ACQUISITION OF DESIGN-RELEVANT KNOWLEDGE WITHIN THE DEVELOPMENT OF SHEET-BULK METAL FORMING

Sebastian Röhner[1], Thilo Breitsprecher[1] and Sandro Wartzack[1]
(1) University of Erlangen-Nuremberg

ABSTRACT
The increasing requirements on technical products represent a growing challenge for the manufacturing engineering. This challenge will be met by the development of a new manufacturing technology called sheet-bulk metal forming. For the early consideration of the full potential of sheet-bulk metal forming in a design process, a design engineer has to know the process limitations as soon as possible. Hence, the objective has to be to acquire design-relevant knowledge already in the early phases of process development and to maintain this knowledge simultaneously to the further development of the process. These are the declared aims of the self-learning engineering assistance system that will carry out the acquisition and maintenance of knowledge owing to its self-learning aspect. In this article, within an evaluation of knowledge acquisition methodologies, data mining was identified as a possibility for the realization of the self-learning aptitude. The potential of data mining was shown by its application on simulation data to acquire design-relevant knowledge.

Keywords: Knowledge Acquisition, Data Mining, Sheet-Bulk Metal Forming

1 INTRODUCTION
In recent years, the requirements on technical products in automobile sector increased regarding user-specific flexibility, functionality, space- or weight-savings and will still increase. This fact represents a growing challenge for the manufacturing engineering which will be intensified by simultaneous demand for cost-effective and time-efficient production as well as for economization of energy and resources. This challenge will be met by the development of a new manufacturing technology called sheet-bulk metal forming which will unite the advantages of sheet and bulk metal forming processes to go far beyond the limitations of each process [1].
Precondition for the fast realization of this new manufacturing technology in industrial practice is that design engineers know the process limitations of this technology early to make full use of its potential. Today´s state of the art is to acquire design relevant knowledge only after the completed development of manufacturing process or technology, respectively. But the objective has to be to realize acquisition and maintenance of design-relevant knowledge contemporaneous to the development of the manufacturing technology to enable design engineers to integrate new design possibilities resulted from the new technology. This objective will be pursued with the development of a self-learning engineering assistance system that will support design engineers during a design process regarding production-oriented design. For the analysis of a product regarding its manufacturability, corresponding knowledge has to be acquired and implemented in the assistance system. Furthermore, this knowledge must be maintained to avoid aging of the assistance system. The demand for the maintenance of knowledge implicates that knowledge acquisition has to be carried out at each stage of further development of sheet-bulk metal forming. In summary, the development of the self-learning engineering assistance systems addresses the well-known challenge of knowledge acquisition in the field of expert systems.
This paper reports about the acquisition of design-relevant knowledge within the development of the new manufacturing technology sheet-bulk metal forming. It starts with a description of sheet-bulk metal forming (chapter 2), and evaluates knowledge acquisition methodologies according to its deployment within the development of sheet-bulk metal forming (chapter 3). It continues with the application of data mining on simulation data for knowledge acquisition (chapter 4). In this context, data mining was carried out following the CRISP-DM process (CRoss Industry Standard Process for

Data Mining), i.e. from the definition of the data mining goal over the data preparation to the modeling and evaluation of data mining methods. Finally, a conclusion and outlook are presented.

2 SHEET-BULK METAL FORMING

The manufacturing technology “sheet-bulk metal forming” (SBMF) will be developed within the transregional collaborative research centre 73 (TCRC 73), in which three German universities are involved. This technology will unite the advantages of sheet and bulk metal forming processes to manufacture geometrically complex parts with variants and functional elements from thin sheet metal through forming. The objective is to manufacture these high-precision elements with close geometrical tolerances in which the geometrical details of the variants are in the range of the sheet thickness. The variants to manufacture are carriers and gearings derived from synchronizer rings and seat slide adjusters. The manufacturing of such variants out of sheet metals requires the overlapping or the sequence of 2- and 3-axis strain and stress states. To realize this, various sheet and bulk metal forming processes have to be combined [1]. For the development of SBMF processes, the process combinations “deep drawing – upsetting “, “deep drawing – extrusion“ and “cutting – deep drawing” will be investigated within TCRC 73.

In this paper, the process combination “deep drawing – extrusion” will be analyzed to acquire design-relevant knowledge. This process combination aims at the manufacturing of a part similar to synchronizer rings (Figure 1). In a first step of the process development, merely the extrusion of the variant “tooth” was investigated by simulations to identify influence factors of the extrusion operation regarding the manufacturability of the teeth. Therefore, a reference process was developed forming a ring of teeth beginning with a blank. For the investigation, a three-dimensional FE-model was built to perform several simulation studies varying parameters like blank thickness, friction factor or tool design. As a result of the ensuing sensitivity analysis, knowledge about the manufacturing of the teeth could be gained. For example, the mould filling depends on the availability of a flow-restriction as well as on the ratio of blank thickness to friction factor. The design of the punch, however, does not influence the mould filling [1]. This kind of knowledge can be assigned especially to process knowledge, which is very important for the design of an operating procedure.

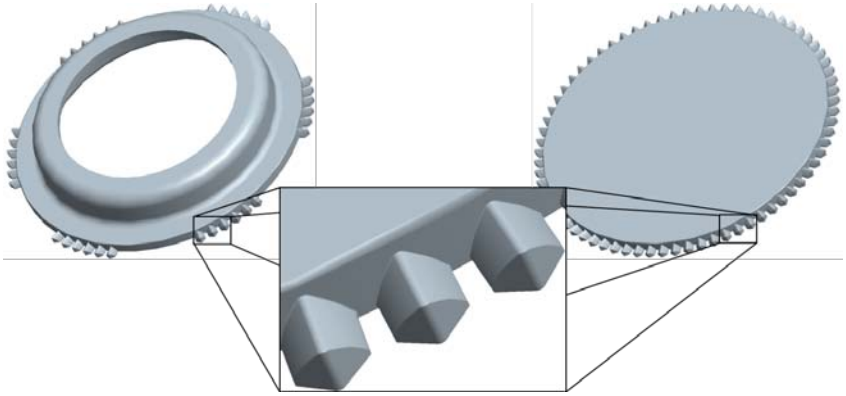

Figure 1. Demonstrator for the process combination “deep drawing – extrusion”

3 THE CHALLENGE WITHIN KNOWLEDGE ENGINEERING

The development and maintenance of knowledge based systems, in literature also called expert or assistance systems, is described as knowledge engineering. The bottleneck of this process represents the initial phase “knowledge acquisition” [2, 3, 4].

3.1 Methodologies of knowledge acquisition

In literature, the term “knowledge acquisition” is discussed controversially. Some authors equate knowledge acquisition merely with knowledge elicitation, whereas other authors regard knowledge acquisition as a process consisting of knowledge elicitation, knowledge interpretation or analysis and knowledge operationalization. However, all authors are in agreement with the fact that knowledge acquisition can be carried out in different ways: directly, indirectly or automatically. The direct knowledge acquisition is based on the dialogue between experts and intelligent knowledge acquisition tools, whereas the indirect knowledge acquisition is a knowledge engineer driven method and bases on

the dialogue between a knowledge engineer and an expert. Automatic knowledge acquisition methods extract knowledge from documents, instructions, diagrams, data bases and so forth without the intervention of experts or knowledge engineers. The methods usually originate from fields like artificial intelligence, statistics or machine learning [6, 7].

3.2 Challenge: knowledge acquisition

In the following, the previous introduced knowledge acquisition methodologies will be evaluated on the basis of five criterions:

Identification of knowledge sources

In general, one challenge represents the identification, i. e. the determination, examination and characterization, of knowledge sources. This applies especially for the acquisition of person-bound knowledge, because it is not obvious who has what kind of knowledge. Following questions arise: Who is consciously or unconsciously competent and who is consciously or unconsciously incompetent? The identification of data-based knowledge is comparatively straight forward, because there are several methods and techniques for the understanding, (see chapter 4.2), preparation (see chapter 4.3) and analyzing of data (see chapter 4.4) [5].

Availability of knowledge sources

A further general problem is the availability of knowledge sources. At first, mainly the availability of experts appears as such one because they first of all have to fulfill their daily business. In addition, the time window for knowledge acquisition can be reduced by organizational and geographical reasons. On the part of automatic knowledge acquisition, however, also the challenge regarding availability of knowledge sources can occur because data can be accessible for a limited time, the knowledge underlying the data is confidential or the data format is not available in a convenient way [4, 5].

Willingness and motivation of experts

Basic conditions for acquisition of person-bound knowledge represent the willingness and motivation of experts regarding transferring and sharing of their knowledge. These conditions can be a problem as a result of the missing understanding concerning the necessity of knowledge management, lack of assurance concerning their knowledge quality, fear of interchangeability and plagiarism [4, 8, 9].

Verbalization and formularization of knowledge

A further challenge concerning the knowledge acquisition from experts is the characteristic of knowledge being either tacit or explicit. Tacit knowledge is bounded by knowledge carriers, unconscious, context-specific, hard to formulate, to transfer and consequently to store. In comparison, explicit knowledge is easy to formulate, to transfer, to represent in analytical formulas or rules and to store in documents, tables or data bases. Owing to the general assumption that the majority of knowledge is tacit and merely the minority is explicit, the acquisition of expert knowledge represents a great challenge [4, 8, 10].

Role of knowledge engineer

The influence of the knowledge engineer on the acquisition process is revealed by the fact alone that he should be replaced by intelligent acquisition tools to solve the communication problem with the domain expert. This communication problem can result in interpretation errors and consequently in a building of an error-prone knowledge base. Reasons for the communication problem are the missing understanding of the problem or application domain, the missing knowledge about the terminology of the expert and also the difficulty of knowledge verbalization and formularization. Furthermore, the success of the knowledge acquisition depends on the experience and the social skills of the knowledge engineer, too. For instance, it is important to know different acquisition techniques to use them target-oriented. The variation of these techniques influences the result of knowledge acquisition likewise [4, 9].

3.3 Knowledge acquisition within TCRC 73

For the early integration or consideration of the full potential of SBMF in a design process, a design engineer has to know its process limitations as soon as possible. For this purpose, the objective has to be on the one hand to acquire design-relevant knowledge already in the early phases of process development and on the other hand to update or to maintain this knowledge simultaneously to the further development of the process. This intention intensifies the great challenge of knowledge acquisition owing to the resulting increase of time pressure. As a consequence of this increased time pressure the methods of the direct and indirect knowledge acquisition won’t be applied during the development of SBMF. The reason for this is that its methods compared to the methods of automatic knowledge acquisition can be considered as time-consuming as well as cost-intensive due to the

necessity of one or more knowledge carriers. Moreover, this time- and cost-factor increases during the process of indirect knowledge acquisition due to the need of one or more knowledge engineers. Furthermore, these methodologies will not be used within TCRC 73 because of the difficulty regarding the verbalization and formularization of knowledge as well as the influence of the knowledge engineer on the acquisition result.
The carrying out of an automatic knowledge acquisition is justified by the fact that the development of a new manufacturing technology requires the performing of numerical and experimental series of experiments. For this, simulation models and experimental setups will be generated by experts with the application of their knowledge, which is composed of theoretical and heuristic knowledge. This knowledge appears with regard to SBMF for example in the selection of the

- type of manufacturing process
- type of lubrication
- number of armament
- type of surface coating or type of surface structures
- types of semifinished parts

Therefore, the data emerged from parameter studies results from the knowledge and experience of the production engineer and represents his knowledge in an implicit way because the data of each individual simulation or experiment contain information like the setup of the simulation model, the target product geometry or the results of the performed forming process. The automatic extraction of this tacit knowledge from data and the transformation in explicit knowledge can be carried out by data mining. Apart from an automatic knowledge acquisition, data analysis via data mining enables the maintenance of knowledge. For this, at each further development of SBMF the resulting data has to be added to the consisting data stock and afterwards data mining has to be applied to extract new knowledge.

3.4 Knowledge acquisition via data mining

Introduction to data mining

Data mining was developed in the 1990s because of the desire to ensure efficient and accurate analyses despite rapidly increasing data volumes. Following [11] and [12] data mining corresponds with an iterative process for the automatic extraction of novel and valid information and knowledge from data stocks, which is potentially useful for decision making and problem solving. Within data mining, various methods from fields like artificial intelligence, statistics or machine learning are applied.
In the last two decades several data mining processes like KDD (Knowledge Discovery in Databases) [13] or CRISP-DM (CRoss Industry Standard Process for Data Mining) [14] were developed. In this article, a data mining process will be carried out in reference to the CRISP-DM process (see Figure 2).

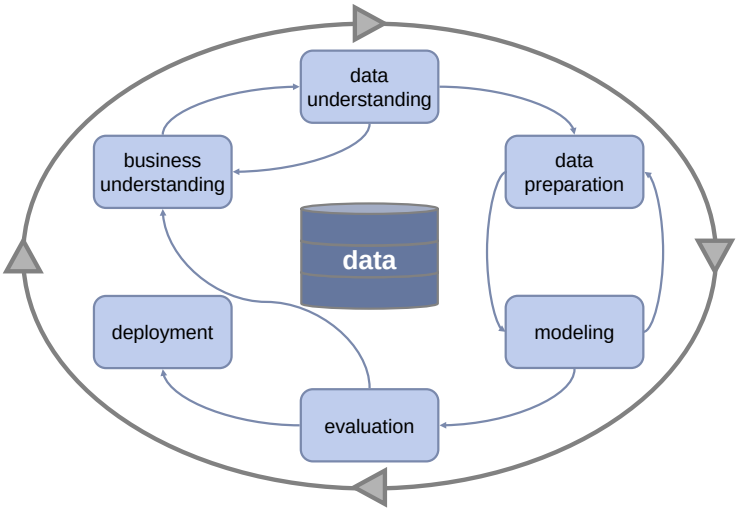

Figure 2. Phases of the CRISP-DM process according to [14]

This process model represents a general approach for carrying out data mining projects and includes all phases of a project, its tasks and its underlying relationships. The CRISP-DM process is subdivided into six phases, in which its sequence isn't linear and rigid. The iterative character of the CRISP-DM process is symbolized by the arrows and the outer ellipse. "Business understanding" represents the initial phase and focuses on determining the project objectives and data mining goals. The next two phases aim at getting a first insight into data and at the creation of the final dataset from the initial raw data, which builds the basis for the development of one or several data mining models during the "modeling"-phase. The quality of these models will be assessed in the "evaluation"-phase regarding the project objectives. If these objectives are satisfactory solved by one data mining model, then this model can be deployed. If none data mining model fulfill the requirement sufficiently, then the previous data mining phases have to be reviewed and in the worst case the whole data mining process has to be repeated again [14].

Data mining in engineering

Data mining and its underlying methods from artificial intelligence, statistics or machine learning can be used for different tasks in engineering. For example, these methods will be applied for fracture forecast in cold forming operations [15], for determination of optimal process parameters in cutting [16], for grouping of car models regarding its crash behavior [17] or for prediction of marine propeller behavior depending on design parameters [18]. A more comprehensive review of different data mining applications in manufacturing can be found in [19].

4 DATA MINING ON SIMULATION DATA

For the early consideration of SBMF in a design process, design-relevant knowledge has to be acquired simultaneously to the process development of SBMF. This intention can be carried out by data mining. Precondition for data mining is the availability of data. This demand will be fulfilled within TCRC 73 because the development of operation procedures requires performing of numerical and experimental parameter studies and therefore a huge amount of data emerges.

In this paper, the simulation data of the process combination "deep drawing – extrusion" will be analyzed for the acquisition of design-relevant knowledge using RapidMiner as data mining software. This data arose by a parameter study on the basis of a three-dimensional FE-model, varying merely the geometric parameters of the teeth. For reducing the computation time, just a 10° sector of the demonstrator was modeled justified by the rotational symmetry of the demonstrator (Figure 3). Theoretically regarding the symmetry condition, the modeling of a half tooth would be sufficient, but therefore the possibility concerning the verification of a faultless model would be dropped. A model can be recognized as faultless if both teeth are shaped nearly identically and the stress and strain values of both teeth are equal to each other. Furthermore, the modeling of the tool design as rigid bodies reduced the computation time. The friction conditions in the FE-model were defined differently. Between the contact bodies blank and die a friction factor of 0.3 was defined. The definition of the friction factor between the contact bodies blank and punch represents a special characteristic because the punch was partitioned into two sectors to influence the mould filling and the punch force. The sector "punch 1" was assigned with a friction factor of 0.05 for the increase of the mould filling, whereas the sector "punch 2" was assigned with a friction factor of 0.3 for the reduction of the punch force as well as for the increase of the mould filling. The semifinished part was a circular blank with a thickness of 2 mm and with the material characteristics of DC04.

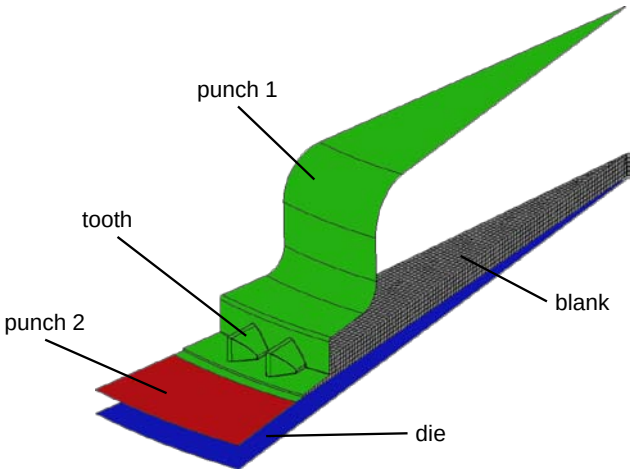

Figure 3. Setup of the FE-model

4.1 Business understanding

The initial phase “business understanding” (compare Figure 2) focuses on understanding the project objectives to convert these into an equivalent data mining problem. In this paper, the business objective is to acquire design-relevant knowledge from simulation data. For the definition of the data mining goal, it has to be clarified, what is meant by design-relevant knowledge. In the author’s opinion, this acquired knowledge should enable the designer to make statements about the manufacturability of a product. For instance, possible approaches for the evaluation of the manufacturability represent the prediction of the failure behavior of the blank or the prediction of the mould filling degree of the teeth. In this paper, the prediction of the total equivalent plastic strain (TEPS) will be used to make statements regarding the failure behavior of the workpiece. Because the blank was assigned with the material characteristics of DC04, the value of TEPS may not be larger than 2.5 to ensure the manufacturability of the teeth. The data mining goal, therefore, is to build data mining models to predict the target variable TEPS on the basis of input variables like geometry parameters or material characteristics.

4.2 Data understanding

The “data understanding”-phase comprises the collection of data and the familiarization with data. The data to analyze came from the FE-models and their underlying CAD-models. From CAD-models, the geometry data can be derived, and from FE-models, the process data like friction factors, the punch travel or the material characteristics can be obtained. The determination of the maximum TEPS was carried out by analyzing the simulation results at the punch travel of 1.4 mm. As in Figure 4 depicted, the maximum TEPS does not appear in the teeth region (ellipses marked with 1). The region with the highest strain, however, will not be considered in the following, as this region will be removed after the SBMF process, like in the conventional production of synchronizer rings. The region of interest represents the tooth root. As it is shown in Figure 4, both teeth roots possess almost similar gradients regarding the TEPS, except the nodes of the zoomed out region. These can be considered as outliers because these only occur at one tooth. Therefore, the TEPS´s of these nodes may not be used for the maximum TEPS. As a result of the outliers, the identification of the maximum TEPS is not straight forward. Thus, a higher number of nodes in the relevant region of the tooth root was always evaluated according to the region 2 in Figure 4. The individual TEPS´s of the nodes were summarized and afterwards averaged to calculate the maximum TEPS of the FE-model.

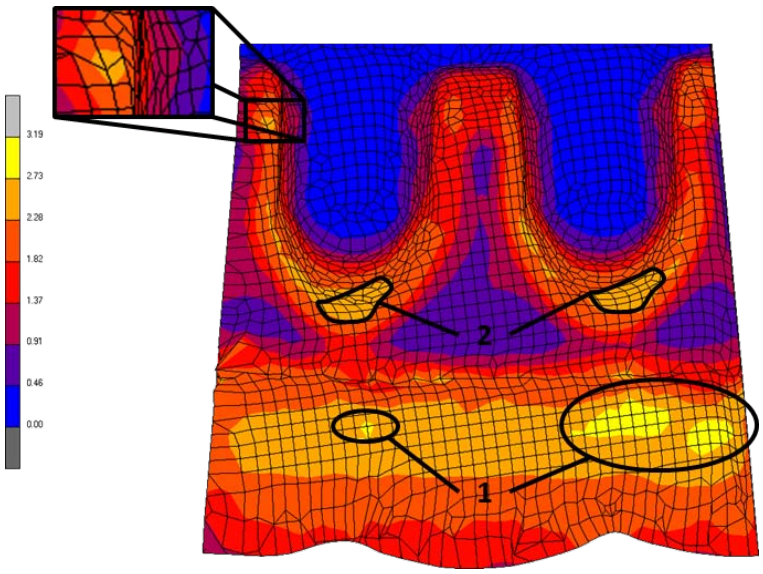

Figure 4. Determination of the maximum total equivalent plastic strain

The familiarization with data includes activities like the determination of the distribution of variables, the relation between variables or the completeness of data. This can be carried out by data visualization techniques or data reports, for example. To get a quick overview about the characteristics of all input variables and target variables, it is convenient to make a meta data view (Table 1). Such a table contains information like the data type, the mean value, the standard deviation, or the number of missing values of the variables.

Table 1. Meta data view of the initial raw simulation data

Name	***Unit***	***Type***	***Mean***	***Standard deviation***	***Range***
BETA_SZ	°	real	55.833	3.128	[52.500 ; 60.000]
B_SZ_1	mm	real	2.500	0.410	[2.000 ; 3.000]
D_BL	mm	real	94.000	0.000	[94.000 ; 94.000]
D_SZ_1	mm	integer	88.000	0.000	[88.000 ; 88.000]
D_SZ_2	mm	integer	96.000	0.000	[96.000 ; 96.000]
FRICT_1	-	real	0.300	0.000	[0.300 ; 0.300]
FRICT_2	-	real	0.100	0.000	[0.100 ; 0.100]
H_DD	mm	real	10.000	0.000	[10.000 ; 10.000]
H_SZ_1	mm	real	2.2000	0.000	[2.200 ; 2.200]
H_SZ_2	mm	real	3.300	0.000	[3.300 ; 3.300]
H_SZ_3	mm	real	0.300	0.000	[0.300 ; 0.300]
H_SZ_4	mm	real	0.200	0.000	[0.200 ; 0.200]
L_SZ_1	mm	real	2.750	0.251	[2.500 ; 3.000]
MAT	-	nominal	-	-	-
PUN_TR	mm	integer	1.400	0.000	[1.400 ; 1.400]
R_D_1	mm	real	0.300	0.000	[0.300 ; 0.300]
R_DD_1	mm	integer	2.000	0.000	[2.000 ; 2.000]
R_DD_2	mm	integer	4.000	0.000	[4.000 ; 4.000]
R_SZ_1	mm	integer	6.000	0.000	[6.000 ; 6.000]
R_SZ_2	mm	real	0.600	0.246	[0.300 ; 0.900]
R_SZ_3	mm	real	0.367	0.170	[0.200 ; 0.600]
R_SZ_4	mm	real	0.050	0.000	[0.050 ; 0.050]
TEMP	°C	integer	20.000	0.000	[20.000 ; 20.000]
TEPS	-	real	2.480	0.262	[1.903 ; 3.196]
THICK_BL	mm	integer	2.000	0.000	[2.000 ; 2.000]
VEL	mm/s	real	0.500	0.000	[0.500 ; 0.500]

4.3 Data preparation

The phase “data preparation” serves to generate a final data set from the initial raw data and consists of different strategies and techniques. For the decrease of computation time and for the increase of the predictive performance, it is appropriate to identify irrelevant, redundant and noisy variables. For this,

there are different approaches like wrapper and filter. Filter approaches use general characteristics of data like the standard deviation to evaluate and select variables. Filters operate independently of any modeling techniques. Contrary to this, wrappers evaluate variables by using accuracy or error estimates provided by the applied modeling technique. Within this paper, the reduction of the number of input variables was performed by a filter, which removes all input variables with a standard deviation equal to zero. Consequently, in the modeling phase just the input variables were considered which were varied within the simulation study. Owing to the application of this filter, the number of input variables decreased from 24 to 5 (see Table 2).

Table 2. meta data view of the final data set

Name	***Unit***	***Type***	***Mean***	***Standard deviation***	***Range***
BETA_SZ	°	real	55.833	3.128	[52.500 ; 60.000]
B_SZ_1	mm	real	2.500	0.410	[2.000 ; 3.000]
L_SZ_1	mm	real	2.750	0.251	[2.500 ; 3.000]
R_SZ_2	mm	real	0.600	0.246	[0.300 ; 0.900]
R_SZ_3	mm	real	0.367	0.170	[0.200 ; 0.600]
TEPS	-	real	2.480	0.262	[1.903 ; 3.196]

Further useful techniques for the preparation of data are the detection of anomalies, the cleaning of data from missing values and the transformation of data. Some modeling techniques cannot handle anomalies or missing values. Therefore, either another modeling technique has to be chosen or the data set has to be prepared. In the case of missing values there are two different approaches: Deletion of the corresponding data objects or prediction of the missing values via a predictive modeling technique. The transformation of data like discretization or binarization of data can be necessary, if data mining algorithm will be applied which can handle only particular types of data. For example, artificial neural network cannot handle discrete input variables. If the dimensionality reduction had not been performed within this data mining project, the value of the attribute “material” would have been transformed from nominal to integer.

4.4 Modeling

In the “modeling”-phase various modeling techniques are selected and applied. Normally, several techniques exist for the same data mining task. The given data mining task corresponds to a regression task because TEPS is a continuous target variable. For this, RapidMiner offers different modeling techniques like artificial neural networks, k-nearest neighbor, methods of regression analysis, support vector machines, rule learners or regression trees.

Modeling techniques for regression

Within this paper, following modeling techniques were applied:

Artificial neural networks: [20]

Artificial neural networks (ANNs) try to simulate biological neuronal systems. In this article, a multilayer ANN was built for the prediction of the maximum TEPS. Such a network comprises an input layer, one or more hidden layers and an output layer. Each layer is made up of neurons, in which the neurons of one layer are only interconnected via weighted links to the neurons in the next layers. The adaption of the ANN to a given data mining task can be carried out by the change of the network topology, i. e. by the change of the number of the hidden layers and the neurons as well as by the adjustment of the weights.

k-nearest neighbor: [20]

The *k-nearest neighbor* (kNN) algorithm is an instance-based learning technique, which represents each instance as a data point in a p-dimensional space, where p is the number of attributes. Via kNN the target attribute of a new data point will be calculated by the average of the target variable values of its k neighbors. In this context, the nearest neighbor of an instance is determined by a distance function like Euclidean distance, Manhattan or city-block metric.

Regression analysis: [12]

Regression analyses are methods used to model the relationship between a dependent target variable y and p independent input variables x_p. A well-known method is the *multiple linear regression* (LinReg), which represents the previously mentioned relationship as follows:

$$y = w_0 + w_1 x_1 + w_2 x_2 + ... + w_p x_p \quad (1)$$

where $w_0, w_1, ..., w_p$ are the regression coefficients. These coefficients are calculated during the training process by minimizing an error measure like the sum of squares error (SSE):

$$SSE = \sum_{i=1}^{n} \left(y_i - y_i^{'} \right)^2_1 \quad (2)$$

where n is the number of instances in the training set, y_i is the actual and $y_i^{'}$ the predicted value of the target variable. For fitting non-linear relationships, a *polynomial regression* (PolReg) can be used. This regression analysis method is a special case of LinReg because of m-order terms:

$$y = w_0 + w_1 x_1 + w_2 x_1^2 + ... + w_p x_p^m \quad (3)$$

Performance estimation of modeling techniques

Regression tasks are characterized by the partition of the given dataset into a training and test set. The training set is used for building predictive models, whereas the test set is used for its performance estimation. For splitting a dataset, different methods exist:

Holdout method:

In the *holdout method*, a given dataset is subdivided into two disjoint sets, the training and test set. The partition of the dataset will be carried out by a random sampling in which the usual split is 50-50 or two-thirds for training or one-third for testing. To improve the estimation of the model performance, the *holdout method* will be repeated k times, which is called *stratified holdout* [21] or *random subsampling* [11, 12, 22]. The resulting estimations will be summarized and averaged.

k-fold cross-validation:

The *k-fold cross-validation* splits the dataset randomly into k approximately equal disjoint subsets. The following learning procedure is executed r-times in which per procedure one subset is used for testing and the remainder is used for training. Finally, the resulting k error estimates are summarized and averaged to the total error estimate [11]. According to [12] and [21], the recommended value for k is 10. For a more reliable error estimate, a *stratified k-fold cross-validation* can be applied to reduce the effect of the random partition of the dataset [21, 22].

Bootstrap:

The *bootstrap* approach bases on random sampling with replacement, i.e. an instance, once chosen, can be selected again for the training set. A widely used approach of bootstrapping is *0.632 bootstrap*. Within this approach, the original dataset, consisting of n instances, is sampled n times to get the training set with n data objects. As a result of the replacement, the original dataset contains instances, which are not in the training set. These instances will be used for testing. The probability to select an instance each time is $1/n$, whereas the probability to not select this each time is $1-1/n$. For a sufficiently large data set, the test set will comprise $100 \cdot (1-1/n)^n$ = 36.8 % of instances and the training set will be 63.2 % of them [11, 12, 22]. The *0.632 bootstrap* estimate is according to [16] defined as

$$e = 0.632 \cdot e_{test} + 0.368 \cdot e_{train} \quad (4)$$

where e_{test} is the error estimate of the test set and e_{train} is the resubstitution error on the instances in the training set.

According to [21], for a small dataset the standard estimation technique is the *stratified k-fold cross-validation*. This was also confirmed from [22] by the study of *cross-validation* and *bootstrap* for accuracy estimation and model selection. Therefore, within this paper a *10 times 10-fold cross-validation* was used for the performance estimation of the various modeling techniques.

Performance measures for regression techniques

For the assessment of regression models, a couple of performance measures can be used. The measures applied in this paper and their underlying formulas are depicted in Table 3.

Table 3. Performance measures for regression techniques [21]

Performance measure	***Formula***
Mean-squared error	$\frac{\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2}{n}$
Mean absolute error	$\frac{\sum_{i=1}^{n}\left\|y_i - y_i^{'}\right\|}{n}$
Relative absolute error	$\frac{\sum_{i=1}^{n}\left\|y_i - y_i^{'}\right\|}{\sum_{i=1}^{n}\left\|y_i - \bar{y}\right\|}$, where $\bar{y} = \frac{1}{n}\sum_i y_i$
Correlation coefficient	$\frac{S_{yy'}}{\sqrt{S_y S_{y'}}}$, where $S_{yy'} = \frac{\sum_i \left(y_i - \bar{y}\right)\left(y_i^{'} - \bar{y}^{'}\right)}{n-1}$, $S_y = \frac{\sum_i \left(y_i - \bar{y}\right)^2}{n-1}$, $S_{y'} = \frac{\sum_i \left(y_i^{'} - \bar{y}^{'}\right)^2}{n-1}$

The characteristic of *mean-squared error* compared to the *mean absolute error* is that this tends to exaggerate the effect of outliers, whereas the *mean absolute error* does not do this. In some cases, the assessment via preceding measures is not strong enough, thus a relative measure like the *relative absolute error* has to be taken into account. For the estimate of the statistical correlation between the real target variable y_i and the predicted target variable $y_i^{'}$, the *correlation coefficient* can be applied. This coefficient ranges from "+1" through "0" to "-1". A coefficient value of "+1" means a perfect correlation, whereas a coefficient value of "0" means no correlation. A perfect negative correlation is characterized by a coefficient value of "-1".

Assessment of applied modeling techniques

During the modeling process, RapidMiner offers a high number of setting possibilities for each modeling technique. Within this paper, the default settings of RapidMiner were used for each technique. For instance, an ANN possesses one hidden layer and will be trained by the backpropagation algorithm. The number of neurons in the hidden layer depends on the number of input and output variables and will be calculated by following formula:

$$\text{number of hidden neurons} = (\text{number of input variables} + \text{number of output variables})/2 + 1 \quad (5)$$

Therefore, the number of hidden neurons is 4.

The result of the different regression techniques on simulation data is shown in Table 4. LinReg is the best according to all 4 metrics because it has the smallest value for each error measure and the largest *correlation coefficient*. The performance of 3NN and ANN is open to dispute because 3NN is better than ANN according to error measures, whereas 3NN is worse than ANN according to the *correlation coefficient*. PolReg represents the worst modeling technique because compared to the other modeling techniques shows larger error measures and a smaller *correlation coefficient*.

Table 4. Performance measures or applied modeling techniques

Performance measure	***3NN***	***ANN***	***LinReg***	***PolReg***
Mean-squared error	0.035	0.041	0.024	0.110
Mean absolute error	0.152	0.167	0.127	0.256
Relative absolute error	6.22 %	6.99 %	5.22 %	10.53 %
Correlation coefficient	0.686	0.725	0.792	0.229

4.5 Evaluation

The predictive models, created in the previous step, have to be evaluated according to the business objectives. This can be done, by applying the created models on new simulation data and by comparison of the real and predicted TEPS. Within this paper, the attributes of 5 new instances are analyzed. In Figure 5, the corresponding prediction results of each data mining model are compared

with the real TEPS. Eye-catching is that partially the predicted TEPSs vary strongly compared to the real TEPSs. This is particularly true for the predicted values of PolReg. Furthermore, the bar chart in Figure 5 indicates, that no applied model is capable to predict the TEPSs best and that the prediction quality from test data to test data varies differently strong. One reason for this is the determination of the maximum TEPS in the FE-models because always an averaged TEPS was taken as the maximum TEPS. Hereby, a slight error has crept in the value of maximum TEPS, which can differ between the individual simulation models.

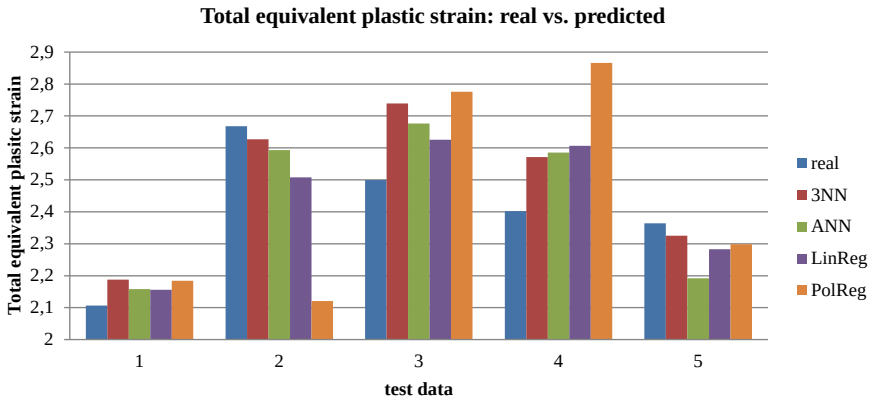

Figure 5. Comparison of the real and predicted TEPS

To increase the prediction quality of the individual models, the previous steps of the data mining process have to be reviewed and some optimization in the different phases have to be carried out. Especially, the “data preparation”- and the “modeling”-phase are predestined for this. In the “data preparation”-phase, for example, another filter or wrapper method for the selection of input variables can be more effective. Furthermore, the creation of a new set of input variables from the original raw data can improve the prediction quality because this new set can capture the important information much more effectively. Regarding the “modeling”-phase, other modeling techniques like regression trees or support vector machines can be applied, the setting possibilities of the modeling techniques can be optimized by evolutionary algorithm or the modeling techniques can be combined. Within this article, no optimization loops were carried out.

4.6 Deployment

In general, the creation of a data mining model is not the end of a data mining project. The created model has to be provided for an user in a convenient way, for example by a report or by the integration of the data mining model in a software program. Regarding the self-learning engineering assistance system, the best model has to be implemented in this system. For the provision of actual knowledge about the process limitations of SBMF, the data mining process has to be repeated at each further development of the manufacturing technology and the new data mining model has to be implemented.

5. CONCLUSSION AND OUTLOOK

This research was concerned with the knowledge acquisition during the development of a new manufacturing technology, called sheet-bulk metal forming. Outgoing from a description of the methodologies of knowledge acquisition, these methodologies were evaluated with regard to its deployment during the development of a manufacturing technology to acquire design-relevant knowledge. Result of this evaluation was that in the author´s opinion an automatic knowledge acquisition via data mining is most suitable for doing this. The potential of data mining was shown by its applying on simulation data to predict the total equivalent plastic strain on the basis of geometry parameters. Starting with the definition of the data mining goal over the data preparation to the modeling and evaluation of data mining models, the entire data mining process according to the CRISP-DM process was described and carried out. The data mining models, built during this process, based on *artificial neural network*, *k-nearest neighbor*, *multiple linear regression* and *polynomial regression*. During the “evaluation”-phase, these models were applied on new data to compare the real and the predicted total equivalent plastic strain. In this context, it was found out, that partially the prediction accuracy varied strongly and that no model was capable to predict the target variable best.

A reason for this was the error-prone determination of the maximum total equivalent plastic strain in the FE-models because averaged values were taken as maximum values to minimize the effect of outliers in the FE-model.
Within the performed data mining process, no optimization loops were carried out to improve the prediction quality of each data mining model. Future work has to be the development of a specific data mining process model for sheet-bulk metal forming. In this context, the both phases "data preparation" and "modeling" have to be investigated carefully to guarantee the modeling as well as the selection of the best prediction model. Furthermore, this process development has to deal with the question, in which way the acquired knowledge should be represented to implement this in the self-learning engineering assistance system.
In summary, data mining represents a promising approach to face the well-known bottleneck "knowledge acquisition" in the field of expert systems. On the part of engineering design, the employment of data mining in a design process enables the consideration of new manufacturing technologies already in early stages of its development. Precondition for this, however, is the availability of data which contain information like the setup of the simulation model, the target product geometry or the result of the performed forming process.

ACKNOWLEDGEMENT

This work was supported by the German Research Foundation (DFG) within the scope of the Transregional Collaborative Research Centre on sheet-bulk metal forming (SFB/TR 73).

REFERENCES

[1] Merklein, M., Koch, J., Schneider, T. and Oppelt, S., Manufacturing of Complex Functional Components with Variants by Using a new Metal Forming Process – Sheet-Bulk Metal Forming, in Proc. *Conference on Material Forming, ESAFORM*, Brescia, April 2010
[2] Feigenbaum, E. A., The *fifth generation – artificial intelligence and Japan´s computer challenge to the world*. 1984 (New American Library, New York)
[3] Hayes-Roth, F., *Building expert systems*. 1983 (Addison-Wesley, London)
[4] Welbank, M., *A review of knowledge acquisition techniques for expert systems*. 1983 (Martlesham Consultancy Services British Telecom Research Laboratories, Ipswich)
[5] Stokes, M.; *Managing engineering knowledge. MOKA: methodology for knowledge based engineering applications*. 2001 (Professional Engineering Publishing, London)
[6] Neale, I. N., First generation expert systems – a review of knowledge acquisition methodologies, in *The knowledge engineering review*, 1988, 3(2) pp.105-145
[7] Kim, J. and Courtney, J. F., A Survey of Knowledge acquisition Techniques and Their Relevance to Managerial Problem Domains, in *Decision Support Systems*, 1988, 4(3) pp. 269-284
[8] Hart, A., The role of induction in knowledge elicitation, in *Expert Systems*, 1984, 2(1) pp.24-28
[9] Fellers, J., Key factors in knowledge acquisition, in *ACM SIGCPR Computer Personnel*, 1987, 11(1) pp.10-24
[10] Gaines, B. R., An overview of knowledge-acquisition and transfer, in *International Journal of Man-Machine Studies*, 1987, 26(4) pp.453-472
[11] Tan, P., Steinbach, M. and Kumar, V., *Introduction to Data Mining*, 2006 (Pearson Education, Boston).
[12] Vercellis, C., *Business Intelligence: Data Mining and Optimization for Decision Making*, 2009 (John Wiley & Sons, Chichester).
[13] Fayyad, U., Piatetsky-Shapiro, and Smyth, P., From Data Mining to Knowledge Discovery in Databases, in *AI Magazine*, 1996, 17 (3) pp. 37-54
[14] Chapman, P. et al, *CRISP-DM 1.0 – Step by Step data mining guide*, http://www.crisp-dm.org/CRISPWP-0800.pdf
[15] Lorenzo, R. D., Ingarao, G., Micari, F., On the use of artificial intelligence tools for fracture forecast in cold forming operations, in *Journal of Materials Processing Technology*, 2006, 177 (1-3) pp. 315-318
[16] Tansel, I. N., Ozcelik, B., Bao, W. Y., Chen, P., Rincon, D., Yang, S. Y. and Yenilmez, A., Selection of optimal cutting conditions by using gonns, in *International Journal of Machine Tools and Manufacture*, 2006, 46 (1) pp. 26-35
[17] Kuhlmann, A., Vetter, R.-M, Lübbing, C. and Thole, C.-A., Data mining on crash simulation, in *Machine Learning and Data Mining in Pattern Recognition*, 2005, 3587 pp. 558-569
[18] Reich, Y. and Barai, S. V., A methodology for building neural networks models from empirical

engineering data, in *Engineering Applications of Artificial Intelligence*, 2000, 13 (6) pp. 685-694
[19] Choudhary, A. K., Harding, J. A. and Tiwari, M. K., Data mining in manufacturing: a review based on the kind of knowledge, in *Journal of Intelligent Manufacturing*, 2009, 20 (5) pp. 501-521
[20] Mitchell, T. M., *Machine Learning*, 1997 (McGraw-Hill, New York)
[21] Witten, I. H. and Eibe, F., *Data Mining – Practical Machine Learning Tools and Techniques*, 2005 (Morgan Kaufmann, San Francisco).
[22] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, in Proc. *Proceedings of the Fourteenth International on Artificial Intelligence, IJCAI*, 2010, pp. 1137-1143

Contact: Sebastian Röhner
University of Erlangen-Nuremberg
Chair of Engineering Design
Martensstraße 9
Erlangen, 91058
Germany
Tel: Int +49 9131 8527286
Fax: Int +49 09131 8527988
Email: roehner@mfk.uni-erlangen.de
URL: www.mfk.uni-erlangen.de

Dipl.-Ing. Sebastian Röhner is a research associate at the Chair of Engineering Design. He studied mechanical engineering at the University of Erlangen-Nuremberg; his main focus is knowledge-based engineering.
Dipl.-Ing. Thilo Breitsprecher is a research associate at the Chair of Engineering Design. He studied mechanical engineering at the University of Erlangen-Nuremberg; his main focus is knowledge-based engineering.
Prof. Dr.-Ing. Sandro Wartzack is head of the Chair of Engineering Design at the Friedrich-Alexander-University of Erlangen-Nuremberg, where he also received his PhD in 2000 and studied mechanical engineering some years before. After finalization of his PhD Study he was employed in an international automotive supplier company as director of virtual product development. Currently, he is member of the Design Society, honorary member of TechNet Alliance, member of Berliner Kreis, member of VDI, member of the scientific committee of ICED, DESIGN Conference and CIRP Conference on Computer Aided Tolerancing, expert advisory board member of the conference 'plastics+simulation', Editor of the Symposium DFX, as well as SIG Leader of the SIG 'Decision Making' in the Design Society.