ARCHIVES
Of
FOUNDRY ENGINEERING
Published quarterly as the organ of the Foundry Commission of the Polish Academy of Sciences

ISSN (1897-3310)
Volume 11
Issue 2/2011
113–122

23/2

Data acquisition in modeling using neural networks and decision trees

R. Sika [a], **Z. Ignaszak** [b,]*
[a] Division of Foundry, Poznań University of Technology, Piotrowo 3, 61-138 Poznań, Poland
[b] Division of Foundry, Poznań University of Technology, Piotrowo 3, 61-138 Poznań, Poland
*Corresponding author e-mail: zenon.ignaszak@put.poznan.pl

Otrzymano 11.04.2011; zaakceptowano do druku 26.04.2011

Abstract

The paper presents a comparison of selected models from area of artificial neural networks and decision trees in relation with actual conditions of foundry processes. The work contains short descriptions of used algorithms, their destination and method of data preparation, which is a domain of work of Data Mining systems. First part concerns data acquisition realized in selected iron foundry, indicating problems to solve in aspect of casting process modeling. Second part is a comparison of selected algorithms: a decision tree and artificial neural network, that is CART (Classification And Regression Trees) and BP (Backpropagation) in MLP (Multilayer Perceptron) networks algorithms. Aim of the paper is to show an aspect of selecting data for modeling, cleaning it and reducing, for example due to too strong correlation between some of recorded process parameters. Also, it has been shown what results can be obtained using two different approaches: first when modeling using available commercial software, for example Statistica, second when modeling step by step using Excel spreadsheet basing on the same algorithm, like BP-MLP. Discrepancy of results obtained from these two approaches originates from a priori made assumptions. Mentioned earlier Statistica universal software package, when used without awareness of relations of technological parameters, i.e. without user having experience in foundry and without scheduling ranks of particular parameters basing on acquisition, can not give credible basis to predict the quality of the castings. Also, a decisive influence of data acquisition method has been clearly indicated, the acquisition should be conducted according to repetitive measurement and control procedures. This paper is based on about 250 records of actual data, for one assortment for 6 month period, where only 12 data sets were complete (including two that were used for validation of neural network) and useful for creating a model. It is definitely too small portion in case of artificial neural networks, but it shows a scale of danger of unprofessional data acquisition.

Key words: Computer aided foundry processes, data acquisition, Data Mining, decision trees, neural networks

1. Introduction

One of the biggest challenges in foundry industry is production of castings with quality defined by client (designer), which is included in acceptance conditions. Possible casting defects, both fixable and unfixable, at different intensity levels must be referred to these conditions. Procedures of analysis of defects apperance causes, frequently informal and spontaneous, require using software tools. Methods from area of soft modeling [1] have been proven being more and more useful for solving these matters. Examples of these methods are decision trees (DT) and artificial neural networks (ANNs) [2], that can realize such analysis in a more effective way. This is an approach named Data Mining, well known in other industry branches.

[1] A type of mathematical modeling. Unlike hard modeling, physical nature effects is not taken into consideration, thanks to this it can be applied to any manufacturing process, where nature of occuring effects is not known. Model is considered as an approximate, empirical description of dependences between input and output quantities of the process [1].

[2] Opinions on the name differ. There is a significant and growing number of scientist specializing in neural networks (like prof. Tadeusiewicz) which thinks that it is more precise to use the term “neurolike networks”, which way of operation is based on real neural networks, emphasizing that artificial networks are much simpler.

In this work, basing on cooperation with significant, nationwide recognized iron foundry, first effects of global approach to data acquisition and applying algorithms from group of DT and ANNs have been summarized. Demanded goal function is the quality classification of final product, which qualifies the casting to one of two classes: conform final product and nonconform final product. Disposing a full range of available information (data) about technological parameters of the process (input variables) and information about the amount of faulty products regarding to selected assortment and date of production (output variables), it is possible to build a model basing on that collection of teaching data.

This collection should be numerous and diverse, assuring representative group of records needed to build a classifying model (classifier). Its effects are the classifying rules, named, due to graphical form, the decision tree. Depending on applied Data Mining method or type of algorithm, form of results for interpretation comes in high variety. Variability of type and amount of input data and structure of prepared model has to be taken into account too, considering also the problem of its undertraining or overtraining. [3].

In the paper, a comparison of effectiveness of selected DT and ANNs algorithms has been presented. These are the CART (Classification And Regression Trees) and BP (Backpropagation) in MLP (Multi Layer Perceptron) networks algorithms. Production process of selected ductile iron casting in automatic machines has been taken into consideration, with two methods applied: “side by side” approach, using Excel spreadsheet and using Statistica® by StatSoft© – leader in area of statistical analysis software and Data Mining systems.

2. Identification of manufacturing processes course and data selection for analysis

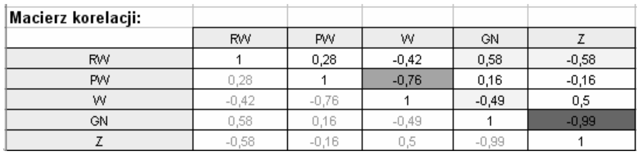
Casting processes, considering data acquisition methodology proposed by authors, can be divided chronologically to casting related subprocesses. The following groups of processes has been proposed:

- producing of molding material,
- pouring the metal into the mold,
- melting and metalurgical evaluation of the cast iron,
- final quality check of the casting, considering customer acceptance conditions.

The following parameters have been recorded:

- time of tapping from furnace to pouring ladle,
- initial and final temperature of pouring the metal into mold [°C],
- number of sample for chemical analysis,
- time of making the test using spectrometer,
- quantity of alloying elements: C, Si, Mn, P, S, Cr, Ni, Al, Cu,
- carbon equivalent value CE = %C+1/3*(%Si+%P),
- time of measurement of molding material parameters,
- molding material parameters: green sand strength RW, permeability PW, humidity W, bulk density GN, compactibility Z,
- defects units [%].

Basing on calculated correlation coefficients of molding material parameters a correlation matrix, showing their mutual interaction, has been built.



Macierz korelacji:					
	RW	PW	W	GN	Z
RW	1	0,28	-0,42	0,58	-0,58
PW	0,28	1	-0,76	0,16	-0,16
W	-0,42	-0,76	1	-0,49	0,5
GN	0,58	0,16	-0,49	1	-0,99
Z	-0,58	-0,16	0,5	-0,99	1

Fig. 1. Results of correlation between parameters of molding material

Basing on this stage of analysis, a conclusion can be formulated that correlations between molding material parameters allow to eliminate some of the parameters in studies. Thus, it has been proposed to reduce the number of tests of some parameters of the molding material, which would allow to spare time and focus on the more significant parameters – these, that mostly testify the stability of green sand quality.

Identification of relations between manufacturing processes and parameters measured in the foundry was a stage leading to discover expected dependancies (a graph of ranks has been made). Without this knowledge could lead to incoherent results of modeling. Identification of data collections, which consumed a significant portion of time, helped to estimate stability of particular processes (still without conclusions about end result of modeling).

An obvious fact is noteworthy – the manufacturing data collection, originating from casting process, significantly differs from non-industrial sets of data regarding homogeneity and type. These non-industrial data sets use mostly so-called weak scales, like binary input variables (“buy a house” – “do not buy a house”) [5].

Designing models of classification basing on actual manufacturing data (ductile iron casting) were dependant on: model of classifying decision tree CART and artificial neural network model MLP (with BP algorithms). Both approaches have been compared.

Build of CART and MLP models has been started with preparation full records containing parameters mentioned earlier from training data set.

Indication of set of training cases requires substantiation. In castings production, defects of different nature and with most probable causes of occurance appear, yet still there are such defects that result from superposition of accepted deviations of parameters of particular processes. Foundries have their own classifications of individual casting defects types. For example, in iron foundry where described study was realized, defects products are assigned to groups from D01 to D26 and they have quite conventional reference to polish standard describing casting defects.

From the days in six month production period, when analyzed assortment has been produced, it has been an attempt to select these cases, where prevailing number of casting defects were of similar nature (meaning that these defects may have been caused by deviation of parameters referring to the same technological processes): sand holes, drop in a mold, cold shut. Aim of authors is also to analyse the possibility of neural networks and decision trees application in reference to selected, predominant groups of casting defects. However, this subject exceeds the framework of this paper.

3. CART Algorithm – decision tree. Variants of the method.

Basing on the set of measured technological data from six month period on the basis of statistical averages (assuming that collection of particular parameters has the character of Gaussian distribution), expected values and value deviations of these parameters for selected assortment have been determined. Such a procedure is justified, because tolerances of technological parameters were not unequivocally assigned to selected assortment and all measured parameters taken into consideration. Arithmetic average and standard deviation were used to define acceptable values X_i, minimal and maximal in the set. It has been acknowledged that such approach represents the character of deviations of process parameters X_i in a sufficient way.

After estimation of above mentioned values, values of statistical estimators summarizing each considered day have been compared with recommended values. Following rule has been adopted – if estimators of statistical distribution: arithmetic average, standard deviation from the sample and range were contained in recommended intervals, such parameter was assigned with "Good" mark. In the opposite case, if statistical estimator value for given parameter deviated from recommended value[3], the assigned mark was "Wrong". Juxtaposition of marks of three estimators allowed to assign a cumulative mark to given, measured parameter X_i as input variable, according to rules presented in table 1.

Table 1. Adopted rule of assigning mark (discrete value) to a set of statistical estimators for given parameter

	X_i	X_i	X_i	X_i	X_i
X_{av}	**GOOD**	**GOOD**	**WRONG**	**WRONG**	**WRONG**
Б	**GOOD**	**WRONG**	**GOOD**	**WRONG**	**WRONG**
min-max	**GOOD**	**WRONG**	**GOOD**	**GOOD**	**WRONG**
MARK	**GOOD**	**AVERAGE**	**AVERAGE**	**WRONG**	**WRONG**

According to adopted assumptions, input variable X_i has total mark in a form of one of three decision classes:

- „Good" only when all three statistical estimators have „Good" mark,
- "Average" in two cases: when arithmetical average is marked "Good" and two other estimators "Wrong" or when any estimator is marked "Wrong",
- "Wrong" when two or more statistical estimators were marked with "Wrong".

Quality variable (parameter referring to quality) was classified as „Good" for amount of defects products lower than 9%, in other cases it was assigned with "Wrong" mark.

According to above mentioned rules, X_i parameters were assigned with marks, which were the input variables. Ten selected days have been considered, when selected assortment was casted and ten records of training set have been obtained. This limit was caused by the fact that from available data for six months, only ten records with complete X_i entries could have been chosen.

Only full set of variables (Wrong, Average, Average, Average, …) can be the base for further concluding and one of the rules of casting process classifying in foundry X is as following: *it can be assumed that* **IF** (W {*good, wrong*}) **AND** (Si {*good*}) **AND** (Z {*wrong, good*}) **THEN** (product {*faulty*}).

Used algorithm was CART, which is based on optimal division in node (or root) of tree. Selection of optimal division is carried out by finding the maximum $\Phi(s\,|\,t)$ value, by moving through all possible divisions in node t [4].

Model was decided to be tested using Statistica software and also "manually", using spreadsheet, performing subsequent stages of division with Microsoft® Excel™ software package. It turns out that depending on a priori assumptions made during modeling, analysis give divergent results and, from the technological point of view, can suggest absurd results. On purpose, tree model made with Statictica package is presented first.

Two different tree models made "manually" are also presented. The differences are caused by selecting a priori the greatest $\Phi(s\,|\,t)$ value. Maximum values after first division indicate Si and W, which means that intuitively one of parameters Si or W should be selected as the first to take part in the division ("goes into" the root). In first case, Si parameter was selected and it resulted in entirely different (and from technological point of view unreliable) tree structure than when selecting W parameter. Statistica software selected W automatically, which is the right choice in the opinion of authors, if it comes to relations of process parameters. It resulted in tree structure identical with "manual" solution.

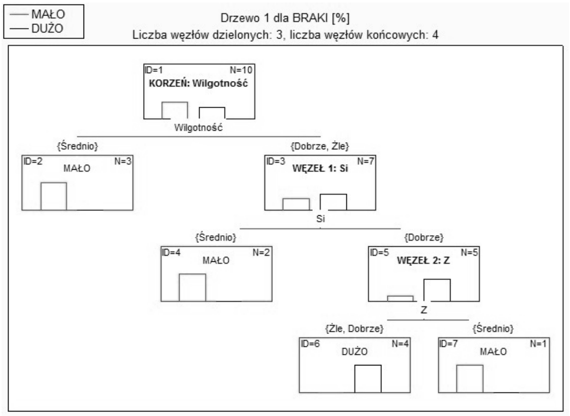


Fig. 2. Diagram of CART decision tree prepared using Statistica 9.1

[3] Following rule was adopted: value of average x_i diverges, if it is not contained within the interval $X_{av\ recommended}$ ± $б_{recommended}$; standard deviation diverges if it significantly exceeds the value of $б_{recommended}$, range min-max diverges if it goes beyond the limits of the interval T_{min} – T_{max}.

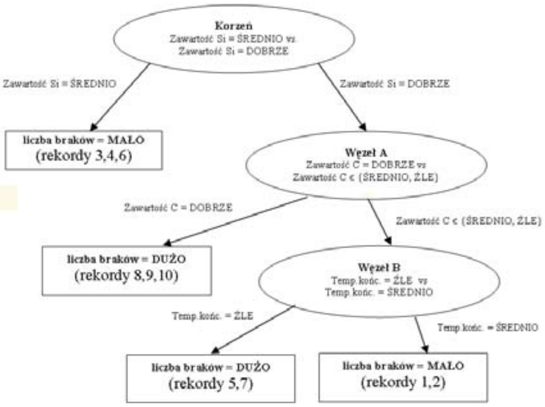


Fig. 3. Diagram of CART decision tree prepared „manually" (Microsoft™ Excel). Variant I – first division by „Si"

When observing many data records only for spectrometric study, a conclusion can be drawn about acceptable stability of these parameters. Situation is different if it comes to humidity. Too low humidity influences other technological parameters of mass, like permeability and compactibility, which as it turns out also takes part in creating tree structure.

From a technological point of view some divisions may seem slightly illogical. Both models shown that if values of variables W, Si and Z are "Good", then number of defects products is low. However, this study needs to be repeated for greater number of records, because the number of 10 is definitely too low. Tree suggests that if W and Si are marked "Average", then amount of faulty products is low, but on the other side, for W and Si marked respectively "Good" or "Wrong" and "Good" another divisions are generated. From the set of training records a conclusion can be drawn that such division (despite some records with "good" humidity and "good" silicon content) is influenced by other parameters: Z, initial and final temperature (they do not take part in tree structure). Chart of parameters weights in Statistica indicates, that still final temperature is more important than humidity (fig. 5).

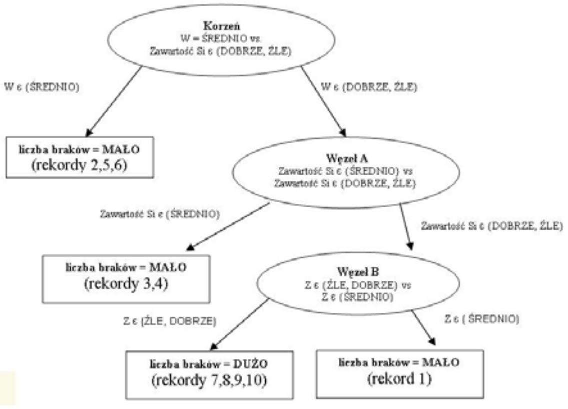


Fig. 4. Diagram of CART decision tree prepared „manually" (Microsoft™ Excel). Variant II – first division by „W"

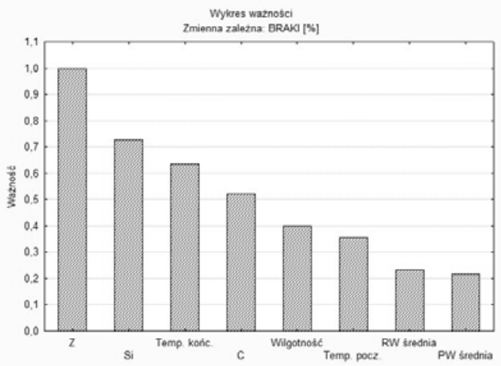


Fig. 5. Chart presenting significance of parameters taking part in tree generation – Statistica 9.1

As a summary, it can be stated that any seemingly small change of input parameter X_i has significant influence on the form of decision tree. In case of presented assortment, the most important division parameters are: humidity - W, silicon - Si and compactibility - Z. With different choice of first division, the parameters are: silicon - Si, carbon - C and final temperature.

Division obtained using Statistica software has to be treated as credible, although it has been acquired basing only on ten records. Following summarizing postulates can be formulated:

- number of training data records should be increased at least by three times (in presented division of tree, if looking at set of training records, it turns out that final temperature has bigger influence on the faulty parts amount, although CART algorithm suggested C and Si elements as more important in division)
- quality evaluation and classifying products as "Good" and "Wrong" should refer to one, specific defect (of course an assumption should be made that any person who classifies casting defect does it in a professional way, according to generally adopted rules and knowledge about the defect origin, which requires permanent training of quality inspectors)
- giving parameters marks using basic statistical estimators, additionally with taking into account acceptable tolerance limits for minimum/maximum values, for example according to ranges defined by process engineers – specialists (then a table of possible divisions may change significantly) for representative sample; however, such sample should contain at least over a dozen complete records giving answer "Small amount of faults" and the same number of records giving answer "Large amount of faults"[4].

4. BP algorithm in MLP artificial neural network

As some literary sources say, ANNs are very primitive attempt to simulate non-linear training which occurs in neural net-

[4] In other study there has been an assumption, that fault percentage is „large", when it exceeds 10%. It has been noted that amount of faulty products for assortment B in one of considered days was 10,6%. This is when the idea emerged to create two decision trees using the same algorithm (CART), but for the first tree treat 10,6% as „Small" amount of faulty products [%] and for the second as „Large". Analysis gave surprisingly divergent results.

works that can be found in nature. Imitation of human nerve cell, artificial neuron gathers input signals (x_i) from preceding neurons (or from data set) and creates a new value out of it (for example using function of sum). Obtained result is an input for activation function, which subsequently returns output signal (y) (basing on: [4, 15]).

Neural networks structure, unlike presented earlier decision tree models, is much harder to interpret. In previous chapter, to select variables for modeling, discrete (descriptive) parameters have been used, with names: *good, average, wrong*. According to this rule, variable might have been described only with one of above-mentioned values (classes). In case of ANNs, individual parameters are described mostly using variables of continuous character (in ANNs discrete values can be used, but they should be very expertly coded later, for example in form of flags). Moreover, presented CART decision trees were characterized with binary divisions (although they are not the only possible), which is not a case for neural networks. More precise comparison of both methods can be found in next chapter.

Disposing the same technological data, it has been decided to design a model of artificial neural network, with output neuron (y) to return a continuous value from interval 0 to 1, which after conversion could be classified as (used also in case of decision trees) quality parameter – amount of faulty products. Analyses shown underneath were conducted also for identical assortment of castings and parameters – input data – also originated from ten days with high and low defects numbers of products (data set identical as in decision tree model). Also in case of ANNs, build of the model was started with preparation of data set (detailed description of this stage was omitted).

Full set of normalized values accurately to 4 decimal places was placed in cumulative table. Ordinal number is an equivalent of number of training record from decision tree model. Each column contains values of interval [0,1], where 0 is the minimum and 1 is the maximum value of the parameter in the set originating from considered ten days.

In the study, one-way, multi-layer neural network (MLP – Multi Layer Perceptron) was used. Basing on above mentioned assumptions, it has been decided to build a model containing 8 neurons of input layer – equivalents of particular technological parameters, 3 neurons of hidden layer – according to rule, that greater number of neurons of this layer makes training process much more difficult (in turn too small amount generalizes the end result too much) and 1 neuron of output layer, which is the parameters of goal variable – classification of the casting. Authors realize the fact, that it is recommended to experiment with more/less hidden layers number. However, much more training records are needed for this and for the study, only small amount of records were available. It can be a premise for further, more detailed research in area of ANNs, for example in one year period. Testing of such models should be preceded by properly organized data acquisition procedures.

Input neurons, which are the process parameters: Tp, Tk, C, Si, etc. have been assigned with numbers (1,2, …, 8) to simplify the notation of performed calculations and for more comprehensible reading (fig. 6)

Having full set of input data with weights, according to neural network training algorithm described earlier, a *net* connection function was applied and then sigmoid activation function was used. In that way the input signals from hidden layer neurons A,B C and from neuron Z was obtained, which is the outcome of network operation after first run (fig. 7). Acquired value was compared with expected (appropriate) value for this record from training data set. On that basis an error (difference between expected value and output value) has been calculated and neural network manipulated weighs in a way to decrease the sum of these errors squares (so-called cumulative error) to the lowest possible value [4].

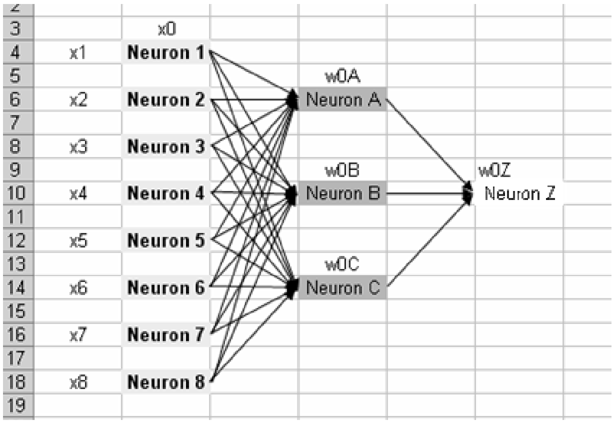
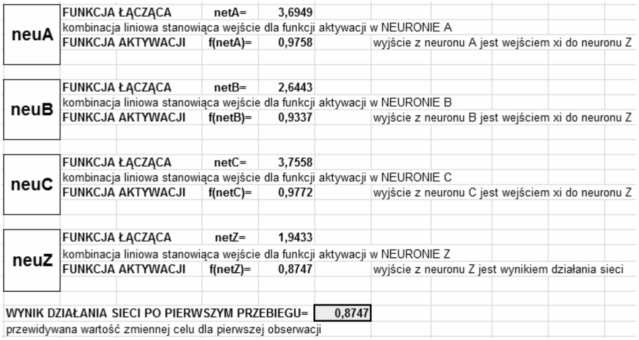


Fig. 6. Diagram of artificial neural network for selected assortment



neuA	FUNKCJA ŁĄCZĄCA	netA=	3,6949	
	kombinacja liniowa stanowiąca wejście dla funkcji aktywacji w NEURONIE A			
	FUNKCJA AKTYWACJI	f(netA)=	0,9758	wyjście z neuronu A jest wejściem xi do neuronu Z
neuB	FUNKCJA ŁĄCZĄCA	netB=	2,6443	
	kombinacja liniowa stanowiąca wejście dla funkcji aktywacji w NEURONIE B			
	FUNKCJA AKTYWACJI	f(netB)=	0,9337	wyjście z neuronu B jest wejściem xi do neuronu Z
neuC	FUNKCJA ŁĄCZĄCA	netC=	3,7558	
	kombinacja liniowa stanowiąca wejście dla funkcji aktywacji w NEURONIE C			
	FUNKCJA AKTYWACJI	f(netC)=	0,9772	wyjście z neuronu C jest wejściem xi do neuronu Z
neuZ	FUNKCJA ŁĄCZĄCA	netZ=	1,9433	
	kombinacja liniowa stanowiąca wejście dla funkcji aktywacji w NEURONIE Z			
	FUNKCJA AKTYWACJI	f(netZ)=	0,8747	wyjście z neuronu Z jest wynikiem działania sieci

WYNIK DZIAŁANIA SIECI PO PIERWSZYM PRZEBIEGU= 0,8747
przewidywana wartość zmiennej celu dla pierwszej obserwacji

Fig. 7. Outcome of network operation after first run

At second run, value of network „output" was 0,8529 (at first run f(netz) = 0,8747) and prognosis error was decreased by 0,075 (0,7651215 - 0,689956).

Thereafter, the activities were repeated eight times, each time using updated weights obtained from previous training and from input signals representing subsequent records of training data set. All weights (after each run) were compiled in one table and range of changes was presented in graphical form (fig. 8-10)

It turns out that weights of connections between hidden layer neurons revealed significant drop of value, which reached its minimum at 9th run. Next, 10th run caused dynamical increase in weights values, so it can be supposed that in that place a local minimum of the error occured (again with the restriction about small number of training records).

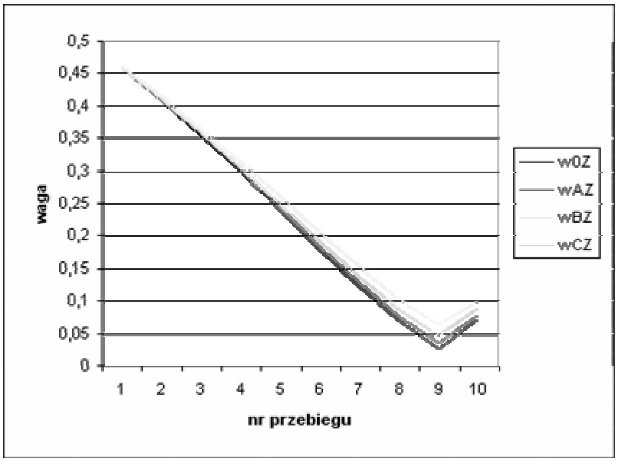


Fig. 8. Diagram presenting changes of weights of connections for output layer neuron

Simultaneously it has been observed that weights of connections between input and hidden layer neurons have undergone slight decreases (fig. 9), only after enlarging the scale of the diagram change of weights could be seen (fig. 10). Therefore, a conclusion can be drawn that for these connections process of training should be carried out again multiple times, to set the weights on stable level or another neural network model should be analysed (for example with more/less hidden layers). Because of limited number of training records it was impossible to perform another runs. Authors are working towards more disciplined data acquisition in this foundry and introducing inspection of record completeness, what in period preceding described study using CART and BP-MLP algorithms was not of great importance.

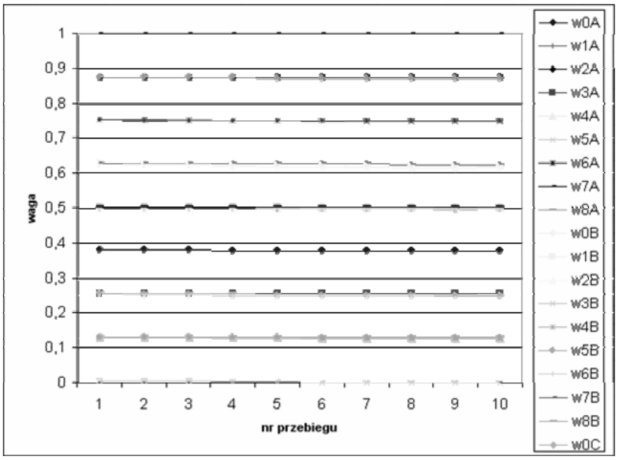


Fig. 9. Diagram presenting changes of weights of connections for hidden and input layer neurons

Further inspection included checking what was weight correction influence on prognosis error. On fig. 11 prognosis errors for each run were put together, also the cumulative prognosis error was calculated (its value is 4,3996). Can it be stated that after 9th run process of neural network training reached the "stop" condition? Regardless of reaching the minimum value of error for training data set, neural networks do not ensure finding optimal solution. Frequently the algorithm can stop training at local minimum, which represents good but not always optimal solution. It can not be unequivocally determined how would prognosis error change with another trials of training without having greater number of training records.

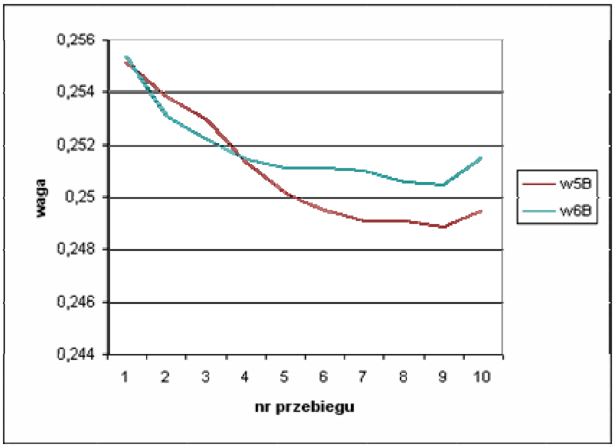
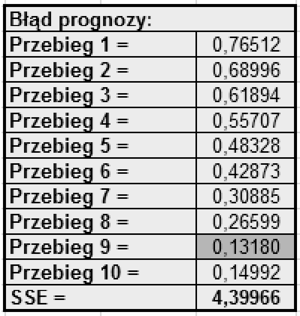


Fig. 10. Weights of connections 5B and 6B in enlarged scale



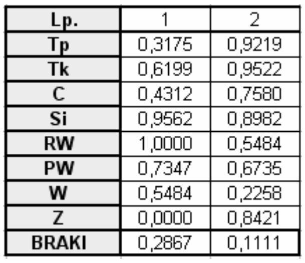
Błąd prognozy:	
Przebieg 1 =	0,76512
Przebieg 2 =	0,68996
Przebieg 3 =	0,61894
Przebieg 4 =	0,55707
Przebieg 5 =	0,48328
Przebieg 6 =	0,42873
Przebieg 7 =	0,30885
Przebieg 8 =	0,26599
Przebieg 9 =	0,13180
Przebieg 10 =	0,14992
SSE =	**4,39966**

Fig. 11. Weights. Prognosis error for ten subsequent runs.

Basing on 10 runs, weights of connections have been determined. An assumption can be made, that data set containing over a 100 cases of training data would be sufficient. Although the results might be not credible enough, basing on them the study have been decided to be continued, to bring the procedures to the final concluding stage.

Next step of inspecting acknowledged model of neural network is creating a test set, on which newly found weights will be tested. To this end, records with normalized input data were isolated again and network operation has been tested basing on constant weights, determined during training process. It has been decided to test the network for two new records of data, assigned with numbers 1 and 2 (fig. 12).

Activation function of neuron Z returned consecutively values 0,65 and 0,6527 as answers (outputs) of neural network on given parameters. After data denormalization (action opposite to normalization[5]) these values have been converted in a way to make them represent the percentage of faulty parts: Y1 denorm. = 31,55 %, Y2 denorm. = 31,67 %.



Lp.	1	2
Tp	0,3175	0,9219
Tk	0,6199	0,9522
C	0,4312	0,7580
Si	0,9562	0,8982
RW	1,0000	0,5484
PW	0,7347	0,6735
W	0,5484	0,2258
Z	0,0000	0,8421
BRAKI	0,2867	0,1111

Fig. 12. Input data of test set records

Afterwards, both values have been compared with actual faulty parts level for both test records. These values are as following: Y1 = 15,2 %, Y2 = 7,3 %.

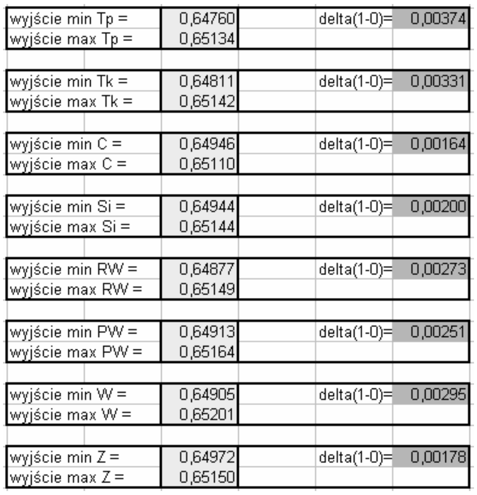
There are considerable differences between actual data and values foreseen by MLP network. It does not mean that modeling with ANNs basing on manufacturing data was unsuccessful. Because of too low amount of training records model could not establish weights on desired level, what was suggested earlier anyway. Weights determined randomly changed their values slightly after training, for example: 0,3 → 0,2972 and 0,7 → 0,6994, when with the complete training process (reaching "stop" condition) changes should be more distinct, for example 0,3 → 0,75 and 0,7 → -0,1.

Additionally, network performance was tested using another technique – sensitivity analysis, which allows to estimate relative influence of each attribute on the network outcome [4]. To perform the analysis, new observation $X_{i\ average}$ has been isolated, with value of each attribute equal to average of all records from training data set.

In further stage, value of each parameter has been changed from minimum to maximum and analysed how it will affect the outcome of network performance, with other parameters left unchanged (equal to average).

Each row accounts for the model input (for example, for the parameter Tp set on minimum level, the values are as following: 0; 0,6892; 0,6276; 0,5104; 0,6011; 0,5148; 0,4505; 0,4421) and obtained difference between output for maximum value of parameter and output for minimum value of parameter is a measure of influence of each attribute on network final outcome. Acquired results for all 16 measurements (study was carried out for 8 parameters) along with calculated difference between min-max outputs are shown on fig. 13.

According to conducted sensitivity analysis, parameters Tp, Tk and W (initial temperature, final temperature and humidity) have the highest influence on network performance, what can be seen on the fig. 13. The result of the analysis differs from the CART decision tree model for the same assortment, although humidity is common for both. Moreover, chart of importance of parameters taking part in generating tree in Statistica program indicates, that also Tp parameter is important (check fig. 5). Choice of such parameters is more justified from the technological point of view.



wyjście min Tp =	0,64760	delta(1-0)=	0,00374
wyjście max Tp =	0,65134		
wyjście min Tk =	0,64811	delta(1-0)=	0,00331
wyjście max Tk =	0,65142		
wyjście min C =	0,64946	delta(1-0)=	0,00164
wyjście max C =	0,65110		
wyjście min Si =	0,64944	delta(1-0)=	0,00200
wyjście max Si =	0,65144		
wyjście min RW =	0,64877	delta(1-0)=	0,00273
wyjście max RW =	0,65149		
wyjście min PW =	0,64913	delta(1-0)=	0,00251
wyjście max PW =	0,65164		
wyjście min W =	0,64905	delta(1-0)=	0,00295
wyjście max W =	0,65201		
wyjście min Z =	0,64972	delta(1-0)=	0,00178
wyjście max Z =	0,65150		

Fig. 13. Results of sensitivity analysis

This method from Data Mining domain used in the foundry industry, on the basis of own past experience and better organized procedures for acquisition data sets gives a broader possibilities, for example as using a few of output neurons in the above described method.

Created model with defined weights (constant) and given output parameters taken into consideration is able to identify the casting as good or faulty. Result – output of neural network is obtained in a form of continuous variable, so user can determine the value of goal variable (in this case: defective products).

5. Presentation of possibilities and comparison of DT and ANNs methods

In the paper, an attempt has been made to show practical application of decision tree and artificial neural network models. Methods based on different algorithms of data preparation and program operation. In case of classifying decision trees, input set was divided into decision classes and each of them was described with appropriate parameter (using discrete variable – GOOD, AVERAGE, WRONG). It is worth mentioning, that regressive decision trees also exist and they utilize continuous (quantitative) variables, but for the purpose of this work classifying trees were used. In case of ANN, input data were not divided into classes, normalized values have been used instead.

Application of discrete variables in case of decision trees gives clear and comprehensible results, because model classifies records to classes determined by user. For artificial neural networks, there are no direct procedures to convert weights of network to decision rules. Resulting from classifying decision trees intuitive rules are additional advantage of these models, because of more intelligible way of pointing out existing relations (for example in a form of *if* A = yes *and* B = yes *then* C = yes). In presented study, much easier to build decision tree models al-

[5] Denormalization was carried out according to assumption: *foreseen value = network result * range + minimum*.

lowed to determine which parameters influenced the process and to what degree. Model allowed to foresee casting quality classification basing on input variables described with three statistical estimators.

Model based on neural network did not require dividing input data set into classes. Actual (continuous) values of measured parameters have been used. In case of decision tree it has been decided to limit the number of input parameters because of the risk of model overtraining and also to avoid too complex calculations. This limit was not necessary for neural network, but to make the comparison of methods and obtained results possible, it has been also implemented.

ANNs allow to model greater number of outputs, for example determining the number of faulty products if the pouring temperature is higher than 1380°C. Advantage of neural network used in presented study is the possibility of classifying the goal variable in form of continuous number variable, which gives precise results, like faulty products amount = 3,6% (unlike in decision tree model, where product is classified basing on a priori defined classes). However, attribute coding has also some disadvantages, if it comes to labour-consumption, because of, for example, data normalization and then denormalization of results to the primal state, comprehensible for production engineer.

Both methods show some similarities. Both models allowed to isolate and indicate process parameters that had the highest influence on casting quality. In case of decision trees it resulted from determining the optimal division and for neural networks from sensitivity analysis.

Before selecting individual Data Mining model, deep and thorough analysis of considered manufacturing process is recommended. The choice should be affected among others by: expected accuracy, degree of complexity (time consumption), offered and predicted possibilites. Comparison of possibilites basing on authors experience and criteria is contained in table 2.

Table 2. Comparison of possibilities of decision trees and artificial neural networks

Criteria of model evaluation	**CART**	**MLP**
Possibility of classifying input variable basing on input parameters	++	+
Possibility of foreseeing an accurate value of output variable	0	++
Possibility of determining the most significant parameters	++	++
Possibility of using several „outputs”	0	++
Low time consumption of calculations	+	-
Clear, comprehensible results	++	-
Possibility of using unlimited number of parameters	-	++
Possibility of modeling with availability of low quantity data sets	0	-
Modeling using continuous variables	0	++
Noisy data resistance	-	+

Legend:
++ meets criterion very well
+ meets criterion well
0 meets or does not meet criterion, depend model used
- does not meet criterion

6. Summary

Application of modern Data Mining tools, already functioning with success for some time in other branches of industry, in production engineering for process modeling, created many new perspectives for manufacturing processes in industrial production plants. Although in this area exploration of data is in its primary stage of development, a dynamical expansion of these studies is predicted. Examples can be found in works listed at the end of the paper [13,17].

In this paper authors tried to show possible results that can be obtained when modeling manufacturing processes using selected Data Mining tools. Presented models of decision tree and artificial neural network belong to so-called soft modeling methods, which means that there are no rigid, imposed rules and algorithms of dealing with these models. In case of conducted research and analyses it has been attempted to show, that used models need to be in significant part matched with the character of considered , particular process. In effect, it is not possible to copy these methods directly from another production plant, only some generalizations may be helpful (compare in: [13,16]).

Additionally, during the study authors ascertained that there are major divergences between character of research conducted in area of manufacturing – among other things with non-homogenous and high level of “impurity” of data understood [14] as measured process parameters, and other fields of application of DT and ANNs. It turned out that complexity of manufacturing processes significantly influences the results of modeling, which were non-reliable from a technological point of view (for example defined “proper” technological parameters of the process did not always guarantee obtaining good quality castings). Simultaneously, cases occured where for the same input data entirely different results were obtainer (decision tree and neural network).

These observations confirm complicated and multiaspect character of data exploration and attempts at testing Data Mining methods in conditions of full manufacturing platform made in this work will be further developed.

Summarizing, while testing data sets during modeling, mostly Excel spreadsheet was used to carry out analyses using available formulas and charts. Statistica 9.1 software was also very helpful, a comparison could be made between “manual” method (step by step algorithm) and built-in model. Minor deviations between acquired results occurred, which could be a result of individual assumptions adopted during modeling with “manual” method.

In presented study only ten element data set was available, plus two element set for MLP model validation. It has been explained, how it decreases the probability of obtaining valuable results. Extending the data set of considered days could give better results, although, as mentioned earlier, too numerous data set and “high” expectations could lead to overtraining the model.

Authors devoted several years to master the procedures of data acquisition in foundries [2,6-12,15] and they treat this work as the beginning of further research and analyses on using credible data to model and make prognoses of casting quality using Data Mining methods, taking into consideration possibly the most complete set of foundry process parameters.

Authors experience indicates that specialized companies implementing Data Mining systems want to offer as universal services as possible. That is why, when offering the system to particular foundry, the preferred way of foundry staff participation in the system is merely consultation. How known examples proved, chance of correct implementation is definitely lower in this case. Also, there are known foundries that develop such systems from the scratch in cooperation with external companies. In such cases, effects of implementation are undoubtedly beneficial.

Acknowledgements

The paper was written within the scope of supervisor research grant N508 526538.

Literature

[1] Ignaszak Z., Virtual Prototyping in foundry. Poznan University of Technology, Poznań 2002 (in Polish).

[2] Ignaszak Z., The specification and examples of on-line validation methods needed for quality forecasting systems for industrial castings [in:] Innovations in castiing part III, Edited by J. Sobczak, Institut of Casting, Cracow 2009 (in Polish).

[3] Perzyk M., Kochański A., The Possibility of artificial neural networks application for casting processes modeling", Soldification of Metals and Alloys, No. 38, 1998 (in Polish)

[4] [Larose T., Knowledge discovering from data. Introduction to Data Mining, PWN, Warsaw 2006 (in Polish)

[5] Łapczyński M., Classification trees in customer satisfaction and loyalty studies. Statsoft 2003 [w:] http://www.statsoft.pl/czytelnia/marketing/drzewa.pdf (in Polish).

[6] Sika R., Ignaszak Z., Acquisition and preliminary preparation of non-homogenous data needed to Data Mining systems on the example of foundry industry, Archives of Production Engineering and Automation, Poznan University of Technology, Poznan 2009 (in Polish).

[7] Sika R., Ignaszak Z., Data Mining in the foundry industry - problems recording and collection of non-homogenous data. Conference Modeling of Casting and Foundry Processes – Śrem 2008 (in Polish).

[8] Sika R., Ignaszak Z., Data Analysis – system to optmalization quality of production processes in foundry. User guide. Poznań – Leszno 2009 (in Polish, non published).

[9] Sika R., Ignaszak Z., After implementation KonMas-final program - use to cast production process analyze - W6 department - Foundry in Śrem, International Symposium – Modeling of casting and foundry processes, Poznań – Śrem, October 2006.

[10] Sika R., Study on the SAP R/3 system structure and possibility of adapting to the management and quality control in Śrem Foundry, Master's thesis under the direction of Z. Ignaszak, Poznan University of Technology, 2006 (in Polish)

[11] Ignaszak Z., Sika R., Exploration system for selected production data and its testing in the foundry, Archives of Production Engineering and Automation, Poznan University of Technology, Posen 2008 (in Polish).

[12] Ignaszak Z., Sika R., Application of Data Mining systems for compilation of data in virtualization systems explored in the design and manufacturing process control in found, Unpublished work, Poznan University of Technology, 2009 (in Polish)

[13] Perzyk M., Data mining in the foundry. The potential, problems and projects. Presentation at the XI International Symposium on Modeling of Casting and Foundry Processes, October 26-27, 2008, Poznan-Srem (Poland)

[14] Wyrozumski T., How do I make the data were clean? Proceedings PLOUG IX Conference, October 2003, Kościeliskom Poland (in Polish)

[15] R. Sika, Z. Ignaszak, Quality Assurance in the foundry industry. Acquisition and preliminary development of heterogeneous, Archives of Mechanical Engineering and Automation, 2009

[16] Jakubski J., Dobosz, M., Application of neuronal networks for quality control of moulding sand. XXXIII Conference of National Foundryman Day, Faculty of Foundry Engineering Science and Technology, Krakow, 2009,

[17] Perzyk M., Soroczyński A., Comparison of selected tools for generation of knowledge for foundry production, Archives of Foundry Engineering, Vol.8, Issue 4/2008.