# A support architecture to MDA contribution for data mining

Fatima Meskine*

LITIO Laboratory,
University Oran1,
Ahmed Ben Bella, BP 1524,
El-M'Naouer, Oran, Algeria
and
Department of Computer Science,
Hassiba Benbouali University of Chlef,
B.P 78C, Ouled Fares,
02180 Chlef, Algeria
Email: f.meskine@univ-chlef.dz
*Corresponding author

Safia Nait-Bahloul

LITIO Laboratory,
University Oran1,
Ahmed Ben Bella, BP 1524,
El-M'Naouer, Oran, Algeria
Email: nait-bahloul.safia@univ-oran.dz

**Abstract:** The data mining process is the sequence of tasks applied to data, in order to discover relations between them to have knowledge. However, the data mining process lacks a formal specification that allows it to be modelled independently of platforms. Model driven architecture (MDA) is an approach for the development of software systems, based on the use of models to improve their productivity. Several research works have been elaborated to align the MDA approach with data mining on data warehouses, to specify the data mining process in a very high level of abstraction. In our work, we propose a support architecture that allows positioning these researches in different abstraction levels, on the basis of several criteria; with the aim to identify strengths for each level, in term of modelling; and to have a clear visibility on the MDA contribution for data mining.

**Keywords:** data mining; model driven architecture; MDA; data warehouses; UML profiles; data multidimensional model; transformation.

**Reference** to this paper should be made as follows: Meskine, F. and Nait-Bahloul, S. (2020) 'A support architecture to MDA contribution for data mining', *Int. J. Data Mining, Modelling and Management*, Vol. 12, No. 2, pp.207–236.

**Biographical notes:** Fatima Meskine is a Doctoral student at the Computer Science Department of the Faculty of Exact and Applied Science at the University of Oran 1, and a member of the Laboratory of Information and Information Technology of Oran (LITIO) approved in 2009. She is also an

Assistant Professor at the Computer Science Department of the Faculty of Exact Sciences and Computer Science at the Hassiba Benbouali University of Chlef. Before beginning her scientific research career, she was a Computer Science Engineer for ten years, where she worked at the Computer Centre of Abdelhamid Ibn Badis University, Mostaganem, and at other Algerian companies. She was also a trainer at the University Agency of Francophonie (AUF).

Safia Nait-Bahloul obtained her Doctorate in Computer Science from the University of Oran1. Since 2011, she has been a member of the LITIO Laboratory at the University of Oran1, which was accredited in 2009. She manages a research team in the LITIO Laboratory on data engineering and web technology. Since 2008, she has also been responsible for an Academic Master's in Information Systems and Web Technology. Her research focuses on advanced aspects of databases, web technology and unsupervised classification. Her work has been published in several journals and conference proceedings. She supervised several doctoral and masterate candidates and undergraduate projects in the field of information research, clustering, MDA and security (access control).

## 1 Introduction

Data mining (Han and Kamber, 2001) is the process of exploring and analysing data to extract knowledge models, in order to make proactive and knowledge-based decisions. Data mining tools include a set of techniques, implemented by specific algorithms, and giving access to different data organisations (flat files, databases, data warehouses, ..., etc.). However, we note the absence of a model or consensus for specifying the data mining process independently of platforms. The main problem is that the data mining process is not modelled before its realisation; it is generally specific to platforms. Modelling the data mining process is of major utility. On the one hand, it makes it easier for users to specify data mining analyses in a certain level of abstraction, and on the other hand, it allows the data mining process to have a modelling standard that follows and guides it.

The problem of the lack of abstraction in the data mining process was mentioned by Gonzalez-Aranda et al. (2008) and Pardillo et al. (2008). Pardillo et al. (2008) show the different models and existing standards for data mining process, and their limitations to provide intuitive artefacts to specify the data mining process, the authors then propose to apply the model driven architecture (MDA) (OMG, 2014a; Blanc, 2005) for data mining, to enable analysts to model their analyses easily and independently of platforms. The choice of the MDA approach is justified by its ability to provide standard formalisms that make it possible to specify any IT system, regardless of platforms, by using models and transformation of models.

In the literature, we find only the researches of: Zubcoff and Trujillo (2005, 2006, 2007) and Zubcoff et al. (2007, 2009, 2008) that propose to model the data mining process in a very high level of abstraction, by using MDA formalisms, and by considering data mining process only on data warehouses (multidimensional data) (Inmon, 2002). Our work then consists in presenting these researches in the context of the MDA approach, from the defined meta-models, to the implementations, through the transformations. We classify them according to a support architecture that we propose

while highlighting the strengths of each one, and determining the differences between them; in order to show the contribution of the MDA approach for data mining.

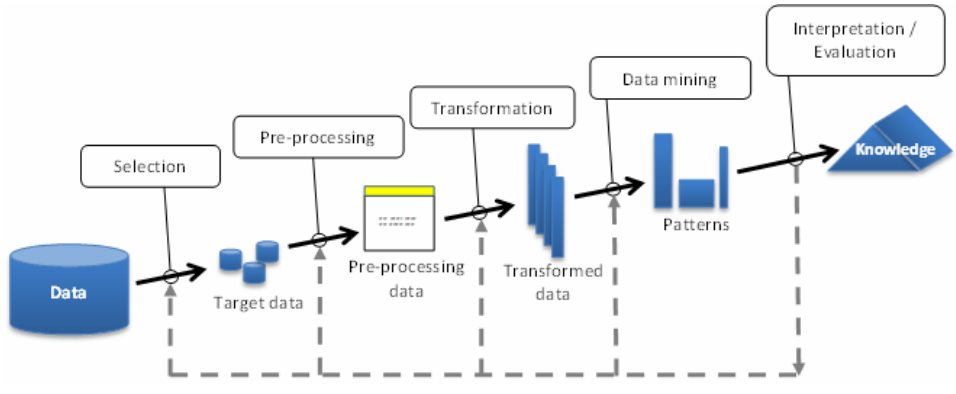
Our paper is organised as follows: In Section 2 we define the data mining process and the problematic. Section 3 presents the data mining process on data warehouses. Section 4 presents the related work. Section 5 introduces the fundamentals of the MDA approach. Section 6 presents the MDA-Data mining key works concerned by our classification study. Section 7 develops our support architecture that according to which, and following relevant criteria we classify them in Section 8. In Section 9 we discuss the results of our classification study, followed by a more general discussion in Section 10, and Section 11 concludes our work.

## 2 Data mining process: problem definition

Data mining (Han and Kamber, 2001) is a field of information technology that finds several applications in different other fields. In their entitled study 'Data mining from 1994 to 2004: an application-orientated review', Chen and Liu (2005) present four main areas in which data mining is applied. We find its application even in software engineering (Taylor et al., 2010). However data mining lacks a framework that specifies its use by practitioners, given its exploratory nature of data. This leads us to approach data mining from a process point of view, and then find ways to specify it and to model it.

The data mining process (or data mining analysis) is the sequence of tasks, organised in steps, allowing a user (analyst) to discover the desired knowledge from a data set. Several models of data mining process exist in the literature (Kurgan and Musilek, 2006). Each of these models proposes a set of steps for the data mining process, whether in the academic or the industrial field. For simplicity, we give the steps defined in the study of Fayyad et al. (1996) (Figure 1).

**Figure 1** An overview of the data mining process steps (see online version for colours)




*Source:* Fayyad et al. (1996)

The data to be analysed are generally seen as a collection of examples (instances, samples, items, objects or tuples), each example is characterised by a features vector, called attributes (Frawley et al., 1992). A data mining technique groups together a set of

algorithms that can be used to explore the data for a specific purpose. We quote among the data mining techniques: classification analysis (Mitchell, 1997), clustering analysis (Jain and Dubes, 1988), analysis of the association (association rules) (Agrawal et al., 1993) and time series analysis (Bowerman and O'Connell, 1993; Esling and Agon, 2012; Han and Kamber, 2001). The results are patterns of knowledge. Their evaluation allows the analyst to determine the fate of the data mining process, stopping it if the results are satisfactory according to some criteria, or going back to one of the previous steps (Figure 1). The user then intervenes in all stages of the process to:

1 Select the data under analysis from some dataset.

2 Select data mining attributes from these data.

3 Choose the data mining technique.

4 Choose the algorithm corresponding to this technique.

5 Set the algorithm parameters.

6 Choose the evaluation criteria.

To perform these tasks, the analyst can use the graphical user interface (GUI) of the data mining platform, or write code in a language offered by the platform. The user therefore has many choices to make. However, he does not have the means to formalise them independently of the platform. The WEKA platform (The University of Waikato, 2018) for example, is the most popular data mining platform, especially in the academic researches; it does not have tools to model the performed data mining analyses. Usually these analyses are described textually, or by the images of the windows platform (David et al., 2013). As a result, the data mining process lacks an abstract formalism that describes it more generally, making it more productive.

## 3 Data mining process on data warehouses

The data mining process can be applied to different data organisations (Frawley et al., 1992; Kumar and Tamilarasi, 2013), but since our classification study crosses with data warehouses, we present in the following the data mining process on data warehouses.

Inmon (2002, p.31) define a data warehouse as a subject-oriented, time-variant, integrated, non-volatile data collection in support of management's decisions. So, the structure of data warehouses is based on multidimensional model. This one sees the data as a cube, in multiple dimensions. The cube represents the fact. A fact is the subject (the central theme) on which the data warehouse is built. Dimensions are the criteria on which this subject is seen. Logical schemas that represent the multidimensional model are based on the relational model (Codd, 1970). We cite the star schema, the constellation schema, and the snowflakes schema (Han and Kamber, 2001). Data warehouses are at the heart of the IT field, which is business intelligence (Rausch et al., 2013). This includes data warehousing query tools, which the best known is online analytical processing (OLAP) (Han and Kamber, 2001) and data mining (Han and Kamber, 2001; Wang and Wang, 2008). In terms of platforms, data warehousing tools market offers a multitude of platforms and software that allow different companies to design and explore data warehouses. Typically, it is the providers of database management systems (DBMS) that

enable integrating, both the data warehousing and data mining tools into their solutions. Among these platforms, we mention Oracle Corporation (2018a) and Microsoft SQL Server (Microsoft Corporation, 2018a). To perform the data mining process on data warehouses, the analyst performs the same tasks mentioned in Section 2. Only that, he is restricted by the multidimensional data model.

## 4 Related work

Many researches propose models for data mining but aligning it with data warehouses. Yan and Li-li (2007) presents a data mining model based on an enterprise logistics information system, where heterogeneous data are integrated into a data warehouse following to multidimensional model or another, data mining techniques are then applied in an appropriate manner to the company's logistical management. Chen et al. (2004) propose an architecture that integrates data mining process as a company's business service, using the web services technology (Alonso et al., 2004) and common warehouse meta-model (CWM) standard (OMG, 2003), to build a metadata warehouse from an existing data warehouse. Chaoji et al. (2016) consider the data mining process as an important business process for enterprises, and show an end-to-end modelling building pipeline, which summarises the pathway of the data mining process in the reality, regardless of the data organisation. These models (Chaoji et al., 2016; Chen et al., 2004; Yan and Li-li, 2007) can be seen as flowcharts showing the flow of the data mining process steps outside of its specification. We quote the model of Wasilewska and Menasalvas (2008) which is an abstract mathematical model aiming to understand the data mining process including the data pre-processing, but it cannot be used directly by practitioners.

At the level of domain standards, such as cross industry standard process for data mining (CRISP-DM) (Smart Vision Europe Ltd., 2015), or predictive model markup language (PMML) (DMG, 2018), they describe amply the data mining process, only they have limits to specify it by abstract models (Pardillo et al., 2008).

Hofmann and Tierney (2003) present a detailed study of the participation of human resources in a large-scale data mining project, which can take over the management of the project. The authors define a life cycle for the data mining process, and the involvement of human resources in this cycle, according to their role. However, Gonzalez-Aranda et al. (2008) deplore the lack of a methodology for conducting large data mining projects, because of the lack of abstraction in the definition of the data mining process. The authors propose then a methodology for planning a data mining project, considering the data mining process steps and its life cycle, but without proposing a formal specification.

In this same vision, Pardillo et al. (2008) propose to consider the realisation of the data mining process as a software engineering process, and the authors propose to apply the MDA approach (OMG, 2014a; Blanc, 2005) to formalise all the data mining process with the underlying data. The authors have even proposed model transformation architecture to perform a data mining analysis. Unfortunately, this study did not go beyond the stage of the proposal.

Earlier researches applied MDA approach for data warehouses specification and design (Luján-Mora et al., 2006; Maté and Trujillo, 2014; Mazón et al., 2005, 2006;

Mazón and Trujillo, 2008). Other aspects of data warehouses were aligned with the MDA approach later (El Akkaoui et al., 2011; Taktak et al., 2017). These researches (El Akkaoui et al., 2011, Luján-Mora et al., 2006; Maté and Trujillo, 2014; Mazón et al., 2005, 2006; Mazón and Trujillo, 2008; Taktak et al., 2017) only considers data warehouses without supporting the data mining process. We quote the CWM (OMG, 2003), which is an MDA approach standard. It allows the exchange of metadata, to solve the problem of data integration in data warehouses, but it does not provide technical modelling pieces (artefacts) to specify a data mining analysis.

## 5 Model driven architecture

MDA (OMG, 2014a; Blanc, 2005) is an approach adopted by the OMG (2018a) for the specification and the interoperability of informatics systems, based on the object-oriented paradigm (Booch, 2004) for software design. In this section, we give an overview of the MDA approach fundamentals, which are necessary for our classification study.

### *5.1 The four-layer meta-model hierarchy of MDA approach*

MDA approach (OMG, 2014a; Blanc, 2005) is based on the formal use of models, which will be used initially to model the system, then by successive transformations to generate the code. This allows separating the business features of the system, of its implementation details. In order to express the models, MDA relies on the use of formalisms, standardised by the OMG (2018a). According to its specification [OMG, (2011a), p.17] these formalisms are organised on four layers of abstraction. The highest abstraction level of the MDA approach is the M3 layer, in which we find the meta-object facility (MOF) (OMG, 2011a, 2018b). This defines the UML meta-model (OMG, 2011a, 2011b), which is in the M2 layer. The UML meta-model provides the modelling elements and the abstract syntax of the use of these elements in the unified modelling language (UML) (OMG, 2011b), which is in the layer M1. The lowest level of this hierarchy is the M0 layer, which gathers all the applications of the real world, instances of the models which are written in the UML language (layer M1).

### *5.2 Models of the MDA approach*

Blanc (2005, p.3) describes the three types of MDA models, which are: computation independent model (CIM), platform independent model (PIM), and platform specific model (PSM). In the most cases, these models are written in the UML language (OMG, 2011b), in the form of UML diagrams, the best known of which is the class diagram, in which a class describes a set of objects that share the same features, constraints, and semantics specifications. CIM models, PIM and PSM relate to each stage of a software development life cycle. The CIM is the model for specifying user needs and requirements; the PIM is the design model of the application regardless of the implementation platform, and the PSM is the implementation code model. MDA establishes links between these models by performing automatic transformations. The model transformations are an automated process that converts a model to another. They are written in standardised language, formed of a set of rules. Query/view/transformations (QVT) (OMG, 2015) is the standard to design transformations from one

model to another, for example from a CIM to a PIM, or from a PIM to a PSM. MDA also allows a transformation of a model to a source code, for example from a PSM to a Java code (Oracle Corporation, 2018b). MOFM2T (OMG, 2008) is the standard to design this type of transformation. This last transformation allows the generation of the final application in a given platform.

### *5.3 UML profiles*

To adapt to specific domains, the UML meta-model can be extended by UML profiles (OMG, 2011a, 2011b). The central concept of UML profiles is the stereotype. OMG (2011a, p.192) states that a stereotype defines how an existing meta-class in the UML meta-model can be extended, and it activates the use of terminology or domain notation, instead of or in addition to the one used for the extended meta-class. Meta-classes are meta-modelling elements organised in packages according to their semantics in the UML meta-model. In general, a package makes it possible to group different modelling elements under the same namespace. As the class in the class diagram, a stereotype has a name, and may have properties, called tagged values. An icon (graphical item) can also represent it. The stereotype must always be related to the meta-class it extends. Relationships between stereotypes can also be established, and are the same relations, defined between classes of objects in class diagrams. The aggregation relation is a non-symmetrical association between two stereotypes, which means that one stereotype is a part of another. Generalisation is a taxonomic relationship between a more general stereotype and a specific stereotype. In addition to stereotypes, a UML profile can contain object classes and constraints. Constraints are expressed in the textual standardised language object constraint language (OCL) (OMG, 2014b). The designed UML profile for a given domain can be used to produce models for that domain. These models are still UML diagrams, but labelled by the stereotypes defined in the designed UML profile.

## 6 Formal approaches for modelling the data mining process

We present in the following the unique researches that proposed the MDA formalisms for modelling the data mining process, and which are concerned by our classification study. These are the works of Zubcoff and Trujillo (2006, 2007) and Zubcoff et al. (2007, 2009), which define UML profiles adapted to four data mining techniques; and the work of Zubcoff et al. (2008), which proposes an MDA transformation for the data mining (Table 1).

The UML profiles of the data mining techniques (Table 1) consider that the data mining technique is applied to the multidimensional data, because the authors start from the idea of specifying the data mining as a part of the specification of the data warehouse requirements. So, these research works (Table 1) are in the context of data warehouses. The multidimensional data (data warehouse) are formalised according to the UML profile of Luján-Mora et al. (2006). We must note that the UML profiles of the data mining techniques (Table 1) are completed by conceptual models and implementations of data mining analysis, according to different case studies.

**Table 1** Presentation of the MDA-data mining works retained for our classification study

| *Research works* | *Data mining technique* | *MDA formalisms* | *MDA standards* | *Data meta-model* |
|---|---|---|---|---|
| Zubcoff and Trujillo (2006) | Classification (Mitchell, 1997) | UML profile | UML 2.0 (OMG, 2005b) | Multidimensional data UML profile (Luján-Mora et al., 2006) |
| Zubcoff and Trujillo (2007) | Association rules (Agrawal et al., 1993) | UML profile | UML 2.0 (OMG, 2005b) | |
| Zubcoff et al. (2007) | Clustering (Jain and Dubes, 1988) | UML profile | UML 2.1.1 (OMG, 2007) | |
| Zubcoff et al. (2009) | Time series analysis (Bowerman and O'Connell, 1993) | UML profile | UML 2.1.1 (OMG, 2007) | |
| Zubcoff et al. (2008) | | Model to text transformation | MOFM2T 1.0 (OMG, 2008) | |

The last reference of Table 1 (Zubcoff et al., 2008) is a model to text transformation, allowing transforming a conceptual model of data mining analysis derived from the UML profile of the time series analysis technique (Zubcoff et al., 2009) to a code in the DMX language (Microsoft Corporation, 2016) of the SQL Server Analysis Services platform (Microsoft Corporation, 2018b) (Table 1).

In the literature, we also find researches that use or propose the MDA approach for data mining (Pardillo et al., 2008; Cuzzocrea et al., 2011). These researches are based on the works of Table 1, so we discuss them at the end of our classification study. The UML profile of Zubcoff and Trujillo (2005) is a UML profile for the association rules technique, but we retain the UML profile of Zubcoff and Trujillo (2007) for the same technique, because it is an improved version (Table 1).

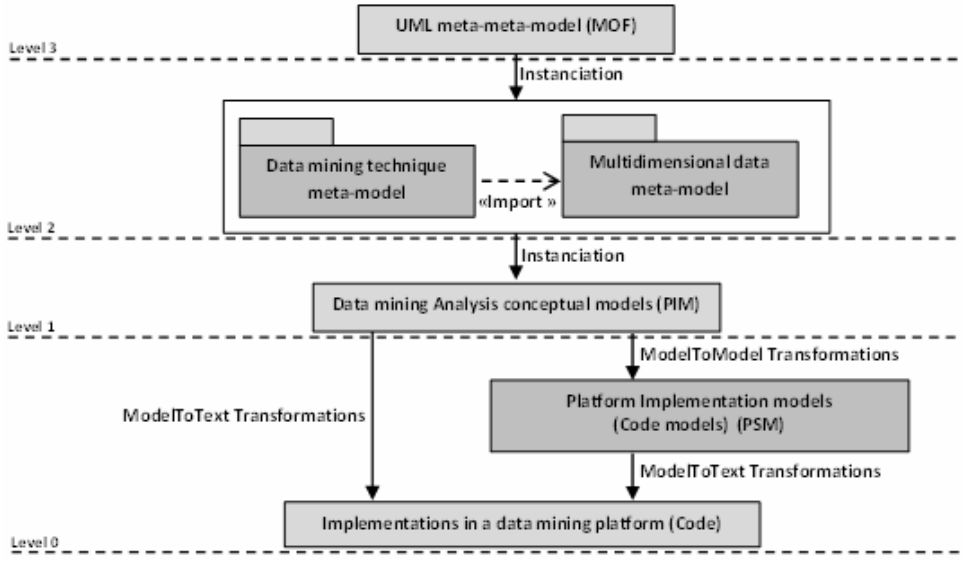
However, we notice that these researches works (Table 1) each concern a data mining technique. Our goal then is to bring them together in the context of the data mining process, and classify them as part of the MDA approach, in order to show the contribution of MDA approach for data mining.

## 7 Support architecture for the MDA-data mining works classification

Based on the fundamentals of the MDA approach (Section 5), we propose a support architecture that we use to help us to study and classify the works of Table 1 (Figure 2).

The level 3 of our architecture is the MOF (OMG, 2011a, 2018b), from which the UML meta-model is derived, which is also an instance of it (Figure 2). The UML meta-model (OMG, 2005b, 2007) has been extended by the UML profiles of Table 1. We consider these profiles as data mining techniques meta-models and multidimensional data meta-model (Table 1). The multidimensional data meta-model (Luján-Mora et al., 2006) possesses the elements that are used by the meta-model of the underlying data mining technique; it can be the meta-model of classification (Zubcoff and Trujillo, 2006), association rules (Zubcoff and Trujillo, 2007), clustering (Zubcoff et al., 2007) or time series analysis (Zubcoff et al., 2009). The two meta-models (data mining technique meta-model and multidimensional data meta-model) are in the level 2 of our support architecture, we specify the relationship between them by the association 'import' [OMG, (2011b), p.112] (Figure 2).

**Figure 2** Support architecture for the classification of the works of Table 1




The analysis conceptual models developed in the works of Table 1, are instantiated from the meta-models defined in level 2. According to our architecture, these models are independent of any platform, and are considered as PIMs. They belong to level 1 of our architecture (Figure 2).

We place the PSM between level 1 and level 0 of our architecture (Figure 2). Because on the one hand, it is a model; and on the other hand, it is implementation-specific.

Level 0 of our architecture represents all the data mining analyses that were performed in the works of Table 1 (Figure 2).

Since the work of Zubcoff et al. (2008) (Table 1) proposes an MDA transformation for data mining, we introduce in our architecture the MDA transformations, which are of two types:

- The model to model transformations, that we name 'ModelToModel', and which we place between the PIM and the PSM (Figure 2, right part).
- The model to code transformations, that we name 'ModelToText', and which we place between: the PIM and the code (Figure 2, left part), and the PSM and the code (Figure 2, right part).

## 8 Study and classification of the MDA-data mining works via the support architecture

To classify the works in Table 1, we start with the level 2 of our architecture (Figure 2), which contains the data mining techniques meta-models and the multidimensional data meta-model. We then go to the level 1 that matches the instantiated conceptual models. Finally, we discuss the various performed implementations in level 0 of our architecture.

### 8.1 Level 2 of the support architecture

The meta-models of level 2 are UML profiles, described by stereotypes, classes, associations and tagged values. Restrictions that are specific to a data mining technique (constraints) are expressed in OCL language, version 2.0 (OMG, 2006).

In the general case, the description of UML profiles is in two forms one is visual (a set of diagrams), and the other is textual. In the designed meta-models, we find both visual and textual descriptions, except the clustering meta-model (Zubcoff et al., 2007) where the authors were contented only with the visual description. Zubcoff et al. (2009) used the textual description model of the CORBA UML profile 1.0 (OMG, 2002) and UML testing profile (UTP 1.0) (OMG, 2005a).

#### 8.1.1 Criteria choice for the study of the data mining techniques meta-model

To study the data mining techniques meta-models (Table 1), we chose six different criteria that we subdivided into two categories:

a Criteria related to the data of the data mining process:

- Relationship between the data and the data mining techniques.
- Data mining attributes specification.
- Data selection specification.

b Criteria related to the data mining technique of the data mining process:

- Specification of the data mining process results.
- Specification of the technique parameters.
- Specification of the data mining technique.

These criteria are in relation with the data mining process that we defined in Section 2. So, we use the level 2 of our support architecture to discuss the designed meta-models according to these criteria.

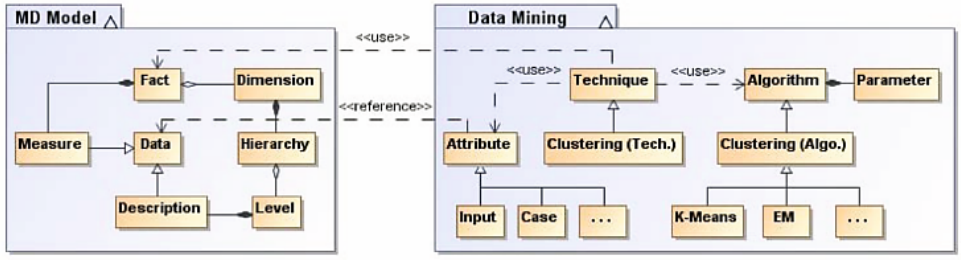
#### 8.1.2 Criteria related to the data of the data mining process

##### 8.1.2.1 Relationship between the data and the data mining techniques

The relation between the data and the data mining techniques is represented by the relation between the data mining technique meta-model and the multidimensional data meta-model. In our support architecture this relation is specified by the ‘import’ association which by definition, allows to the data-mining technique meta-model to use the modelling elements of the multidimensional data meta-model, and it’s the same choice adopted by Zubcoff et al. (2009).

Zubcoff et al. (2007) use two associations: ‘reference’ and ‘use’ (OMG, 2011a, 2011b), between elements of the data mining package (the one of the data mining technique) and elements of the MD model package (the one of multidimensional data). This specification is given in a separate model named “the multidimensional data clustering domain model” (Figure 3).

**Figure 3** Overview of the multidimensional data clustering domain model (see online version for colours)




*Source:* Zubcoff et al. (2007)

Zubcoff and Trujillo (2007), define an OCL constraint that dictates the allowed modelling elements in an instance of the association rules meta-model, among which the modelling elements of the multidimensional data meta-model. Zubcoff and Trujillo (2006) do not specify this relation.

We will see later, that it is important to define the relationship between the two meta-models at this level, to be able to specify the use of the data mining attributes in the data mining technique.

### *8.1.2.2 Data mining attributes specification*

Data mining attributes are chosen from the data set under analysis, which will be in our case multidimensional data. In the data mining techniques meta-models (Table 1), these attributes are specified by stereotypes that are named according to the roles of these attributes in the data mining analysis. We distinguish four roles (Table 2):

- In input: These attributes are used as input for the analysis.
- To predict: Are the attributes whose values are predicted by the analysis.
- In input and to predict: These are the attributes that have two roles: as input and to predict.
- Case: Is a role attribute, which is necessary only when the analysis is applied on multidimensional data, because it must designate the group of attributes from the data warehouse, which are involved in the data mining analysis.

The clustering technique (Zubcoff et al., 2007) does not require the 'predict' role for these attributes, nor the 'in input and to predict' role for the time series analysis technique (Zubcoff et al., 2009), so these roles are not specified for these data mining techniques meta-models (Table 2).

The stereotypes of Table 2 have been generalised by super stereotypes (Table 3). We think that this generalisation relation make the meta-models more expressive.

**Table 2** Names of stereotypes specifying the roles of the data mining attributes

| *Attributes roles* | *Stereotypes names* | | | |
|---|---|---|---|---|
| | *In input* | *To predict* | *In input and to predict* | *Case* |
| Zubcoff and Trujillo (2006) | I | P | IP | C |
| Zubcoff and Trujillo (2007) | I | P | IP | C |
| Zubcoff et al. (2007) | Input | - | - | Case |
| Zubcoff et al. (2009) | AsInput | AsPredict | - | AsCase |

**Table 3** Super stereotypes and meta-classes of data mining attributes roles stereotypes

| *Attributes roles stereotypes* | *Iconic definition* | *Super stereotype* | *Meta-class* |
|---|---|---|---|
| Zubcoff and Trujillo (2006) | No | MiningAttribute | Property |
| Zubcoff and Trujillo (2007) | No | MiningAttribute | Property |
| Zubcoff et al. (2007) | Yes | Attribute | Class |
| Zubcoff et al. (2009) | Yes | - | Usage |

From the meta-classes of the attributes roles stereotypes point of view, we notice that they are different, from one meta-model to another (Table 3). In the classification meta-model (Zubcoff and Trujillo, 2006) and the association rules meta-model (Zubcoff and Trujillo, 2007) the data mining attributes which are coming from the multidimensional data are seen as proprieties of the data mining technique where the use of the meta-class *property*. In the clustering meta-model (Zubcoff et al., 2007), they are considered as a class apart, from where the use of the meta-class *class*, but we can see the link between the data mining attributes and the multidimensional data in the domain of multidimensional data for clustering model (Figure 3). Note that the meta-classes *property* and *class* are part of the *Kernel* package of the UML meta-model (OMG, 2007). However, in the time series analysis meta-model (Zubcoff et al., 2009) the data mining attributes stereotypes extend the meta-class *usage* of the package *dependencies* of the UML meta-model (OMG, 2007), which means that these attributes are used by the data mining technique. We think that the choice of the meta-class *usage* is due to the association 'import' between the multidimensional meta-model and the data mining technique meta-model. As a final remark, we see that the meta-models of clustering (Zubcoff et al., 2009) and time series analysis (Zubcoff et al., 2009) use the icons to identify roles stereotypes of the data mining attributes, in addition to their names (Tables 2 and 3).

### *8.1.2.3 Data selection specification*

Data selection is the task that filters the data set under analysis, so that only interesting objects (items) are taken. Filters are conditions on data; which allow selecting these objects. In the classification meta-model (Zubcoff and Trujillo, 2006) and the association rules meta-model (Zubcoff and Trujillo, 2007), filters are specified as a tagged value of the stereotype that represents the data mining technique parameters. In the time series meta-model (Zubcoff et al., 2009), they are specified by a stereotype named *filter* and defined by an icon, it extends the meta-class *constraint*. We see filters as constraints on data and not a property of the data mining technique, in which the second choice is the most suitable. In the clustering meta-model (Zubcoff et al., 2007) they are not specified.

### *8.1.3 Criteria related to the data mining technique of the data mining process*

#### *8.1.3.1 Specification of the results of the data mining process*

After applying a data mining technique on the data, the results are visualised to analysts. The specification of the data mining process results is only present in the meta-models of classification (Zubcoff and Trujillo, 2006) and association rules (Zubcoff and Trujillo, 2007). In which, the results are specified by stereotype extending the meta-class *class* (Table 4). To know that, the results in both association rules technique and classification technique are in the form of rules.

**Table 4** Specification of the results of the data mining process

| *Data mining process results* | *Modelling elements* | *Name* | *Meta-class* |
|---|---|---|---|
| Zubcoff and Trujillo (2006) | Stereotype | CMResults | Class |
| Zubcoff and Trujillo (2007) | Stereotype | ARMResults | Class |
| Zubcoff et al. (2007) | - | - | - |
| Zubcoff et al. (2009) | - | - | - |

#### *8.1.3.2 Technique parameters specification*

Each data mining technique regroups a set of algorithms. The parameters concern both the data mining technique and the algorithms. They are chosen in an appropriate way to the analysis that one wishes to apply on the data. In the four meta-models (Table 5), the parameters are specified in relation to the technique and not in relation to a specific algorithm. Only in the meta-model of classification (Zubcoff and Trujillo, 2006), the algorithm is considered as a technique parameter, and thus specified.

**Table 5** Specification of the technique parameters

| *Parameters* | *Modelling elements* | *Name* | *Meta-class* | *Initialisation in the meta-model* |
|---|---|---|---|---|
| Zubcoff and Trujillo (2006) | Stereotype | CMSettings | Class | No |
| Zubcoff and Trujillo (2007) | Stereotype | ARMSettings | Class | No |
| Zubcoff et al. (2007) | Class | Settings | - | Yes |
| Zubcoff et al. (2009) | Class | Settings | - | Yes |

We notice that in the classification meta-model (Zubcoff and Trujillo, 2006) and the association rules meta-model (Zubcoff and Trujillo, 2007), the authors have chosen the modelling element *stereotype* in which the technique parameters are grouped as a tagged values. These stereotypes extend the meta-class *class* (Table 5). In the clustering meta-model (Zubcoff et al., 2007) and the time series analysis meta-model (Zubcoff et al., 2009), the parameters were considered as properties of the modelling element *class* instead of *stereotype* (Table 5). In this case, the meta-model is seen as a specification of an instance, and the parameters were initialised by default values in the meta-model (Table 5). We think this change is due to the elimination of the stereotype that represents the results in these meta-models.

### 8.1.3.3 Data mining technique specification

In the meta-models of Table 1, the data mining technique is also represented by a modelling element of the UML profile. We show in Table 6 the used stereotypes and their properties.

**Table 6** Specification of the data mining techniques

| *Technique* | *Stereotype name* | *Meta-class* | *Tagged values* | *Iconic definition* |
|---|---|---|---|---|
| Zubcoff and Trujillo (2006) | CMModel | Class | Attributes roles stereotypes | No |
| | CMM | Model | Classes | No |
| Zubcoff and Trujillo (2007) | ARMModel | Class | Attributes roles stereotypes | No |
| | ARMM | Model | Classes | No |
| Zubcoff et al. (2007) | Clustering | InstanceSpecification | - | No |
| Zubcoff et al. (2009) | TimeSeriesAnalysis | InstanceSpecification | - | Yes |

In the classification meta-model (Zubcoff and Trujillo, 2006) and the association rules meta-model (Zubcoff and Trujillo, 2007), the data mining technique is specified by two stereotypes. One extends the meta-class *class* and the other extends the meta-class *model* (Table 6). If we take stereotypes that extend the meta-class *model*: *CMM* and *ARMM*, we notice that they are considered as stereotypes that represent the model instance. Because we see that, the *ARMM* stereotype is used to allow classes that must be present in an instance of the meta-model (Zubcoff and Trujillo, 2007). Even, the tagged values of these two stereotypes are classes (Table 6).

If we take the stereotypes that extend the meta-class class: *CMModel* and *ARMModel* (Table 6), we see that these stereotypes represent the entire specification of the data mining analysis. In the classification meta-model (Zubcoff and Trujillo, 2006) for example, the stereotype *CMModel* aggregates the stereotypes that represent the parameters *CMSettings* and the results *CMResults*, and it has as tagged values the roles stereotypes of the data mining attributes. The same goes for the association rules meta-model (Zubcoff and Trujillo, 2007), except that, the roles stereotypes of the data mining attributes are not visually represented as tagged values in the stereotype *ARMModel*, but specified by OCL constraints, and there are no aggregation relationships between the stereotype *ARMModel* and stereotypes that represent the parameters *ARMSettings* and the results *ARMResults*, but a dependency relation. This was visually represented in the diagram and by OCL constraints.

In the clustering meta-model (Zubcoff et al., 2007) and the time series analysis meta-model (Zubcoff et al., 2009), the data mining technique is specified by a single stereotype which has the same name as the data mining technique, and making extend the meta-class *InstanceSpecification* (Table 6). The stereotype that represents the parameters was substituted by a class, and the stereotype that represents the results was eliminated. The only difference is the use of the icon to identify the technique stereotype in the time series analysis meta-model (Zubcoff et al., 2009) (Table 6).

We can understand that it is the change of the data mining technique that has required the revision in the two last meta-models (Table 6), since in association rules and classification techniques the results are in the form of rules.

### *8.1.4 Points of view*

At this level of our architecture (level 2), we distinguish three parts in the specification of the data mining process:

- A common part which concerns: the specification of the roles of the data mining attributes (input, prediction and case), and the specification of the filters on the data. This part does not change according to the data mining technique, and can be unified for all the studied meta-models.
- A part that concerns the relationship between the data mining attributes and the data mining technique. In the studied meta-models, we notice that this relation is expressed by the meta-class of the roles stereotypes of the data mining attributes. It remains whether this requirement must be unified for all meta-models or not. Because according to the studied meta-models, this relationship depends on the results nature of the data mining technique.
- Another part, which is own to the data mining technique, and concerns the parameters specification, because they are specific to it.

## *8.2 Level 1 of the support architecture*

In the works of Table 1, the authors propose real examples of data mining analyses (case studies) to design conceptual models for these analyses. These conceptual models are class diagrams, instantiated from the studied meta-models and the meta-model of the multidimensional data (level 2 of our architecture). The principle is that for a single conceptual model of multidimensional data, we can design several conceptual models of analysis, as needed. We consider these conceptual models as PIMs, in the level 1 of our architecture (Table 7). In this level, we study these PIMs in order to see how the differences and similarities between the meta-models in level 2 are concretised in the PIMs.

**Table 7** Synthesis of exemplary PIMs for the four data mining techniques meta-models

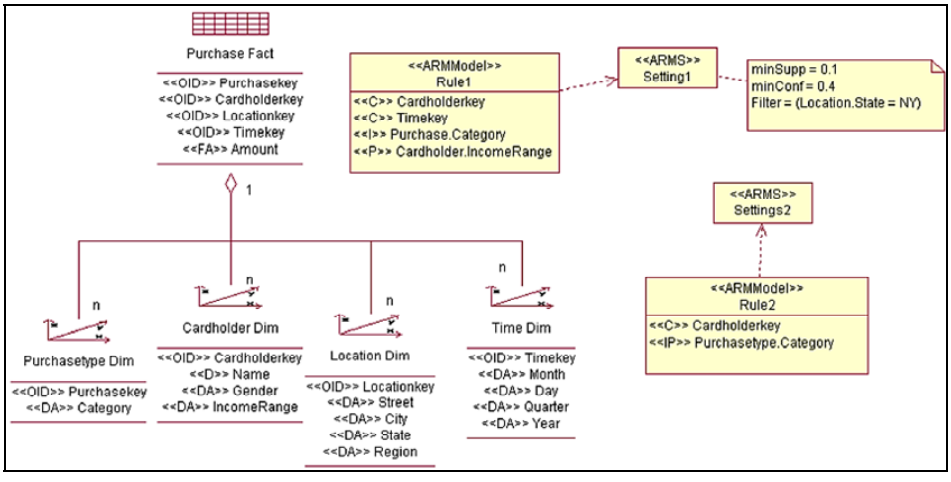
| *Techniques meta-models* | *Number of PIMs* | *Examples of analyses (case study)* | *Multidimensional data PIM positioning* |
|---|---|---|---|
| Zubcoff and Trujillo (2006) | 1 | Marine areas characteristics | With the analysis PIM |
| Zubcoff and Trujillo (2007) | 2 | Basket market | With the analysis PIM |
| Zubcoff et al. (2007) | 1 | Purchases by credit card | Separated from the analysis PIM |
| Zubcoff et al. (2009) | 2 | Fish captures | Separated from the analysis PIM |

### 8.2.1 PIMs analysis for association rules and classification techniques

#### 8.2.1.1 Case study 1: association rules

In the PIM analysis of the association rules technique (Zubcoff and Trujillo, 2007), the authors propose to analyse links between items in the customer shopping cart. This information is stored in a credit cards company, which has customers (cardholders) who use their credit cards in different categories of stores. The data are stored in a data warehouse whose model (PIM) is instantiated from the multidimensional meta-model of Luján-Mora et al. (2006).

**Figure 4** PIM of the example analysis of the association rules technique (see online version for colours)




*Source:* Zubcoff and Trujillo (2007)

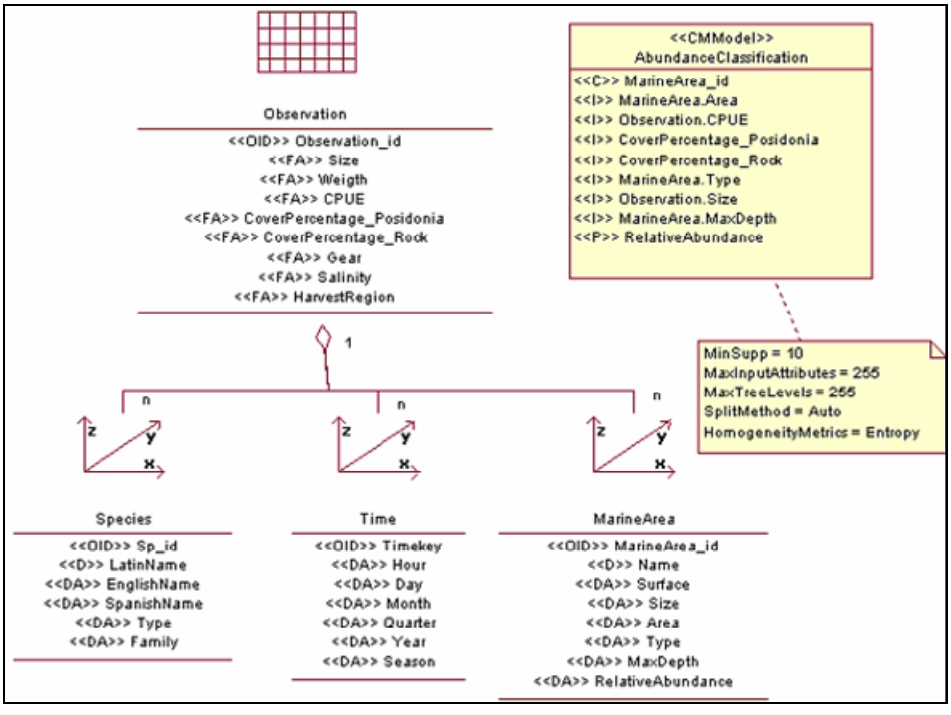
The left part of Figure 4 shows the example of the data conceptual model that includes: the class representing the fact table: *purchase fact*, and the four classes that represent the dimensions tables: *PurchaseType*, *time*, *cardholder* and *location*. For this data PIM, two different analyses PIMs are instantiated on two different data groups. These PIMs are represented by the two instances *Rule1* and *Rule2* of the stereotype *ARMModel* from the association rules technique meta-model (Zubcoff and Trujillo, 2007). The data mining attributes are represented as properties of these two classes *Rule1* and *Rule2*, and stereotyped according to their roles in both analyses. They are also taken from the data PIM present in the same schema. The two instances of the stereotype *ARMS: Setting1* and *Setting2* represent the parameters of the association rules technique, used for both *Rule1* and *Rule2* data mining analyses. In the *Setting1* class, we even see the use of a filter on the data (Figure 4).

#### 8.2.1.2 Case study 2: classification

The classification analysis PIM (Zubcoff and Trujillo, 2006) is designed for data concerning the characteristics of marine areas, to obtain a classification of *relative abundance* by *marine area type* (protected or not), *catch per unit effort* (CPUE), *Posidonia* and *rock cover percentages*, *total marine area*, and *maximum depth*. These

data are part of the multidimensional data, modelled in the same analysis PIM. Figure 5 shows the PIM of the classification analysis.

**Figure 5** PIM of the example analysis of the classification technique (see online version for colours)




*Source:* Zubcoff and Trujillo (2006)

### *8.2.1.3 Comparison*

Considering the study of the two scenarios, we see that there are similarities between the two PIMs of: association rules analysis (Figure 4), and classification analysis (Figure 5). To model an analysis, the stereotype representing the data mining technique and that extend the meta-class *class* is instantiated (Table 6). This stereotype is instantiated as a class, having as properties the attributes chosen for the analysis, labelled by the names of the roles stereotypes of these attributes (Table 2). Note that roles stereotypes of data mining attributes extend the meta-class *property* in both meta-models of association rules and classification (Table 3). As a result, the multidimensional data PIM, and the analysis PIM are juxtaposed in the same diagram (Figures 4 and 5), since the data mining attributes are taken from the multidimensional data under analysis. As a last similarity, we notice that the stereotype representing the results of the analysis is not instantiated, whereas it is present in the two meta-models (Table 4). However, some differences are also seized in the two PIMs that come from differences between the two meta-models (level 2 of our architecture). In the classification analysis, the instantiation of the stereotype representing parameters is different from that in association rules analysis. It is made by giving the parameters fixed values, and recitals as a note related to the instance

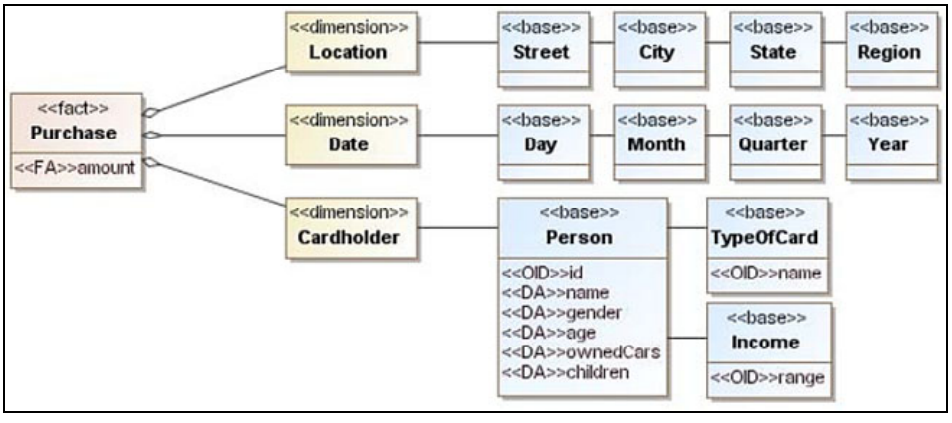
of the stereotype representing the technique (Figure 5), because in the meta-model, the stereotype of those parameters is in aggregation relationship with the one of the technique (level 2 of our architecture). In association rules analysis, the parameters are represented by a class in dependency relation with the instance of the technique stereotype (Figure 4), according of course to the meta-model (level 2 of our architecture). Concerning the parameters, we think that modelling adopted in association rules analysis is better than that adopted in classification analysis, as it enables better readability of PIMs.

### *8.2.2 PIMs analysis for time series and clustering technique*

#### *8.2.2.1 Case study 1: clustering*

The case study of the clustering technique meta-model (Zubcoff et al., 2007) concerns the analysis of purchases by credit cards. The multidimensional data are modelled in a separated PIM (Figure 6) instantiated from the meta-model of Luján-Mora et al. (2006).

**Figure 6** The multidimensional data PIM of the clustering technique analysis example (see online version for colours)




*Source:* Zubcoff et al. (2007)

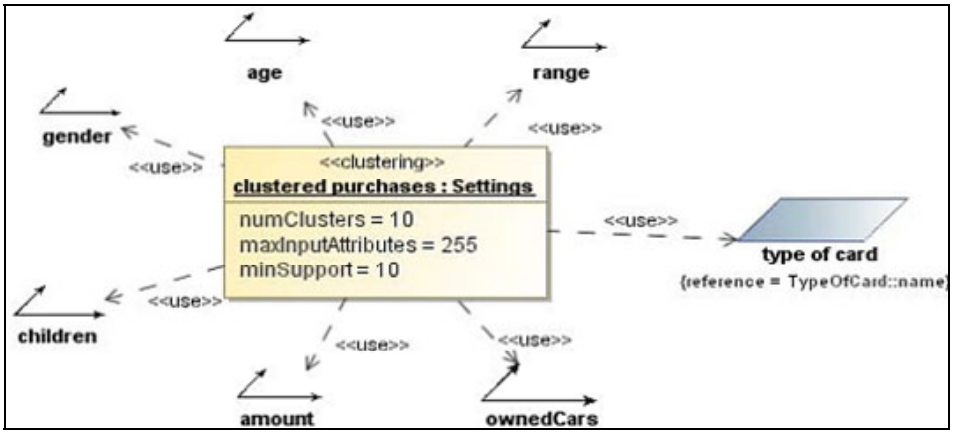
On these multidimensional data, a clustering analysis clusters purchases by credit card using the attributes that characterises the cardholder: *age*, *gender*, *name*, *children*, *range*, and the attribute *amount* (Figure 6).

In the PIM of this analysis (Figure 7), we see the input attributes which are represented by the icon of the stereotype *input*. Furthermore, the attribute *type of card* is considered as case, and is represented by the icon of the stereotype *case*. The clustering analysis is represented by instantiating the class which represents the parameters: *settings*. The values of the parameters *numClusters*, *maxInputAttributes* and *minSupport* are given in the model (the rest of parameters have default values) (Figure 7).

The dependency relationship, labelled 'use', taken from the domain of multidimensional data for clustering (Figure 3) is used to link between: the *class clustered purchases: settings*, which represents the parameters of the clustering technique, and the classes that represent the data mining attributes, according to their role in the data mining analysis. Note that the iconic representation of stereotypes is used only

for the *input* and *case* stereotypes, in the clustering meta-model (level 2 of our architecture).

**Figure 7** PIM of the example analysis of the clustering technique (see online version for colours)



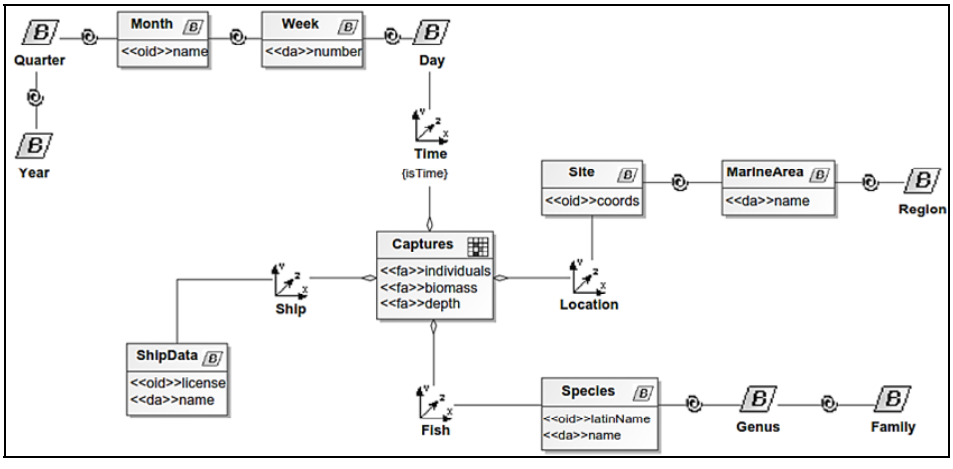

*Source:* Zubcoff et al. (2007)

### *8.2.2.2 Case study 2: time series analysis*

In this case study (Zubcoff et al., 2009), two different time series analyses are shown:

- Depths prediction: Prediction of the depth of entire captures (any fish) by marine area and by week.
- Individual's prediction: Prediction of the total number of carps captured per month.

**Figure 8** The multidimensional data PIM of the time series analysis technique example



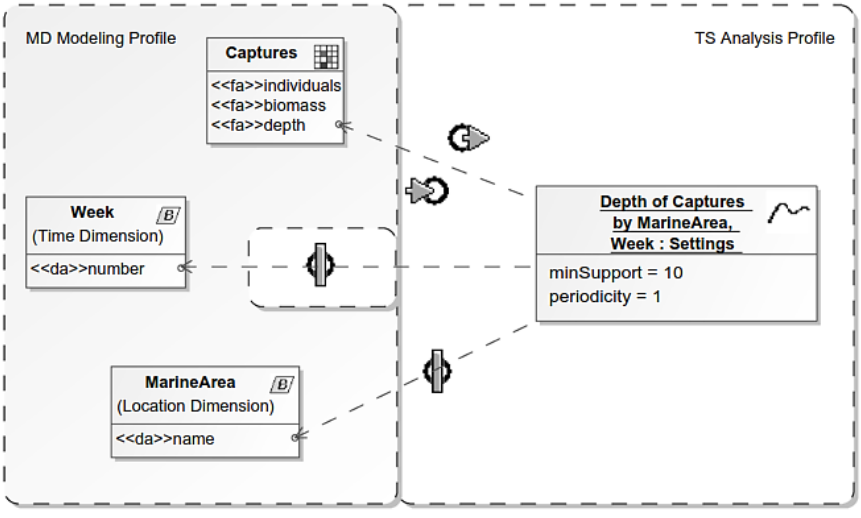

*Source:* Zubcoff et al. (2009)

The data under analysis are about fish captures, and are modelled by a PIM instantiated from the meta-model of multidimensional data of Luján-Mora et al. (2006). This PIM

represents the central fact table: *captures*, surrounded by different dimension tables: *time*, *location*, *fish*, and *ship* (Figure 8).

In the PIM of the 'depth prediction' analysis (Figure 9), we see the class: *depth of captures* by *MarineArea*, *week: settings* with the parameters values, as proprieties of this class. The selected data mining attributes, are taken from the multidimensional data PIM, with their graphical representation in this PIM.

**Figure 9** PIM of the example of the time series analysis technique




*Source:* Zubcoff et al. (2009)

The role of each of these attributes in the analysis is shown on the dependency relationship between this class and the classes representing these attributes, by labelling each relationship with the stereotype icon of the concerned data mining attribute, according to its role (Table 8).

**Table 8** Stereotypes icons in the time series analysis meta-model

| *Stereotype icon* | *Stereotype name* | *Represents* |
|---|---|---|
| | TimeSeriesAnalysis | Data mining technique |
| | AsInput | Input data mining attributes |
| | AsPredict | Predict data mining attributes values |
| | AsCase | Case data mining attributes |
| | Filter | Data filters |

The same goes for the second PIM concerning the prediction of the total number of carps captured per month, only we note the use of the data filter, because it concerns only carps.

#### 8.2.2.3 Comparison

From the study of these examples, we can detect similarities and differences. Regarding the similarities, we note the use of icons to identify stereotypes in the PIMs of the time series analysis, and the PIM of the clustering analysis (Figures 9 and 7), also the PIMs are generated by instantiating the class representing parameters (Table 5) stereotyped by the data mining technique icon in time series analysis, and the data mining technique name in the clustering (Table 6). Also, the PIM of the multidimensional data is given separately from the PIM of the analysis in the two case studies. The differences are in the representation of the data mining attributes. In the PIM of clustering analysis, data mining attributes are taken from the multidimensional data, and labelled by the icons of the roles stereotypes of the data mining attributes (Table 3). They are then linked to the parameters class of the clustering technique by the *use* relation, coming from the domain of multidimensional data for clustering (Figure 3). In time series analysis PIMs, the icons that represent the roles of the data mining attributes (Table 3) are placed on the dependency relationship between the class that represents the parameters, and the classes of the multidimensional data PIM, used as data mining attributes. We think that this is the consequence of linking the stereotypes of data mining attributes to the meta-class *usage* in the time series analysis meta-model (Table 3).

We think that the PIMs of the time series analysis are better than the PIM of the clustering analysis, from the point of view of the legibility of the PIMs, because it best represents the role of data mining attributes in the analysis.

### 8.2.3 Points of view

We find that there are two factors, which caused the main differences between the PIMs of these analyses:

- The definition of the relationship between the technique meta-model and the multidimensional data meta-model.
- The meta-classes of the roles stereotypes of the data mining attribute.

And it is two factors that made the multidimensional data PIM with the analysis PIM in the association rules and classification techniques, and separated from the analysis PIM in the classification and time series analysis. More generally, and considering the study of the two levels, we think that the definition of the relationship between the data mining technique meta-model, and the multidimensional data meta-model determines the meta-classes of the stereotypes roles of the data mining attributes, which in turn determines the specification of the data mining technique in the meta-model.

## 8.3 Level 0 of the support architecture

The level 0 of our architecture concerns the implementation of the studied PIMs (level 1 of our architecture). The objective behind the study of this level is to show how these PIMs have been implemented, and on which platforms.

In the works of Table 1, the analysis PIMs, were implemented under Microsoft SQL Server 2005 Analysis Services (Microsoft Corporation, 2018b), and using three different tools:

- the GUI
- written code in data mining extensions (DMX) language (Microsoft Corporation, 2016)
- model to code transformation (OMG, 2008).

DMX language (Microsoft Corporation, 2016) allows writing data mining models under the platform SQL Server Analysis Services (Microsoft Corporation, 2018b), without using the GUI, and only for multidimensional data.

Because they are platform-specific, we consider all these implementations as PSMs that produce data mining analysis in level 0 of our support architecture. We then distinguish two types of PSMs: PSM without transformation and PSM with transformation.

### *8.3.1 PSM without transformation*

We resume the PSMs without transformation in Table 9.

**Table 9** Implementation of the different PIMs (PSMs)

| *PSMs* | *Implementation tools* | | *Data mining platforms* | *Data under analysis* |
|---|---|---|---|---|
| | *Graphical interface* | *DMX code* | | |
| Zubcoff and Trujillo (2006) | | × | SQL Server Analysis Services 2005 (Microsoft Corporation, 2018b) | Multidimensional (data warehouses) |
| Zubcoff and Trujillo (2007) | × | | | |
| Zubcoff et al. (2007) | No indication | | | |
| Zubcoff et al. (2009) | | × | | |

From Table 9, we find that all analyses PIMs (level 1 of our architecture) have been implemented under the same platform. This does not allow to assess the independence degree of these models against thereof. In addition, the SQL Server 2005 Analysis services (Microsoft Corporation, 2018b) implements the technique without considering a specific algorithm to it. In all case studies, the data warehouses under analysis (Table 9) are pre-implemented in the SQL Server platform (Microsoft Corporation, 2018a), according to the designed PIMs of the multidimensional data (level 1 of our architecture). In time series analysis, the multidimensional data PIM has been translated into a star schema (Zubcoff et al., 2009), which we also consider as a PIM. In the remaining case studies the data warehouses logical schema is not indicated. Note that the designed multidimensional data PIMs can be translated into different schemas of data warehouses (star, constellation and snowflake).

### *8.3.2 PSM with transformation*

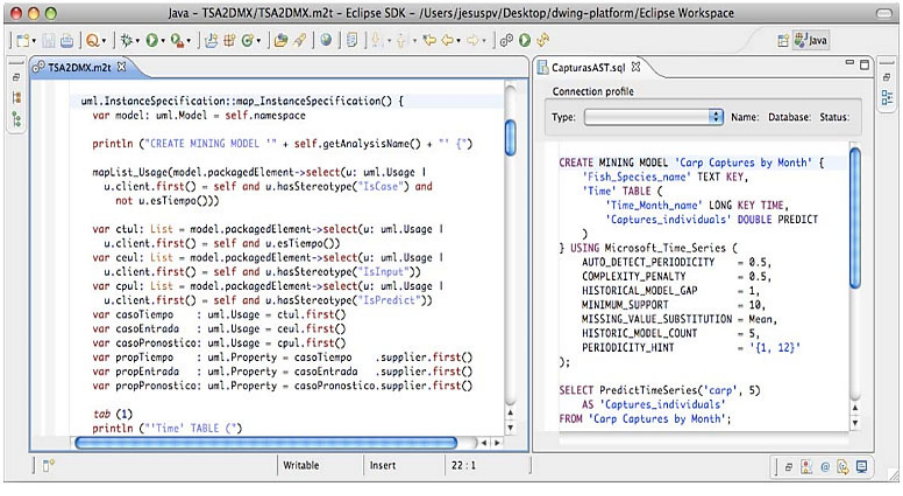
Zubcoff et al. (2008) propose a transformation approach, based on:

- The multidimensional data meta-model (level 2 of our architecture).
- The time series analysis meta-model (level 2 of our architecture).

- A model to text transformation that automatically generates analysis code, written in DMX language (Microsoft Corporation, 2016) from the PIMs of the time series analysis (level 1 of our architecture).

The transformation code (Figure 10) is implemented in the Eclipse Platform (The Eclipse Foundation Open Source Community Website, 2018) using the MOFScript plug-in which is an implementation of the MDA standard: MOFM2T (OMG, 2008).

**Figure 10** Example of a MOFM2T transformation and the generated code (see online version for colours)




*Source:* Zubcoff et al. (2008)

The transformation (Figure 10) uses the elements of the time series analysis meta-model and the multidimensional data meta-model (level 2 of our architecture), and it makes links between these elements and the syntactic elements of the DMX language. Then, each specified attribute in the PIM is transformed into an attribute in DMX, according its role. The parameters technique specified in the PIM of the time series analysis are converted into their counterpart in DMX. However, parameters may not be specified in the PIM, so they are added explicitly later in the implementation code. Once the transformations are properly established, the MOFScript Engine can interpret them, to transform a time series analysis PIM to a DMX code.

This transformation is a 'ModelToText' transformation, between level 1 and level 0 of our support architecture (Figure 2, left part). It has as a source a time series analysis PIM, and as target a code in the DMX language (Microsoft Corporation, 2016). We consider this code as a PSM with transformation.

### *8.3.3 Points of view*

Comparing what has been done in this level with our support architecture, we notice that the only realised transformation is a 'ModelToText' transformation, which concerns the left part of our support architecture, between level 1 and level 0, since it is a transformation of a PIM to a code (Figure 2). This only proposed transformation is restricted only to the PIMs of the time series analysis, and the DMX language (Microsoft

Corporation, 2016). By its nature, it cannot be applied to other platforms, as well as to other data mining techniques. We think that one reason is the difference in the definition of the studied meta-models (level 2 of our architecture). We also note that the transformation concerns only the analysis PIMs, and not the multidimensional data PIMs, because the data warehouse is implemented beforehand.

Apart from this transformation, the data mining analysis PIMs (level 1 of our architecture) were implemented directly in the SQL Server Analysis Services platform (Microsoft Corporation, 2018b), without MDA transformations (Table 9). The data mining analyses resulting from these implementations are classified in level 0 of our architecture, as instances of the PIMs realised in the level 1 of our architecture.

## 9 Discussion of the results of our classification study

We find that the works of Table 1 concern effectively the data mining process (Section 2), and are well within the MDA approach. Beginning with the development of meta-models for each technique, in the form of UML profiles, which will provide the necessary artefacts to generate analyses conceptual models (PIMs). These PIMs can be directly implemented on a platform, or transformed into a code, so we discuss in this section two points:

- The realised MDA approach steps in the works of Table 1.
- The data mining process artefacts in the works of Table 1.

These two points consolidate the contribution of the MDA approach for data mining. Our discussion is based on our study guided by the support architecture.

### *9.1 MDA Approach performed steps*

We present in Table 10 the MDA approach steps, carried out in the works of Table 1.

**Table 10** The taken MDA approach steps in the works of Table 1

| *MDA approach steps* | *Zubcoff and Trujillo (2006)* | *Zubcoff and Trujillo (2007)* | *Zubcoff et al. (2007)* | *Zubcoff et al. (2009, 2008)* |
|---|---|---|---|---|
| Meta-model visual specification | × | × | × | × |
| Meta-model textual specification | | × | | × |
| PIMs conception. | × | × | × | × |
| 'ModelToModel' transformation | | | | |
| PSMs conception | | | | |
| Implementation without transformation | × | × | × | × |
| Implementation with 'ModelToText' transformation | | | | × |
| Total number | 3 | 4 | 3 | 5 |

We note that all the defined meta-models have been instantiated by PIMs that have been implemented. No 'ModelToModel' transformation was given for the four meta-models of data mining techniques. Because none of the five studied approaches used a code model to implement the designed PIMs, i.e., a true PSM as specified by the MDA approach. So, comparing Table 10 with our support architecture (Figure 2), we find that between level 1 and level 0, only the left part has been realised, and only for a single data mining technique: time series analysis.

## *9.2 Data mining process artefacts*

We detect in the studied meta-models of Table 1 the technical parts of the data mining process modelling; these are the artefacts of the data mining process (Table 11).

**Table 11** The artefacts of the data mining process in the works of Table 1

| *Data mining process artefacts* | *Zubcoff and Trujillo (2006)* | | *Zubcoff and Trujillo (2007)* | | *Zubcoff et al. (2007)* | | *Zubcoff et al. (2009)* | |
|---|---|---|---|---|---|---|---|---|
| | *Meta-model* | *Model* | *Meta-model* | *Model* | *Meta-model* | *Model* | *Meta-model* | *Model* |
| Data mining attributes | × | × | × | × | × | × | × | × |
| Technique parameters | × | × | × | × | × | × | × | × |
| Data filters | × | | × | | | | × | × |
| Results | × | | × | | | | | |
| Missing values | | | | | | | × | |
| Total number | 4 | 2 | 4 | 2 | 2 | 2 | 4 | 3 |

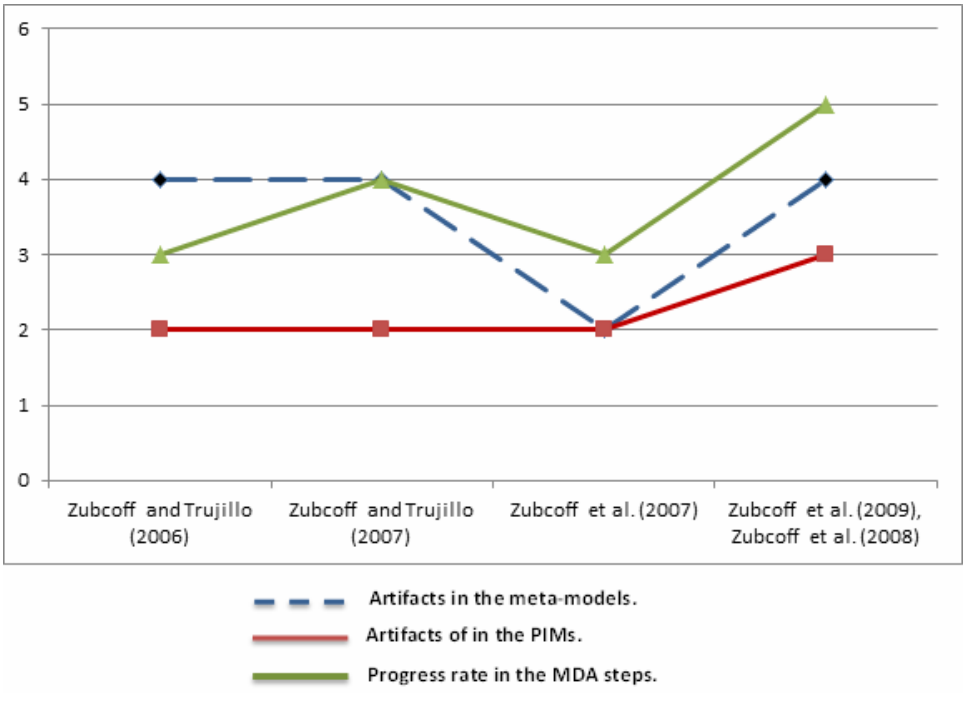
We note that the artefacts 'data mining attributes' and 'techniques parameters' are provided by all the studied meta-models even they are specified in different ways. Namely, the data can only be multidimensional, this is imposed by:

- Writhen constraints in OCL (OMG, 2006), for the meta-models of Zubcoff and Trujillo (2007) and Zubcoff et al. (2009).
- The definition of the multidimensional data domain for clustering in Zubcoff et al. (2007).
- Introducing the role 'case' for the data mining attributes, in the all studied meta-models.

We note that the specified parameters in the studied meta-models are general parameters, and that there is no separation between the technique parameters and the algorithm parameters. At the level of the instantiated models, instantiation examples have given for

each meta-model. We note that the artefact 'results' is given by the two meta-models of association rules and classification but its use has not been shown in the generated models (Table 11). We think that the results of any data mining analysis are response to choices made at the conceptual models, so we propose that it be replaced with the artefact 'evaluation' to allow analysts to select the criteria of the results evaluation (Section 2), in the level of models. There is only one meta-model (Zubcoff et al., 2009) that provides the artefact 'missing values treatment' (Table 11) which allows to specify the type of processing of these values, in the data pre-processing step of the data mining process, but its use has not been shown in the instantiated models. In the graph (Figure 11) we summarise the comparison of the works of Table 1.

**Figure 11** Comparison summary of the works of Table 1 (synthesis of Tables 10 and 11) (see online version for colours)




The graph shows that the time series analysis technique has benefited the most from the MDA approach (Zubcoff et al., 2008, 2009). However, its meta-model dates back to 2009, and since then no other meta-model has been proposed in the literature. We quote the work of Cuzzocrea et al. (2011) which proposes the composition of conceptual models of data mining analysis, concerning clustering, classification, and time series analysis techniques. These proposed models (Cuzzocrea et al., 2011), can be classified in level 1 of our architecture, as data mining analysis PIMs. Only these analysis PIMs concern several data mining techniques, whereas they are instantiated from the time series analysis meta-model (Zubcoff et al., 2009), hence the importance of integrating these data mining techniques meta-models.

## 10 General discussion

The results of our classification study show the interest of our support architecture, in the sense that we can have a clear vision of what has been done in the proposed approaches, in terms of the data mining process modelling, as part of the MDA approach. This allowed us to show the contribution of the MDA approach for data mining. It is obvious at first sight, that the conceived data mining techniques meta-models are different, because they concern different data mining techniques, but the data mining process is the same regardless of the technique, and what is shown by our study. For this purpose, we have been able to detect the modelling artefacts of the data mining process.

Pardillo et al. (2008) propose an architecture named "Model transformation architecture for model-driven data mining." This architecture details the transition between level 1 and level 0 of our support architecture (Figure 2) for any data organisation.

Our support architecture can be adapted to other areas as well, in particular for the study of MDA data warehouses (multidimensional data) applications (Luján-Mora et al., 2006; Maté and Trujillo, 2014; Mazón et al., 2005, 2006; Mazón and Trujillo, 2008; El Akkaoui et al., 2011).

## 11 Conclusions

In this work, we have established MDA support architecture for data mining on multidimensional data, we have listed all the research works of data mining that are related to MDA approach, and we have positioned them in different levels of abstraction according to this architecture. Our work presents also, a detailed and exhaustive study, guided by our support architecture, of these works, with the aim of identify differences and strong points, in order to see clearly the MDA approach interest for data mining.

We note that the proposed UML profiles, really concern the data mining process, even if they are initially seen as UML profiles for data mining techniques. We have shown that the MDA approach can help to specify the data mining process in a very high level of abstraction, by integrating these meta-models, and by enriching them to support as many platforms as possible. This will help design a transformation model that allows analysts to automatically generate analyses, for any data mining technique, and any platform. Our future work is to enrich one of the studied meta-models, and adapt it directly to the data mining process.

## References

Agrawal, R., Imielinski, T. and Swami, A. (1993) 'Mining association rules between sets of items in large databases', in *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM, Washington, DC, USA, pp.207–216.

Alonso, G., Casati, F., Kuno, H. and Machiraju, V. (2004) *Web Services Concepts, Architectures and Applications*, 1st ed., Springer-Verlag, Berlin, Heidelberg.

Blanc, X. (2005) *MDA en action Ingénierie logicielle guidée par les modèles*, Eyrolles, Paris.

Booch, G. (2004) *Object-Oriented Software Analysis and Design with Applications*, 3rd ed., Addison Wesley Longman, Redwood City, CA.

Bowerman, B.L. and O'Connell, R.T. (1993) *Forecasting and Time-series: An Applied Approach*, 3rd ed., South-Western College Pub, Cincinnati, OH.

Chaoji, V., Rastogi, R. and Gourav, R. (2016) 'Machine learning in the real world', *Proceedings of the VLDB Endowment*, Vol. 9, No. 13, pp.1597–1600.

Chen, Y., Chi, C. and Yin, J. (2004) 'Data mining service based on MDA', in *AWCC 2004: Proceedings of the Advanced Workshop on Content Computing*, Springer-Berlin, Heidelberg, ZhenJiang, JiangSu, China, pp.297–302.

Chen, Y.S. and Liu, X. (2005) 'Data mining from 1994 to 2004: an application-orientated review', *International Journal of Business Intelligence and Data Mining*, Vol. 1, No. 1, pp.4–21.

Codd, E.F. (1970) 'A relational model of data for large shared data banks', *Communications of the ACM*, Vol. 13, No. 6, pp.377–387.

Cuzzocrea, A., Mazón, J.N., Trujillo, J. and Zubcoff, J. (2011) 'Model-driven data mining engineering: from solution-driven implementations to 'composable' conceptual data mining models', *International Journal of Data Mining Modelling and Management*, Vol. 3, No. 3, pp.217–251.

Data Mining Group (DMG) (2018) *PMML 4.3 – General Structure* [online] http://dmg.org/pmml/v4-3/GeneralStructure.html (accessed 30 March 2018).

David, S.K., Saeb, A. and Al Rubeaan, K. (2013) 'Comparative analysis of data mining tools and classification techniques using WEKA in medical bioinformatics', *Computer Engineering and Intelligent System*, Vol. 4, No. 13, pp.28–38.

El Akkaoui, Z., Zimànyi, E., Mazón, J.N. and Trujillo. J. (2011) 'A model-driven framework for ETL process development', in *DOLAP '11: Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP*, ACM, Glasgow, Scotland, UK, pp.45–52.

Esling, P. and Agon, C. (2012) 'Time-series data mining', *ACM Computing Surveys*, Vol. 45, No. 1, Article No. 12.

Fayyad, U., Piatetsky-Shapiro, G. and Padhraic, S. (1996) 'Knowledge discovery and data mining: towards a unifying framework', in *KDD '96: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, AAAI Press, pp.82–88.

Frawley, W.J., Piatetsky-Shapiro, G. and Matheus, C.J. (1992) 'Knowledge discovery in databases: an overview', *AI Magazine*, Vol. 13, No. 3, pp.57–70.

Gonzalez-Aranda, P., Menasalvas, E., Millan, S., Ruiz, C. and Segovia, J. (2008) 'Towards a methodology for data mining, project development: the importance of abstraction', in Lin, T.Y. et al. (Eds.): *Data Mining: Foundations and Practice Studies in Computational Intelligence*, Springer, Berlin Heidelberg, pp.165–178.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*, 1st ed., Morgan Kaufmann Publishers, San Francisco, CA.

Hofmann, M. and Tierney, B. (2003) 'The involvement of human resources in large scale data mining projects', in *ISICT '03: Proceedings of the 1st International Symposium on Information and Communication Technologies*, Trinity College, Dublin, Ireland, pp.103–109.

Inmon, W.H. (2002) *Building the Data Warehous*e, 3rd ed., Wiley, New York, NY.

Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, 1st ed., Prentice-Hall, Englewood Cliffs, New Jersey.

Kumar, D.V. and Tamilarasi, A. (2013) 'An effective approach to mine relational patterns and its extensive analysis on multi-relational databases', *International Journal of Data Mining, Modelling and Management*, Vol. 5, No. 3, pp.277–297.

Kurgan, L.A. and Musilek, P. (2006) 'A survey of knowledge discovery and data mining process models', *The Knowledge Engineering Review*, Vol. 21, No. 1, pp.1–24.

Luján-Mora, S., Trujillo, J. and Song, I.Y. (2006) 'A UML profile for multidimensional modeling in data warehouses', *Data & Knowledge Engineering*, Vol. 59, No. 3, pp.725–769.

Maté, A. and Trujillo, J. (2014) 'Tracing conceptual models evolution in data warehouses by using the model driven architecture', *Computer Standards & Interfaces*, Vol. 36, No. 5, pp.831–843.

Mazón, J.N. and Trujillo, J. (2008) 'An MDA approach for the development of data warehouses', *Decision Support Systems*, Vol. 45, No. 1, pp.41–58.

Mazón, J.N., Pardillo, J. and Trujillo J. (2006) 'Applying transformations to model driven data ware-houses', in *DaWaK 2006: Proccedings of 8th International Conference on Data Warehousing and Knowledge Discovery*, Springer-Berlin Heidelberg, Krakow, Poland, pp.13–22.

Mazón, J.N., Trujillo, J., Serrano, M. and Piattini, M. (2005) 'Applying MDA to the development of data warehouses', in *DOLAP '05: Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP*, ACM, Bremen, Germany, pp.57–66.

Microsoft Corporation (2016) *Data Mining eXtensions (DMX) Reference* [online] http://msdn.microsoft.com/en-us/library/ms132058.aspx (accessed 30 March 2018).

Microsoft Corporation (2018a) *SQL Server 2016* [online] https://www.microsoft.com/en-us/sql-server/sql-server-2016 (accessed 30 March 2018).

Microsoft Corporation (2018b) *SQL Server Analysis Service* [online] http://technet.microsoft.com/en-us/library/ms175609(v=sql.90).aspx (accessed 30 March 2018).

Mitchell, T.M. (1997) *Machine Learning*, 1st ed., McGraw Hill, New York, NY.

Object Management Group (OMG) (2002) *UML Profile for CORBA*, *Version 1.0* [online] http://www.omg.org/spec/CORP/1.0/ (accessed 30 March 2018).

Object Management Group (OMG) (2003) *Common Warehouse Metamodel (CWM)*, *Version 1.1*, OMG Document [online] http://www.omg.org/spec/CWM/1.1/PDF (accessed 30 March 2018).

Object Management Group (OMG) (2005a) *UML Testing Profile*, *Version 1.0* [online] http://www.omg.org/spec/UTP/1.0/ (accessed 30 March 2018).

Object Management Group (OMG) (2005b) *Unified Modeling Language (UML)*, *Version 2.0* [online] http://www.omg.org/spec/UML/2.0/ (accessed 30 March 2018).

Object Management Group (OMG) (2006) *Object Constraint Language (OCL)*, *Version 2.0* [online] http://www.omg.org/spec/OCL/2.0/ (accessed 30 March 2018).

Object Management Group (OMG) (2007) *Unified Modeling Language (UML)*, *Version 2.1.1* [online] http://www.omg.org/spec/UML/2.1.1/ (accessed 30 March 2018).

Object Management Group (OMG) (2008) *MOF Model To Text Language (MOFM2T)*, *Version 1.0* [online] http://www.omg.org/spec/MOFM2T/1.0/ (accessed 30 March 2018).

Object Management Group (OMG) (2011a) *Unified Modeling Language (UML) Infrastructure*, *Version 2.4.1*, OMG Document [online] http://www.omg.org/spec/UML/2.4.1/Infrastructure/PDF (accessed 25 December 2016).

Object Management Group (OMG) (2011b) *Unified Modeling Language (UML) Superstructure*, *Version 2.4.1*, OMG Document [online] http://www.omg.org/spec/UML/2.4.1/Superstructure/PDF (accessed 25 December 2016).

Object Management Group (OMG) (2014a) *Model Driven Architecture ® (MDA®): The MDA Guide Rev 2.0*, OMG Document [online] http://www.omg.org/cgi-bin/doc?ormsc/14-06-01 (accessed 16 March 2018).

Object Management Group (OMG) (2014b) *Object Constraint Language (OCL)*, Version 2.4 [online] http://www.omg.org/spec/OCL/2.4/ (accessed 30 March 2018).

Object Management Group (OMG) (2015) *Query/View/Transformation (QVT)*, Version 1.2 [online] http://www.omg.org/spec/QVT/1.2/ (accessed 30 March 2018).

Object Management Group (OMG) (2018a) [online] http://www.omg.org/ (accessed 30 March 2018).

Object Management Group (OMG) (2018b) *Meta Object Facility (MOF) Core* [online] http://www.omg.org/spec/MOF/ (accessed 30 March 2018).

Oracle Corporation (2018a) [online] https://www.oracle.com/index.html (accessed 30 March 2018).

Oracle Corporation (2018b) *Java* [online] https://www.java.com/ (accessed 30 March 2018).

Pardillo, J., Mazon, J.N, Zubcoff, J. and Trujillo, J. (2008) 'Towards a model driven engineering approach of data mining', in *IADIS 2008: Proceedings of IADIS European Conference on Data Mining, International Association for Development of the Information Society*, Amsterdam, Netherlands, pp.144–147.

Rausch, P., Sheta, A.F. and Ayesh, A. (2013) *Business Intelligence and Performance Management: Theory, Systems, and Industrial Applications*, 1st ed., Springer-Verlag, London.

Smart Vision Europe Ltd. (2015) *CRISP-DM: Cross Industry Standard Process for Datamining* [online] http://crisp-dm.eu/ (accessed 30 March 2018).

Taktak, S., Alshomrani, S., Feki, J. and Zurfluh G. (2017) 'The power of a model-driven approach to handle evolving data warehouse requirements', in *MODELSWARD 2017: Proceeding of International Conference on Model-Driven Engineering and Software Development*, Porto, Portugal, pp.169–181.

Taylor, Q., Giraud-Carrier, C. and Knuston, C.D. (2010) 'Applications of data mining in software engineering', *International Journal of Data Analysis Techniques and Strategies*, Vol. 2, No. 3, pp.243–257.

The Eclipse Foundation Open Source Community Website (2018) [online] https://eclipse.org/ (accessed 30 March 2018).

The University of Waikato (2018) *Weka 3: Data Mining Software in Java* [online] https://www.cs.waikato.ac.nz/ml/weka/ (accessed 12 November 2018).

Wang, H. and Wang, S. (2008) 'A knowledge management approach to data mining process for business intelligence', *Industrial Management & Data Systems*, Vol. 108, No. 5, pp.622–634.

Wasilewska, A. and Menasalvas, E. (2008) 'Data preprocessing and data mining as generalization', in Lin, T.Y. et al. (Eds.): *Data Mining: Foundations and Practice Studies in Computational Intelligence*, Springer, Berlin, Heidelberg, pp.469–484.

Yan, C. and Li-li, Q. (2007) 'The research of universal data mining model system based on logistics data warehouse and application', in *ICMSE 2007: Proceedings of the 2007 International Conference on Management Science & Engineering*, IEEE, Harbin, China, pp.280–285.

Zubcoff, J. and Trujillo, J. (2005) 'Extending the UML for designing association rule mining models for data warehouses', in *DaWaK 2005: Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery*, Springer-Berlin, Heidelberg, Copenhagen, Denmark, pp.11–21.

Zubcoff, J. and Trujillo, J. (2006) 'Conceptual modeling for classification mining in data warehouses', in *DaWaK 2006: Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery*, Springer-Berlin, Heidelberg, Krakow, Poland, pp.566–575.

Zubcoff, J. and Trujillo, J. (2007) 'A UML 2.0 profile to design association rule mining models in the multidimensional conceptual modeling of data warehouses', *Data & Knowledge Engineering*, Vol. 63, No. 1, pp.44–62.

Zubcoff, J., Pardillo, J. and Trujillo, J. (2007) 'Integrating clustering data mining into the multidimensional modeling of data warehouses with UML profiles', in *DaWaK 2007: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery*, Springer-Berlin, Heidelberg, Regensburg, Germany, pp.199–208.

Zubcoff, J., Pardillo, J. and Trujillo, J. (2009) 'A UML profile for the conceptual modelling of data-mining with time-series in data warehouses', *Information and Software Technology*, Vol. 51, No. 6, pp.977–992.

Zubcoff, J., Pardillo, J., Mazon, J.N. and Trijillo, J. (2008) 'Integrating the development of data mining and data warehouses via model-driven engineering', *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos*, Vol. 2, No. 1, pp.75–86.