we need more information that we have in information system. The links between information system and relations might be useful in data preprocessing.

The information system (1) describes the static nature of the system. In practical applications we have to deal with dynamics of the system. Orłowska [7] introduced the term *dynamic information system*:

$$D = (X, A, V, \rho, T, R), \tag{2}$$

where:

T – is a nonempty set whose elements are called moments of time,

R – is a order on the set T (here we assume linear order),

X – is the finite and nonempty set of objects or observations,

A – is finite and nonempty set of attributes,

$V = \bigcup_{a \in A} V_a$, V_a is a set values of attribute $a \in A$, called the domain of a,

ρ – is an information function: $\rho : A \times X \times T \to V$.

Orłowska in [7] have considered dynamic information system in context of a logic. In this paper, we wiil use this system as a base for formulation of the temporal data mining problems.

2.2. The preprocessing

The data sources of a data mining process might be described by the set of dynamics information systems:

$$\Sigma = \{D^1, D^2, \ldots, D^l\}. \tag{3}$$

The data sources have to be transformed into the input dynamic information system IT that is needed for data mining models:

$$IT = \mathcal{P}(\Sigma),$$

where:
IT – is an input dynamic information system,
Σ – is a set of source information systems,
$\mathcal{P}$ – is a process of preprocessing.

The IT is defined as:

$$IT = (X, F, V, \rho, T, R), \tag{4}$$

where:
T – is a nonempty set whose elements are called moments of time,
R – is a order on the set T (here we assume linear order),
X – is the finite and nonempty set of objects or observations,
F – is finite and nonempty set of features of the objects,
$V = \bigcup_{f \in F} V_f$, V_f is a set values of feature $f \in F$, called the domain of f,
ρ – is an information function: $\rho : F \times X \times T \to V$.

A process of preprocessing can be defined by the set of preprocessing steps (Fig. 2). The preprocessing step can be defined as:

$$N^i = (PN^i, SN^i, ID^i, ON^i, OD^i),$$

where:
N^i – ith preprocessing step,
PN^i – the set of steps that are the predecessors of the step N_i,
SN^i – the set of successors of the step N_i,
ID^i – the set of input dynamic information systems for the step N_i,
ON^i – the operator of the step,
OD^i – the set of output dynamic information systems of the step N_i.

The dynamic information system $IT \equiv OD^i_j$ for selected step N^i. ON^i belong to set of operators:

$$\mathcal{O} = (O_1, O_2, \dots, O_k).$$

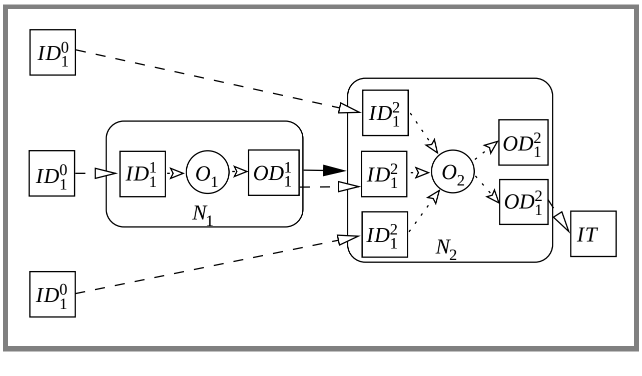

Fig. 2. A process of preprocessing – an example.

We might have the basic sets of operators on physical level like: projection, selection, etc. However, the preprocessing phase requires a broad knowledge about the data and methods of data transformation. In data mining we need advanced systems for preprocessing that will allow to store and reuse the knowledge about this phase. MiningMart [6] is an example of the system dedicated to preprocessing.

2.3. Modeling – the model building

After execution of the preprocessing step we have an dynamic information system IT that might be used for building a model. A model might have various forms. We can write that model $\mathcal{M}$ is build on the base of the dynamic information system IT:

$$IT \Rightarrow \mathcal{M}.$$

In this paper we restrict our models to feature based models. Feature base modeling assumes that objects are described by a set of features and the models find dependencies between features or predict unknown values.

We have to define the training, test, evaluation and scoring dynamic information systems. These information systems are equivalent to the sets defined in [2]. The training dynamic information system is used for preliminary model building. The test dynamic information system is used for refining the model. The performance of the model is tested by using evaluation dynamic information system. The model is applied to the score dynamic information system (Fig. 3).

The training, test, evaluation and scoring information systems are defined as follows:

$$IT^{\{id\}} = (X^{\{id\}}, F^{\{id\}}, V^{\{id\}}, \rho^{\{id\}}, T^{\{id\}}, R), \tag{5}$$

where:
id = Train – for a training dynamic information system,
id = Test – for a test dynamic information system,
id = Eval – for an evaluation set dynamic information system,
id = Score – for a scoring set dynamic information system,
$T^{\{id\}}$ – is a nonempty set whose elements are called moments of time,
R – is a order on the set T (here we assume linear order),
$X^{\{id\}} \subset X$ – is the finite and nonemty set of objects or observations,
$F^{\{id\}} = F$ – is finite and nonempty set of features of the objects,
$V^{\{id\}} = V = \bigcup_{f \in F} V_f$, V_f is a set values of feature $f \in F$, called the domain of f,
$\rho^{\{id\}}$ – is an information function:
$\rho^{\{id\}} : F \times X^{\{id\}} \times T \to V$.

The sets of objects fulfill the following condition:

$$X = X^{Train} \cup X^{Test} \cup X^{Eval} \cup X^{Score}.$$

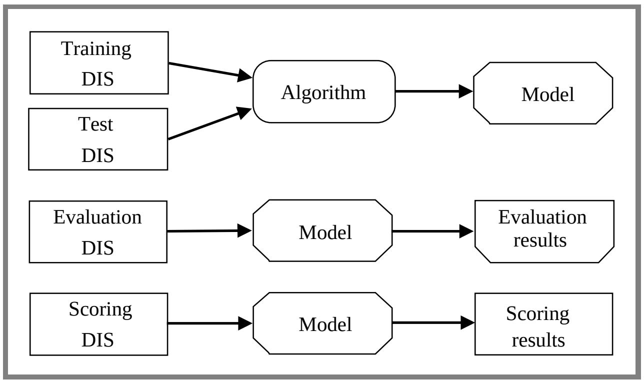

Fig. 3. The main components of a data mining process (DIS is a dynamic information system).

2.4. The data mining models

2.4.1. The classification models

There is a set of predefined m classes of the objects:

$$C = \{C_1, C_2, \dots, C_m\}.$$

These classes divide the set X^{id} into m subsets:

$$X^{id} \mapsto \{X^{id}_{C_1}, X^{id}_{C_2}, \dots, X^{id}_{C_m}\},$$

$$X^{id}_{C_i} \subset X^{id}, X^{id}_{C_i} \cap X^{id}_{C_j} = \emptyset \text{ for } i \neq j, \ \bigcup_i X^{id}_{C_i} = X^{id}.$$

The classification model assigns for each object its category. Let us consider the selected f_C feature of the object (where $C \in FC$, FC – is the index of a feature that identify the class), called class feature, and the subsets of input features f_I ($I \in FI$, $FI = F \setminus FC$, FI – is the index set of input features, F – is the index set of all object features).

The model is defined as follows:

$$\rho(f_C, x_i, t) = M_C(\rho(f_{k1}, x_i, t), \rho(f_{k2}, x_i, t), \dots, \rho(f_{kk}, x_i, t)),$$

where:
x_i – is an object identifier,
$t \in T^{id}$ – is a moment of time,
$k1, k2, \dots kk \in FI$.

2.4.2. The clustering based models

There is a set of objects X^{id} of a dynamic information system IT^{id} and the similarity measure between objects $x_i, x_j \in X^{id}$, $i \neq j$:

$$\varphi(x_i, x_j).$$

The clustering algorithms divide the set of objects into m subsets of similar objects (based on the similarity measure):

$$X^{id} \mapsto_{\varphi(x_i, x_j)} \{X^{id}_{S_1}, X^{id}_{S_2}, \dots \cup X^{id}_{S_m}\},$$

$$X^{id}_{S_i} \subset X^{id}, X^{id}_{S_i} \cap X^{id}_{S_j} = \emptyset \text{ for } i \neq j, \ \bigcup_i X^{id}_{S_i} = X^{id}.$$

Each of the clusters has the corresponding identifier:

$$S = \{S_1, S_2, \dots, S_m\}.$$

For a huge data set we have to find the clusters of objects for a training set and then we build a classification model that can be applied for a scoring set. Let us consider the selected f_S feature of the object (where $S \in FS$, FS – is the index of cluster feature), called cluster feature, and the subsets of input features f_I ($I \in FI$, $FI = F \setminus FC$, FI – is the index set of input features, F – is the index set of all object features).

The model is defined as follows:

$$\rho(f_S, x_i, t) = M_S(\rho(f_{k1}, x_i, t), \rho(f_{k2}, x_i, t), \dots, \rho(f_{kk}, x_i, t)),$$

where:
x_i – is an object identifier,
$t \in T^{id}$ – is a moment of time,
$k1, k2, \dots kk \in FI$.

2.4.3. The estimation models

The estimation model is used for finding the unknown values of the target feature that depend on some input data. Let us consider the set of objects X^{id} of a dynamic information system IT^{id}, the selected unknown feature of the object f_O (where $O \in FO$, FO – is the index of target (output) feature), called target feature, and the subsets of input features f_I ($I \in FI$, $FI = F \setminus FO$, FI – is the index set of input features, F – is the index set of all object features).

The model is defined as follows:

$$\rho(f_O, x_i, t) = M_E(\rho(f_{k1}, x_i, t), \rho(f_{k2}, x_i, t), \dots, \rho(f_{kk}, x_i, t)),$$

where:
x_i – is an object identifier,
$t \in T^{id}$– is a moment of time,
$k1, k2, \dots kk \in FI$.

2.4.4. The predictive models

The prediction model is used for finding the unknown values of the target that depend on some input historical data. The time is important in this model. Let us consider the set of objects X^{id} of a dynamic information system IT^{id}, the selected unknown feature of the object f_O (where $O \in FO$, FO – is the index of target (output) feature), called target feature, and the subsets of input features f_I ($I \in FI$, $FI = F \setminus FO$, FI – is the index set of input features, F – is the index set of all object features).

$$\begin{aligned}\rho(f_O,\mathit{client\_id},t_{churn}) = M_P\big(&\rho(f_1,\mathit{client\_id},t_1),\rho(f_1,\mathit{client\_id},t_2),\rho(f_1,\mathit{client\_id},\dots),\rho(\mathit{client\_id},f_1,t_T),\\ &\rho(f_2,\mathit{client\_id},t_1),\rho(f_2,\mathit{client\_id},t_2),\rho(f_2,\mathit{client\_id},\dots),\rho(f_2,\mathit{client\_id},t_T),\\ &\dots,\\ &\rho(f_6,\mathit{client\_id},t_1),\rho(f_6,\mathit{client\_id},t_2),\rho(f_6,\mathit{client\_id},\dots),\rho(f_6,\mathit{client\_id},t_T),\\ &\dots\big).\end{aligned}$$

The model is defined as follows:

$$\begin{aligned}\rho(f_O,x_i,t_p) = M_E(&\rho(f_{k1},x_i,t_1),\rho(f_{k1},x_i,t_2),\rho(f_{k1},x_i,\dots),\\ &\rho(f_{k2},x_i,t_1),\rho(f_{k2},x_i,t_2),\rho(f_{k2},x_i,\dots),\\ &\dots,\\ &\rho(f_{kk},x_i,t_1),\rho(f_{kk},x_i,t_2),\rho(f_{kk},x_i,\dots)),\end{aligned}$$

where:
x_i – is an object identifier,
$T^{id} = \{t_1,t_2,\dots,t_T,t_p\}$,
$t_p = t_T + \zeta,\ \zeta > 0$,
t_p – is the prediction time,
$k1,k2,\dots kk \in FI$.

2.4.5. The association rules

Let us consider the set of objects X of a dynamic information system IT, the set of the identifiers of the rules $N = \{1,2,\dots,m\}$, the selected subset of features of the object FP_i (where $FP_i \subset F$, $i \in N$, F – is the index set of all object features), and the subsets of features of object $FQ_i = F \setminus FP_i$.

The association rules are defined as follows:

$$\begin{aligned}&P_i(\rho(f_{l1},x_{l1},t_{l1}),\rho(f_{l2},x_{l2},t_{l2}),\dots,\rho(f_{ll},x_{ll},t_{ll})) \Rightarrow\\ &Q_i(\rho(f_{r1},x_{r1},t_{r1}),\rho(f_{r2},x_{r2},t_{r2}),\dots,\rho(f_{rr},x_{rr},t_{rr})),\end{aligned}$$

where:
$i \in N$,
$f_{\dots}$ – is a feature of the object,
$x_{\dots}$ – is an object identifier,
$t_{\dots} \in T$– is a moment of time,
$l1,l2,\dots ll \in FP_i \quad \forall i \in N$,
$r1,r2,\dots rr \in FQ_i \quad \forall i \in N$.

3. An example of the model formulation

One of the main problems that have to be solved by marketing departments of telecommunications operator is a long-term relationship. They have found the way of convincing current clients to continue using the services. The methods that predicts the set of customers who are going to leave the operator might be a significant tool that improves the marketing campaigns [10, 12].

The telecommunication operator is storing a lot of information about the clients in the databases. At the detail level they have switch recordings in the form of call detail records (CDR). This information is useful for billing but can not be directly used for churn analysis. Therefore, this detailed information should be aggregated and additional data should be added. Table 1 shows a subset of the data for churn analysis.

Table 1
The features that describes the clients

Client	f_1	f_2	f_3	f_4	f_5	f_6	…	churn
id	t_1	t_1	t_1	t_1	t_1	t_1		t_c
1273	20	300	50	30	25	1	…	Y
2234	100	400	100	20	30	10	…	N
…	…	…	…	…	…	…	…	…

There are the following features of the clients in the Table 1:

- f_1 – remaining binding days,
- f_2 – total amount billed,
- f_3 – incoming calls,
- f_4 – outgoing calls within the same operator,
- f_5 – outgoing calls to other mobile operator,
- f_6 – international calls,
- and others.

The training information system is defined as follows:

$$IT^{Train} = (X^{Train},F^{Train},V^{Train},\rho^{Train},T^{Train},R), \qquad (6)$$

where:
$T^{Train} = t_1,t_2,\dots,t_T,t_{churn}$, $t_{churn} = t_c = t_T + \zeta,\ \zeta > 0$,
X^{Train} – is the finite and nonemty set of clients,
F^{Train} – is finite and nonempty set of features of the objects,
$V^{Train} = \bigcup_{f \in F} V_f$, V_f, is a set values of feature $f \in F$, called the domain of f,
ρ^{Train} – is an information function:
$\rho : F^{Train} \times X^{Train} \times T^{train} \to V^{train}$.

A predictive model has been selected for a churn modeling. The “churn” feature has been selected as a target feature ($f_O = churn$), the indexes of the input features f_I belongs to the set $FI = \{1,2,3,4,5,\dots\}$.

The model is defined as follows – see the top of this page.

4. Conclusions

In this paper there is an overview of complex telecommunications problems modeling. We have applied the definiotion of the dynamic information system for a formal description of the preprocessing as well as model definition. We have stressed the importance of the preprocessing step in a data mining process. An example of churn model formulation has been provided. The presented approach might be stimulating for a development of various temporal data mining models.

References

[1] J.-L. Amat, "Using reporting and data mining techniques to improve knowledge of subscribers; applications to customer profiling and fraud management", *J. Telecommun. Inform. Technol.*, no. 3, pp. 11–16, 2002.

[2] M. J. Berry and G. S. Linoff, *Mastering Data Mining. The Art and Science of Customer Relationship Management*. Wiley, 2000.

[3] S. Greco, B. Matarazzo, and R. Słowiński, "Rough sets theory for multicriteria decision analysis", *Eur. J. Oper. Res.*, vol. 129, pp. 1–47, 2001.

[4] J. J. Lee and R. Ben-Natan, *Integrating Service Level Agreements. Optimizing Your OSS for SLA Delivery*. Indianapolis, Indiana: Wiley, 2002.

[5] R. Mattison, *Data Warehousing and Data Mining for Telecommunications*. Boston, London: Artech House, 1997.

[6] K. Morik and M. Scholz, "The MiningMart approach to knowledge discovery in databases" in *Handbook of Intelligent IT*, Ning Zhong and Jiming Liu, Eds. IOS Press, 2003

[7] E. Orłowska, "Dynamic information systems", *Ann. Soc. Math. Polon., Ser. IV: Fundam. Informat.*, vol. 5, no. 1, pp. 101–118, 1982.

[8] Z. Pawlak, "Rough sets", *Int. J. Inform. Comput. Sci.*, vol. 11, pp. 341–356, 1982.

[9] Z. Pawlak, *Systemy informacyjne. Podstawy teoretyczne*. Warszawa: WNT, 1983.

[10] Z. Pawlak, "Rough set theory and its applications", *J. Telecommun. Inform. Technol.*, no. 3, pp. 7–10, 2002.

[11] M. Shawa, C. Subramaniama, G. Tana, and M. Welgeb, "Knowledge management and data mining for marketing", *Decis. Supp. Syst.*, vol. 31, no. 1, pp. 127–137, 2001.

[12] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call detail to churn prediction: a data mining approach", *Expert Syst. Appl.*, vol. 23, pp. 103–112, 2002.

Janusz Granat received his M.Sc. in control engineering (1996) and his Ph.D. (1997) in computer science from the Warsaw University of Technology. He holds a position as an Assistant Professor at the Warsaw University of Technology, and is the leader of a research group on applications of decision support systems at the National Institute of Telecommunications in Warsaw. He lectured decision support systems and various subjects in computer science. His scientific interests include data mining, modeling and decision support systems, information systems for IT management. Since 1988 he has been cooperating with IIASA. He contributed to the development of decision support systems of DIDAS family and the ISAAP module for specifying a user preferences. He has been involved in various projects related to data warehousing and data mining for telecommunication operators. He is involved in EU MiningMart project.

e-mail: J.Granat@itl.waw.pl.
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Institute of Control and Computation Engineering
Warsaw University Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland