8

Knowledge in Imperfect Data

Andrzej Kochanski, Marcin Perzyk and Marta Klebczyk
Warsaw University of Technology
Poland

1. Introduction

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 these 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) an increase in the accuracy of the model. An additional motivation for data preparation is that it offers a possibility of arriving at a deeper understanding of the process under modeling, including the understanding of the significance of its most important parameters.

The literature pertaining to data preparation (Pyle, 1999, 2003; Han & Kamber, 2001; Witten & Frank, 2005; Weiss & Indurkhya, 1998; Masters, 1999; Kusiak, 2001; Refaat, 2007) discusses various data preparation tasks (characterized by means of numerous methods, algorithms, and procedures). 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 their own specific character, different from that of 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.

The information pertaining to how time-consuming data preparation is appears in the works by many authors. The widely-held view expressed in the literature is that the time devoted to data preparation constitutes considerably more than a half of the overall data exploration time (Pyle, 2003; McCue, 2007). A systematically conducted data preparation can reduce this time. This constitutes an additional argument for developing the data preparation methodology which was proposed in (Kochanski, 2010).

2. A taxonomy of data preparation

Discussing the issue of data preparation the present work uses the terms *process, stage, task,* and *operation*. Their mutual relations are diagrammatically represented in Fig. 1 below. The data preparation process encompasses all the data preparation tasks. Two stages may be distinguished within it: the introductory stage and the main stage. The stages of the process involve carrying out tasks, each of which, in turn, involves a range of operations. The data preparation process always starts with the introductory stage. Within each stage (be it the introductory or the main one) different tasks may be performed. In the introductory stage, the performance of the first task (the choice of the first task is dictated by the specific nature of the case under analysis) should always be followed by data cleaning. It is only after data cleaning in the introductory stage that performing the tasks of the main stage may be initiated. In the main stage, the choice of the first task, as well as ordering of the tasks that follow are not predetermined and are dependent on the nature of the case under consideration. As far as the data collected in real-life industrial (production) processes are concerned, four tasks may be differentiated:

- 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 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.

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.

In Fig. 1, within each of the four tasks the two stages of the process are differentiated: the introductory stage of data preparation (in the diagram, this represented as the outer circles marked with broken lines) and the main stage of data preparation (represented as inner circles marked with solid lines). These two separate stages, which have been differentiated, have quite different characters and use quite different tools. The first stage - that of the introductory data preparation - is performed just once. Its performance in particular tasks is limited to a single operation. This operation is always followed by the task of data cleaning. The stage of the introductory data cleaning is a non-algorithmized stage. It is not computer-aided and it is based on the knowledge and the experience of the human agent preparing the data.

The second - that is, the main - stage is much more developed. It can be repeated many times at each moment of the modeling and of the analysis of the developed model. The repetition of the data preparation tasks or the change in the tools employed in particular tasks is aimed at increasing the accuracy of the analyses. The order in which the tasks are carried out may change depending upon the nature of the issue under analysis and the form of the collected data. According to the present authors, it seems that it is data cleaning which should always be performed as the first task. However, the remaining tasks may be

performed in any order. In some cases different tasks are performed in parallel: for instance the operations performed in the task of data integration may be performed simultaneously with the operations involved in data transformation, without deciding on any relative priority. Depending on the performed operations, data reduction may either precede data integration (attribute selection) or go at the very end of the data preparation process (dimensionality selection).

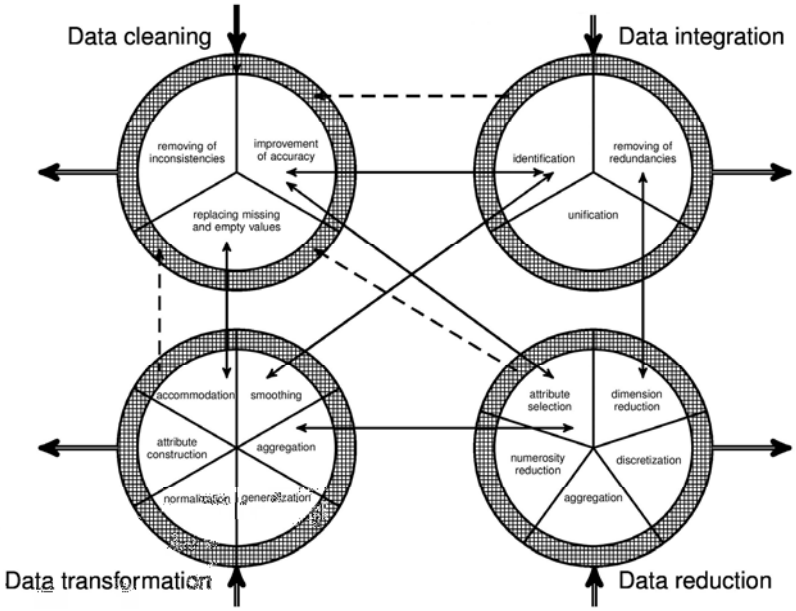


Fig. 1. Tasks carried out in the process of data preparation with the distinction into the introductory and the main stage (Kochanski, 2010)

3. Task of data preparation

A task is a separate self-contained part of data preparation which may be and which in practice is performed at each stage of the data preparation process. We can distinguish four tasks (as represented in Fig. 1): data cleaning, data transformation, data integration, and data reduction. Depending upon the stage of the data preparation process, a different number of operations may be performed in a task – at the introductory stage the number of operations is limited. These operations, as well as the operations performed in the main stage will be discussed below, in the sections devoted to particular tasks.

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 can be initiated. This is motivated by the need to avoid any computer-aided operations on raw data.

3.1 Data cleaning

Data cleaning is a continuous process, which reappears at every stage of data exploration. However, it is especially important at the introductory data preparation stage, among others, because of how much time these operations take. At this phase, it involves a one-time data correction, which resides in the elimination of all kind of errors resulting from human negligence at the stage of data collection. The introductory data cleaning is a laborious process of consulting the source materials, often in the form of paper records, laboratory logs and forms, measuring equipment outprints, etc., and filling in the missing data by hand. The overall documentation collected at this stage may also be used in the two other operations of this task: in accuracy improvement and in inconsistency removal.

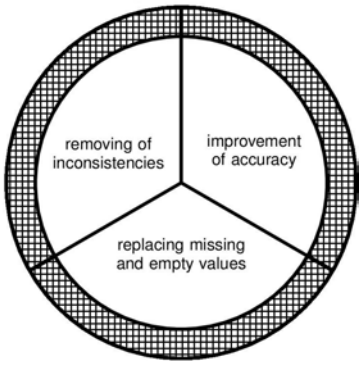


Fig. 2. Operations performed in the data cleaning task (with the introductory stage represented as the squared area and the main stage represented as the white area)

As shown in Fig. 2, the data cleaning task may involve three operations, the replacement of the missing or empty values, the accuracy improvement, and the inconsistency removal, which in the main stage employ algorithmized methods:

- replacement of missing or empty values employs the methods of calculating the missing or empty value with the use of the remaining values of the attribute under replacement or with the use of all attributes of the data set;
- accuracy improvement is based on the algorithmized methods of the replacement of the current value with the newly-calculated one or the removal of this current value;
- inconsistency removal most frequently employs special procedures (e.g. control codes) programmed in data collection sheets prior to the data collecting stage.

In what follows, these three operations will be discussed in detail because of their use in the preparation of the data for modeling.

a. The replacement of missing or empty values

The case of the absent data may cover two kinds of data: an empty value – when a particular piece of data could not have been recorded, since - for instance - such a value of the

measured quantity was not taken into consideration at all, and a missing value – when a particular piece of data has not been recorded, since - for instance - it was lost. The methods of missing or empty value replacement may be divided into two groups, which use for the purpose of calculating the new value the subset of the data containing:

either

- only the values of the attribute under replacement (simple replacement)

or

- either the specially selected or all the values of the collected set (complex replacement).

The methods of simple replacement include all the methods based on statistical values (statistics), such as, for instance the mean value, the median, or the standard deviation. The new values which are calculated via these methods and which are to replace the empty or the missing values retain the currently defined distribution measures of the set.

The complex replacement aims at replacing the absent piece of data with such a value as should have been filled in originally, had it been properly recorded (had it not been lost). Since these methods are aimed at recreating the proper value, a consequence of their application is that the distribution measures of the replacing data set may be and typically are changed.

In complex replacement the absent value is calculated from the data collected in the subset which contains selected attributes (Fig. 3a) or selected records (Fig. 3b) and which has been created specifically for this purpose. The choice of a data replacement method depends on the properties of the collected data – on finding or not finding correlations between or among attributes. If there is any correlation between, on the one hand, selected attributes and, on the other, the attribute containing the absent value, it is possible to establish a multilinear regression which ties the attributes under analysis (the attribute containing the absent value and the attributes which are correlated with it). In turn, on this basis the absent value may be calculated. An advantage of this method is the possibility of obtaining new limits, that is, a new minimal and maximal value of the attribute under replacement, which is lower or higher than the values registered so far. It is also for this reason that this method is used for the replacement of empty data.

When there is no correlation between the attribute of the value under replacement and the remaining attributes of the set, the absent value is calculated via a comparison with the selected records from the set containing complete data. This works when the absent value is a missing value.

The choice of a data replacement method should take into consideration the proposed data modeling method. Simple replacement may be used when the modeling method is insensitive to the noise in the data. For models which cannot cope with the noise in the data complex data replacement methods should be used.

It is a frequent error commonly made in automatized absent data replacement that no record is made with respect to which pieces of the data are originally collected values and which have been replaced. This piece of information is particularly important when the model quality is being analyzed.

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	0,44264	0,33500	0,25427	0,35923	0,16962	0,00716	0,48662	0,32565
2	0,22330	0,28883	0,83385	0,47689	0,70585	0,37254	0,93544	0,56954
3	0,27493	0,33005	0,58590	0,34546	0,01733	0,31112	0,85849	0,45207
4	0,98221	0,48329	0,60996	0,63946	0,36832	0,47123	0,20616	0,87007
5	0,13867	0,04751	0,82575	0,95047	0,74808	0,51304	0,12866	0,92690
6	0,19544	0,64661	0,54804	0,54047	0,19131	0,03098	0,49135	0,71928
7	0,30852	0,16328	0,55166	0,54092	0,54074	0,04924	0,58695	0,66732
8	0,00041	0,25978	0,31942	0,45177	0,03746	0,01925	0,14991	0,46064
9	0,19001	0,25658	0,05029	0,15835	0,33636		0,80948	0,15902
10	0,96184	0,46663	0,95992	0,54820		0,50451	0,27177	0,28748
11	0,66559	0,95384	0,84559	0,76149	0,79613	0,99048	0,73015	0,97383
12	0,44050	0,31420	0,95545	0,69524	0,19512	0,47188	0,56926	0,95379
13	0,34064	0,64198	0,61337	0,27064	0,18290	0,48872	0,95916	0,69138
14	0,54202	0,40815	0,69619	0,49253	0,33158	0,64028	0,37219	0,78562
15	0,19946	0,10397	0,62387	0,90537	0,85044	0,66453	0,30543	0,80890
16	0,41862	0,48597	0,93570	0,73022	0,73567	0,42406	0,68604	0,82964
17	0,06957	0,49556	0,57802	0,45966	0,56085	0,35513	0,91340	0,75443
18	0,45747	0,98770	0,43052	0,36385	0,56939	0,14738	0,62208	0,65532
19	0,09304	0,95937	0,44208	0,61713	0,06287	0,65605	0,26071	0,67018
20	0,07758	0,17060	0,62618	0,93290	0,19343	0,88051	0,74191	0,94996

(a)

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	0,44264	0,33500	0,25427	0,35923	0,16962	0,00716	0,48662	0,32565
2	0,22330	0,28883	0,83385	0,47689	0,70585	0,37254	0,93544	0,16954
3	0,27493	0,33005	0,58590	0,34546	0,01733	0,41112	0,85849	0,45207
4	0,98221	0,48329	0,60996	0,63946	0,36832	0,47123	0,20616	0,87007
5	0,13867	0,04751	0,82575	0,95047	0,74808	0,10304	0,12866	0,82690
6	0,19544	0,64661	0,54804	0,54047	0,19131	0,03098	0,49135	0,71928
7	0,30852	0,16328	0,81166	0,54092	0,54074	0,84924	0,58695	0,66732
8	0,00041	0,25978	0,51942	0,45177	0,03746	0,10925	0,14991	0,46064
9	0,19001	0,25658	0,05029	0,15835	0,33636	0,59725	0,80948	0,15902
10	0,96184	0,46663	0,95992	0,54820		0,50451	0,27177	0,28748
11	0,66559	0,95384	0,84559	0,76149	0,79613	0,99048	0,73015	0,97383
12	0,44050	0,31420	0,95545	0,69524	0,19512	0,47188	0,56926	0,95379
13	0,34064	0,64198	0,14337	0,27064	0,18290	0,48872	0,95916	0,09138
14	0,94202	0,40815	0,69619	0,49253	0,33158	0,64028	0,37219	0,48562
15	0,19946	0,10397	0,32387	0,90537	0,85044	0,36453	0,30543	0,80890
16	0,91862	0,48597	0,93570	0,53022	0,73567	0,52406	0,18604	0,29636
17	0,06957	0,49556	0,07802	0,45966	0,56085	0,35513	0,31340	0,35443
18	0,45747	0,98770	0,43052	0,36385	0,56939	0,14738	0,62208	0,65532
19	0,09304	0,95937	0,14208	0,61713	0,06287	0,65605	0,26071	0,67018
20	0,07758	0,17060	0,62618	0,93290	0,19343	0,88051	0,74191	0,54996

(b)

Fig. 3. (a) Selected attributes (X_3, X_8) will serve in calculating the absent value in attribute X_6 (b) Selected records (4, 14, 16) will serve in calculating the absent value in attribute 10

b. Accuracy improvement

The operation of accuracy improvement, in the main stage of data preparation, is based on algorithmized methods of calculating a value and either replacing the current value (the value recorded in the database) with the newly-calculated one or completely removing the current value from the base. The mode of operation depends upon classifying the value under analysis as either noisy or an outlier. This is why, firstly, this operation focuses on identifying the outliers within the set of the collected data. In the literature, one can find reference to numerous techniques of identifying pieces of the data which are classified as outliers (Alves & Nascimento, 2002; Ben-Gal, 2005; Cateni et al., 2008; Fan et al., 2006; Mohamed et al., 2007). This is an important issue, since it is only in the case of outliers that their removal from the set may be justified. Such an action is justified when the model is developed, for instance, for the purpose of optimizing an industrial process. If a model is being developed with the purpose of identifying potential hazards and break-downs in a process, the outliers should either be retained within the data set or should constitute a new, separate set, which will serve for establishing a separate pattern. On the other hand, noisy data can, at best, be corrected but never removed – unless we have excessive data at our disposal.

In the literature, there are two parallel classifications of the outlier identification methods. The first employs the division into one-dimension methods and multidimension methods. The second divides the relevant methods into the statistical (or traditional) ones and the ones which employ advanced methods of data exploration.

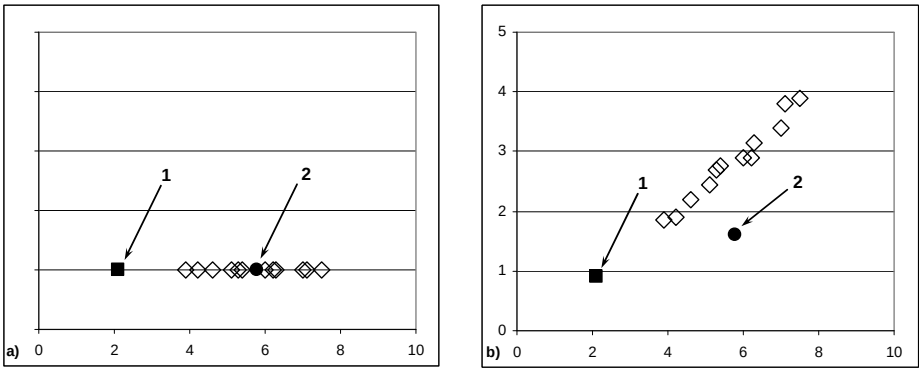


Fig. 4. Outliers a) one-dimension, b) multidimension; 1, 2 outliers – a description in the text

The analysis of one-dimension data employs definitions which recognize as outliers the data which are distant, relative to the assumed metric, from the main data concentration. Frequently, the expanse of the concentration is defined via its mean value and the standard error. In that case, the outlier is a value (the point marked as 1 in Fig. 4a) which is located outside the interval (X_m – k* 2SE, X_m + k* 2SE), where: SE – standard error, k – coefficient, X_m – mean value. The differences between the authors pertain to the value of the k coefficient and the labels for the kinds of data connected with it, for instance the outliers going beyond the boundaries calculated for k = 3,6 (Jimenez-Marquez et al., 2002) or the outliers for k = 1,5 and the extreme data for k = 3 (StatSoft, 2011). Equally popular is the method using a box plot (Laurikkala et al., 2000), which is based on a similar assumption concerning the calculation of the thresholds above and below which a piece of data is treated as an outlier.

For some kind of data, for instance those coming from multiple technological processes and used for the development of industrial applications, the above methods do not work. The data of this kind exhibit a multidimensional correlation. For correlated data a one-dimension analysis does not lead to correct results, which is clearly visible from the example in Fig. 4 (prepared from synthetic data). In accordance with the discussed method, Point 1 (marked as a black square) in Fig. 4a would be classified as an outlier. However, it is clearly visible from multidimensional (two-dimensional) data represented in Fig. 4b that it is point 2 (represented as a black circle) which is an outlier. In one-dimension analysis it was treated as an average value located quite close to the mean value.

Commonly used tools for finding one-dimension outliers are statistical methods. In contrast to the case of one-dimension outliers, the methods used for finding multidimensional outliers are both statistical methods and advanced methods of data exploration. Statistical methods most frequently employ the Mahalanobis metric. In the basic Mahalanobis method

for each vector x_i the Mahalanobis distance is calculated according to the following formula (1) (Rousseeuw & Zomeren, 1990):

$$MD_i = \left(\left(\mathbf{x}_i - T(\mathbf{X}) \right) \mathbf{C}(\mathbf{X})^{-1} \left(\mathbf{x}_i - T(\mathbf{X}) \right)^t \right)^{1/2} \qquad \text{for i} = 1, 2, 3, \ldots n \tag{1}$$

where: $T(\mathbf{X})$ is the arithmetic mean of the data set, $\mathbf{C}(\mathbf{X})$ is the covariance matrix.

This method takes into account not only the central point of the data set, but also the shape of the data set in multidimensional space. A case with a large Mahalanobis distance can be identified as a potential outlier. On the basis of a comparison with the chi-square distribution, an outlier can be removed as was suggested in (Filzmoser, 2005). In the literature, further improvements of the Mahalanobis distance method can be found, for example the one called the Robust Mahalanobis Distance (Bartkowiak, 2005). The outlier detection based on the Mahalanobis distance in industrial data, was performed in e.g. (Jimenez-Marquez et al., 2002). As far as data exploration is concerned, in principle all methods are used. The most popular ones are based on artificial neural networks (Alves & Nascimento, 2002), grouping (Moh'd Belal Al- Zgubi, 2009), or visualization (Hasimah et al., 2007), but there are also numerous works which suggest new possibilities of the use of other methods of data mining, for example of the rough set theory (Shaari et al., 2007). An assumption of methodologies employing data exploration methods is that the data departing from the model built with their help should be treated as outliers.

c. Inconsistency removal

At the introductory stage inconsistencies are primarily removed by hand, via reference to the source materials and records. At this stage it will also encompass the verification of the attributes of the collected data. It is important to remove all redundancies at this stage. Redundant attributes may appear both in basic databases and in databases created via combining a few smaller data sets. Most often, this is a consequence of using different labels in different data sets to refer to the same attributes. Redundant attributes in merged databases suggest that we have at our disposal a much bigger number of attributes than is in fact the case. An example of a situation of this kind is labeling the variables (columns) referring to metal elements in a container as, respectively, *element mass*, *the number of elements in a container*, and *the mass of elements in the container*. If the relevant piece of information is not repeated exactly, for example, if instead of the attribute *the mass of elements in the container* what is collected is *the mass of metal in the container*, then the only problem is increasing the model development time. It is not always the case that this is a significant problem, since the majority of models have a bigger problem with the number of records, than with the number of attributes. However, in the modeling of processes aimed at determining the influence of signal groups, an increase in the number of inputs is accompanied with an avalanche increase in the number of input variable group combinations (Kozlowski, 2009). It should also be remembered that some modeling techniques, especially those based on regression, cannot cope with two collinear attributes (Galmacci, 1996). This pertains also to a small group of matrix-based methods (Pyle, 1999).

At the main stage of data preparation inconsistency removal may be aided with specially designed procedures (e.g. control codes) or tools dedicated to finding such inconsistencies (e.g. when the correlations/interdependencies among parameters are known).

3.2 Data integration

Because of the tools used in knowledge extraction it is necessary that the data be represented in the form of a flat, two-dimensional table. It is becoming possible to analyze the data recorded in a different form, but the column-row structure – as in a calculation sheet - is the best one (Pyle, 2003). The task of data integration may be both simple and complicated, since it depends on the form in which the data is collected. In the case of synthetic data, as well as in appropriately prepared industrial data collecting systems this is unproblematic. However, the majority of industrial databases develop in an uncoordinated way. Different departments of the same plant develop their own databases for their own purposes. These databases are further developed without taking into consideration other agencies. This results in a situation in which the same attributes are repeated in numerous databases, the same attributes are labeled differently in different databases, the same attributes have different proportions of absent data in different bases, the developed databases have different primary keys (case identifiers), etc. The situation gets even worse when the databases under integration have been developed not within a single plant but in more distant places, for example in competing plants.

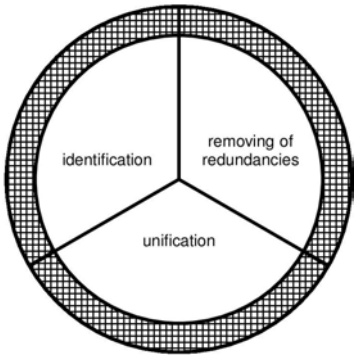


Fig. 5. The operations performed in the data integration task (with the introductory stage represented as the squared area and the main stage represented as the white area)

As represented in Fig 5, data integration consists of three operations. The introductory stage of data integration focuses on two of them: identification and redundancy removal. Both these operations are closely connected with one another. The first operation – identification – is necessary, since the task of data integration starts with identifying the quantity around which the database sets may be integrated. Also, at the data integration introductory stage the removal of obvious redundancies should be performed. Their appearance may result, for instance, from the specific way in which the industrial production data is recorded. The results of tests conducted in centrally-operated laboratories are used by different factory departments, which store these results independently of one another, in their own databases. The result of data integration performed without introductory analysis may be that the same quantities may be listed in the end-product database a number of times and, moreover, they may be listed there under different labels. A person who knows the data under analysis will identify such redundant records without any problem.

At the main stage, unlike in other data preparation tasks, the integration operations are only performed with the aid of algorithmized tools and are still based on the knowledge about the process under analysis. At this stage, the operations represented in Fig. 5 are performed with the following aims:

- identification – serves the purpose of identifying the attributes which could not have been identified at the introductory stage, for instance, in those cases in which the label does not explain anything,
- redundancy removal – just like the attribute selection in the data reduction task, which is characterized below, uses algorithmized methods of comparing the attributes which have been identified only in the main stage with the aim of removing the redundant data,
- unification – this is performed so that the data collected in different sets have the same form, for instance the same units.

The operation of identification uses methods which make it possible to identify the correlation between the attribute under analysis and other identified process attributes. In this way it is possible to establish, for instance, what a particular attribute pertains to, where the sensor making the recorded measurement could have been installed, etc.

The second performance of the redundancy removal operation pertains only to attributes which have been identified only in the second, main stage of data integration. As far as this second performance is concerned, redundancy removal may be identified with the attribute selection operation from the data reduction task and it uses the same methods as the attribute selection operation.

Unification is very close to the data transformation task. In most cases it amounts to transforming all the data in such a way that they are recorded in a unified form, via converting part of the data using different scales into the same units, for instance, converting inches into millimeters, meters into millimeters, etc. A separate issue is the unification of the data collected in different sets and originating from measurements employing different methods. In such cases one should employ the algorithms for converting one attribute form into another. These may include both analytical formulas (exact conversion) and approximate empirical formulas. When this is not possible, unification may be achieved via one of the data transformation operations, that is, via normalization (Liang & Kasabov, 2003).

The main principle that the person performing the data integration task should stick to is that the integrated database should retain all the information collected in the data sets which have been integrated. Seemingly, the suggestion, expressed in (Witten & Frank, 2005), that data integration should be accompanied by aggregation, is misguided. Aggregation may accompany integration, but only when this is absolutely necessary.

3.3 Data transformation

Data transformation introductory stage usually amounts to a single operation dictated by data integration, such as aggregation, generalization, normalization or attribute (feature) construction. Performing aggregation may be dictated by the fact that the data collected in multiple places during the production process may represent results from different periods:

a different hour, shift, day, week or month. Finding a common higher-order interval of time is necessary for further analyses. The same reasons make necessary the performance of generalization or normalization. Generalization may be dictated, for instance, by the integration of databases in which the same attribute is once recorded in a nominal scale (e.g. ultimate tensile strength - ductile iron grade 500/07) and once in a ratio scale (ultimate tensile strength – UTS = 572 MPa). We can talk about normalization in the introductory stage only when we understand this term in its widest sense, that is, as a conversion of quantities from one range into those of another range. In that case, such a transformation as measurement unit conversion may be understood as normalization at the introductory stage.

Data transformation encompasses all the issues connected with transforming the data into a form which makes data exploration possible. At the introductory stage it involves six operations represented in Fig. 6:

- smoothing – this resides in transforming the data in such a way that local data deviations having the character of noise are eliminated. Smoothing encompasses, among others, the 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 converting the collected data containing the measurements of the registered process quantity into higher-order quantities, for instance via their discretization;
- normalization – this resides in the rescaling (adjustment) of the data to a specified, narrow range, for instance, from 0.0 to 1.0;
- attribute (feature) construction – this resides in mathematical transformations of attributes (features) with the aim of obtaining a new attribute (feature), which will replace in modeling its constituent attributes;
- 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).

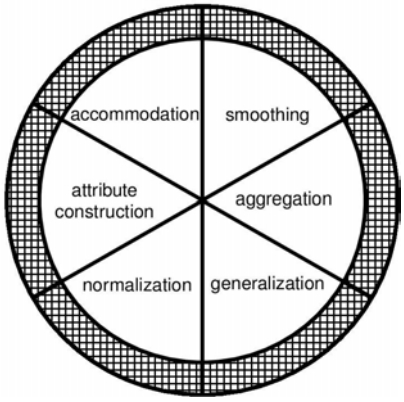


Fig. 6. Operations performed in the data transformation task (with the introductory stage represented as the squared area and the main stage represented as the white area).

Smoothing is an operation which follows from the assumption that the collected data are noisy with random errors or with divergence of measured attributes. In the case of industrial (also laboratory) data such a situation is a commonplace, unlike in the case of business data (the dollar exchange rate cannot possibly be noisy). Smoothing techniques are aimed at removing the noise, without at the same time interfering with the essence of the measured attributes. Smoothing methods may be of two kinds: (a) those which focus on comparing the quantity under analysis with its immediate neighborhood and (b) those which analyze the totality of the collected data.

Group (a) includes methods which analyze a specified lag or window. In the former case, the data are analyzed in their original order, while in the latter they are changed in order. These methods are based on a comparison with the neighboring values and use, for instance, the mean, the weighted moving mean, or the median. Group (b) includes methods employing regression or data clustering. The method of loess – local regression could be classified as belonging to either group.

In the first group of methods (a) a frequent solution is to combine a few of techniques into a single procedure. This is supposed to bring about a situation in which further smoothing does not introduce changes into the collected data. A strategy of this kind is resmoothing, in which two approaches may be adopted: either 1) smoothing is continued up to the moment when after a subsequent smoothing the curve does not change or 2) a specified number of smoothing cycles is performed, but this involves changing the window size. Two such procedures, 4253H and 3R2H were discussed in (Pyle, 1999). A method which should also be included in the first group (a) is the PVM (peak – valley – mean) method.

Binning, which belongs to group (a) is a method of smoothing (also data cleaning) residing in the creation of bins and the ascription of data to these bins. The data collected in the respective bins are compared to one another within each bin and unified via one of a few methods, for example, via the replacement with the mean value, the replacement with the median, or the replacement with the boundary value. A significant parameter of smoothing via binning is the selection of the size of bins to be transformed. Since comparing is performed only within the closest data collected in a single bin, binning is a kind of local smoothing of the data.

As was already mentioned above, the second important group of smoothing methods (b) are the techniques employing all the available data. Among others, this includes the methods of regression and of data clustering.

Fig.7 represents the smoothing of the laboratory data attribute (ultimate compressive strength) of greensand in the function of moisture content. The data may be smoothed via adjusting the function to the measured data. Regression means the adjustment of the best curve to the distribution of two attributes, so that one of these attributes could serve for the prediction of the other attribute. In multilinear regression the function is adjusted to the data collected in the space which is more than two-dimensional.

An example of smoothing with the use of data clustering techniques was discussed in (Kochanski, 2006). In the data which is grouped the outliers may easily be detected and either smoothed or removed from the set. This procedure makes possible defining the range of variability of the data under analysis.

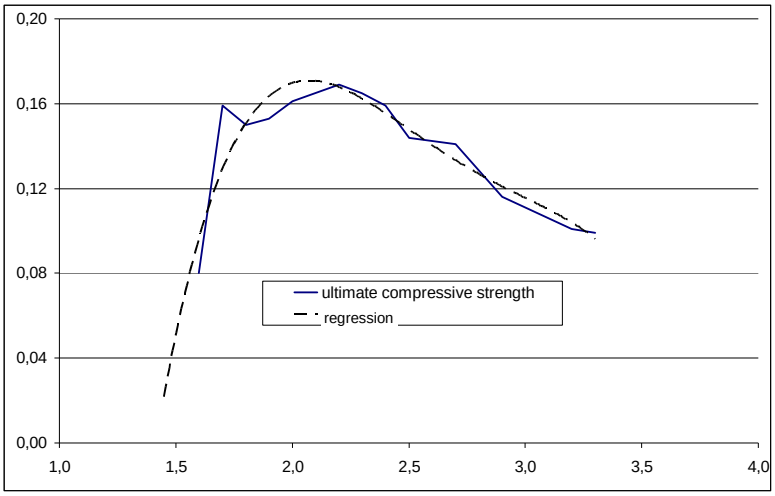


Fig. 7. The regression method applied to the industrial data: the ultimate compressive strength of greensand in the function of moisture content (the authors' own research)

In the case of industrial data the issue of smoothing should be treated with an appropriate care. The removal or smoothing of an outlier may determine the model's capability for predicting hazards, break-downs or product defects. The industrial data often contain quantities which result from the process parameter synergy. The quantity which may undergo smoothing is, for instance, the daily or monthly furnace charge for the prediction of the trend in its consumption, but not the form moisture content for the prediction of the appearance of porosity.

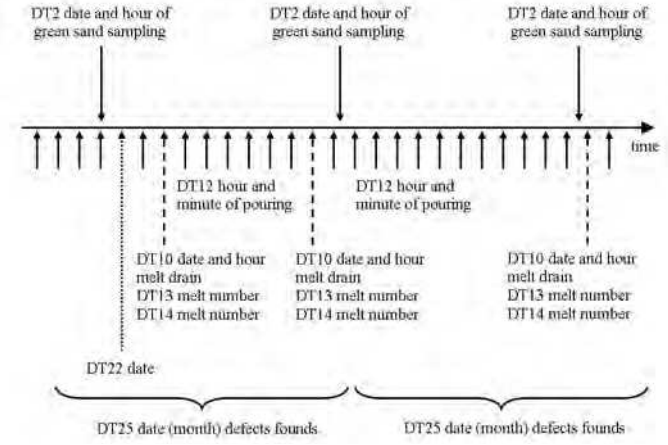


Fig. 8. Measurements performed with different frequency and encompassing different time periods: from a measurement encompassing 1 minute to a measurement encompassing the data from an entire month; (the data are collected in different places: DTxx – a symbol for the process technical documentation) (Research report, 2005)

In manufacture processes, when different operations are performed in different places and records are collected in separate forms, a frequent situation is the impossibility of integrating the separate cases into a single database. The absence of a single key makes data integration impossible. (Research report, 2005) discusses data integration with the use of aggregation. Data aggregation (represented as parentheses) and data location on the time axis, which makes possible further integration, are diagrammed in Fig. 8.

Generalization in data preparation is an operation which is aimed at reducing the number of the values that an individual attribute may take. In particular, it converts a continuous quantity into an attribute which takes the value of one of the specified ranges. The replacement of a continuous quantity with a smaller number of labeled ranges makes easier grasping the nature of the correlation found in the course of data exploration. Generalization applies to two terms: discretization and notional hierarchy.

Discretization is a method which makes possible splitting the whole range of attribute variation into a specified number of subregions. Discretization methods in the two main classifications are characterized in terms of the direction in which they are performed or in terms of whether or not in feature division they employ information contained in the features other than the one under discretization. In the first case, we talk about bottom-up or top-down methods. In the second case, we distinguish between supervised and unsupervised methods.

The notional hierarchy for a single feature makes possible reducing the number of the data via grouping and the replacement of a numerical quantity, for instance the percentage of carbon in the chemical composition of an alloy, with a higher-order notion, that is, *a low-carbonic* and *a high-carbonic alloy*. This leads to a partial loss of information but, in turn, thanks to the application of this method it may be easier to provide an interpretation and in the end this interpratation may become more comprehensible.

There are many commonly used methods of data normalization (Pyle, 1999; Larose, 2008). The effect of performing the operation of normalization may be a change in the variable range and/or a change in the variable distribution. Some data mining tools, for example artificial neural networks, require normalized quantities. Ready-made commercial codes have special modules which normalize the entered data. When we perform data analysis, we should take into consideration the method which has been employed in normalization. This method may have a significant influence on the developed model and the accuracy of its predictions (Cannataro, 2008; Ginoris et al., 2007; Al Shalabi et al., 2006).

The following are the main reasons for performing normalization:

- a transformation of the ranges of all attributes into a single range makes possible the elimination of the influence of the feature magnitude (their order 10, 100, 1000) on the developed model. In this way we can avoid revaluing features with high values,
- a transformation of the data into a dimensionless form and the consequent achievement of commensurability of a few features makes possible calculating the case distance, for instance with the use of the Euclid metric,
- a nonlinear transformation makes possible relieving the frequency congestion and uniformly distributing the relevant cases in the range of features,
- a nonlinear transformation makes it possible to take into consideration the outliers,

- data transformation makes possible a comparison of the results from different tests (Liang & Kasabov, 2003).

The following may be listed as the most popular normalization methods:

- min-max normalization – it resides, in the most general form, in the linear transformation of the variable range in such a way that the current minimum takes the value 0 (zero), while the current maximum takes the value 1 (one). This kind of normalization comes in many variants, which differ from one another, among others, with respect to the new ranges, for instance (-1,1); (-0,5, 0,5);
- standarization – it resides in such a transformation in which the new feature has the expectation mean value which equals zero and the variance which equals one. The most frequently used standardization is the Z-score standardization, which is calculated according to the following formula (2):

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

where: x – feature under standarization, μ – mean value, σ – standard population deviation;

- decimal normalization, in which the decimal separator is moved to the place in which the following equation (3) is satisfied:

$$X^1 = \frac{X}{10^j} \tag{3}$$

where: j is the smallest integer value, for which $Max\ (|X^1|) < 1$.

A weakness of the normalization methods discussed above is that they cannot cope with extreme data or with the outliers retained in the set under analysis. A solution to this problem is a linear – nonlinear normalization, for example the soft-max normalization or a nonlinear normalization, for example the logarithmic one (Bonebakker, 2007).

As was mentioned above, the nonlinear normalization makes it possible to include into the modeling data set extreme data or even outliers, as well as to change the distribution of the variable under transformation. The logarithmic transformation equations given in (4a,b) and the tangent transformation equation in (5) were employed in (Olichwier, 2011):

$$X_i^{'} = w \cdot \lg(c + X_i) \tag{4a}$$

$$X_i^{'} = w \cdot \lg(1 + cX_i) \tag{4b}$$

where: w, c - coefficients

$$X_i^{'} = \frac{1}{2}[\tanh(k \cdot (\frac{X_i - \mu}{\sigma}) + 1] \tag{5}$$

where: k - coefficient, μ – mean value, σ – standard population deviation.

The use of the logarithmic transformation made it possible to locally spread the distributions of the selected parameters. The original skew parameter distribution after the

transformation approached the normal distribution. One effect of the transformation in question was a visible increase in the prediction quality of the resulting model.

The data in the databases which are under preparation for mining may be collected in many places and recorded in separate sets. As a result of integrating different data sources it is possible to perform their transformation, for instance in the domain of a single manufacture process. The data represented by two or more features may be replaced with a different attribute. A classical example of this is the replacement of the two dates defining the beginning and the end of an operation, which were originally recorded in two different places, with a single value defining the duration of the operation, for instance the day and the hour of molding, as well as the day and the hour of the mould assembly are replaced with the mould drying time. A deep knowledge of the data makes possible using mathematical transformations which are more complex than a mere difference. As a consequence, new, semantically justified feature combinations are created. An example of this was discussed in (Kochanski, 2000).

The knowledge of the properties of the algorithm employed in data mining is a different kind of reason behind the creation of new attributes. In the case of algorithms capable only of the division which is parallel to the data space axis (for instance in the case of the majority of decision trees) it is possible to replace the attributes pertaining to features characterized with linear dependence with a new attribute which is the ratio of the former attributes (Witten & Frank, 2005).

It is possible to perform any other transformations of attributes which do not have any mathematical justification but which follow from the general world knowledge: the names of the days of the week may be replaced with the dates, the names of the elements may be replaced with their atomic numbers, etc.

The last encountered way of creating attributes is the replacement of two attributes with a new attribute which is their product. A new attribute created in this way may have no counterpart in reality.

The common use of accommodation suggests that it should be considered as one of the issues of data transformation. What is used in data mining are databases which have been created without any consideration for their future specific application, that is, the application with the use of selected tools or algorithms. In effect, it is often the case that the collected data format does not fit the requirements of the tool used for mining, especially in a commercial code. The popular ARFF (attribute - relation file format) makes use only of the data recorded in a nominal or interval scale. The knowledge gathered in the data recorded in a ratio scale will be lost as a result of being recorded in the ARFF format.

The calculation spreadsheet Excel, which is popular and which is most frequently used for gathering the industrial data, may also be the cause of data distortion. Depending upon the format of the cell in which the registered quantity is recorded, the conversion of files into the txt format (which is required by a part of data mining programs) may result in the loss of modeling quality. This is a consequence of the fact that a defined cell contains the whole number which was recorded in it but what is saved in file conversion is only that part of this number which is displayed (Fig.9).

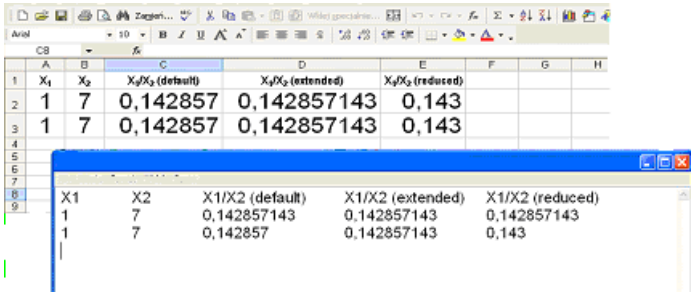


Fig. 9. Data conversion – the conversion of a file from the xls to the txt format brings about a partial loss of information (the authors' own work)

3.4 Data reduction

At the introductory stage of data preparation this task is limited to a single operation – attribute selection – and it is directly connected with the expert opinion and experience pertaining to the data under analysis. It is only as a result of a specialist analysis (for example, of brainstorming) that one can remove from the set of collected data those attributes which without question do not have any influence on the features under modeling.

The aim of the main stage data reduction techniques is a significant decrease in the data representation, which at the same time preserves the features of the basic data set. Reduced data mining makes then possible obtaining models and analyses which are identical (or nearly identical) to the analyses performed for the basic data.

Data reduction includes five operations, which are represented in Fig. 10:

- attribute selection – this resides in reducing the data set by eliminating from it the attributes which are redundant or 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;
- numerosity reduction – this is aimed at reducing the data set via eliminating the recurring or very similar cases;
- discretization – this resides in transforming a continuous variable into a limited and specified number of ranges;
- aggregation – this resides in summing up the data, most frequently in the function of time, for instance from a longer time period encompassing not just the period of a single shift but e.g. the period of a whole month.

The collected industrial database set may contain tens or even hundreds of attributes. Many of these attributes may be insignificant for the current analysis. It is frequently the case that integrated industrial databases contain redundant records, which is a consequence of the specific way in which the data is collected and stored in a single factory in multiple places. Keeping insignificant records in a database which is to be used for modeling may lead not only to – often a significant – teaching time increase but also to developing a poor quality model. Of course, at the introductory stage, a specialist in a given area may identify the

group of attributes which are relevant for further modeling, but this may be very time-consuming, especially given that the nature of the phenomenon under analysis is not fully known. Removing the significant attributes or keeping the insignificant ones may make it very difficult, or even impossible, to develop an accurate model for the collected data.

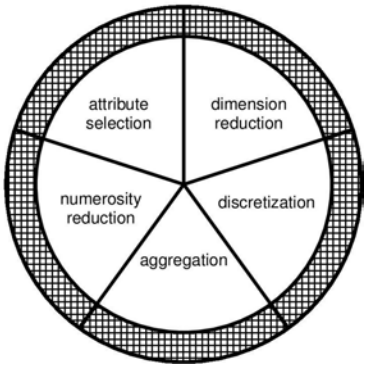


Fig. 10. The operations performed in the data reduction task (with the introductory stage represented as the squared area and the main stage represented as the white area)

At the main stage, attribute selection is performed without any essential analysis of the data under modeling, but only with the help of selection algorithms. This is supposed to lead to defining the intrinsic (or effective) dimensionality of the collected data set (Fukunaga, 1990) The algorithms in question are classified as either *filters* or *wrappers*, depending on whether they are employed as the operation preceding the modeling proper (filters) or as the operation which is performed alternately with the modeling (wrappers).

Attribute selection is performed for four reasons: (a) to reduce the dimensionality of the attribute space, (b) to accelerate the learning algorithms, (c) to increase the classification quality, and (d) to facilitate the analysis of the obtained modeling results (Liu et al., 2003). According to the same authors, it may – in particular circumstances – increase the grouping accuracy.

Dimensionality reduction is the process of creating new attributes via a transformation of the current ones. The result of dimensionality reduction is that the current inventory of attributes $X\{x_1,x_2,\ldots,x_n\}$ is replaced with a new one $Y\{y_1,y_2,\ldots,y_m\}$ in accordance with the following equation (6):

$$Y = F(x_1, x_2, \ldots, x_n) \tag{6}$$

where F() is a mapping function and $m < n$. In the specific case $Y_1 = a_1x_1 + a_2x_2$, where a_1 and a_2 are coefficients (Liu & Motoda, 1998). The same approach, which takes into consideration in its definition the internal dimensionality, may be found in (Van der Maaten et al., 2009). The collected data are located on or in the neighborhood of a multidimensional curve. This curve is characterized with m attributes of the internal dimensionality, but is placed in the n-dimensional space. The reduction frees the data from the excessive dimensionality, which is important since it is sometimes the case that $m<<n$.

The operation of dimensionality reduction may be performed either by itself or after attribute selection. The latter takes place when the attribute selection operation has still left a considerable number of attributes used for modeling. The difference between dimensionality reduction and attribute selection resides in the fact that dimensionality reduction may lead to a partial loss of the information contained in the data. However, this often leads to the increase in the quality of the obtained model. A disadvantage of dimensionality reduction is the fact that it makes more difficult the interpretation of the obtained results. This results from the impossibility of naming the new (created) attributes. In some publications attribute selection and dimensionality reduction are combined into a single task of data dimensionality reduction[1] (Chizi & Maimon, 2005; Fu & Wang, 2003; Villalba & Cunningham, 2007). However, because of the differences in the methods, means, and aims, the two operations should be considered as distinct.

Discretization, in its broadest sense, transforms the data of one kind into the data of another kind. In the literature, this is discussed as the replacement of the quantitative data with the qualitative data or of the continuous data with the discrete data. The latter approach is the approach pertaining to the most frequent cases of discretization. It should be remembered, however, that such an approach narrows down the understanding of the notion of discretization. This is the case since both the continuous and the discrete data are numerical data. The pouring temperature, the charge weight, the percent element content, etc. are continuous data, while the number of melts or the number of the produced casts are discrete data. However, all the above-mentioned examples of quantities are numbers. Discretization makes possible the replacement of a numerical quantity, for instance, the ultimate tensile strength in MPa with the strength characterized verbally as high, medium, or low. In common practice ten methods of discretization method classification are used. These have been put forward and characterized in (Liu et al., 2002; Yang et al., 2005). A number of published works focus on a comparison of discretization methods which takes into account the influence of the selected method on different aspects of further modeling, for instance, decision trees. These comparisons most frequently analyze three parameters upon which data preparation – discretization exerts influence:

- the time needed for developing the model,
- the accuracy of the developed model,
- the comprehensibility of the model.

Comparative analyses are conducted on synthetic data, which are generated in accordance with a specified pattern (Ismail & Ciesielski, 2003; Boullé, 2006), on the widely known data sets, such as Glass, Hepatitis, Iris, Pima, Wine, etc. (Shi & Fu 2005, Boullé, 2006; Ekbal, 2006; Wu QX et al., 2006; Jin et al., 2009; Mitov et al., 2009), and on production data (Perzyk, 2005).

As has been widely demonstrated, the number of cases recorded in the database is decisive with respect to the modeling time. However, no matter what the size of the data set is, this time is negligibly short in comparison with the time devoted to data preparation. Because of that, as well as because of a small size, in comparison, for instance, with the business data sets, numerosity reduction is not usually performed in industrial data sets.

[1]Data Dimensionality Reduction DDR, Dimension Reduction Techniques DRT

The operation of aggregation was first discussed in connection with the discussion of the data transformation task. The difference between these two operations does not reside in the method employed, since the methods are the same, but in the aims with which aggregation is performed. Aggregation performed as a data reduction operation may be treated, in its essence, as the creation of a new attribute, which is the sum of other attributes, a consequence of this being a reduction in the number of events, that is, numerosity reduction.

4. The application of the selected operations of the industrial data preparation methodology

For the last twenty years, many articles have been published which discuss the results of research on ductile cast iron ADI. These works discuss the results of research conducted with the aim of investigating the influence of the parameters of the casting process, as well as of the heat treatment of the ductile iron casts on their various properties. These properties include, on the one hand, the widely investigated ultimate tensile strength, elongation, and hardness, and, on the other, also properties which are less widely discussed, such as graphite size, impact, or austenite fraction. The results discussed in the articles contain large numbers of data from laboratory tests and from industrial studies. The data set collected for the present work contains 1468 cases coming both from the journal-published data and from the authors' own research. They are characterized via 27 inputs, such as: the chemical composition (characterized with reference to 14 elements), the structure as cast (characterized in terms of 7 parameters), the features as cast (characterized in terms of 2 parameters), the heat treatment parameters (characterized in terms of 4 parameters), as well as via 11 outputs: the cast structure after heat treatment, characterized in terms of the retained austenite fraction and the cast features (characterized in terms of 10 parameters). 13 inputs were selected from the set, 9 of them characterizing the melt chemical composition and 4 characterizing the heat treatment parameters. Also, 2 outputs were selected – the mechanical properties after treatment – the ultimate tensile strength and the elongation. The set obtained in this way, which contained 922 cases, was prepared as far as data cleaning is concerned, in accordance with the methodology discussed above.

Prior to preparation, the set contained only 34,8 % of completely full records, containing all the inputs and outputs. The set contained a whole range of cases in which outliers were suspected.

For the whole population of the data set under preparation outliers and high leverage points were defined. This made possible defining influentials, that is, the points exhibiting a high value for Cook's distance. In Fig 11, which represents the elongation distribution in the function of the ultimate tensile strength, selected cases were marked (a single color represents a single data source). The location of these cases in the diagram would not make it possible to unequivocally identify them as outliers. However, when this is combined with an analysis of the cases with a similar chemical composition and heat treatment parameters, the relevant cases may be identified as such.

Fig. 12 below represents the generated correlation matrix of the collected data. The high level of the absent data in the database is visible, among others, in the form of the empty values of the correlation coefficient – for instance, the database contains no case of a simultaneous appearance of both *Al contents* and *participation of graphite nodules*.

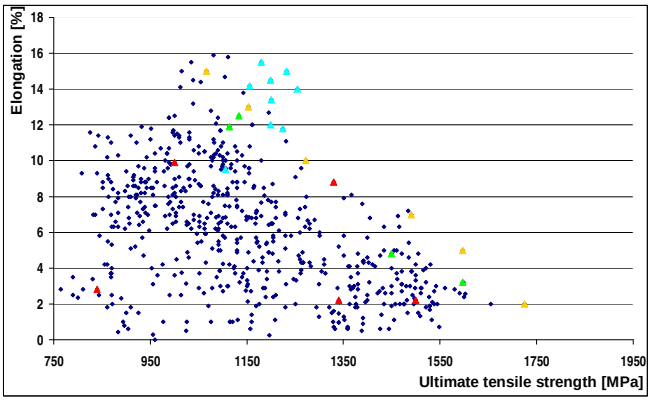


Fig. 11. The elongation distribution in the function of the ultimate tensile strength. The colored pints represent cases identified as outliers (the authors' own work)]

With the use of the methodology discussed above, the database was filled in, which resulted in a significant decrease in the percentage of the absent data for the particular inputs and outputs. The degree of the absent data was between 0,5 and 28,1% for inputs and 26,9% for the UTS and 30,6% for the elongation. The result of filling in the missing data was that the degree of the absent data for inputs was reduced to the value from 0,5 to 12,8%, and for strength and elongation to, respectively, 1,3% and 0,8%. Fig. 13 represents the proportion of the number of records in the particular output ranges "before" and "after" this replacement operation. Importantly, we can observe a significant increase in the variability range for both dependent variables, that is, for elongation from the level of 0÷15% to the level of 0÷27,5%, and for strength from the level of 585÷1725MPa to the level of 406÷1725MPa, despite the fact that the percentage of the records in the respective ranges remained at a similar level for the data both "before" and "after" the filling operation. Similar changes occurred for the majority of independent variables – inputs. In this way, via an increase in the learning data set, the reliability of the model was also increased. Also, the range of the practical applications of the taught ANN model was increased, via an increase in the variability range of the parameters under analysis.

The data which was prepared in the way discussed above were then used as the learning set for an artificial neural network (ANN). Basing on the earlier research [4], two data sets were created which characterized the influence of the melt chemical composition and the heat treatment parameters (Tab. 1) of ductile cast iron ADI alloys on the listed mechanical properties. The study employed a network of the MLP type, with one hidden layer and with the number of neurons hidden in that layer equaling the number of inputs. For each of the two cases, 10 learning sessions were run and the learning selected for further analysis was the one with the smallest mean square error.

A qualitative analysis of the taught model demonstrated that the prediction quality obtained for the almost twice as numerous learning data set obtained after the absent data replacement was comparable, which is witnessed by a comparable quantitative proportion of errors obtained on the learning data set for both cases (Fig. 14).

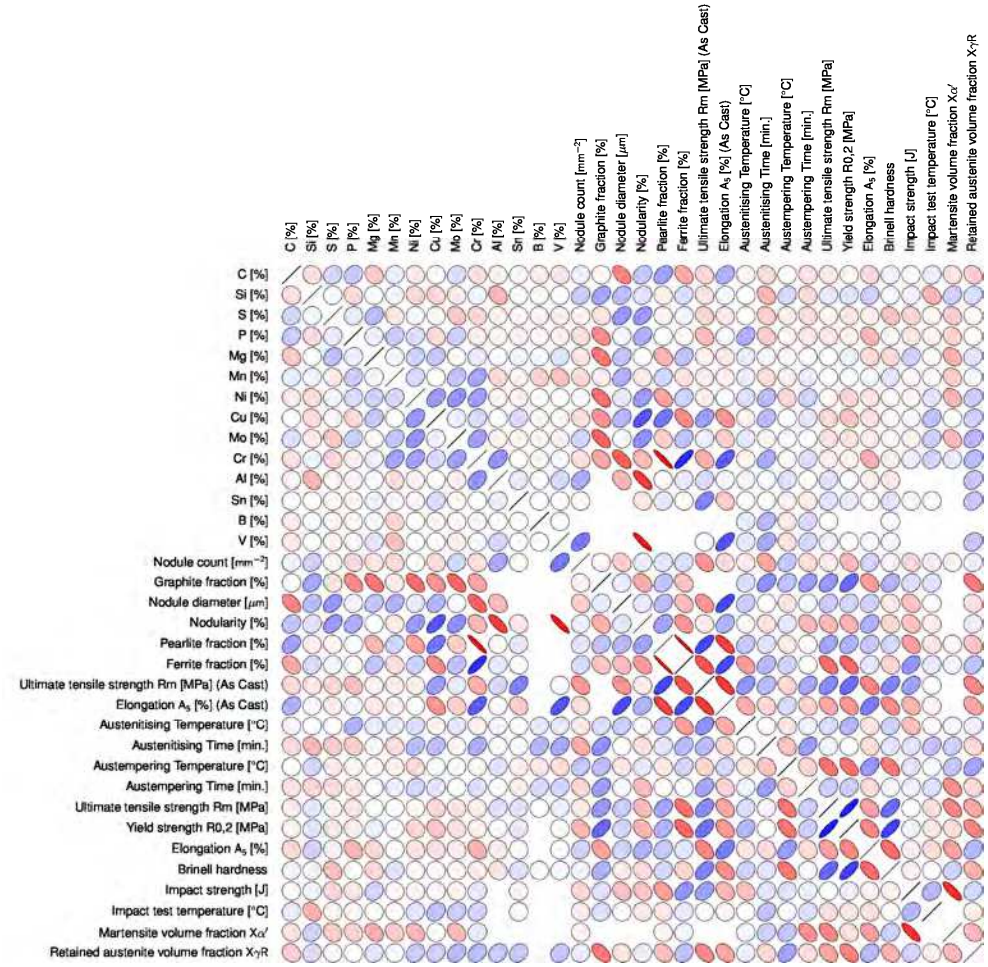


Fig. 12. The input and output correlation matrix. The blue color represents the positive correlation while the red color represents the negative one; the deformation size correlation, of the circle (ellipsoidality) and the color saturation indicate the correlation strength (Murdoch & Chow, 1996)

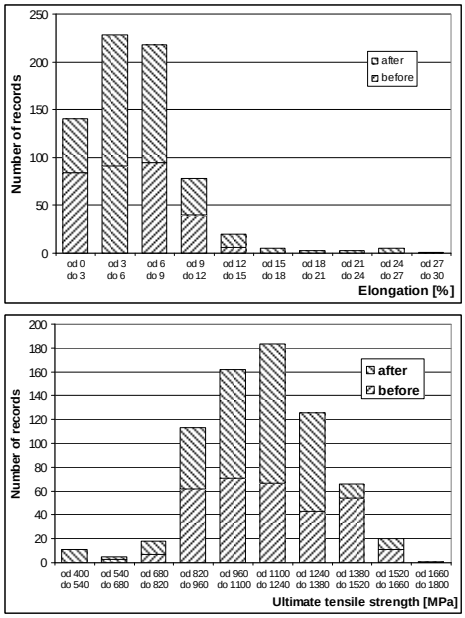


Fig. 13. The proportions of the numbers of records in the specified ranges of the output variables a) elongation [%], b) ultimate tensile strength [MPa] before and after replacing the missing and empty values

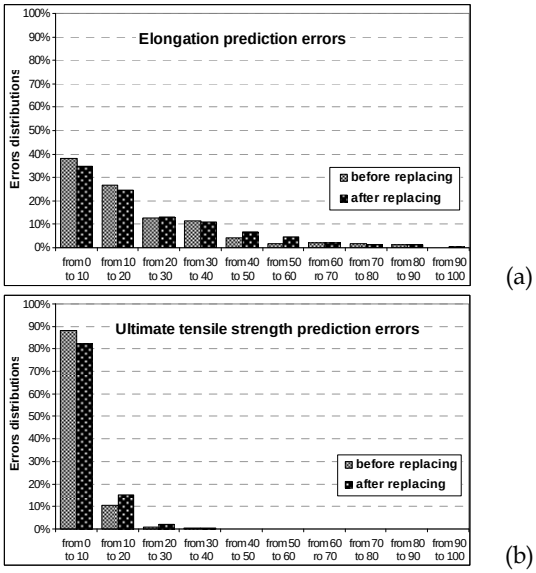


Fig. 14. Output prediction error distribution a) elongation [%], b) ultimate tensile strength [MPa]

A further confirmation of this comes from a comparative analysis of the real results and the ANN answers which was based on correlation graphs and a modified version of the correlation graph, taking into account the variable distribution density (Fig. 15 and 16). In all cases under analysis we can observe that the ANN model shows a tendency to overestimate the minima and to underestimate the maxima. This weakness of ANNs was discussed in (Kozlowski, 2009).

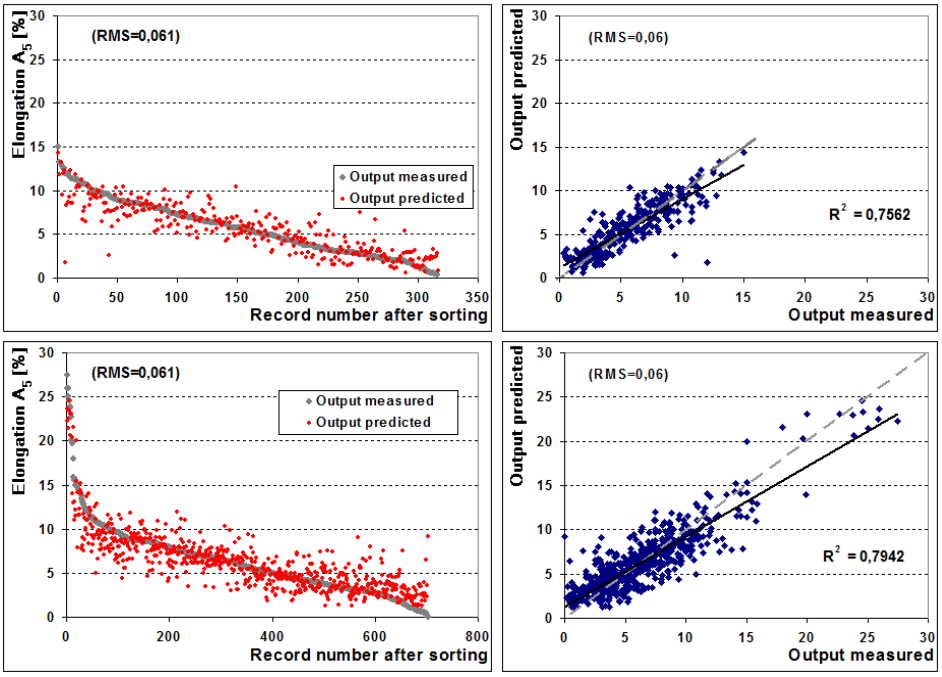


Fig. 15. The distribution of the real results and the predictions of the elongation variable for a) the data set without replacement; b) the data set after replacement

The analyzed cases suggest that an ANN model taught on a data set after absent data replacement exhibits similar and at the same high values of the model prediction quality, that is, of the mean square error, as well as of the coefficient of determination R^2. In the case of predicting A_5, the more accurate model was the ANN model based on the data set after replacement (for R_m the opposite was the case). However, the most important observable advantage following from data replacement and from the modeling with the use of the data set after replacement is the fact that the ANN model increased its prediction range with respect to the extreme output values (in spite of a few deviations and errors in their predictions). This is extremely important from the perspective of the practical model application, since it is the extreme values which are frequently the most desired (for instance, obtaining the maximum value R_m with, simultaneously, the highest possible A_5), and, unfortunately, the sources – for objective reasons – usually do not spell out the complete data, for which these values were obtained. In spite of the small number of records characterizing the extreme outputs, an ANN model was successfully developed which can

make predictions in the whole possible range of dependent variables. This may suggest that the absent data was replaced with appropriate values, reflecting the obtaining general tendencies and correlations. It should also be noted that the biggest prediction errors occur for the low values of the parameters under analysis (e.g.: for A_5 within the range 0÷1%), which may be directly connected with the inaccuracy and the noise in the measurement (e.g.: with a premature break of a tensile specimen resulting from discontinuity or non-metallic inclusions).

The application of the proposed methodology makes possible a successful inclusion into data sets of the pieces of information coming from different sources, including also uncertain data.

An increase in the number of cases in the data set used for modeling results in the model accuracy increase, at the same time significantly widening the practical application range of the taught ANN model, via a significant increase, sometimes even doubling, of the variability range of the parameters under analysis.

Multiple works suggested that ANN models should not be used for predicting results which are beyond the learning data range. The methodology proposed above makes possible the absent data replacement and therefore, the increase of the database size. This, in turn, makes possible developing models with a significantly wider applicability range.

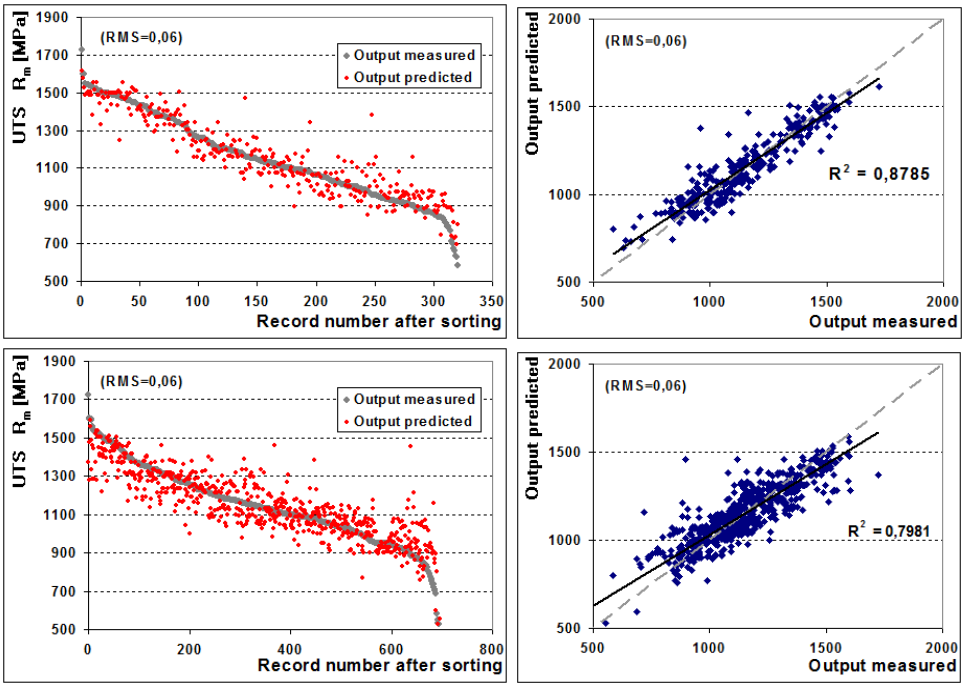


Fig. 16. The distribution of the real results and the predictions of the UTS (ultimate tensile strength) variable for a) the data set without replacement; b) the data set after replacement

5. Conclusion

The proposed methodology makes possible the full use of imperfect data coming from various sources. Appropriate preparation of the data for modeling via, for instance, absent data replacement makes it possible to widen, directly and indirectly, the parameter variability range and to increase the set size. Models developed on such sets are high-quality models and their prediction accuracy can be more satisfactory.

The consequent wider applicability range of the model and its stronger reliability, in combination with its higher accuracy, open a way to a deeper and wider analysis of the phenomenon or process under analysis.

6. References

Agre G. & Peev S. (2002). On Supervised and Unsupervised Discretization, *Cybernetics and information technologies*, Vol. 2, No. 2, pp. 43-57, Bulgarian Academy of Science, Sofia

Al Shalabi L. & Shaaban Z. (2006). Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, *Proceedings of the International Conference of Dependability of Computer Systems (DEPCOS-RELCOMEX'06)*, ISBN: 0-7695-2565-2

Al Shalabi L.; Shaaban Z. & Kasasbeh B. (2006). Data Mining: A Preprocessing Engine, *Journal of Computer Science* 2 (9); pp. 735-739, ISSN 1549-3636

Alves R.M.B. & Nascimento C.A.O. (2002). Gross errors detection of industrial data by neural network and cluster techniques, *Brazilian Journal of Chemical Engineering*, Vol. 19, No. 04, pp. 483 - 489, October - December 2002

Bartkowiak A. (2005). Robust Mahalanobis distances obtained using the 'Multout' and 'Fast-mcd' methods, *Biocybernetics and Biomedical Engineering*, Vol. 25, No. 1, pp. 7-21,

Ben-Gal I. (2005). Outlier detection, *In Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, edited by Maimon O. and Rockach L., Kluwer Academic Publishers, ISBN 0-387-24435-2

Bensch M.; Schröder M.; Bogdan M. & Rosenstiel W. (2005). Feature Selection for High-Dimensional Industrial Data, *ESANN2005 proceeding – European Symposium on Artificial Neural Networks,* Bruges, Belgium, ISBN2-930307-05-6, 27-29 April, 2005

Bonebakker J. L. (2007). Finding representative workloads for computer system design, *Sun Microsystems*, ISBN/EAN: 978-90-5638-187-5

Boullé M. (2006). MODL: A Bayes optimal discretization method for continuous attributes, *Journal Machine Learning*, Vol. 65, No. 1, pp. 131 – 165, ISSN 0885-6125 (Print) 1573-0565 (Online)

Cannataro M. (2008). Computational proteomics: management and analysis of proteomics data, *Briefings in Bioinformatics*, Vol. 9, No. 2, pp. 97-101

Cateni S.; Colla V. & Vannucci M. (2008). Outlier Detection Methods for Industrial Applications in Advances and Robotics, *Automation and Control*, edited by Aramburo J. and Trevino A.R., Publisher InTech, ISBN 978-953-7619-16-9

Chizi B. & Maimon O. (2005). Dimension reduction and feature selection, *In Data Mining and Knowledge Discovery Handbook*, Ch. 5, pp. 93–111, edited by: Maimon O., Rokach L., *Springer Science+Business Media*, ISBN 978-0-387-24435-8 (Print) 978-0-387-25465-4 (Online)

Chuanng-Chien Ch.; Tong-Hong L. & Ben-Yi L. (2005). Using correlation coefficient in ECG waveform for arrhythmia detection, *Biomedical Engineering - Applications, Basis & Communications*, 2005(June), 17, 147-152

Ekbal A. (2006). Improvement of Prediction Accuracy Using Discretization and Voting Classifier, *18th International Conference on Pattern Recognition*, Vol. 2, pp: 695 – 698

Fan H.; Zaïane O.R.; Foss A. & Wu J. (2006). A Nonparametric Outlier Detection for Effectively Discovering Top-N Outliers from Engineering Data, *Advances in Knowledge Discovery and Data Mining*, *Lecture Notes in Artificial Intelligence*, 2006, Vol. 3918

Filzmoser P. (2005). Identification of multivariate outliers: A performance study, *Austrian Journal of Statistics*, Vol. 34, No. 2, pp. 127-138,

Fu X. & Wang L.(2003). Data Dimensionality Reduction With Application to Simplifying RBF Network Structure and Improving Classification Performance, *IEEE Transactions on systems, man and cybernetics – part B: Cybernetics*, Vol. 33, No. 3, June 2003, pp. 399 – 409

Fukunaga K. (1990). Introduction to Statistical Pattern Recognition, 2nd ed., *Academic Press*, San Diego, ISBN 0-12-269851-7

Galmacci G. (1996). Collinearity detection in linear regression models, *Computational Economics*, Vol. 9, pp. 215-227

Ginoris Y.P.; Amaral A.L.; Nicolau A.; Coelho M.A.Z. & Ferreira E.C. (2007). Raw data pre-processing in the protozoa and metazoa identification by image analysis and multivariate statistical techniques, *Journal of chemometrics*, Vol. 21, pp. 156–164

Han J. & Kamber M. (2001). Data Mining. Concepts and Techniques, Morgan Kaufmann Publisher, ISBN 1-55860-489-8

Hasimah Hj M.; Abdul Razak H. & Azuraliza Abu B.(2007). Pixel-based Parallel-Coordinates technique for outlier detection in Cardiac Patient Dataset, *Proceedings of the International Conference on Electrical Engineering and Informatics,* Institut Teknologi Bandung, Indonesia June 17-19, 2007

Ismail M.K. & Ciesielski V. (2003). An empirical investigation of the impact of discretization on common data distributions, *Proc. of the Third International Conference on Hybryd Intelligent Systems (HIS'03)*, edited by Abraham A., Koppen M., Franke K., pp.692-701, IOS Press

Jimenez-Marquez S.A.; Lacroix C. & Thibault J.(2002). Statistical Data Validation Methods for Large Cheese Plant Database, *Journal of Dairy Science*, Vol. 85, No. 9, 2081–2097, American Dairy Science Association, 2002

Jin R.; Breitbart Y. & Muoh Ch.(2009).Data discretization unification, *Journal Knowledge and Information Systems*, Vol. 19, No.1, (April, 2009), pp. 1 – 29, ISSN 0219-1377 (Print) 0219-3116 (Online)

Kochanski A. (2000)., Combined methods of ductile cast iron modeling, *III Polski Kongres Odlewnictwa, Zbior Materialow*, str. 160-165, Warszawa (in Polish)

Kochanski A. (2006). Aiding the detection of cast defect causes. Polish Metallurgy 2002 – 2006, *Komitet Metalurgii Polskiej Akademii Nauk*, red. K. Swiatkowski, ISBN 83-910-159-4-4, Krakow (in Polish)

Kochanski A. (2010). Data preparation, *Computer Methods in Materials Science*, Vol. 10, No. 1, pp. 25-29

Kozlowski J. (2009). Aiding the foundry process control with the use of advanced artificial neural network analysis methods, *PhD thesis, (in Polish)*, Warsaw University of Technology, Faculty of Production Engineering

Kusiak A. (2001). Feature Transformation Methods in Data Mining, *IEEE Transactions on Electronics Packaging Manufacturing*, Vol. 24, No. 3, July 2001

Larose D. T. (2005). Discovering Knowledge in Data. An Introduction to DATA MINING, John Wiley & Sons, Inc.

Laurikkala, J.; Juhola M.& Kentala E. (2000). Informal Identification of Outliers in Medical Data, *Proceeding in the Fifth International Workshop on Intelligent Data Analysis on Medicine and Pharmacology IDAMAP-2000,* Berlin

Liang Goh & Kasabov, N.(2003). Integrated gene expression analysis of multiple microarray data sets based on a normalization technique and on adaptive connectionist model, *Proceedings of the International Joint Conference on Neural Networks*, Vol. 3, pp. 1724 – 1728, 20-24 July, 2003

Liu H., Hussain F.; Tan C.L. & Dash M. (2002). Discretization: An Enabling Technique, Data Mining and Knowledge Discovery, Vol. 6, No. 4, pp. 393–423, Kluwer Academic Publishers, ISSN: 1384-5810 (Print) 1573-756X (Online)

Liu H. & Motoda H.(1998). Feature extraction, construction and selection: a data mining perspective, Kluwer Academic Publishers, ISBN 0-7923-8196-3,print. 2001

Liu H.; Motoda H. & Yu L. (2003). Feature Extraction, Selection, and Construction, In: *The Handbook of Data Mining* ed. Nong Ye, Lawrence Erlbaum Associates, ISBN 0-8058—4081-8

Loy Ch. Ch.; Lai MW. K. & Lim Ch. P. (2006). Dimensionality reduction of protein mass spectrometry data using Random Projection, *Lecture Notes in Computer Science*, Vol. 4233, ISBN 978-3-540-46481-5

Van der Maaten L.; Postma E. & Van den Herik J. (2009) *Dimensionality Reduction: A Comparative Review*, preprint

Masters T. (1993). Practical Neural Network Recipes in C++, Academic Press Inc.

McCue C.(2007). Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis, Butterworth-Heinemann, ISBN 0750677961, 9780750677967

Mitov I.; Ivanova K.; Markov K.; Velychko V.; Stanchev P. & Vanhoof K. (2009). Comparison of discretization methods for preprocessing data for Pyramidal Growing Network classification method, *New Trends in Intelligent Technologies Supplement to International Journal - Information Technologies and Knowledge* Volume 3

Moh'd Belal Al- Zgubi (2009). An Effective Clustering-Based Approach for Outlier Detection, *European Journal of Scientific Research*, ISSN 1450-216X Vol.28 No.2, pp.310-316

Mohamed H. Hj. ; Hamdan A.R. & Bakar A.A. (2007). Pixel-based Parallel-Coordinates technique for outlier detection in Cardiac Patient Dataset, *Proceedings of the*

International Conference on Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia June 17-19, 2007

Murdoch, D.J. & Chow, E.D. (1996). A graphical display of large correlation matrices. The American Statistician 50, 178-180

Olichwier J. (2011). The influence of data preparation on the accuracy of the modeling with the use of the laboratory and literature ADI ductile cast iron data, *MSc thesis, (in Polish)*, supervised by A. Kochanski, Warsaw University of Technology, Faculty of Production Engineering

Perzyk M. ; Biernacki R.& Kochanski A.(2005). Modelling of manufacturing processes by learning systems: the naive Bayesian classifier versus artificial neural networks, *Journal of Materials Processing Technology*, Elsevier, Vol. 164-165, pp. 1430-1435

Pyle D. (1999). Data Preparation for Data Mining, Morgan Kaufmann Publisher, ISBN 1-55860-529-0

Pyle D. (2003). Data Collection, Preparation, Quality, and Visualization, In: *The Handbook of Data Mining* edited by Nong Ye, Lawrence Erlbaum Associates, ISBN 0-80584-081-8

Research raport KBN Nr 003353/C.T08-6/2003 (in Polish)

Refaat M. (2007). Data Preparation for Data Mining Using SAS, Morgan Kaufmann Publisher, ISBN 13-978-0-12-373577-5

Rousseeuw P.J. & Zomeren B.C. van (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, Vol. 85, No. 411, pp. 633-651, American Statistical Association

Saeys Y.; Inza I. & Larrañaga P. (2007). A review of feature selection techniques in bioinformatics, *Bioinformatics*, Vol. 23, No. 19, pp. 2507–2517

Shaari F.; Abu Bakar A. & Razak Hamdan A. (2007). On New Approach in Mining Outlier, *Proceedings of the International Conference on Electrical Engineering and Informatics Institut Teknologi Bandung*, Indonesia, June 17-19, 2007

Shi H.& Fu J-Z.(2005). A global discretization method based on rough sets, *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 August, 2005

StatSoft (2011) web page http://www.statsoft.pl/textbook/stathome.html

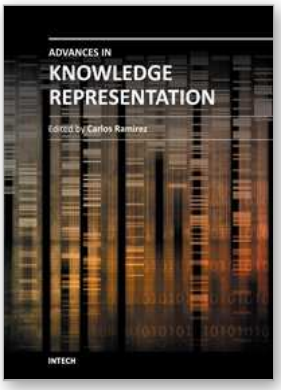
Villalba S. D. & Cunningham P. (2007). An Evaluation of Dimension Reduction Techniques for One-Class Classification, *Technical Report UCD-CSI-2007-9,* University College Dublin, , August 13th, 2007

Weiss S. M. & Indurkhya N. (1998). Predictive Data Mining: a practical guide, Morgan Kaufmann Publisher, ISBN 1-55860-403-0

Witten I.H. & Frank E.(2005). Data Mining. Practical Machine Learning Tools and Techniques, 2nd ed., Elsevier Inc., ISBN-13: 978-0-12-088407-0

Wu QX.; Bell D.; McGinnity M.; Prasad G.; Qi G. & Huang X. (2006). Improvement of Decision Accuracy Using Discretization of Continuous Attributes, *Lecture Notes in Computer Science*, Vol. 4223, Publisher Springer Berlin / Heidelberg, ISSN 0302-9743 (Print) 1611-3349 (Online)

Yang Y.; Webb G.I. & Wu X. (2005). Discretization Methods, In: *Data Mining and Knowledge Discovery Handbook*, Ch. 6, pp. 113–130, edited by: Maimon O., and Rokach L., Springer Science+Business Media, ISBN 978-0-387-24435-8 (Print)



Advances in Knowledge Representation

Edited by Dr. Carlos Ramirez

ISBN 978-953-51-0597-8
Hard cover, 272 pages
Publisher InTech
Published online 09, May, 2012
Published in print edition May, 2012

Advances in Knowledge Representation offers a compilation of state of the art research works on topics such as concept theory, positive relational algebra and k-relations, structured, visual and ontological models of knowledge representation, as well as detailed descriptions of applications to various domains, such as semantic representation and extraction, intelligent information retrieval, program proof checking, complex planning, and data preparation for knowledge modelling, and a extensive bibliography. It is a valuable contribution to the advancement of the field. The expected readers are advanced students and researchers on the knowledge representation field and related areas; it may also help to computer oriented practitioners of diverse fields looking for ideas on how to develop a knowledge-based application.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821