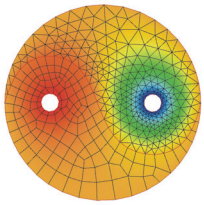
Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE

Informatyka w Technologii Materiałów

Vol. 8, 2008, No. 2



APPLICABILITY OF DECISION TREES IN MANUFACTURING INDUSTRY

MARCIN PERZYK, ROBERT BIERNACKI, ARTUR SOROCZYŃSKI

Warsaw University of Technology, Institute of Materials Processing, Narbutta 85, 02-524 Warszawa
Corresponding Author: M.Perzyk@acn.waw.pl (M. Perzyk)

Abstract

In manufacturing companies large amounts of data are collected and stored, related to designs, products, equipment, materials, manufacturing processes etc. Utilization of that data for improvement of product quality and lowering manufacturing costs requires extraction of knowledge from the data, in the form of appropriate conclusions, rules, relationships and procedures. Data mining provides tools and methodologies for semi-automated extraction of that type of knowledge. It is a multidisciplinary field, rapidly growing in recent years, and used mainly in business, medicine, social sciences. Applications to manufacturing and design on a large scale are relatively seldom.

In the present work some important manufacturing-related problems are characterized, from the standpoint of benefits from application of data mining methods. In the second part of the paper some selected results of the authors' studies and research are presented, showing performance of decision trees in solving important typical problems in manufacturing industry.

Key words: data mining, manufacturing processes, parameter significance, statistical methods, artificial neural networks

1. INTRODUCTION

In majority of manufacturing companies large amounts of data are collected and stored, related to designs, products, equipment, materials, manufacturing processes etc. Utilization of that data for improvement of product quality and lowering manufacturing costs requires extraction of knowledge from the data, in the form of appropriate conclusions, rules and procedures. This can be facilitated by methods offered by a relatively new, interdisciplinary field, called data mining (DM), which includes methodologies and tools from several disciplines such as database systems, visualization, statistics and learning systems. DM is rapidly growing in recent years, however until now, it has been used mainly in business, medicine and social sciences. Applications to manufacturing and design on a large scale are relatively seldom (Kusiak, 2006; Harding et al. 2006; Wang 2007; Perzyk, 2006; Perzyk et al., 2007).

The knowledge obtained by DM techniques from recorded past data can be in various forms, such as verbally expressed logic rules, classification systems suitable for classification of new objects without explicit rules presentation, regression models used for prediction of new values of real-type variables, relative significances of input variables, characteristics of groups (clusters) based on of regularity appearing in the data, association rules etc.

The most important type of DM tools are those utilizing learning systems, such as decision trees, Bayesian classifiers, rough sets theory based systems, artificial neural networks, MARSplines etc. In general, the models used in DM can be parametric or non-parametric. Non-parametric models differ from parametric models in that the model structure is not

 ISSN 1641-8581

specified a priori but is instead determined from data. The non-parametric models are essentially more suitable for knowledge discovery as the nature of the relationships hidden in the data is usually not known. The decision tress (DTs), also called classification trees, are probably the most often used models of that type, due to their simplicity, low computational costs and ease of graphical and verbal presentation of knowledge.

The paper consists of two main parts. In chapter 2 some important manufacturing-related problems are characterized, from the standpoint of benefits from application of DM methods. In chapter 3, selected results of the present authors research, revealing important features and performances of DT in solving important typical problems in manufacturing industry, are presented.

2. POTENTIAL APPLICATIONS OF DM TOOLS IN MANUFACTURING

The DM methods can be helpful in solving problems appearing at all main stages and concerning various aspects of manufacturing processes, i.e. design, control, running and quality assurance. Some of them are discussed below.

Designing of the manufacturing processes and tooling in contemporary industry is assisted by advanced computer tools, covering simulation software, expert systems based on knowledge acquired form human experts as well as the knowledge extracted by the semi-automated DM methods. In general, the designing aids, both conventional ones like formulas, procedures, data bases etc., and the advanced knowledge systems, play particularly important role at the initial stage of the designing process. The proper choice of the manufacturing process alternative in that phase allows reduction of number of design versions and, consequently, the number of necessary corrections resulting from simulation and/or floor tests. Exemplary problems related to selection of the manufacturing process alternatives in foundry industry, for which the knowledge obtained by DM methods can significantly contribute to the right decision making, can be found in (Perzyk et al., 2008).

Detection of irregularities, appearing as excessive variations of manufacturing process, is usually carried out with the Statistical Process Control methods (SPC). It is important that SPC includes not only the detection of the irregularity occurrence but also analysis of the case and usage of its results to determination and elimination of the cause. The fundamental tools of SPC are control charts which are used to detect the statistical process instabilities. The most often used type of control charts is the Shewhart's sample mean, for which several rules of interpretation have been formulated, permitting identification of the excessive variation type or its general cause (see, e.g Hill, 2007). In the present authors' opinion, the classification tools could be also helpful for this identification. The idea is to replace the information about specific sequences of the sample points by a set of actual values of recent sample means, which would be input (independent) variables for a classification tree, while the output will be the type of variation (general cause). The primary training could be performed using artificially generated data sets, including various realizations of the typical cases. The final induction of the tree would be based on data including real cases, recorded in a given enterprise.

Discovery of root causes of manufacturing process irregularities, leading to deteriorating product quality, is undoubtedly one of the most important tasks which could be performed with a use of the DM techniques, particularly learning systems. One of the possibilities is to build a regression model of

Table 1. *Applicability of knowledge obtained with DM methods for solving manufacturing problems*

		Type of knowledge			
		Prediction with regression models	Significance of process parameters	Classification of cases	
				Verbal logic rules	Classification models (implicit)
Type of task	Process design	+++	-	+++	+
	Process control	++	++	+++	+
	Detection of irregularities (excessive variations) in process	-	-	+	+++
	Determination of root causes of process irregularities (quality deterioration)	-	+++	++	++

the process, in which the input (independent) variables would be widely understood process parameters and the output (dependent) variable should characterize the process quality. An analysis of the model would indicate the most significant input variables, i.e. those which affect the output in the largest extent – these are the most probable causes of the quality drop. The input variables can be factors related to material, machine, man, organization, environment etc while the quality can be defined by the product's property level (e.g. strength) or fraction of defective parts.

The summary of applications of particular types of knowledge at the main stages of manufacturing processes, are presented in a symbolic form, in Table 1.

3. ASSESSMENT OF DECISION TREES APPLICABILITY AND PERFORMANCE

As a result of extensive literature studies, including recently published handbooks and research reports as well as the authors' own research, an overview of DT application areas has been elaborated and summarized in Table 2. The term „precision" denotes ability of flexible and precise fitting the model to data; „availability" is related to software and includes also its user-friendliness and pricing.

In the following sections the authors' research results are presented. Most of the computations regarding decision trees were carried out using MineSet™ software package in version 100M, provided by a US company Purple Insight. Details concerning the research methodology and other necessary comments are given at the beginning of each section.

3.1. Regression-type modeling

DTs are basically classification models which means, that the independent variable has to be of discrete type, i.e. nominal (often called categorical) or ordinal. In the present work, continuous output variables were converted to ordinal values before tree induction by assigning their actual values to appropriate order numbers of uniformly arranged intervals. Discrete responses of classification tree models, expressed as the intervals' order numbers, were converted to continuous values by assigning central values of the appropriate intervals. The number of intervals used for the conversions was 10 in most cases.

Although the DTs are the most popular tree-type non-parametric models used in DM, another models of that class, dedicated for modeling relationships between continuous variables, called regression trees, have been also developed and are incorporated in commonly available software. These models are also included in the present research.

Numerous successful applications of artificial neural networks (ANNs) in modeling of various manufacturing processes reported in recent years indicate that neural regression models are generally recognized to be outstandingly flexible and therefore capable of reflecting the unknown dependencies existing in a recorded data. This is true, despite the fact that neural models cannot be considered as non-parametrical, as a particular structure of the network and a form of the activation function must be assumed. The MLP-type ANNs were also considered in the present work and their performance was compared with that of classification and regression trees. Details concerning ANNs can be found in (Perzyk et al., 2007).

Two types of data were used for testing of the regression type modeling: simulated data sets, with assumed hidden relationships, and real data, collected in the foundry industry. The simulated data sets were generated in the following way. First, an analytical formula of the type Y = f(X1, X2, …) was assumed. Then for random values for independent variables X1, X2, … the dependent variable Y was calculated. Finally, a Gaussian-type noise with maximum deviations ±20% was imposed on the

Table 2. *Assessment of DT as a support for solving production problems*

		Decision tree features		
Type of knowledge		Applicability	Precision	Availability
Prediction with regression models		satisfactory	satisfactory	good
Significance of process parameters		good	poor	satisfactory
Classification of cases	Verbal logic rules	good	good	good
	Classification models	good	good	good

independent variables. All the values were normalized within 0 – 1 interval. 1200 records for each data set were generated in that way; of those 1000 records were used for training the models and 200 records were used for testing their predictive capabilities. Several different basic formulas were assumed, however, only results for the following two are presented below: Y = X1+2·X2+3·X3+4·X4+5·X5, i.e. with no interactions between input variables, and Y=X1·X2+X3+X4+X5, with strongly interacting variables X1 and X2. The motives of the particular forms of the above formulas are related with the significance analysis of the input variables, presented in the next section, where the same data sets are used. The industrial data set correlates chemical composition of ductile cast iron with its tensile strength, obtained as a result of the melting process; some more details can be found in (Perzyk & Kochański, 2001).

All the results are summarized in Fig. 1 in a form of the prediction error bars. It can be seen that for the simulated data sets classification and regression trees are significantly less accurate than ANNs. This observation is true for the both subsets: the training data, which implies lesser flexibility of the tree type models and the new (testing) data which indicates their lesser predictive capability. It is also worth noticing that the regression trees do not perform remarkably better in modelling regression-type relationships than the classification trees. The latter can be even more accurate if a larger number of ordinal values is assumed.

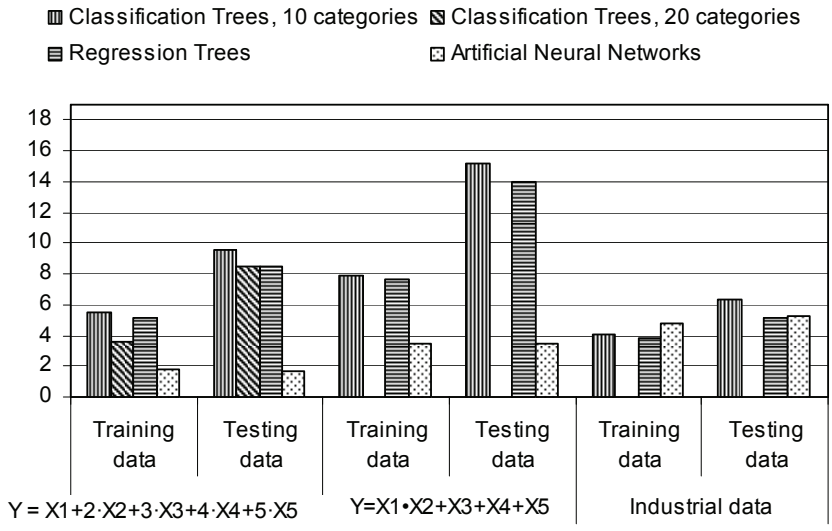


Fig. 1. Comparison of % average output prediction errors for classification trees, regression trees and ANNs

For the industrial data set the errors for all the three models appeared to be very similar. Analysis of the results presented in Fig. 1 indicates that the ANNs errors are higher and the tree-type models errors are smaller, compared to the simulated data. This could be a result of the fact that the relationships hidden in the data are much more complex, compared to the simulated data, and that the non-parametric models perform relatively better in such cases.

3.2. Significances of process input variables

There are two main approaches to extraction of valuable information from regression or classification models (Etchells & Lisboa, 2006). One is called 'pedagogical' and treats the model as a black-box, i.e. uses a specially designed interrogation procedure to obtain the desired information. Another approach is called 'decompositional' and is based on an analysis of the model's parameters. For ANNs it was found that the methods for finding relative importances of independent variables based on the network weights values, i.e. the decompositional approach, were not satisfactory (Perzyk et al., 2003).

For DTs, a qualitative evaluation of input variables significances is simple: variables appearing in the tree structure (or logic rules resulting from it) are significant (Wang, 2007; Huang & Wu, 2006). Analysis of the DT structure and parameters can be also used for qualitative or semi-quantitative estimations of the variables significances (Huang & Wu, 2006). Some commercial software packages suggest measures for quantitative ranking of independent variable importances based on various DT purity definitions (Vanderberg & Motroni, 2007). These methods qualify as 'decompositional' ones.

In the present work the relative significance factors based on the 'pedagogical' approach were tested for all the three models mentioned in the previous section, i.e. classification and regression trees and ANNs. The same factor definition, previously developed and tested by the present authors for ANNs, was assumed for all the models: the significance factor for a single input is defined as the maximum difference

of the output, which can be obtained by changing the value of the analyzed input. These values are calculated repeatedly a number of times for other variables set at random levels. Thus, the magnitude of resulting scatter of the significance factor of the given input reflects the possible interactions with the other variables. All the significances thus obtained are normalised by dividing them by the value obtained for the most significant variable. Details concerning computational aspects of the above methodology can be found in (Perzyk et al., 2007).

In Fig. 2 comparisons of the relative significance factors of input variables, obtained from the three regression models for the two simulated data sets using the above methodology, are presented. The values found from classification and regression tress, essentially reflect the expected tendencies, however, their values are remarkably less accurate then those obtained form ANNs.

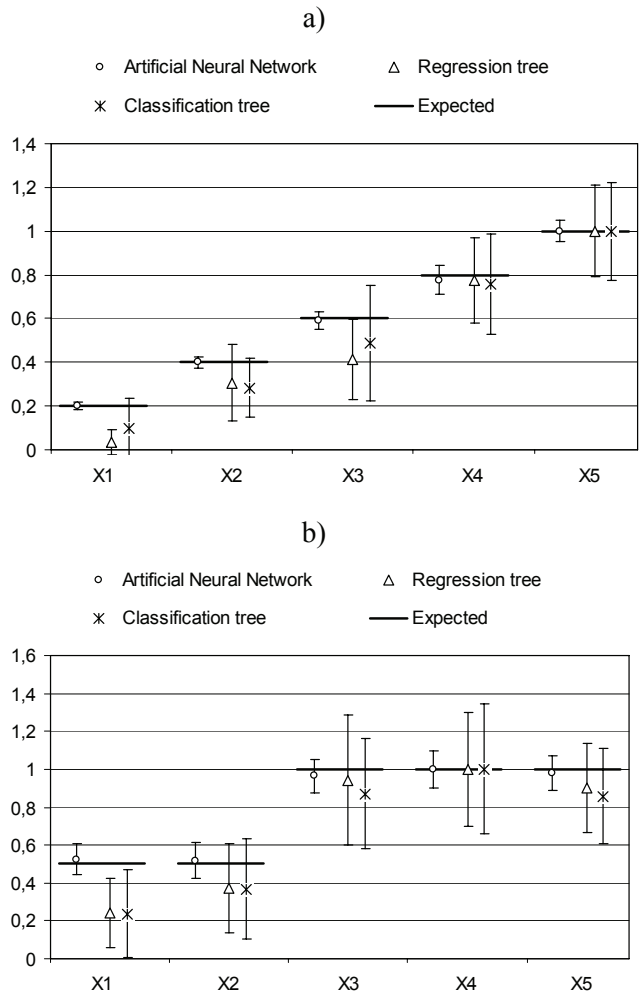


Fig. 2. *Comparison of relative significance factors obtained from regression models using various learning systems for the simulated data sets obtained from the basic formulas: a) Y = X1 + 2·X2 + 3·X3 + 4·X4 + 5·X5, b) Y = X1·X2 + X3 + X4 + X5*

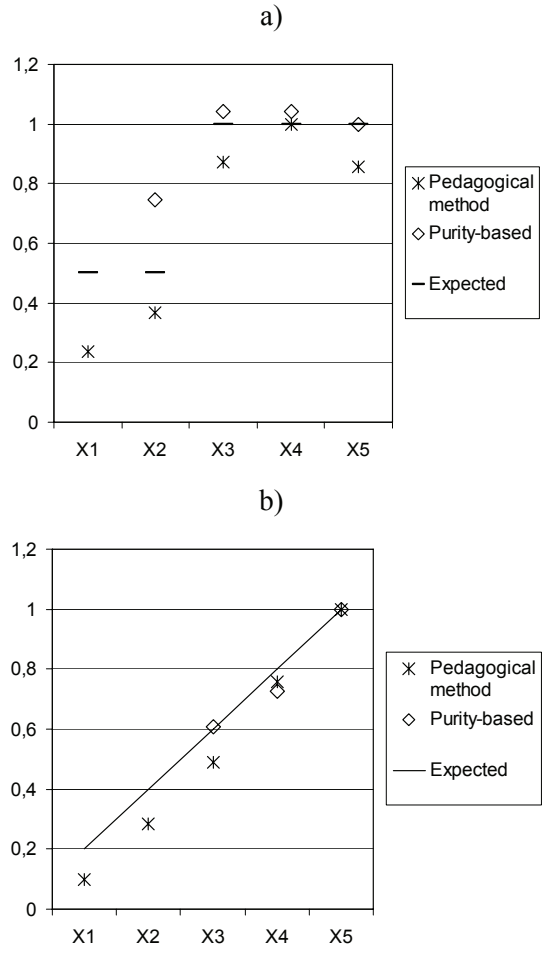


Fig. 3. *Comparison of relative significance factors obtained from classification trees by two methodologies: the pedagogical method proposed in the present work and the purity-based method, for simulated data sets obtained form two basic formulas: a) Y = X1 + 2·X2 + 3·X3 + 4·X4 + 5·X5, b) Y = X1·X2 + X3 + X4 + X5*

Dispersions of relative significance factors are observed for all three models. However, negligible scatters for variables with no interactions (all variables in Fig. 2a and X3, X4 and X5 in Fig. 2b) are observed for ANNs only, while the both tree-type models evidently reveal non-existent interactions between input variables.

In Fig. 3 comparisons of the relative significance factors of single variables obtained form classification tress using two methodologies are presented: the above 'pedagogical' methodology, proposed by the present authors, and the purity-based technique (Vanderberg & Motroni, 2007), also normalised in respect to the maximum values. The latter provides non-zero values only for those input variables which are found as important. The three values obtained for the simulated data set with no interactions (Fig. 3a) are fairy satisfactory. However, the results presented in Fig. 3b indicate, that purity-based significances are correct only for the non-interacting variables

(X3, X4, and X5). The significances of the first two variables, interacting between each other and essentially equally important, are far from expectations. The value for X2 is much too high, while the variable X1 was recognised as not important by the purity-based technique. Such a wrong qualification of a manufacturing process variable could lead to hazardous reduction of the control procedures for that variable.

In the opinion of the present authors, the above inaccuracies and incorrectness can be only partly attributed to the lower prediction accuracies of the classification and regression trees, compared to ANNs (as shown in Fig. 1) and should be further investigated.

For the industrial data (Fig. 4) all the regression models pointed at copper as the most significant alloying element, which agrees well with the industrial practice (copper is used to obtain pearlitic structure of the ductile cast iron, necessary for high strength grades the alloy).

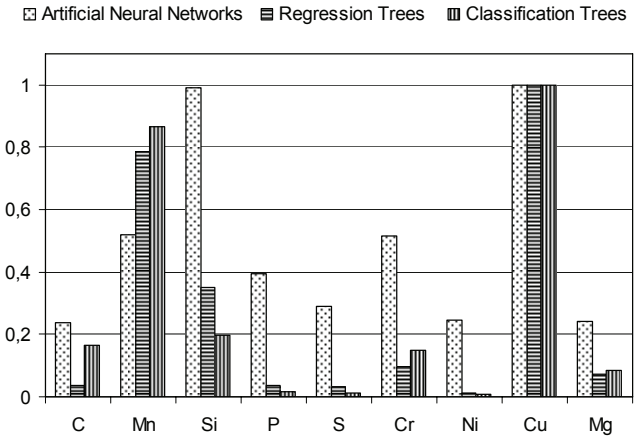


Fig. 4. *Relative significance factors for the industrial data set (tensile strength of ductile cast iron vs its chemical composition as defined by 9 elements)*

Different predictions form different models were obtained for other elements, however, in case of the least significant variables, such as C, Ni, and Mg, all models were also fairly conformable. It should be noticed, that these elements are generally not insignificant for the strength and the obtained results merely indicate that they were not important in that data set (probably their variability was too small to significantly influence the cast iron properties). However, the control of these elements could be possibly reduced in that particular plant.

3.3. Testing of logic rules generation from decision trees

3.3.1. Data sets

For the preliminary evaluation of decision trees as the engineering knowledge extraction tools, the data records were obtained as readouts from a nomograph published in the professional literature related to foundry technology (Holzmuller & Wlodawer, 1953). This nomograph, comprising a semi-empirical knowledge, is widely used for calculation of the feeding shrinkage of grey cast iron castings and determination of appropriate dimensions of risers. The fundamental decision which should be made in designing of rigging systems for that kind of castings is whether the application of a riser is necessary. The so called riserless design can be appropriate when the iron expansion, which occurs during the solidification period, is capable of compensation its shrinkage, which takes place during cooling of the liquid phase, i.e. when the overall volume change (called imprecisely shrinkage) will be positive. The volume changes appearing during cooling and solidification of grey cast iron castings depend on:

- pouring temperature (superheating of the alloy), affecting mainly the liquid contraction,
- cooling rate of the casting dependent mainly on its massiveness and defined by solidification modulus,
- chemical composition of cast iron (defined by the fractions of two groups of elements: carbon and summary fraction of silicon and phosphorus),

It is worth noticing that the relationships between shrinkage and the above variables are not independent on each other, e.g. only massive castings can be poured from lower temperatures. In general, the complexity of the problem results in that exact, analytical methods of calculation of shrinkage and risers are not available.

Number of readouts of the nomograph made for various combinations of all input variables was 191. The continuous output variable values (shrinkage S) were converted to nominal (discrete) ones, expressed by classes. Two versions of the output variable classifications were assumed:

Version 1: two values: „riser not required" (if $S \geq 0$) and „riser required" (if $S < 0$).

Version 2: three values defining the necessity of use and type of the riser: „ not required" (if $S \geq 0$),

„small" (if -1%<S<0) and „large" (if S<-1%). When the riser volume is relatively small, it is usually cost ineffective to apply the exothermic sleeves, while for large riser volumes the sleeves are commonly used. That type of classification would be therefore helpful in making decision concerning both the necessity of riser and its type.

Finally, two test data sets were obtained, each of four real value inputs (carbon contents C, %, summary content of silicon and phosphorus Si+Pi, %, solidification modulus mo, cm, pouring temperature tpor, ^{o}C) and with one output, in the form of the above defined two types of nominal values. That type of data sets can be considered, in a certain extent, as examples of real, noisy data sets obtained in industrial conditions. On the other hand, they express the hidden relationships about which there is much known, thus permitting better interpretation of the results of testing the trees and rules induction.

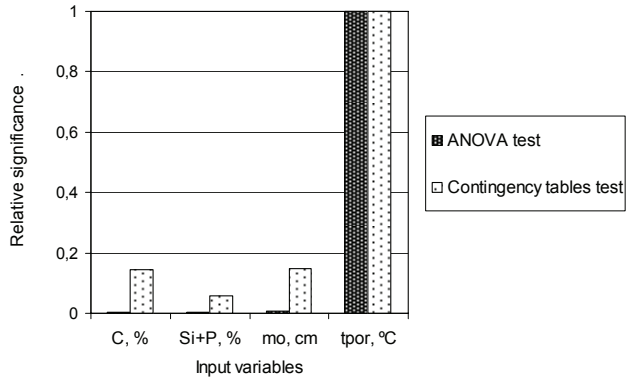


Fig. 5. *Relative significances of input variables for output variable 'cast iron shrinkage S' obtained by statistical non-parametric methods*

A preliminary statistical analysis of significance of input variables (in respect of the shrinkage S) was carried out; detailed methodology can be found in (Perzyk & Kozłowski, 2006). The results, presented in Fig. 5, show that unquestionably predominant influence on the shrinkage has the pouring temperature, while the other variables are much less significant.

3.3.2. Results of trees and rules generation

In Fig. 6 a graphic representation of the decision tree for *Version 1* data set is presented, obtained for the MineSet software default settings (pessimistic pruning at confidence level 0,7). The sizes of the three bars appearing in each node reflect numbers of records (cases) of the data set: directed to left branch, right branch and split in the node (base bar). In Table 3 the information available for the tree shown in Fig. 6 is presented.

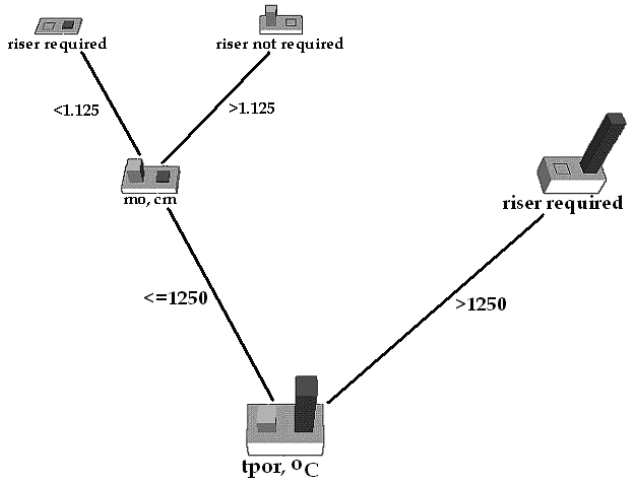


Fig. 6. *Illustration of decision tree generated by MineSet software for Version 1 data set*

The simplest form of the verbal rule equivalent to the generated tree will be:

„If a casting modulus is larger then 1.125cm and a pouring temperature is lower then 1250^{o}C, than a riser is not necessary, else a riser is necessary".

This result is in agreement with expectations based on foundry experience.

Table 3. *Logic relationships and numerical values corresponding the classification tree shown in Fig. 3 (Version 1 data set), obtained for pessimistic pruning at confidence level 0,7*

Path to node – logic rule left part	Splitting variable in node or leaf (**result of classification**)	Notation of branch leading to node	Sizes of output variable values in node (riser not necessary / riser necessary)	% fractions of output variable values in node (riser not necessary / riser necessary)	Node purity
`tpor, $^{o}$C`			41, 150	21.466, 78.534	24.97
`tpor, $^{o}$C`<= 1250	mo, cm	<= 1250	41, 6	87.234, 12.766	44.90
`tpor, $^{o}$C`<= 1250:`mo, cm`<= 1.125	**riser required**	<= 1,125	0, 6	0, 100	100
`tpor, $^{o}$C`<= 1250:`mo, cm`> 1.125	**riser not required**	> 1,125	41, 0	100, 0	100
`tpor, $^{o}$C`> 1250	**riser required**	> 1250	0, 144	0, 100	100

It is worth noticing that the procedure used for tree induction has completely ignored the two less significant variables (defining the cast iron chemical composition), which could essentially be a result of a low precision of the tree model. However, a closer examination of the training data revealed that the sign of shrinkage, deciding about the need of riser application, is a result of the pouring temperature and casting modulus only. In other words, there was no pair of records in which the pouring temperature and casting modulus would be the same and only one or both of two ignored variables would be different, which would have different classes of the output variable. For that kind of data, the tree structure could not be different from the obtained one.

In Table 4 the logic rules equivalent to the classification tree, obtained for *Version 2* data set with the MineSet software default settings (pessimistic pruning at confidence level 0,7), are presented.

The results obtained for the *Version 2* data set (with three nominal values of the output variable, i.e. riser: "not required", "small" and "large") for default settings of the MineSet program are so complex that their presentation in a form of one or a few simple verbal rules is difficult. Nevertheless, using those results for decision making regarding feeding method in any particular case is simple. It is worth noticing that here the tree induction algorithm has also utilized the previously ignored variables of low significance, defining the cast iron chemical composition, which also affect the classification results.

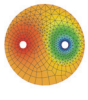
Because of the complexity of the tree obtained for the *Version 2* data set, another method of tree pruning was tried. Instead the MineSet default pessimistic pruning, the cost-complexity criterion at the default level = 0 (tree of minimum cost) was applied. The results are shown in Table 5.

Table 4. *Logic relationships corresponding the classification tree obtained for Version 2 data set for pessimistic pruning at confidence level 0,7*

Path to node – logic rule left part	Splitting variable in node or leaf (**result of classification**)
`tpor, °C	
`tpor, °C`<= 1250	mo, cm
`tpor, °C`<= 1250:`mo, cm`<= 1.125	**small**
`tpor, °C`<= 1250:`mo, cm`> 1.125	**not required**
`tpor, °C`> 1250	tpor, °C
`tpor, °C`> 1250:`tpor, °C`<= 1350	mo, cm
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`<= 2.25	**large**
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25	C, %
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`<= 3.1	**large**
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1	C, %
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5	Si+P, %
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5:`Si+P, %`<= 2.5	C, %
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5:`Si+P, %`<= 2.5:`C, %`<= 3.3	**large**
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5:`Si+P, %`<= 2.5:`C, %`> 3.3	Si+P, %
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5:`Si+P, %`<= 2.5:`C, %`> 3.3:`Si+P, %`<= 1.5	**large**
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5:`Si+P, %`<= 2.5:`C, %`> 3.3:`Si+P, %`> 1.5	**small**
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`<= 3.5:`Si+P, %`> 2.5	**small**
`tpor, °C`> 1250:`tpor, °C`<= 1350:`mo, cm`> 2.25:`C, %`> 3.1:`C, %`> 3.5	**small**
`tpor, °C`> 1250:`tpor, °C`> 1350	**large**

Table 5. *Logic relationships corresponding the classification tree obtained for Version 2 data set for cost-complexity pruning criterion = 0*

Path to node – logic rule left part	Splitting variable in node or leaf (**result of classification**)
`tpor, °C`	
`tpor, °C`<= 1250	mo, cm
`tpor, °C`<= 1250:`mo, cm`<= 1.125	**small**
`tpor, °C`<= 1250:`mo, cm`> 1.125	**not required**
`tpor, °C`> 1250	**large**



For the so simplified tree the verbal decision rule can be also relatively simple, e.g.:

„If a pouring temperature is lower then 1250°C, then for a casting modulus larger than 1,125cm a riser is not required, while for a casting modulus smaller than 1,125cm a small riser is necessary; in all other cases a large riser is needed".

It is worth noticing that for the pruning method based on the cost-complexity criterion a significantly simpler tree was obtained, compared to the pessimistic pruning, which is in agreement with a general tendency for these pruning methods (Quinlan, 1987). In particular, the relatively less significant variables, defining the cast iron chemical composition, were ignored.

4. CONCLUSIONS

The studies presented in the paper allow better understanding the role that application of DM methods can play in designing, control and fault diagnostic of manufacturing processes. Preliminary tests have shown that in regression modeling both classification and regression trees seem to be less accurate compared to ANNs, particularly for simple relationships present in the data. The relative significances of independent variables based on interrogation of trees exhibit non-existing interactions between the variables. The significances based on classification tree node purity measure can be very inaccurate for strongly interacting variables.

Decision trees appeared to be relatively simple and convenient tools for knowledge rules generation, enabling flexibility of the choice between a large number of precise rules and a small amount of rough rules, giving simple hints for decision making in various situations in designing and running manufacturing processes.

REFERENCES

Etchells, T.A., Lisboa, P.J.G., 2006, Orthogonal Search-Based Rule Extraction (OSRE) for Trained Neural Networks: A Practical and Efficient Approach, *IEEE Transactions on Neural Networks*, 17, 374-384.

Harding, J.A., Shahbaz, M., Srinivas, Kusiak, A., 2006, Data mining in manufacturing: A review, *J. Manuf. Sci. Eng. Trans. ASME*, 128, 969–976.

Hill, T., Lewicki, P., 2007, *Statistics Methods and Applications*, StatSoft, Tulsa, OK.

Holzmüller, A., Wlodawer, R., 1953, Zehn Jahre Speiser-Eingguss-Verfahren fur Gusseisen, *Giesserei*, 50, 781–791.

Huang, H., Wu, D., 2006, Product quality improvement analysis using data mining: A case study in ultra-precision manufacturing industry, *Lect. Notes Comput. Sci*., 3614LNAI, 577-580.

Kusiak, A., 2006, Data mining: manufacturing and service applications, *Int. J. Production Research*, 44, 4175–4191.

Perzyk, M., 2006, Data mining in foundry production, *Research in Polish Metallurgy at the Beginning of XXI Century*, ed., Świątkowski K., Committee of Metallurgy of the Polish Academy of Sciences, Kraków, 255-275.

Perzyk, M., Biernacki, R., Kozlowski, J., 2007, Data mining in manufacturing: methods, potentials, limitations, *Proc. Conf. Advances in Production Engineering 2007*, ed., Dąbrowski L., Warsaw, 147-156.

Perzyk, M., Kochański A., 2001, Prediction of ductile cast iron quality by artificial neural networks, J. Mat. Proc. Techn., 109, 305–307.

Perzyk, M., Kochanski, A., Kozlowski, J., 2003, Relative importance of input signals of neural network, *Computer Methods in Materials Science*, 3, 172-179 (in Polish, English abstract).

Perzyk, M., Kozłowski, J., 2006, Comparison of statistical and neural networks-based methods in analysis of significance and interaction of manufacturing processes parameters, *Computer Methods in Materials Science*, 6, 81–93.

Perzyk, M., Soroczyński, A., Biernacki, R., 2008, Possibilities of decision trees applications for improvement of quality and economics of foundry production, submitted to *Archives of Foundry Engineering*.

Quinlan, J. R., 1987, Simplifying decision trees, *Int. J. Man-Machine Studies*, 27, 221-234.

Vanderberg, H., Motroni S., 2007, *PurpleInsight MineSet 3.2™ Reference Guide*, available from www.purpleinsight.com/downloads/docs.shtml.

Wang, K., 2007, Applying data mining to manufacturing: The nature and implications, *J. Intell. Manuf*., 18, 487–495.

STOSOWALNOŚĆ DRZEW DECYZYJNYCH W PRZEMYŚLE WYTWÓRCZYM

Streszczenie

W przedsiębiorstwach produkcyjnych są zbierane i przechowywane duże ilości danych związanych z konstrukcją wyrobów, oprzyrządowaniem, materiałami, procesami technologicznymi itd. Wykorzystanie tych danych do poprawy jakości produkcji i obniżenia kosztów wytwarzania wymaga wydobycia z nich wiedzy w postaci odpowiednich wniosków, reguł, zależności i procedur. Eksploracja danych (data mning) dostarcza narzędzia i metodologie dla półautomatycznego wydobywania tego typu wiedzy. Jest to wielodyscyplinarna dziedzina wiedzy, gwałtownie rozwijająca się w ostatnich latach i stosowana głównie w biznesie, medycynie i naukach społecznych.

W niniejszej pracy scharakteryzowano niektóre ważne problemy związane z wytwarzaniem, z punktu widzenia korzyści ze stosowania metod eksploracji danych. W drugiej części artykułu przedstawiono niektóre wybrane wyniki własnych prac studialnych i badawczych, pokazujące możliwości i zachowanie się drzew decyzyjnych w rozwiązywaniu istotnych, typowych problemów w przemyśle wytwórczym.

Submitted: January 15, 2008
Submitted in a revised form: March 1, 2008
Accepted: March 4, 2008