ARCHIVES OF METALLURGY AND MATERIALS
Volume 52 2007 Issue 2

J. TALAR[*]

DATA MINING METHODS – APPLICATION IN METALLURGY

METODY EKSPLORACJI DANYCH – ZASTOSOWANIE W METALURGII

The objective of the paper is an evaluation of data mining techniques in application to both the analysis of large data sets and the modelling of complex manufacturing processes in the field of metallurgy. The paper presents an idea of the knowledge exploration process from large data sets and the major tasks of data mining. The basics of selected data mining methods are also presented: *k*- means clustering, decision trees, artificial neural networks and Bayesian networks. The second part of the paper presents some results of the application of selected data mining methods in metallurgy. The examples apply to the data analysis as well as modelling and control of metallurgical processes. The results have shown that data mining methods are very useful for both the analysis and the modelling of complex metallurgical processes.

Keywords: data mining, decision trees, artificial neural networks, Bayesian networks, clustering, data filtering, modelling and optimisation of metallurgical processes

Celem pracy jest ocena technik eksploracji danych w zastosowaniu do analizy dużych zbiorów danych oraz modelowania złożonych procesów wytwarzania w obszarze metalurgii. W pracy przedstawiono ideę procesu eksploracji wiedzy z dużych zbiorów danych oraz główne zadania eksploracji danych. Zaprezentowano również podstawy wybranych metod eksploracji danych: klasteryzacja *k*- średnich, drzewa decyzyjne, sieci neuronowe oraz sieci Bayesowskie. Druga część artykułu zawiera wyniki zastosowania wybranych metod eksploracji danych w metalurgii. Przykłady dotyczą analizy danych oraz modelowania i sterowania procesów metalurgicznych. Wyniki pokazały, że metody eksploracji danych są bardzo przydatne do analizy i modelowania złożonych procesów metalurgicznych.

1. Introduction

Data mining (DM) is an interdisciplinary field including various scientific disciplines such as: database systems, statistics, machine learning, artificial intelligence and the others [1]. The main goal of data mining is the knowledge exploration from large data sets. The methods of data mining give the possibility to find the unknown regularities in the accumulated data sets. The experimental data from laboratory tests and observations, as well as the data recorded in industrial conditions may possibly be the source of significant information for modelling, control and optimisation of the complex processes – also in metallurgy. There is the knowledge included in the data sets, and yet it is very often unused because of the difficulties in the analysis of the large data sets. It is very complicated to discover the knowledge from the huge data sets and it is impossible to do it manually. Therefore, there is necessity of the scientific research in the area of both the data analysis and the data mining techniques, which are able to support the analysis of the data in the field of metallurgy and materials science. Data mining is one of the stages of the process of knowledge discovery in databases (KDD). Knowledge discovery process consists of the following steps [1]: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation. Data mining is the essential stage of the knowledge discovery process, where the intelligent methods are applied in order to extract data patterns.

2. Data mining tasks

Data mining gives the possibility to solve many and various types of problems. The major tasks of data mining are: classification and prediction, clustering, association analysis, time series analysis, deviation analysis and the others [1, 3].

[*] FACULTY OF METALS ENGINEERING AND INDUSTRIAL COMPUTER SCIENCE, AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY, 30-059 KRAKÓW, AL. MICKIEWICZA 30, POLAND

2.1. Classification

Classification [1, 4] is the form of data analysis that can be used to extract the models describing important data classes. The classification model is constructed by analysing the objects described by attributes. Each object is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. There are many classification techniques [1], but the most popular methods are: decision trees, artificial neural networks and Bayesian networks. Classification model elaborated on the basis of the training data set, can be used for prediction of the objects classes of which the class label is unknown. Classification is used to predict the discrete and nominal values, while regression is used to predict continuous or ordered values.

2.2. Clustering

Clustering [5] can be defined as the process of partitioning the objects (characterised by attributes) into clusters (groups) such that the objects within the same cluster have a high degree of similarity, while the objects belonging to different clusters have a high degree of dissimilarity. Clustering is an example of unsupervised learning. Unlike classification, clustering does not rely on predefined classes and class-labelled training examples. Clustering algorithms group the objects into clusters on the basis of the criterion of maximisation of the objects similarity inside the cluster and minimisation of the similarity between the various clusters. Many clustering algorithms have been developed [6–8], which belong to two groups: hierarchical clustering and partitioning.

2.3. Association analysis

The goal of association analysis [9, 10] is to find and uncover the hidden relationships in large data set, which can be presented in the form of rules. An association rule is an expression $A \Rightarrow B$, where A and B represent two different sets of items. Let D be a database of transactions, where each transaction $T \in D$ is a set of items, $A \Rightarrow B$ expresses that whenever a transaction T contains A than T probably contains B too. The probability or rule confidence c is defined as the percentage of transactions containing B in addition to A with regard to the overall number of transactions containing A. The support of a rule is the number of transactions containing A (irrespective of presence of B) with respect to the total number of transactions in the database. The idea of association rules originated from the analysis of market-basket data, where rules like “A customer who buys the products x_1 and x_2 would also buy a product y with probability of $c\%$” are found. Another example of an association rule within the area of production technology could be as follows: “If the manufacturing defect A occurs, then the defect B would also occur with the probability of $c\%$”. The task of association analysis is to find the items that frequently appear together. More information about the association rule generating algorithms may be found in [1, 2].

2.4. Time series analysis

The analysis of time series [1, 2] is based on the assumption that successive values in the data file represent consecutive measurements taken at equally spaced time intervals. The time series analysis allows identifying of regularities between variable’s values from different time periods. The main goal of time series analysis is the prediction of future events (future variable’s values) based on known past events (past variable’s values). This type of analysis can be used for modelling of dynamic phenomena, where current value of analysed parameter depends on the previous values of this parameter and/or the previous values of the other process parameters. There are many methods used to model and forecast time series: autoregression methods, univariate ARIMA (autoregressive integrated moving average) modelling, artificial neural networks and the others [2, 11].

3. Review of data mining methods

In this section the selected data mining methods are described: k-means clustering, decision tree method (CART), artificial neural networks (multi-layer perceptron) and Bayesian networks.

3.1. K-means clustering

K-means clustering [4] is a popular technique for partitioning of large data sets. K-means clustering is an algorithm to group objects based on attributes into k number of groups (clusters). The grouping is done by minimising the sum of squares of distances between data and the corresponding cluster centre. The classical k-means algorithm is described as follows:

1. Choose k cluster centres (also called centroids) randomly generated in a domain containing all points,
2. Assign each point to the closest cluster centre,
3. Redefine the cluster centres using the current cluster memberships,
4. If a convergence criterion is not met, go to step 2.

Typical convergence criterion is minimal reassignment of points to new cluster centres. The Euclidean norm d is chosen as the distance measure (in step 2 of algorithm):

$$d(x,c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}, \tag{1}$$

where: n – the number of dimensions, x_i – the coordinate of a point, c_i – the coordinate of the centroid.

The centroid coordinates (in step 3 of algorithm) are calculated as:

$$c(x_i) = \frac{p_1(x_i) + ... + p_m(x_i)}{m}, \tag{2}$$

where: m – the number of the points, which are assigned to the cluster, p_1, p_2, ..., p_m – the points belonging to the cluster, $c(x_i)$, $p_1(x_i)$, ..., $p_m(x_i)$ – the coordinates.

The main advantages of k-means algorithm are its simplicity and speed which allows it to run on large data sets. The result of k-means clustering is that each point belongs to just one cluster. In fuzzy k-means clustering [12, 13] each point has a degree of belonging to clusters, as in fuzzy logic, so the point can be assigned to many clusters simultaneously.

3.2. Decision trees

Decision tree method is one of the classification techniques. This method allows representing the knowledge in a simple form, which is easy for interpretation and understanding for the user. The goal of decision tree induction is to find the set of logic rules "if *premise* then *conclusion*". Depending on the kind of classifying attribute (discrete or continuous), the classification or regression trees are built. In the structure of decision tree there are these attributes, which carry the most information on belonging of the objects to the classes. The criterion, which is used for developing the decision tree, is the maximisation of information gain (or minimisation of heterogeneity in the node). There are many different criteria used for evaluation of information gain for the various decision tree algorithms [14, 15] (e.g. in C4.5 algorithm [16] the measure of entropy is used for evaluation of information gain; in CART algorithm [14] the Gini index is used for evaluation of node heterogeneity). The typical structure of decision tree is shown in Fig. 1.

The CART algorithm [2, 14, 17] (Classification and Regression Trees) is very popular and useful method of knowledge exploration. CART is a method that generates a binary tree through binary recursive partitioning that splits a subset of the data set into two subsets according to the minimisation of heterogeneity criterion. Each split is based on a single variable. Some variables may be used several times while the others may not be used at all.

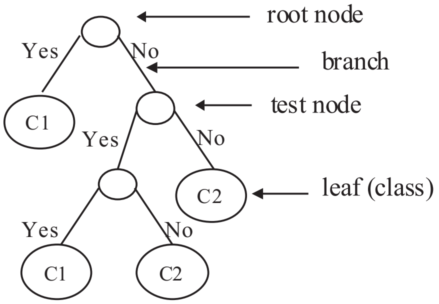

Fig. 1. Scheme of decision tree structure

Let $\Delta i\,(\psi\,(A)\,;S)$ denote the change of node heterogeneity as the result of partitioning of S set by ψ criterion on A attribute. The Gini index is defined as:

$$i(S) = 1 - \sum_{j} p^2\,(j\,|S), \tag{3}$$

where: $p\,(j\,|S)$ is defined as the probability of belonging the object to class j in S set.

The change of node heterogeneity for Gini index criterion can be defined as follows:

$$\Delta_G\,(\psi\,(A)\,;S) = i\,(S) - p_L * i\,(S_L) - p_P * i\,(S_P)\,, \tag{4}$$

where: S_L and S_P – sets determine the partition of initial S set by $\psi(A)$ criterion, $p_L = \frac{|S_L|}{S}$ and $p_P = \frac{|S_P|}{S}$ – probabilities that the object will be classified into one of these sets.

During the construction of decision tree one selects the attribute and such dealing criterion ψ, to maximise the expression Δ_G.

In the case of *regression trees*, the leaves are characterised by two parameters: $\overline{y}$ – value of classifying attribute (equation 5) and σ – standard deviation of cases in the node t (equation 6). The value of classifying attribute is the average value from N cases considered in the node t:

$$\overline{y}(t) = \frac{1}{N(t)} \sum_{n} y_n \tag{5}$$

$$\sigma = \sqrt{\frac{1}{N(t)} \sum_{x \in t} (y_n - \overline{y}(t))^2}. \tag{6}$$

For regression trees, the split rule relies on selection of such $\psi(A)$ criterion, which maximally reduces the classification error in node t:

$$\Delta R\,(\psi\,(A)\,, t) = R(t) - R(t_L) - R(t_P) \tag{7}$$

where:

$$R(t) = \frac{1}{N} \sum_{x \in t} (y_n - \overline{y}(t))^2. \tag{8}$$

The stop criterion for regression tree is reduced to the condition: $N(t) < N_{min}$, where: $N(t)$ is a number of cases considered in node t.

More details about CART algorithm, pruning method and the other decision tree algorithms may be found in [2, 14].

3.3. Artificial neural networks

Artificial neural network (ANN) [18–21] is the computer system, which simulates the functions of human brain. The basis of ANN is that it has many interconnected processing nodes, called neurons, which form the layered configurations. Each computing node in the network is based on the concept of an idealized neuron. An ideal neuron is assumed to respond optimally to the applied inputs. Neural network is a set of such neural nodes, in which the neurons are connected using complex synaptic connections characterized by weight coefficients and every single neuron makes its contribution to the computational properties of the whole system. There are many different types of ANNs: Multi-layer Perceptron, Radial Basis Function Network, Generalized Regression Neural Network.

The Multi-layer Perceptron [22] (MLP) is the most popular architecture of neural networks. The network consists of an input layer, hidden layers (typically one or two hidden layers are used) and an output layer. The architecture of multi-layer perceptron is presented in Fig. 2.

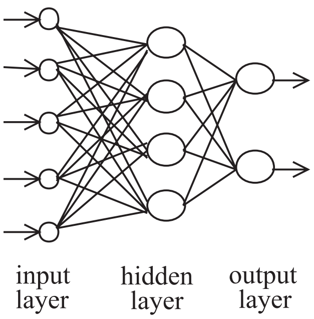

Fig. 2. Scheme of multi-layer perceptron with one hidden layer

The input layer is direct linked to the inputs of the first hidden layer. The output of each neuron in a layer is connected to the inputs of all neurons in the subsequent layer. Signals flow through the network in one direction (from input to output), and this type of network is called a feedforward network. Sigmoidal activation functions for the neurons in the hidden and output layers are the most used. Generally, the logistic function is utilized as follows:

$$f(x_j) = \frac{1}{1 + e^{-x_j}}, \tag{9}$$

where: $f(x_j)$ is the output of neuron j.

This type of network is trained using the supervised learning methods [18]. This means that during training of the network, both the input and corresponding output data are used. The most common is the backpropagation algorithm [23]. An input pattern is applied to the network and an output is generated. This output is compared to the corresponding target output and an error is produced. The error is then propagated back through the network, from output to input, and the network weights are adjusted in such a way as to minimize a cost function, typically the sum of the errors squared. The important feature of the MLP is that this network can accurately represent any continuous non-linear function relating the inputs and outputs. More information about other types of ANNs (both the static and dynamic neural networks) may be found in [23].

3.4. Bayesian networks

The Bayesian Network [18, 21, 24] is directed acyclic graph in which the nodes represent variables and the links represent causal influences among the variable. The strength of an influence is represented by conditional probabilities. Thus, two random variables, X and Y, are represented in a Bayesian network as two nodes in a directed graph. An edge directed from Y to X represents the influence of the node Y, the "parent" node, on the node X, the "child" node. The intensity of the influence of variable Y on variable X is quantified by the conditional probability $P(x \mid y)$, for every possible set of values (x, y).

Let $\mathbf{P}$ be the set of all parent nodes of a node X. Further, let $\mathbf{p}$ be a set of values for all the variables in $\mathbf{P}$ and let x be a value of the variable X. The influence of $\mathbf{P}$ on X can be modelled by any function $\mathbf{F}$ such that $\sum_x \mathbf{F}(x, \mathbf{p}) = 1$ and $0 \leqslant \mathbf{F}(x, \mathbf{p}) \leqslant 1$. The function $\mathbf{F}(x,\mathbf{p})$ provides a numerical quantification for $P(x \mid \mathbf{p})$.

A Bayesian network for a joint probability distribution $P(x_1, x_2, x_3, x_4, x_5)$ is shown in Fig. 3.

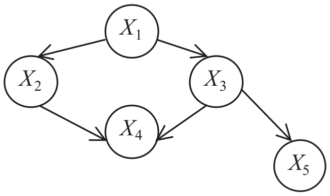

Fig. 3. Example of the Bayesian network

Node X_1, the "root node", is a node without parents whose probability distribution is $P(x_1)$, where the domain of x_1 is the set of values that X_1 takes on with non-zero probability, and is called a priori probability. This probability can be used to represent previous knowledge of the modelled domain. Due to the independencies declared in Fig. 3, the joint probability distribution can be computed as $P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_3)$. Each variable X_i is conditionally independent of all its other predecessors.

4. Examples of application of data mining techniques in metallurgy

In this part of the paper the selected examples of application of data mining methods in metallurgy are described. The problems considered apply to the analysis of measuring data, modelling and optimisation of the processes. The results presented here were obtained using the techniques described in the previous section.

4.1. Data analysis

4.1.1. Clustering of the multidimensional industrial data

Databases which describe the industrial processes can supply important information for modelling and optimisation of the processes. However, the analysis of the industrial data sets is very complicated because of the large number of both the data and the various process parameters, and also because of the different kinds of measurement noise.

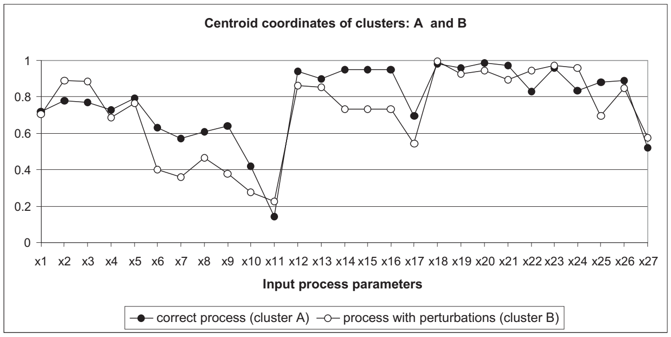

Fig. 4. Comparison of the centroid coordinates of the clusters, which describe the correct process and process with perturbations, for the input process parameters [26]

The subject of the analysis is the copper flash smelting process [21, 25], which is described by numerous parameters (27 input and 19 output parameters). Considered process is very complicated and the values of each parameter can change within wide range. It is possible to determine the ranges of variability of the process parameters using the statistic analysis. However, the permissible limits of all input and output parameters, for which the process will proceed correctly, are unknown. Wide ranges of variability of all input parameters result in many problems in modelling and control of the process.

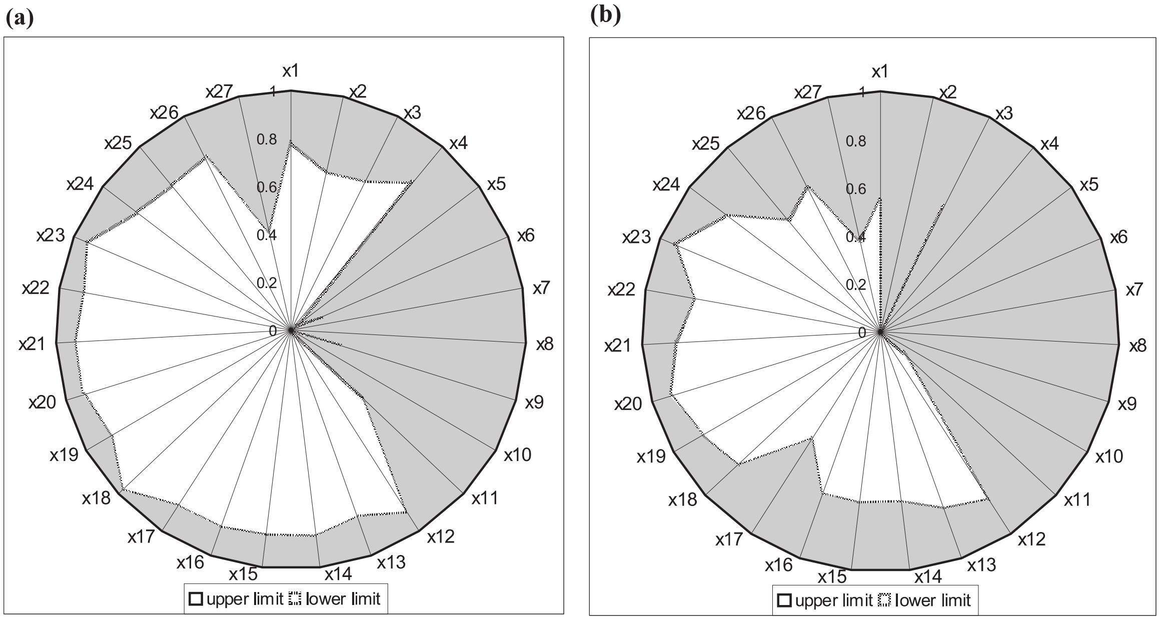

Fig. 5. Comparison of the searching areas (grey colour) for optimisation of the process: (a) – the searching space obtained on the basis of the cluster *A* (correct process); (b) - the searching space obtained on the basis of all data set (without clustering) [26]

The aim of this work was the data analysis of the real industrial process using clustering method [26]. The *k*-means algorithm was applied in order to group the industrial data into sets indicating the correct or incorrect process. The data from the monitoring system of production were used for the analysis of the process. The considered data set contains about 10000 records. Two groups (clusters) were obtained as the result of grouping of the multidimensional data. This clustering gave the possibility of evaluation of the group *A* which describes the correct process and the group *B* which characterizes the process with perturbations. The comparison of the centroid coordinates of the clusters is shown in Fig. 4 (for the normalised data). One may notice in Fig. 4 that the centroid coordinates of these clusters are distant from each other.

The centroid of the cluster, which describes the correct process, can be used as the starting point for the optimisation of the process. This cluster can also be used for the evaluation of the rages of variability of input parameters. The determination of the lower and the upper limits of the correct process can be helpful to constrain of the searching space for the optimisation methods. The comparison of the potential solution area obtained in the result of clustering process and the area of permissible solutions without clustering is shown in Fig. 5.

The searching area obtained in the result of clustering is narrower than the area of potential solutions without clustering, what can be very useful for the optimisation of the process (see section 4.3). The determined upper and lower limits (for the correct process) of the input parameters can be used in creation of instructions for the control of the process.

4.1.2. Filtering of the experimental data

The noisy experimental data are very often useless for the further analysis and modelling of the processes. Therefore, the modelling of the processes must be preceded with the data pre-processing. The data filtering is a very difficult task, because the smoothing procedures can eliminate important information or leave the unnecessary noise.

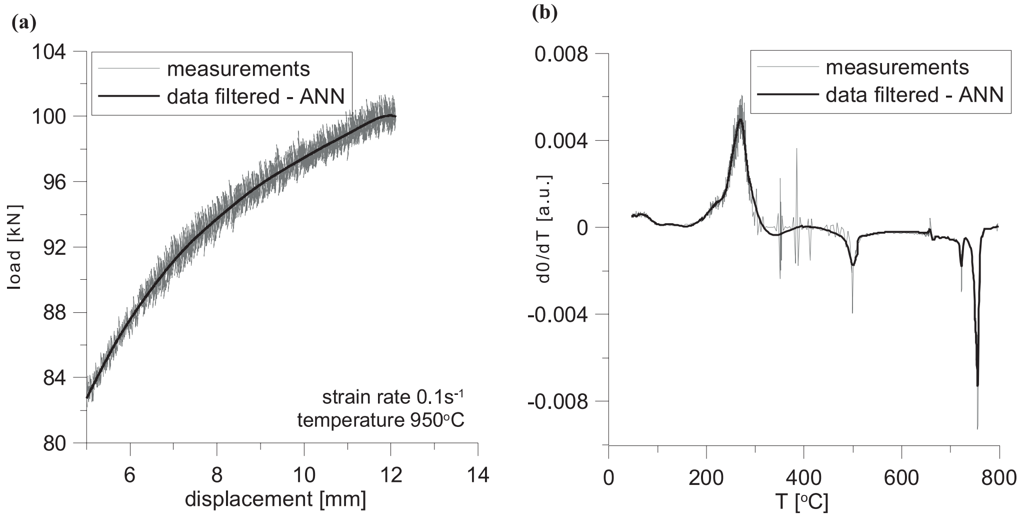

Fig. 6. Comparison of the data from measurements and from ANN filtering [27, 28]

Examples of data filtering using artificial neural networks (ANN) are presented in [27, 28]. The solutions described apply to filtering of the stress-strain data curves [27] and thermomagnetic data curves [28]. There are some results of data filtering using ANNs presented in the Fig. 6. The base of analysis [27] (Fig. 6a) were the data from the plane strain compression tests. On the input of the artificial neural network was the value of displacement, and on the output – the value of load. Next example of data de-noising is the filtering of thermomagnetic data curves [28] (Fig. 6b). The ANN had one input – the temperature of material, and one output – the magnetic moment.

The results obtained have shown that ANNs appear to be very efficient in filtering of the experimental data with the various measurement noise.

4.2. Modelling of the metallurgical processes using data mining methods

Application of data mining methods to modelling of the processes is justified in cases, for which the elaboration of the model is difficult, long time-consuming or even impossible. Data mining techniques based on artificial neural networks, Bayesian networks, decision trees and regression methods were applied to modelling of the complex metallurgical processes. In this section the selected results of modelling of the processes using various data mining techniques are presented.

4.2.1. Identification of the copper flash smelting process

The copper flash smelting process [21, 25] is continuous and very complex metallurgical process. Complicated structure of the process and many input and output parameters produce difficulties in modelling and control of that process. The existing models of the process are either very simplified or based on the FEM models [29–32]. The theoretical models which are based on the thermo-physics of the occurring physical and chemical reactions have many drawbacks because of many simplifications of the process description. They also require a lot of computation time, which makes them useless from the point of view of the control of the process in the real time.

The idea of application of artificial neural networks to modelling of the copper flash smelting process was proposed in [25]. Moreover, the attempts of application of the others data mining methods to modelling of that process are made. Now, the CART algorithm of regression trees was used to identification of the copper flash smelting process. The results of identification of the parameters of the copper flash smelting process using artificial neural networks and decision trees are presented below. The models were elaborated on the base of industrial data. The quality of the models was evaluated by a mean square error Φ, mean error E and a relative error τ defined as follows:

$$\Phi = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{P_c - P_m}{P_m}\right)^2} \qquad (10)$$

$$\mathrm{E} = \frac{\sum_{i=1}^{k}\left(\left|P_c - P_m\right|\right)}{k} \qquad (11)$$

$$\tau = \frac{\left|P_{pom} - P_{obl}\right|}{P_{pom}}, \qquad (12)$$

where: P_c – calculated value of a parameter, P_m – measured value of a parameter, k – a number of the values measured.

Results of prediction of NO_x concentration in gases

The results of prediction of NO_x concentration in gases are presented in Tab. 1. The artificial neural network (multi-layer perceptron) and the regression tree (CART algorithm) were compared with the results of regression analysis.

TABLE 1

Comparison of the results of prediction of the NO_x concentration in exhausts obtained using different methods [33]

Methods	Error Φ	Error E	Percentage of the results within the ranges of error τ					
			±τ	%	±τ	%	±τ	%
Artificial Neural Network	0.084	58.36	0.05	52%	0.10	82%	0.15	93%
Decision Tree (CART algorithm)	0.086	58.02	0.05	52%	0.10	83%	0.15	93%
Non-linear Regression	0.110	72.46	0.05	47%	0.10	74%	0.15	86%
Linear Regression	0.107	73.01	0.05	45%	0.10	73%	0.15	87%

On the base of errors of analysed methods (Tab. 1) it is seen, that artificial neural network and decision tree method have obtained the comparable results. The mean square error Φ for these methods equals about 0.08, what is the satisfactory result. Whereas, the results of linear and non-linear regression methods have had a worse convergence with the results of measurements. The results of prediction of the NO_x concentration in gases using artificial neural networks and CART algorithm are shown in Fig. 7.

In the result of CART algorithm application, the input parameters which are significant for prediction of NO_x concentration in gases were indicated. The input parameters, which appear in the structure of regression tree, are significant for prediction of analysed output parameter. The input parameters which have the greatest influence on the NO_x concentration in gases are: chemical composition of concentrate (contents of S, CaO, Cu, SiO_2, Pb), total concentrate stream, volume of oxygen per concentrate unit and oil consumption. From the assumed estimation criterion of parameters' significance point of view, the other input parameters are not such significant for prediction of NO_x concentration in gases.

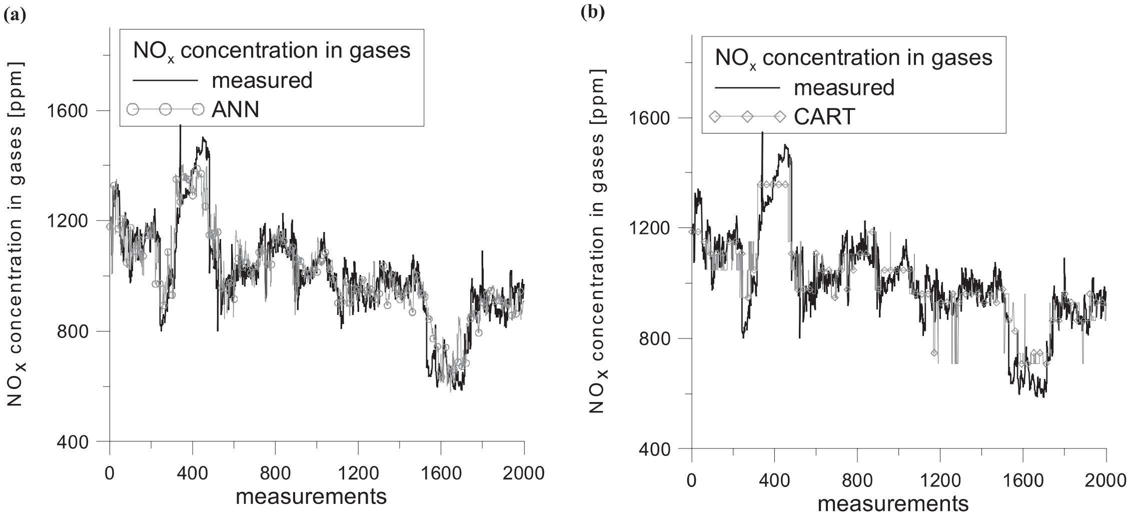

Fig. 7. The results of prediction of NO$_x$ concentration in exhausts using (a) – artificial neural network, (b) – regression tree method [24]

On the basis of elaborated regression tree, the decision rules were generated allowing to represent in overt form the knowledge about analysed problem. An example of decision rule for prediction of NO$_x$ concentration in gases is presented below: *"If* ((S content in concentrate $\leqslant$11.31%) *and* (Cu content in concentrate >26.58%) *and* (total concentrate stream > 97.12 10^3kg/h) *and* (oxygen content in the process blow $\leqslant$ 78.58%)) *then* NO$_x$ concentration in gases = 1186.65 ppm".

The knowledge representation in the form of both the decision trees and rules is very clearly and easy to interpret compared to the artificial neural networks. The idea of ANNs is based on the principle of "black box" action. On the basis of input signals an ANN generates the output signals. In artificial neural networks the knowledge about analysed problem is encoded in weight coefficients of network and this knowledge is not accessible directly in open form. Though, there are some algorithms of extracting rules from artificial neural networks [34]. The application of typical ANNs to modelling of the processes doesn't make it possible to gather the knowledge about analysed phenomenon in an easy and direct way (in comparison with the decision tree method).

Results of prediction of Cu concentration in slag

The next example of evaluation of artificial neural networks and decision tree method in application to modelling of the copper flash smelting process is presented below. The results of prediction of output parameter – Cu concentration in slag – are shown in Tab. 2 and in Fig. 8.

TABLE 2

The results of prediction of Cu concentration in slag

Methods	Error Φ	Error E	Percentage of the results within the ranges of error τ					
			$\pm\tau$	%	$\pm\tau$	%	$\pm\tau$	%
Artificial Neural Network	0.056	0.58	0.05	65%	0.10	92%	0.15	99%
Decision Tree (CART algorithm)	0.051	0.50	0.05	75%	0.10	94%	0.15	99%

The results obtained show that these methods can predict the values of Cu concentration in slag with high accuracy. It confirms that the artificial neural network approach and regression tree method are very useful tools in modelling of chosen output parameters of the copper flash smelting process.

Results of prediction of boiling level in the furnace

The artificial neural network, Bayesian network and rule-based expert system (for which the knowledge base was elaborated using C4.5 algorithm) were applied in order to predict the boiling level in the copper flash smelting furnace. The measurements of the boiling level were divided into two classes: "low" and "high" boiling level. The results of prediction of the boiling level obtained by various classification methods were compared in Tab. 3.

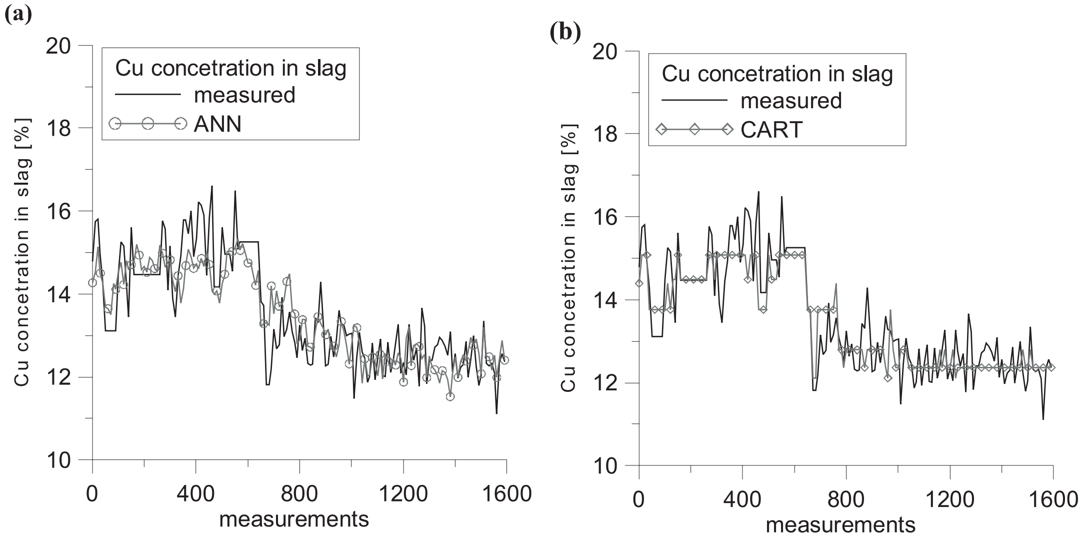

Fig. 8. Comparison of measured and calculated values of Cu concentration in slag: (a) – results of artificial neural network (ANN), (b) – results of regression tree method (CART)

TABLE 3
The results of prediction of the boiling level in the copper flash smelting furnance [21]

Methods	Percentage of correct answers	
	Boiling level	
	low	high
Artificial Neural Network	66%	72%
Bayesian Network	67%	56%
Expert System (C 4.5 algorithm)	87%	36%

The high level of boiling is undesirable. The possibility of prediction of high boiling level is very important from the security and proper work of aggregate point of view. However, there are some parameters, which are significant for prediction of the boiling level, but they are not registered in the databases, so the quality of prediction of this output parameter is not that well. Better results of prediction of the boiling level were obtained by the artificial neural network.

4.2.2. Prediction of cooling time of coils in bell-type annealing furnace

The elaboration of production schedules of cold rolled sheets is very difficult. The most important is the proper planning of bell-type annealing furnaces work, because this department is the bottleneck of production process in cold-rolling mill. The annealing cycles are very long and it is difficult to foresee the ending time of the cycle. This situation causes the disturbances of rhythmical work at furnace department, the lack of synchronization of following processes and the shortage of optimum usage of annealing line.

The models based on artificial neural networks and expert system were proposed to predict the cooling time of coils of cold rolled sheets [35]. These models were worked out on the basis of industrial data. These input parameters, which were considered in prediction of cooling time of coils (for short-term production planning) were, inter alia: the shielding gas temperature under the protective muffle, the surrounding temperature, the charge weight of stack, the average coils width of stack, the average thickness of the sheets of stack, the coils number of stack, the occurrence of the cooling bell, the furnace type. The knowledge base of the expert system was elaborated using the decision tree algorithm. The C4.5 algorithm (the following version of ID3 decision tree algorithm) was applied to create the decision rules, which were implemented in the knowledge base of the expert system. The C4.5 algorithm is the method, which allows to generate the classification trees. The problem of prediction of cooling time is the prediction of continuous variable. Hence, in this case the problem was transformed into the classification problem [35]. The results obtained by artificial neural networks and expert system were compared with the results of linear and non-linear regression analysis. Fig. 9 shows some results of prediction of cooling time of coils in bell-type annealing furnace.

The results obtained by both the artificial neural networks and the expert system (based on the C4.5 algorithm) give the best accuracy of prediction (mean error E – defined by equation (11) – under 2 hours), what is sufficient for scheduling and organizing of production of cold rolling mill. The results of regression models do not give the satisfactory effects.

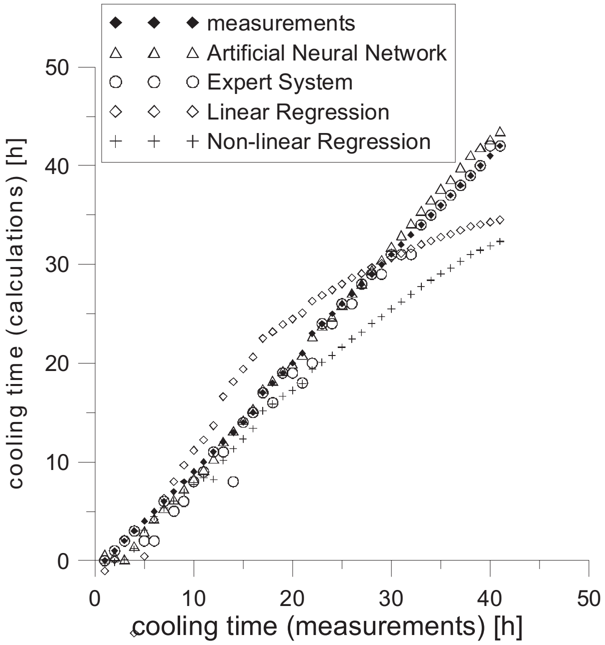

Fig. 9. Comparison of cooling times measured and predicted by various methods (for short-term production planning) [35]

4.3. Optimisation and control of the metallurgical processes based on data mining

The results of data mining can be useful in control of the metallurgical processes. The optimisation of the copper flash smelting process [36] can be an example of using the results of data clustering. In this work, the genetic algorithm (GA) and artificial neural network (ANN) model were used for the optimisation of that process. The optimisation with additional constraints of the input process parameters was applied. These constraints were obtained in the result of clustering of multidimensional industrial data (section 4.1.1).

The ANN model, which allows predicting of the Pb concentration in blister copper, was used in optimisation procedure. The goal of optimisation of the process was the calculation of the values of input process parameters, which minimise the assumed goal function. The error, between the value of output parameter (Pb concentration in blister copper) calculated by ANN and the expected value of this parameter in the real process, was assumed as the goal function. The results of optimisation of the copper flash smelting process may be found in [36]. The 10 process input parameters were assumed as the optimisation variables. The calculations were made for two types of copper concentrates, which significantly differ in Pb content. The low values of goal function obtained by the means of calculations confirm the high effectiveness of the method. The application of the upper and lower limits of input parameters, obtained as the result of data clustering, allows to restrict the searching space for the genetic algorithm. Moreover, the searching of solutions is performed only within the area of permissible limits, which characterise the correct run of the process. The fulfilment of these critical limits by the input parameters makes possible for the aggregate to work correctly and stable, and would also allow to receive the final products of expected properties.

5. Summary

The study presents the selected examples of application of data mining techniques to solving the various problems within the field of metallurgy. Described solutions demonstrate that artificial neural networks and decision tree methods give good results in modelling of complex metallurgical processes, for which there is lack of sufficient theoretical models. Particularly good results were obtained using the multi-layer perceptron and CART algorithm of regression tree generating. Moreover, there are indicated in the paper the possibilities of usage of data clustering method for the analysis of large multidimensional data sets and the possibilities of application of the results of data grouping to optimisation and control of the metallurgical processes. Further investigation will be applied to both the modelling of the dynamic phenomena in the area of metallurgy and materials science (which are dependent on the history of changes of input and output parameters of the process) and the analysis of possibility of using different methods of data mining in metallurgy and materials science.

Acknowledgements

The financial assistance of KBN, project no. 10.10.110.642. The author is very grateful to the board of directors of „Głogów" Copper Plant for making available the data used in this study.

REFERENCES

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers (2001).

[2] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, Press MIT (2001).

[3] D.T. Larose, Discovering Knowledge in Data. An Introduction to DATA MINING, J. Wiley & Sons 2005. Polskie tłumaczenie: Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa 2006).

[4] J.B. McQueen, Some Methods for Classification and Analysis of Multivariate Observations, Proc. of 5th Berkeley Symp. on Mathematical Statistics and Probability **1**, 281-297 (1967).

[5] M.N. Murty, A.K. Jain, P.J. Flynn, Data clustering: a review, ACM Computing Surveys **31**, **3**, 264-323 (1999).

[6] S. Guha, R. Rastogi, K. Shims, Cure: An Efficient Clustering Algorithm for Large Databases, Information Systems, Elsevier **26**, **1**, 35-58 (2001).

[7] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall (1988).

[8] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, New York (1990).

[9] S.K. Pal, P. Mitra, Pattern Recognition Algorithms for Data Mining, Chapman & Hall/CRC (2004).

[10] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, Fast Discovery of Association Rules, in: Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), The MIT Press, 307-328 (1996).

[11] D.J. Berndt, J. Clifford, Finding Patterns in Time Series: A Dynamic Programming Approach, in: Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), The MIT Press, 229-248 (1996).

[12] J.C. Dunn, A. Fuzzy, Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics **3**, 32-57 (1973).

[13] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York (1981).

[14] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall (Wadsworth, Inc.), New York (1984).

[15] J.R. Quinlan, Induction of decision trees, Machine Learning **1**, 81-106 (1986).

[16] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kauffman (1993).

[17] M.A. Kłopotek, Inteligentne wyszukiwarki internetowe, Akademicka Oficyna Wydawnicza Exit, Warszawa (2001).

[18] M.A. Arbib (ed.), The Handbook of Brain Theory and Neural Networks, The MIT Press, London (1995).

[19] S. Osowski, Sieci neuronowe, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa (1996).

[20] R. Tadeusiewicz, Sieci neuronowe, Akademicka Oficyna Wydawnicza RM, Warszawa (1993).

[21] J. Talar, P. Jarosz, J. Kusiak, J. Staszak, Artificial intelligence based modelling of the direct copper flash smelting process, in: Research in Polish metallurgy at the beginning of XXI century, M. Blicharski, K. Fitzner, W. Kapturkiewicz, M. Pietrzyk, J. Kazior (eds.), Committee of Metallurgy of the Polish Academy of Sciences, Akapit, 81-101, Kraków (2006).

[22] J.B. Gomm, G.F. Page, D. Williams, Introduction to neural networks, in: Application of Neural Networks to Modelling and Control, G. F. Page, J. B. Gomm, D. Williams (eds.), Chapmah & Hall, 1-8, London (1993).

[23] M.M. Gupta, L. Jin, N. Homma, Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory, Wiley-IEEE Press (2003).

[24] M.A.P. de Cristo, P.P. Calado, M.D. da Silveira, I. Silva, R. Muntz et al., Bayesian belief networks for IR, International Journal of Approximate Reasoning **34**, **2-3**, 163-179 (2003).

[25] J. Talar, T. Kondek, P. Jarosz, J. Kusiak, J. Dobrzański, J. Staszak, L. Byszyński, Artificial intelligence control system of the copper flash smelting process, Proc. 11th IFSC International Flash Smelting Congress, Bulgaria & Spain, 1-11 (2005).

[26] J. Talar, The possibilities of the clustering method application in analysis of multidimensional data obtained from real industrial process, Proc. CMS'05 Computer Methods and System, R. Tadeusiewicz, A. Ligęza, M. Szymkat (eds.), Oprogramowanie Naukowo-Techniczne, 43-48, Kraków (2005).

[27] J. Talar, Ł. Rauch, J. Kusiak, Filtering of the experimental data using the wavelet analysis and the artificial neural networks, Informatyka w technologii materiałów, Akapit **3**, **3-4**, 180-188 (2003), (in Polish).

[28] Ł. Rauch, J. Talar, T. Žák, J. Kusiak, Filtering of thermomagnetic data curve using artificial neural network and wavelet analysis, Proc. of the International Conference on Artificial Intelligence and Soft Computing 2004, Lecture Notes in Artificial Intelligence **3070**, 1093-1098, Springer-Verlag (2004).

[29] W.G. Davenport, E.H. Partelpoeg, Flash Smelting-Analysis, Control and Optimization, Pergamon Press (1987).

[30] J. Donizak, A. Hołda, Z. Kolenda, Modelowanie matematyczne procesów metalurgicznych-jednostadialny proces zawiesinowy produkcji miedzi hutniczej, Mat. Konf. „Teoria i Inżynieria Procesów Metalurgicznych”, 23-42, Kraków (2003).

[31] J. Donizak, A. Hołda, Z. Kolenda, Zastosowanie deterministyczno-stochastycznego modelu do symulacji zawiesinowego procesu produkcji miedzi hutniczej, Mat. XVIII Zjazdu Termodynamików, Wyd. Polit. Warszawskiej 22, 325-334, (2002).

[32] J. Donizak, A. Hołda, Z. Kolenda, M. Sukiennik, M. Warmuz, G. Szwancyber, J. Garbacki, Z. Gostyński, The Flash Smelting Process Digital Simulation, The IVth International Con-

ference on Non-Ferrous Metals and Alloys, Cracow, 29 (1999).

[33] J. Talar, P. Jarosz, A. Stanisławczyk, J. Kusiak, Modelling of NOx emission in the copper flash smelting process using regression trees method, Rudy i Metale Nieżelazne (2007), (in Polish), (in press).

[34] J.M. Ramirez, Extracting rules from artificial Neural Networks with kernel-based representations, Lecture notes in computer science, Springer **1607**, 68-77 (1999).

[35] J. Talar, Komputerowy model wspomagania operacyjnych decyzji technologicznych w walcowni zimnej blach, PhD thesis, AGH, Kraków (2003), (in Polish).

[36] J. Talar, A. Stanisławczyk, P. Jarosz, J. Kusiak, Z. Gostyński, D. Haze, P. Wroński, Application of the artificial intelligence techniques to optimisation of the copper flash smelting process, Rudy i Metale Nieżelazne **51**, **12**, 736-741 (2006), (in Polish).

Received: 10 March 2007.