PRODUCTION ENGINEERING ARCHIVES

ISSN 2353-5156 (print)
ISSN 2353-7779 (online)

Exist since 4th quarter 2013
Available online at https://pea-journal.eu

Building decision trees based on production knowledge as support in decision-making process

Marcin Matuszny[1]

[1] University of Bielsko-Biala, ul. Willowa 2, 43-309 Bielsko-Biala, Poland
Corresponding author e-mail: marcinmatuszny@wp.pl

Article history
Received 10.03.2020
Accepted 15.04.2020
Available online 30.06.2020

Keywords
decision trees
knowledge identification
production engineering
production knowledge
production processes

Abstract

The article presents sources of production knowledge and thoroughly describes its identification which on the construction of decision trees, and on the construction of knowledge bases for production processes. The problems that arise during the technical preparation of production are briefly characterized and the advanced algorithm with which decision trees can be built is described in detail. A decision tree was built based on real data from the manufacturing company. Decision trees are presented as a method of knowledge representation.

DOI: 10.30657/pea.2020.26.08

JEL: L23, M11

1. Introduction

Decision-making processes have always posed a significant challenge for manufacturing companies, especially in the era of continuous and dynamic changes and ongoing automatization of production processes. Production processes, in their essence, can be optimized by appropriately and effectively tailored decision-making processes. Due to the fact that nowadays, manufacturing companies are facing more complex problems they have to rely on decision trees while making decisions. The construction of decision trees from data is a longstanding discipline. Statisticians attribute the paternity of regression trees to Sonquist and Morgan (since 1963 year) who used it in the process of prediction and explanation. (AID – Automatic Interaction Detection) (Hssina et al., 2014).

Authorities in the field, such as Gorski (Gorski et al., 2016) analyzed configurable products manufacturing processes and noticed that "a measure of flexibility of a manufacturing system is its capability of performing various tasks, as well as time at which it can be prepared for a new task (the shorter the better)". In order for a product configuration to be able to match ever-changing customer needs, it has to be flexible and capable of adapting such production processes that allow using intelligent techniques and supporting data analysis. Application of intelligent techniques has been considered by many authors, e.g. by Uhlmann et. al. (Uhlmann et al., 2017), who are experts in the field of intelligent production systems. Jedrzejewski in the Development of Machine Tool Operational Properties (Jedrzejewski and Kwasny, 2015) discuss intelligent function in diagnosing machine tools. Companies often struggle with the following problem today: how to use the experience and knowledge of current employees to a greater extent and how to analyze the data contained in both electronic and paper documents that are available in the company? The manufacturing process is constantly changing and companies have to think of new ways of maximizing the utilization of experience and knowledge of their employees and conducting proper analysis of the information available for the company and drawing valid conclusions from them. These issues are especially relevant for modern companies, because, on a daily basis, they have to deal with them. Products are offered in many variants produced/ designed to meet a customer's demands (Kutschenreiter-Praszkiewicz, 2018).

The concept of production knowledge, especially the one related to building knowledge bases, and the stages of its acquisition, identification or processing are very closely related to the technical preparation of production. An enterprise aiming to increase its own competitiveness against the background of the market, has to constantly adapt or modify its products to meet the demands of the target market, which, in turn, implies an urgent need to improve the activities that are undertaken as part of the technical preparation of production. Having that in mind manufacturing companies have to think of new ways they can implement to improve the activities in the production process.

One of the basic goals in this respect is the selection of the right technical variant, taking into account the previously adopted decision criteria including:

- production costs (whether fixed or variable) are often the biggest limitation because they have direct impact on the final choice of components,
- product quality,
- the complexity of the structure and its universality,
- dimensional accuracy, often defined by the customer in advance (Matuszny, 2019).

2. Identification of production knowledge

Production knowledge is a very important resource for manufacturing enterprises. Data gathered over the years allow to see changes resulting in streamlining production processes and adapting a given product or service to the requirements of the current market demands.

Knowing what to do is not enough, one has to be able to act using the knowledge they have, and clearly communicate what and how fellow employees need to proceed. Knowledge management can be treated as a set of activities that push processes occurring in the production company in the right direction. Knowledge management deals with locating, acquisition, developing, dissemination, use, and retaining knowledge (Kowalczyk and Nogalski, 2007).

Identification of production knowledge is based on a detailed analysis of the basic problems in the field of technical preparation of production. The aforementioned problems can be divided into two types depending on how they are solved:

- heuristic problems – those whose solution depends on the specific conditions associated with the production company and is based mainly on the experience of employees / designers, who act in this case as experts.
- algorithmic problems, i.e. problems that the algorithm of solutions is known (Paszek, 2011).

Identification of production knowledge regarding problem analysis, where the amount of input data available is not sufficient to be able to use proven algorithms leads to the development of new procedures. The decision-making stages and their problems are presented in the Table no 1 below.

The order of decision stages presented in the Table above is informative, it also results from the selection of appropriate manufacturing activities aimed at obtaining appropriate properties of the products manufactured, which are imposed by the customer. During each of the decision-making stages heuristic problems are determined which at a later stage affect the distribution of knowledge resources for the implementation of specific partial goals. Due to constructing the table as a workflow, it is possible to gradually identify the sets of knowledge required to solve a given problem.

Identification of production knowledge is also based on the use and analysis of relevant sources. Identification of knowledge sources initiated by the need to select appropriate product parameters for a given application can be based on:

- identification of product parameters important for the customer (analysis of requests for proposals),
- identification of product parameters that determine its functionality (analysis of device selection rules),
- identification of product parameters determining the time and costs of its manufacture (analysis of the manufacturing process) (Kustchenreiter-Praszkiewicz, 2012).

Table 1. Stages of knowledge identification in terms of designing production processes

Decision problems	Decision stages
Selection of blank	Identify the basic type of blank
	Selection of appropriate features of the blank
Selection of pre-treatment	Selection of input operations
	Preparation of machining bases
	Selection of initial heat treatment
Selection of basic machining	Determining the required machining methods
	Processing division (roughing / shaping)
	Selection of appropriate technological operations
	Selection of executive positions and proper instrumentation
Selection of finish peel	Selection of abrasive / final cut

Source: author's elaboration on the basis of (Paszek, 2011).

It should also be noted that the resources of production knowledge are most often collected on the basis of an analysis of the construction and technological documentation of machine elements, available in the company's resources and consultation with employees who support experts or act as experts in a given field.

Decision trees are indisputably one of the basic methods of inductive learning of systems. This is influenced, among other things, by their high effectiveness and efficiency, as well as the possibility of a relatively simple program implementation. Considering their easy reading (compared to slightly more complicated, for example: neural networks), for humans, they are also an important part of supporting decision-making processes during production processes.

The use of decision trees is based on the analysis of examples, which are described by a set of attributes, where each of the individual attributes can have a different value. Decision trees are most often represented as directed acyclic graphs, where the edges of the graph are branches, the vertices from which a minimum of one edge comes out are nodes, while the other vertices are called leaves. It is also worth mentioning that there is only one path between different vertices. A decision tree is a classifier which conducts recursive partition over the instance space (Dail and Ji, 2014).

The following are constructions of a decision tree based on a set of examples in the form of an algorithm developed on the basis of available literature~~,~~ which was included in the reference list (Cichosz, 2000).

The first step to start this algorithm is to decide whether the node in question should be

- a leaf ending a tree,
- branching - with a node.

The first selection ends the algorithm, while the second selects the attribute, where, successively, based on the values adopted by the argument, subsequent nodes are created in the same way. The degree of tree expansion depends on the order of attribute selection, which directly affects the quality of the decision tree being created - the smaller the tree are better solution.

The recursive calls of the algorithm should be terminated, the stop criterion corresponds to the appropriate moment of termination, which determines whether the node should be qualified as the final leaf of the tree. Another criterion regarding the selection of the attribute, which is the pillar of the entire algorithm and its correct operation, on the basis of which the set of examples in the node is divided, has a significant impact on the final appearance of the tree. The selection of the appropriate attribute from the available set is streamlined thanks to the introduction of the attribute rating system, which is based on the assumption that the least valuable attribute is the one whose frequency distribution of subsequent selection classes is identical before and after the division of the data set (Rojek, 2017).

Algorithms for constructing decision trees are among the most well-known and widely used of all machine learning methods. Among decision tree algorithms, J. Ross Quinlan's ID3 and its successor, C4.5, are probably the most popular in the machine learning community (Salzberg, 1994).

3. Research methods

C4.5 algorithm, an evolution of ID3, uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. Error-based pruning is performed after the growing phase. C4.5 algorithm can handle numeric attributes. It can also induce from a training set that incorporates missing values by using corrected gain ratio criteria as presented above (Rokach, Maimon, 2008).

Decision tree algorithms begin with a set of cases, or examples, and create a tree data structure that can be used to classify new cases. Each case is described by a set of attributes (or features) which can have numeric or symbolic values. Associated with each training case is a label representing the name of a class. Each internal node of a decision tree contains a test, the result of which is used to decide what branch to follow from that node. For example, a test might ask "is x > 5 for attribute x?" If the test is true, then the case will proceed down the left branch, and if not then it will follow the right branch. The leaf nodes contain class labels instead of tests. In classification mode, when a test case (which has no label) reaches a leaf node, C4.5 classifies it using the label stored there (Salzberg, 1994).

To create decision trees in production processes, a selection criterion based on information increment, based on the entropy change measure, can be used.

In order to calculate this difference it is necessary to:

- calculate entropy E (1) for a set of teaching examples x at the beginning:

$$E(x) = \sum_{i=1}^{k} p(i) \cdot log_2 \frac{1}{p(i)} = -\sum_{i=1}^{k} \frac{n_i}{n} \cdot log_2 \frac{n_i}{n} \quad (1)$$

Where:
p (i) - probability of the event i,
n_i - number of examples describing the i-th object,
n - number of all examples in the training set x.

- in turn select the attribute y, which divides the set of teaching examples x into subsets X1, X2, ..., Xv. Assuming that the subset of Xd contains md elements, then the relative entropy (2) will be:

$$E(y) = \sum_{d=1}^{v} \frac{m_d}{n} E(x_d) \quad (2)$$

- the last step is to choose the attribute for which the calculated information increase is the largest. The increase, of course, consists in calculating the difference between entropy E for the set of teaching examples x at the beginning and relative entropy, according to the formula (3):

$$\Delta E = E(x) - E(y) \quad (3)$$

The above algorithm, used to create decision trees, which in effect are a very important support in the decision-making process, due to the fact that it is not limited to binary divisions, creates a tree whose shape is more diverse. The above algorithm can be expanded with an information increment factor ϑ, which is relatively easy to calculate based on the formula (4):

$$\vartheta t(P) = \frac{\Delta E}{IVt(P)} \quad (4)$$

Where: *IVt(P)* denotes the informative value of the *t* test for the set of examples *P*, given by the formula (5):

$$IVt(P) = \sum_{r \in R_t} -\frac{|Ptr|}{|P|} log \frac{|Ptr|}{|P|} \quad (5)$$

This quantity measures the uniformity with which the t test divides the set of examples P into subsets.

3.1. Research results - the use of decision trees in decision support during production processes

An example of a decision tree is presented below (Fig. 1), using real data from a manufacturing company that deals, among other things, with machining on numerically controlled machines and conventional machines. The company for which the analysis was carried out, deals with the designing and manufacturing of elements, molding for shaping automotive cables, regeneration and fabrication of machine parts, welding of steel, aluminum, acid-resistant, stainless steel load-bearing structures, provides services in the field of locksmith works as well. The company, due to its size and current machinery park, performs the vast majority of its orders based on unit, small- or medium-series production.

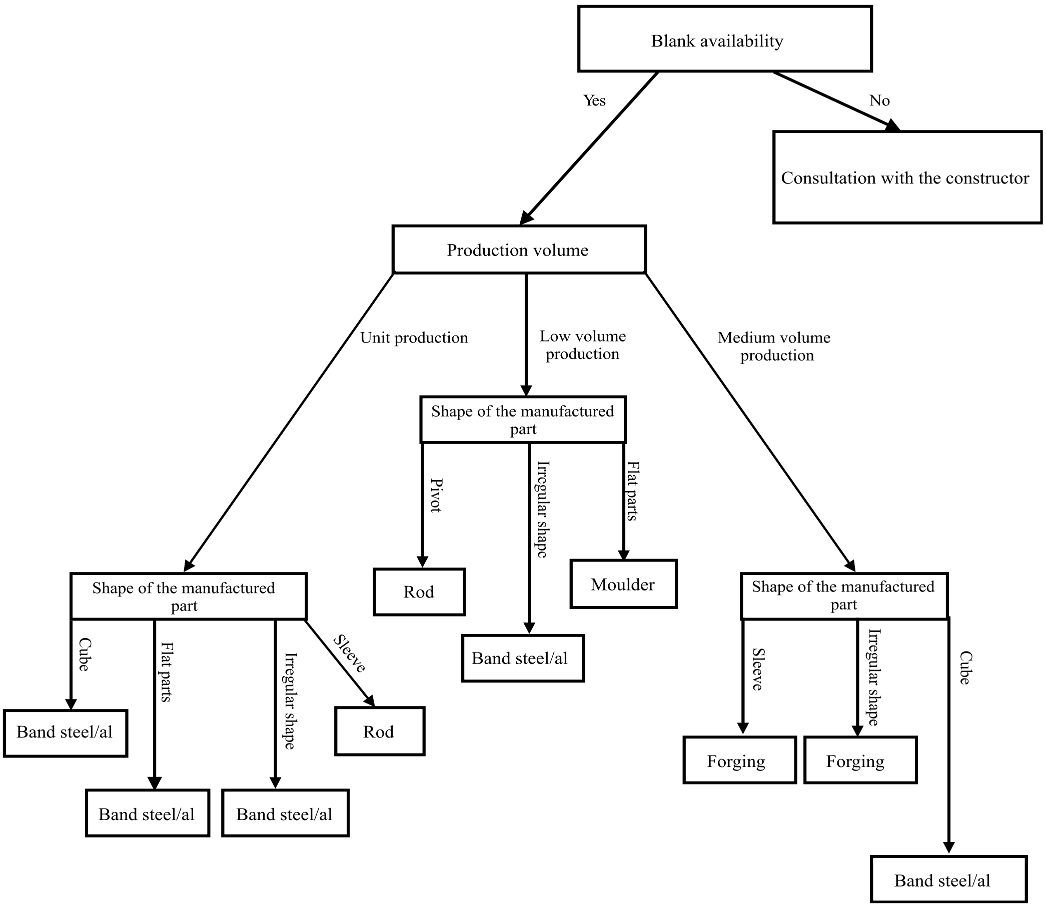

Fig. 1 Sample decision tree

The set of teaching examples prepared by the author (Tab. 2), which were used to generate the following tree, supporting decision making in the selection of the appropriate contained attributes regarding the production volume and the shape produced part.

Table 2. Training set

Type of blank	Blank availability	Production volume	Shape of the manufactured part
Band steel/al	Yes	Unit production	Cube
Rod	Yes	Low volume production	Pivot
Band steel/al	Yes	Unit production	Cube
Rod	Yes	Low volume production	Pivot
Moulder	Yes	Low volume production	Flat parts
Forging	Yes	Medium volume production	Sleeve
Rod	Yes	Unit production	Sleeve
Consultation with the constructor	No	Low volume production	Flat parts
Band steel/al	Yes	Low volume production	Flat parts
Band steel/al	Yes	Unit production	Irregular shape
Forging	Yes	Medium volume production	Irregular shape
Band steel/al	Yes	Low volume production	Irregular shape
Band steel/al	Yes	Medium volume production	Cube

After preparing the training set, it was appropriate to specify the attribute that has the largest increase in information. At this point, it is worth noting that the more uniform the probability distribution is, the greater its information will be. By building a decision tree according to this scheme, i.e. based on the value of the increase of information calculated for each attribute and placing in the root subtree this attribute among those that have not yet been placed in the tree, and which has the largest value of information increase, it is possible to obtain a small decision tree that will be able to classify new records accordingly in a few steps that will be both simple and human readable. In addition, in the case of unwanted tree growth, the C4.5 algorithm gives the possibility of trimming it, thereby increasing the generalization of the assessment of new cases. Pruning begins with the leaves and is analyzed from the bottom up.

To sum up, the next steps of the algorithm can be refined by reducing them in the order listed to:

- the first step is to choose the attribute that most differentiates the output values of attributes from the training set, which works well for applying the selection criterion based on information increment extended by the information increment coefficient, described in more detail in the previous part of the article, sample calculations are shown below for the production volume attribute:

$$E(x) = -\frac{3}{13} \cdot \left(log_2 \frac{3}{13}\right) - \frac{6}{13} \cdot \left(log_2 \frac{6}{13}\right) - \frac{2}{13} \cdot \left(log_2 \frac{2}{13}\right) - \frac{1}{13} \cdot \left(log_2 \frac{1}{13}\right) - \frac{1}{13} \cdot \left(log_2 \frac{1}{13}\right) = 1{,}987 \tag{6}$$

$$E\left(wlk_{prod}\right) = \frac{3}{13} \cdot \left(-\frac{1}{3} log_2 \frac{1}{3} - \frac{2}{3} log_2 \frac{2}{3}\right) + \frac{6}{13} \cdot \left(-\frac{2}{6} log_2 \frac{2}{6} - \frac{2}{6} log_2 \frac{2}{6} - \frac{1}{6} log_2 \frac{1}{6} - \frac{1}{6} log_2 \frac{1}{6}\right) + \frac{4}{13} \cdot \left(-\frac{1}{4} log_2 \frac{1}{4} - \frac{3}{4} log_2 \frac{3}{4}\right) = 1.347 \tag{7}$$

$$\varDelta E = E(x) - E(wlk\_prod) = 1{,}987 - 1{,}347 = 0.640 \tag{8}$$

$$IVwlk_{prod(P)} = -\frac{4}{13} log_2 \frac{4}{13} - \frac{6}{13} log_2 \frac{6}{13} - \frac{3}{13} log_2 \frac{3}{13} = \ 1.526 \tag{9}$$

$$\vartheta wlk\_prod(P) = \frac{0{,}640}{1{,}526} = 0.419 \tag{10}$$

For the attribute *availability of the blank* and the *shape of the manufactured part*, analogous calculations were made:

- in order to create branches for each value of the attribute selected in the first step,
- then divide the cases into subgroups in such a way that they reflect the values in the selected node of the decision tree,
- in the last step, the algorithm should be terminated if the subgroup contains a single node or all subgroups have the same value for the output attribute, while for a subgroup that has not been defined as a leaf, the process should be repeated.

4. Discussion of the results and conclusions

Decision support based on decision trees is a very good solution that can be implemented to manufacturing enterprises, especially larger ones operating on a larger amount of data obtained from production processes and using management systems with a greater degree of integration. Automated rule creation, based on examples, is the right solution for discovering knowledge that is the result of experience of technologists and constructors.

Decision trees are a popular and attractive method of knowledge representation, their indisputable advantage is the ability to represent a number of any hypotheses for individual sets of attributes, as well as the memory and time efficiency of classifying examples and the aforementioned human readability. The main practical problems in the application and induction of decision trees are large training sets and distortions. Missing attribute values are also a problem.

Knowledge acquired in the right way allows to solve decision problems in the area of technical preparation of the production of finished products that can be components of machines. Knowledge bases and systems based on them should effectively support the design of production processes related to knowledge processing, especially in the engineering industry.

Knowledge bases can be integrated with an in-house integrated management system. Considering the specifics of the operation of the aforementioned system, it is possible to associate the effects of its work with many positive changes for the production company, including:

- minimizing production costs,
- optimization towards process and product improvement,
- optimization towards improving the information flow in the company.

Reference

Cichosz, P., 2000. *Learning systems*, WNT, Warszawa.

Dai1, W., Ji, W., 2014. *A MapReduce Implementation of C4.5 Decision Tree Algorithm*, International Journal of Database Theory and Application, 7(1), 49-60.

Gorski, F., Zawadzki, P., Hamrol, A., 2016. *Knowledge based engineering as a condition of effective mass production of configurable products by design automation*, Journal of Machine Engineering, 16(4), 5-30

Hssina, B., Merbouha, A., Ezzikouri, H., Erritali M., 2014. *A comparative study of decision tree ID3 and C4.5 International Journal of Advanced Computer Science and Applications*, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 13-18.

Jedrzejewski, J., Kwasny, W., 2015. *Development of Machine Tool Operational Properties*, Journal of Machine Engineering, 15(1), 5-26

Kowalczyk, A., Nogalski, B., 2007. *Management of knowledge. Concept and tools.*, Print DIFIN, Warszawa.

Kutschenreiter-Praszkiewicz, I., 2012. *Application of knowledge based systems in technical production preparation of machine parts*, Wydawnictwo Naukowe Akademii Techniczno-Humanistycznej w Bielsku-Białej, Bielsko-Biała.

Kutschenreiter-Praszkiewicz, I., 2018. *Machine learning in SMED*, Journal of Machine Engineering, 18(2), 31-40.

Matuszny, M., 2019. *Proccesing and identification of production knowledge for knowledge base build for production processes* Technology, processes and production systems, 3, 115-125.

Paszek, A., 2011. *Construction of knowledge management system in a production company. Part II: Example* Enterprise Management, 1, 35-43.

Rokach, L., Maimon, O., 2008. *Data mining with decision trees*, Singapore, 69, 71-79

Rojek, I., 2017. *Expert system for selection of semi-finished products using the decision trees*, Studies & Proceedings of Polish Association for Knowledge Management, 83, 38-48.

Salzberg, S., 1994. C4.5: *Programs for Machine Learning, Machine Learning*, Kluwer Academic Publishers, 16, 235-240

Uhlmann, E., Hohwieler, E., Geisert, C., 2017. *Intelligent production systems in the era of industrie 4.0 – changing mindsets and business models*, Journal of Machine Engineering, 17(2), 5-24

基于生产知识构建决策树，为决策过程提供支持

關鍵詞
决策树
知识识别
生产工程
生产知识
生产过程

摘要

本文介绍了生产知识的来源，并在决策树的构建和生产过程的知识库的构建中全面描述了其识别。 简要描述了生产技术准备过程中出现的问题，并详细描述了可用于构建决策树的高级算法。 基于制造公司的真实数据构建了决策树。 决策树作为知识表示的一种方法提出