Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE
Informatyka w Technologii Materiałów

Vol. 10, 2010, No. 1

DATA PREPARATION

ANDRZEJ KOCHAŃSKI

Instytut Technik Wytwarzania, Politechnika Warszawska
akochans@wip.pw.edu.pl

Abstract

The paper presents the current state of the art in the area of data preparation. It proposes a complete methodology for industrial data preparation, as well as a nomenclature connected with the proposed methodology. The reasons behind the need to develop a new methodology are explained. The relevant notions, such as process, stage, task, operation, technique or method, are defined. The paper contains a diagrammatic representation of the proposed methodology.

Key words: data mining, data preparation, taxonomy, data cleaning, data integration, data transformation, data reduction, methodology

“Why Data Needs Preparing: *The Business Case*

Although the technical need for data preparation may be clear to the miner, the business benefit is not always so clear. Because data preparation can be expected to take at least 60% of the time (and resources) of any mining project, a business case is always needed to support the time and resources required and to avoid the following scenario: Manager: “How’s that modeling project coming along? Almost three-quarters of the project has elapsed and we need an interim report about how the model is performing!” Miner: “I’m still preparing the data – haven’t started modeling yet. No results so far.”” (Pyle, 2003)

“In fact, the general rule is that the data mining process is 80:20 – 80% preparation and 20% analysis.” (McCue, 2007)

1. NOTIONS AND DEFINITIONS

Data preparation is, in the great majority of published articles and conference papers, a stage of modeling which is treated only cursorily and without detail. This is why this area lacks unequivocal and generally accepted definitions. The present paper proposes a methodology for data preparation and a nomenclature that goes together with this methodology. The *process* of data preparation is understood as all tasks involved in data preparation. Two stages can be distinguished within it: the introductory stage and the main stage. A particular (introductory or main) *stage* can include different tasks. A *task* is a separate self-contained part of data preparation which can be realized in each of the stages. Four kinds of tasks can be differentiated: data cleaning, data integration, data transformation, and data reduction. Each of the tasks consists of one or more operation. An *operation* is a single action performed on the data. The same operation may be involved in different tasks. Each operation may be performed via a number of techniques. A *technique* (or sometimes a *method*, for example grouping) defines a way of performing an operation. An *algorithm* characterizes a technique (a method) in detail.

ISSN 1641-8581

2. A TAXONOMY AND OVERVIEW OF DATA PREPARATION

Data bases collecting a huge amount of information pertaining to real-world processes, for example industrial ones, contain a significant number of data which are imprecise, mutually incoherent, and frequently even contradictory. It is often the case that data bases of this kind often lack important information. All available means and resources may and should be used to eliminate or at least minimize such problems at the stage of data collection. It should be emphasized, however, that the character of industrial data bases, as well as the ways in which such bases are created and the data are collected, preclude the elimination of all errors. It is, therefore a necessity to find and develop methods for eliminating errors from already-existing data bases or for reducing their influence on the accuracy of analyses or hypotheses proposed with the application of data bases. There are at least three main reasons for data preparation: (a) the possibility of using the data for modeling, (b) modeling acceleration, and (c) increase in the accuracy of the model. It should be noted that the absence of a detailed analysis of the data chosen for preparation may lead to the choice of inappropriate tasks, operations and techniques, and – in consequence – to a significant decrease in the quality of the proposed model.

The literature pertaining to data preparation (Han & Kamber, 2001; Kusiak, 2001; Masters, 1996; Pyle, 1999; Pyle, 2003; Refaat, 2007; Weiss & Indurkhya, 1998; Witten & Frank, 2005) discusses various tasks (characterized by means of numerous methods and algorithms). Apparently, however, no ordered and coherent classification of tasks and operations involved in data preparation has been proposed so far. This has a number of reasons, including the following: (a) numerous article publications propose solutions to problems employing selected individual data preparation operations, which may lead to the conclusion that such classifications are not really necessary, (b) monographs deal in the minimal measure with the industrial data, which have its own specificity, different from the business data, (c) the fact that the same operations, are performed for different purposes in different tasks, complicates the job of preparing such a classification.

Figure 1 below represents a schema of the data preparation process. For the data collected in the real-world industrial (production) processes four tasks are distinguished:

- *data cleaning* is used in eliminating any inconsistency or incoherence in the collected data
- *data integration* makes possible integrating data bases coming from various sources into a single

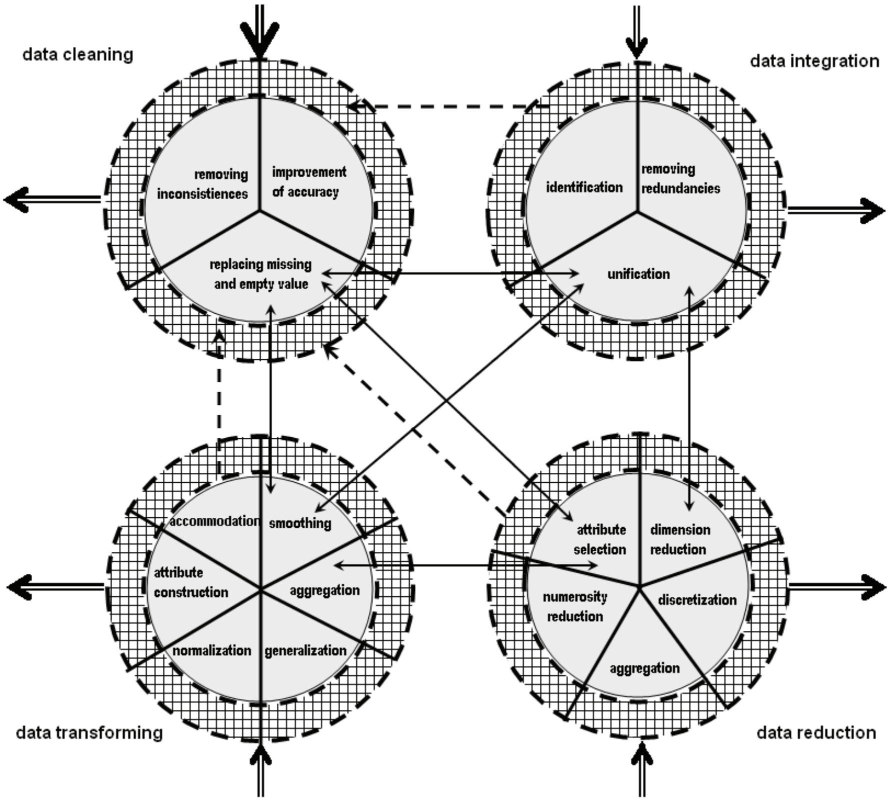

Fig. 1. Tasks realized in the process of data preparation with the distinction into the introductory and the main stage

two-dimensional table, thanks to which algorithmized tools of data mining can be employed

- *data transformation* includes a number of operations aimed at making possible the building of a model, accelerating its building, and improving its accuracy, and employing, among others, widely known normalization or attribute construction methods,
- *data reduction* limits the dimensionality of a data base, that is the number of variables; performing this task significantly reduces the time necessary for data mining.

In each task, two stages are distinguished. The area encompassed by the dashed lines (squared area) represents the introductory stage, while the grey area represents the main stage. Each of the tasks involves one or more operations. An operation is understood here as a single action performed on the data. The same operation may be a part of a few tasks.

3. DATA PREPARATION TASKS

A task is a separate self-contained part of data preparation which may be and which in practice is performed in each stage of the data preparation process. We can distinguish (in accordance with figure 1) four tasks: data cleaning, data transformation, data integration, and data reduction. Depending on the stage of the data preparation process a different number of operations may be performed in a task (also, depending on the kind of process under modeling, not all operations have to be performed in a task).

In *data cleaning* the introductory stage involves a one-time correction of the data residing in the elimination of all kinds of errors resulting from human negligence in the process of data collection. Introductory data cleaning is a laborious process of getting to the source materials, often in the form of paper notes, books and laboratory forms, or printouts from the lab and measuring equipment, of reading them and supplying the lacking data by hand. The introductory stage of *data integration* should involve the elimination of obvious repetitions. Their appearance may result, among others, from the specific nature of registering industrial production data. Introductory *data transformation* usually includes individual operations, whose performance is necessitated by data integration, such as: aggregation, generalization, or attribute construction. Introductory *data reduction* is restricted to attribute selection and is based on the knowledge and experience of the data analyst.

The data preparation process, at the stage of introductory preparation, may start with any task, but after finishing the introductory stage of this task, it is necessary to go through the introductory stage of data cleaning. It is only then when one can perform operations belonging to other tasks, including the operations of the task with which the process of data preparation has started. This procedure follows from the fact that in the further preparation, be it in the introductory or in the main stage, the analyst should have at his disposal possibly the most complete data base and only then make further decisions.

It is only after the introductory stage is completed in all tasks that the main stage, which is discussed in more detail below, can be initiated.

3.1. Data cleaning

Data cleaning in the main stage involves three operations:

- replacement of missing and empty value – this is aimed at enriching the collected data with the missing values representing the absent data,
- accuracy improvement – this involves the elimination of the erroneous data (for example outliers) or the replacement of the collected data with appropriate corrected values,
- inconsistency removal – this is done with the aid of specially designed procedures (for example, control codes) or of tools aimed at discovering such inconsistencies (for example, when functional relations among parameters are known).

3.2. Data integration

Data mining typically requires integrating data sets coming from (collected in) many separate bases. Integrating the data coming from different sources requires performing three operations:

- identification – this involves methods of identifying properties in the data sets to be integrated,
- removing redundancies – this makes possible the elimination of all repetitions and redundancies,
- unification – this leads to format conformability of the integrated data sets and to their meaning identity.

3.3. Data transformation

Data transformation involves all the issues connected with transforming the data into the form which makes possible its exploration. This includes six operations:

- smoothing – this resides in transforming the data with the aim of eliminating local deviations which have the character of a noise. Smoothing includes techniques such as, for example, binning, clustering, or regression;
- aggregation – this resides in summing up the data, most frequently in the function of time, for example, from a longer time period encompassing not just a single shift but a whole month;
- generalization – this resides in the replacement of the collected data containing measurements of a registered process value with a higher-order value, for example via their discretization;
- normalization – this resides in the rescaling (adjustment) of the data to the specified, narrow range, for example from 0.0 to 1.0;
- attribute (feature) construction – this resides in transforming mathematical attributes (features) with the aim of constructing a new attribute (new feature), which replaces in modeling the attributes from which it has been constructed;
- accommodation – this resides in transforming the data into a format used by a specific algorithm or a tool, for example into the ARFF format (Witten & Frank, 2005).

3.4. Data reduction

The operations performed in data reduction are aimed at arriving at a significantly reduced data representation which, at the same time preserves the features of the basic data set. Data reduction includes five operations:

- attribute selection – this resides in reducing the data set by eliminating from it the attributes which have little significance for the phenomenon under modeling,
- dimension reduction – this resides in transforming the data with the aim of arriving at a reduced representation of the basic data, for example in the form of new attributes,
- discretization – this is aimed at replacing the collected data with newly created higher-order values,
- numerosity reduction – this is aimed at eliminating from the data base the recurring or very similar cases,
- aggregation – this is understood as data transformation.

4. SUMMARY

The proposed methodology has the following aims: (a) to arrange and offer a complete list of tasks and operations performed in the process of data preparation, (b) to propose, on the basis of the author's own experience, a particular ordering of tasks and operations performed in the process of data preparation, (c) to give to data preparation the significance it deserves.

Data preparation is time-consuming and hence, reasonable actions taken in this direction make possible finding in a particular project some additional time for the analysis of the results extracted from the model. An ordered taxonomy of data preparation opens the way to identifying the tasks and individual operations which are necessary in each particular case of modeling. In addition it makes possible identifying those operations of data preparation which may bring about an increase in model quality, but which are too time-consuming in relation to the degree of the model quality increase.

In accordance with the proposed methodology, the process of industrial raw data preparation may start with any task (in figure 1, this is represented by the incoming double arrows) in the stage of introductory preparation, but the preferred ordering is to start with data cleaning task. However, the experience suggests that after performing the introductory stage of the first task the task of data cleaning should be performed. Only then the following tasks should be performed (in figure 1 this is represented by single dashed arrows going to data cleaning task). The main stage does not impose any determinate ordering of the tasks to be performed, especially that the individual operations and methods used in them are identical to the operations and methods performed in other tasks (in figure 1 represented as single line arrows going in both directions).

This last aim is dictated by the fact that people who decide to undertake data mining often say “We already have the data – the only thing that we need to do is to copy them from the technologist's computer disc”. And when they are told that such data often require cleaning, correcting, or supplementing and that this brings about certain costs, they react with the following answer “These data were collected and registered by conscientious and careful people – hence there is no need to do anything about them”.

An additional reason is a wish to start anew a discussion – this time, however, a discussion specifically focusing on the preparation of the industrial data in particular, so that people did not voice and support ideas such as the following: “For example, in scientific and engineering applications of data mining, such as weather prediction and industrial process control, the data is gathered from measurement equipments that does not normally result in missing values. When records do contain missing value, it is an indication that something went wrong with the data collection equipment or software, and these records should not be used at all” (Refaat, 2007).

REFERENCES

Han, J., Kamber, M., 2001, *Data Mining. Concepts and Techniques*, Morgan Kaufmann Publisher, San Francisco.

Kusiak, A., 2001, Feature Transformation Methods in Data Mining, *IEEE Transactions on Electronics Packaging Manufacturing*, 24, 3, 214-221.

Larose, D. T., 2008, *Metody i modele eksploracji danych,* Wydawnictwo Naukowe PWN, Warszawa, (in Polish).

Liu, H., Motoda, H., Yu, L., 2003, Feature Extraction, Selection, and Construction, *The Handbook of Data Mining*, ed. Nong Ye, Lawrence Erlbaum Associates, New Jersey.

Masters, T., 1996, *Sieci neuronowe w praktyce, programowanie w języku C++*, Wydawnictwo Naukowo-Techniczne, Warszawa, (in Polish).

McCue, C., 2007, *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*, Butterworth-Heinemann, Burlington.

Pyle, D., 1999, *Data Preparation for Data Mining*, Morgan Kaufmann Publisher, San Francisco.

Pyle, D., 2003, *Data Collection, Preparation, Quality, and Visualization, The Handbook of Data Mining* ed. Nong Ye, Lawrence Erlbaum Associates, New Jersey.

Refaat, M., 2007, *Data Preparation for Data Mining Using SAS*, Morgan Kaufmann Publisher, San Francisco.

Weiss, S. M., Indurkhya, N., 1998, *Predictive Data Mining: a practical guide*, Morgan Kaufmann Publisher, San Francisco.

Witten, I.H., Frank, E., 2005, *Data Mining. Practical Machine Learning Tools and Techniques*, 2nd ed., Elsevier Inc., San Francisco.

PRZYGOTOWANIE DANYCH

Streszczenie

W artykule przedstawiono dotychczasowy stan zagadnienia przygotowania danych. Zaproponowano kompletną metodykę przygotowania danych przemysłowych i związaną z nią terminologię. W pracy wyjaśniono powody, dla których opracowano nową metodykę. Zdefiniowano stosowane pojęcia, takie jak: proces, etap, zadanie, czynność, technika lub metoda. W pracy umieszczono graficzną prezentację zaproponowanej metodyki.

Received: November 4, 2009
Received in a revised form: March 25, 2010
Accepted: March 30, 2010