ARCHIVES of FOUNDRY ENGINEERING

Published quarterly as the organ of the Foundry Commission of the Polish Academy of Sciences

ISSN (1897-3310)
Volume 8
Issue 4/2008

163 – 166

30/4

Comparison of selected tools for generation of knowledge for foundry production

M. Perzyk*, A. Soroczyński
Metal Casting Section, Warsaw University of Technology, Narbutta 85, 02-524 Warszawa, Poland
*Corresponding author. E-mail address: M.Perzyk@acn.waw.pl

Received 11.06.2008; accepted in revised form 08.07.2008

Abstract

Two types of data mining tools, suitable for semi-automatic generation of knowledge in a form of logic rules, are presented in the paper: decision (classification) trees and rough sets theory algorithms. A comparative evaluation of rules obtained by these two methods, used for decision concerning application of feeders for grey iron castings, is performed. Data sets obtained as readouts form a semi-empirical nomograph of Holzmüller and Wlodawer were used for the testing. It was found that both methods lead to similar rules, which are also in agreement with the foundry practice. However, the decision trees were unable to provide some important and reliable rules, which were generated by the rough sets theory algorithm and they can also generate rules which are not supported by the training data.

Keywords: Data mining, Knowledge rules, Decision trees, Rough sets theory, Foundry technology

1. Introduction

In majority of manufacturing companies large amounts of data related to production processes are collected and stored. Utilization of that data for improvement of product quality and lowering manufacturing costs requires extraction of a knowledge from the data, in the form of appropriate conclusions, rules and procedures. This can be facilitated by methods offered by the relatively new, interdisciplinary field called data mining (DM), rapidly growing in recent years. The most valuable DM tools are *computational intelligence* (CI) methods, making possible a semi-automated extraction of useful information from the data sets. One of the most important types of tasks which can be performed with CI tools is generation of knowledge in a form of logic rules. This facilitates a formation of engineering knowledge on the basis of production experience, in a form of design and manufacturing recommendations and relationships.

The aim of the present work was a comparative analysis of two different methods of knowledge extraction in the form of logic (decision) rules having the following general structure:

IF *attribute 1* = … AND *attribute 2* = …, THEN *output* = …

In this notation the sequence of expressions between ‘IF’ and ‘THEN’ is a conditional part or the rule, while its decision part is that appearing after ‘THEN’. Input variables (e.g. process parameters) are often called attributes, which can be of continuous type, i.e. expressed by real numbers (e.g. temperature in ^{0}C), or of a discrete (discontinuous) type. The latter includes ordinal values (e.g. temperature expressed verbally as ‘low’, ‘medium’, ‘high’ or by an ordered finite set of numerical values, e.g. 700.5, 820, 900) as well as nominal values (e.g. a casting can be ‘good’ or ‘bad’, employee’s telephone number can be 23, 11 or 87). If the attributes are of a real type, then the equality signs appearing in the conditional part of the rule are replaced by inequality signs (e.g. Temperature <= 300). The values of output (dependent) variables are always discrete, i.e. of nominal or ordinal types and they designate its decision class. In other words, a conditional rule assigns to a certain combination of attributes’ values one class of the decision variable.

The most widespread tools for extracting logic rules from recorded data, i.e. classification learning systems, are presently decision trees, also known as classification trees, and the methods based on the rough sets theory, developed by Polish researcher

Zdzisław Pawlak. The both approaches are widely treated in the world literature related to DM and CI, and will be only briefly characterized below.

Decision trees are non-parametric classification models, constructed from data by successive splits of the data records (learning examples), starting from the whole set. The splits are made in such a way that in the resulting subsets the classes of the decision variable are possibly homogeneous (preferably identical). The best splitting point is based on one attribute, called splitting variable. This procedure is repeated for successive subsets, which leads to a model structure represented by an oriented graph, reminding a tree. The splitting points are knots of the graph, the first knot is called a core and the lines connecting the knots are called branches or edges. The subsets which are not further divided are called leaves, and they provide results of classification (the dominant class in a leave is decisive). A tree model usually requires a restriction of its size. It is done either in the course of constructing the tree (e.g. by stopping further splits when the assumed minimum number of examples in a knot is achieved) or in a special simplification procedure of an already induced, too complex tree, called pruning. In the latter some knots are replaced by leaves if it does not lead to a significant drop of the model classification accuracy. There are many tree induction algorithms, which differ in the criterion of the class homogeneity in splits and in the criterion of the tree complexity. It is worth noticing, that each route leading from a core to a leaf can be expressed by a logic rule of the previously described structure. Decision trees also allow for evaluation of relative significances of the attributes, based on the so called purity of the splits. The large increments of the class homogeneity resulting from a split based on a given variable indicate its significance.

Rule extraction from the rough set theory requires that not only an output variable, but also all attributes, are of a discrete type. Each discernible learning example (data record) can be basically a rule. Thus obtained set of rules can be usually reduced and the rules can be simplified (i.e. their conditional part can be shortened). This can be done by striking out attributes which do not contribute to classification, i.e. after ignoring them the rule always points at the same class of the output variable, for all input values' combinations in the training data. The rules are evaluated, first of all from the standpoint of uniqueness of the classification. This is expressed by confidence, defined as a ratio of the number of examples in which appears this same combination of attributes values and class variable as in the rule, to the number of examples in which appears that combination of attributes values only (i.e. regardless the output class). Another parameter used for rules evaluation is number (or fraction) of examples compatible with a rule, called rule's support. If it is not possible to obtain from the data set rules of 100% confidence, then some not fully unique rules are utilized, often evaluated on the basis of various combinations of confidence and support values. The rough sets theory also makes possible an easy evaluation of relative significances of the attributes, based on reduction of uniqueness of classification resulting from deleting a given attribute in all rules.

Practical application of the rule induction methods is often difficult because of a lack of satisfactory knowledge about their characteristics and differences in performance. In the literature only a few analyses of that kind can be found (e.g. [1]). In the present work a comparative assessment of rules induced by the above discussed two methods is presented, using the example of decision concerning application of risers for grey iron castings.

2. Methodology

2.1. Data sets

Similarly like in the previous work [2], the data records were obtained as readouts from a nomograph published in the professional literature related to foundry technology [3]. This nomograph encompasses a semi-empirical knowledge and 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 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 inaccurately 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 total fraction of silicon and phosphorus),

The relationships between shrinkage and the above quantities 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 the situation that 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 (attributes) was 191. However, unlike in the previous work [2], now they are treated as ordinal ones, which permitted utilization of identical learning sets for the both rule extraction methods. All values of attributes appearing in the data set are given in Table 1. It is worth noticing that the number of possible combinations of the attributes' values is larger than the number of the training examples, as some combinations the readouts were impossible; such cases correspond to the situations which do not occur in practice.

Table 1.
Values of attributes (input variables values of ordinal type) appearing in the training data set

Content of C, %,	Content of Si+Pi, %	Casting modulus, cm	Pouring temperature, ^{0}C
3.0	1	0.75	1200
3.2	2	1.5	1300
3.4	3	3	1400
3.6	4	6	1500
3.8			

Similarly like in the previous work [2], the continuous output variable (shrinkage S) has been 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≥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≥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 a riser application and its type.

Finally, two test data sets were obtained, each of four ordinal type inputs and one output, in the form of the above defined two kinds 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.

2.2. Software

For induction of classification trees a commercial software package MineSetTM was used. The 'mutual info' splitting criterion (default) was assumed in all tests. Two different pruning criteria were tried: the default 'pessimistic pruning' (confidence level 0,7) and 'cost-complexity pruning' criterion (cost = 0).

For rule extraction based on the rough sets theory the authors' own software was used; it runs in MS Excel environment and utilizes VBA as programming language.

3. Results

For the data in *Version 1* both methods brought relatively small number of rules, all of 100% confidence, in which only two attributes appear: casting modulus and pouring temperature (the other two attributes defining the chemical composition of the alloy were recognized as not important by the both algorithms). All rules for this data version are given in Table 2.

Table 2.
Rules of 100% confidence for data in *Version 1* obtained from decision trees (DT) and rough sets theory (RST)

Rule No.	Method	Attribute (input variable)		Output class variable "Riser"	Rule support
		Modulus, cm	Pouring temperature, ^{0}C		
1	DT&RST	1.5	1200	No	15
2	DT&RST	3	1200	No	13
3	DT&RST	6	1200	No	13
4	DT&RST		1300	Yes	48
5	DT&RST		1400	Yes	48
6	DT&RST		1500	Yes	48
7	**DT**	**0.75**	**1200**	**Yes**	**6**
8	**RST**	**0.75**		**Yes**	**6**

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. The shrinkage variability resulting from the chemical composition was small enough that has not changed the sign of shrinkage, even in a single case. 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 the two ignored variables would be different, leading to different classes of the output variable. For that kind of data, the tree structure and the rough sets theory rules could not be different from the presented ones.

Most of the rules obtained from the two methods coincide, despite the fundamental differences in functioning of the both algorithms. The exceptions are rules 7 and 8, which in fact are related to the situation, when the modulus has the lowest value (equal to 0.75 cm). It can be seen that the rough set theory indicated that this value is satisfactory for the need of the riser use, while the decision tree has added a pouring temperature condition. The fact that the rule 8, including less attributes, has also 100% confidence, means that this condition is redundant. This case is typical for decision trees, in which a splitting variable appearing in the core must also appear in all rules. Similar notice, illustrated by a different example, was made in [1].

For the data set in *Version 2* the both methods have induced remarkably larger number of 100% confidence rules, including also the other two attributes (defining the chemical composition of the cast iron), which made their analysis and practical application more difficult. For decision trees, a significant reduction of amount of rules was obtained automatically by changing the pruning criterion for 'cost – complexity', which is in agreement with a general tendency of these pruning methods [4]. For the rough sets theory based method the significance analysis of the input variables (attributes) was made, in order to check if some of them could be omitted in the training data. In Fig. 1 the results of that analysis made by three different methods, are presented. It can be seen that the first two attributes are remarkably less important and, as such, they have been deleted from the data set. All the rules obtained in the above way for *Version 2*, finally including only two most important attributes, are presented in Table 3.

Similarly like previously, also for data in *Version 2* most of the rules obtained by the two methods appeared to be identical despite the fundamental differences not only in the extraction algorithms but also in the ways in which the number of attributes appearing in the rules was limited. The differences in rules 4 and 5 should be commented similarly like the differences between rules 7 and 8 in *Version 1*, discussed above. However, worth noticing is rule 8 in *Version 2* (see Table 3), obtained from the rough sets theory, which was not induced by the decision tree algorithm. This rule has a 100% confidence and also a relatively high support.

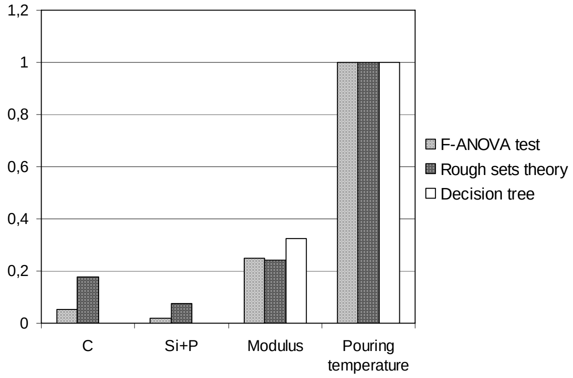

Fig. 1. Relative significances of attributes for the output variable in *Version 2*

It should be added, that the software used for tree induction usually does not calculate the confidence, which is a fundamental parameter in the rough sets theory. In the present case (*Version 2*) the decision tree also included the route corresponding to the rule: „IF *Pouring temperature = 1300*, THEN *Riser = Large*", the confidence of which is equal to 62.5%. Furthermore, the tree induction algorithms can generate rules for which no compatible learning example exists; the confidence of such a rule is indefinite and its support equal to zero.

It should be emphasized, that in all cases presented in Tables 2 and 3, the rules are in agreement with tentative expectations, based on industrial experience concerning design of feeding systems for grey cast iron castings.

Table 3.
Rules of 100% confidence for data in *Version 2* obtained from decision trees (DT) and rough sets theory (RST) after reduction of input variables

Rule No.	Method	Attribute (input variable)		Output class variable "Riser"	Rule support
		Modulus, cm	Pouring temperature, ^{0}C		
1	DT&RST	1.5	1200	Not required	15
2	DT&RST	3	1200	Not required	13
3	DT&RST	6	1200	Not required	13
4	**DT**	**0.75**	**1200**	**Small**	**6**
5	**RST**	**0.75**		**Small**	**6**
6	DT&RST		1400	Large	48
7	DT&RST		1500	Large	48
8	**RST**	**1.5**	**1300**	**Large**	**16**

4. Summary and conclusions

The analysis presented in the paper confirmed usefulness of the both considered methods used for engineering knowledge extraction from industrial data. However, the tests shown some advantages of the algorithms based on rough sets theory over decision trees. They were unable to provide some important and reliable rules, which were generated by the rough sets theory algorithm and they can also generate rules which are not supported by the training data

Further works should be aimed at systematic research on performance of those methods, including artificially generated training data, in which assumed, different relationships would be hidden.

References

[1] A. Kusiak, C. Kurasek, Data mining of Printed-Cicuit Board Defects, IEEE Transactions on Robotics and Automation, vol. 17, No. 2 (2001) 191–196.

[2] M. Perzyk, A. Soroczyński, R. Biernacki, Possibilities of decision trees applications for improvement of quality and economics of foundry production, Archives of Foundry Engineering, vol. 8, No. 1 (2008) 261-268.

[3] A. Holzmüller, R. Wlodawer, Zehn Jahre Speiser-Einguss-Verfahren fur Gusseisen, Giesserei, vol. 50, No. 25 (1963) 781–791.

[4] J. R. Quinlan, Simplifying decision trees, International Journal of Man-Machine Studies, vol. 27, No. 3 (1987) 221-234.