Data Mining Methods for Long-Term Forecasting of Market Demand for Industrial Goods

Bartłomiej Gaweł, Bogdan Rębiasz and Iwona Skalna

Abstract This paper proposes a new method for long-term forecasting of level and structure of market demand for industrial goods. The method employs *k*-means clustering and fuzzy decision trees to obtain the required forecast. The *k*-means clustering serves to separate groups of items with similar level and structure (pattern) of steel products consumption. Whereas, fuzzy decision tree is used to determine the dependencies between consumption patterns and predictors. The proposed method is verified using the extensive statistical material on the level and structure of steel products consumption in selected countries over the years 1960–2010.

Keywords Demand forecasting · Fuzzy decision tree · Clustering · Industrial goods · Data mining

1 Introduction

Forecasting of market demand is one of the most important part of tangible investment appraisal [1]. This paper presents a new method for long-term forecasting. It is based on the concept of analog forecasting, which relies on forecasting the behavior of a given variable by using information about the behavior of another variable whose changes over time are similar, but not simultaneous. The proposed forecasting method combines k-means clustering and fuzzy decision trees into a single framework that operates on historical data. It can be primarily used for

B. Gaweł (✉) · B. Rębiasz · I. Skalna
AGH University of Science and Technology, Kraków, Poland
e-mail: bgawel@zarz.agh.edu.pl

B. Rębiasz
e-mail: brebiasz@zarz.agh.edu.pl

I. Skalna
e-mail: iskalna@zarz.agh.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_1

long-term forecasting of the level and structure (pattern) of demand for products that are traded on industrial markets, i.e., steel industry products, non-ferrus metals industry products, certain chemical industry products, casts, construction materials industry products, energy carriers.

The rest of the paper is organized as follows. In Sect. 2 methods for forecasting apparent consumption of steel products are presented. Section 3 describes methods for building classification models and data clustering. The proposed long-term forecasting method is introduced in Sect. 4. Section 5 discusses industrial application of the proposed method. A comparison of the results with other methods is presented in Sect. 6. The paper ends with concluding remarks and directions for future research.

2 Methods for Forecasting Apparent Consumption of Steel Products

The following methods are used for forecasting apparent consumption of steel products: econometric models, sectorial analysis, trend estimation models, analog methods. In econometric models, the level of apparent consumption of steel products is a function of selected macroeconomic parameters. Typically, gross domestic product (GDP) (see, e.g. [2]), GDP composition, the value of investment outlays, and the level of industrial production are used as exogenous variables [3–9], whereas GDP steel intensity is used as an endogenous variable. Apparent consumption of steel products can also be forecasted by using trend estimation models [10]. However, thus produced forecasts are usually short-term. Analog methods for forecasting apparent consumption of steel products usually assume that indicators characterizing the level and structure of apparent consumption of steel products in a country for which the forecast is drawn up tend to attain the indicators characterizing countries that serve as a comparator. The indicators that are most often compared include consumption of steel products per capita, GDP steel intensity, and the assortment structure of apparent consumption [6].

3 Building Classification Models and Data Clustering

Various data mining methods can be used to build classification models. Among statistical data mining methods, the linear discriminant function method has been widely used to solve practical problems [11]. However, the effectiveness of this method deteriorates when the dependencies between forecasted values and exogenous variables are very complex and/or non-linear [11], which is often the case in practice. Machine learning methods [12–15], such as Bayesian networks, genetic algorithms [16], algorithms for generating decision trees and neural networks are more appropriate in such situations, the two latter being used the most frequently. The strength of neural networks lies in their ability to generalize information contained in analyzed data sets. Decision trees, however, are advantageous over neural networks in the ease of interpretation of obtained results. Rules generated from decision trees, which assign objects to respective classes, are easy to interpret even for users unfamiliar with problems of data mining [17, 18]. This feature is very important, because decision-makers in the industry prefer tools understandable for all participants in a decision-making process and provide easily interpretable results. Nowadays the fuzzy version of decision trees is used increasingly often. This is because available information is often burdened with uncertainty, which is difficult to be captured by classical approach.

Clustering is a data mining method for determining groups (clusters) of objects so that objects in the same group are more similar (with respect to selected attributes) to each other than to objects from other groups. The most popular clustering methods are hierarchical clustering and k-means clustering.

Taking into account the above considerations, from among a very large number of available data mining methods, the k-means and fuzzy version of Iterative Dichotomizer 3 (ID3) were used to develop a method for long-term forecasting of the level and structure of apparent consumption of steel products. The k-means method was used to create consumption patterns, whereas the fuzzy ID3 (FID3) algorithm was used to build a decision tree that assigns specific consumption patterns to predictors.

3.1 The k-Means Clustering

For the sake of completeness, below are reminded the steps of the k-means clustering:

1. Specify k, the number of clusters.
2. Select k items randomly, arbitrarily or using a different criterion. The values of the attributes of chosen items define the centroids of clusters.
3. Calculate the distance between each item and the defined centroids.
4. Separate items into k clusters based on the distances calculated in the third step—assign items to clusters to which they are closest.
5. Determine the centroids of the newly formed clusters.
6. If the stopping criterion is met, end the algorithm; otherwise go to Step 3.

Subsequent iterations are characterized by squared errors function (SES) defined by the formula $SES = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{jS_i}^2$, where $d_{jS_i}^2$ is the distance between the j-th item and the centroid of the i-th cluster, n_i is the number of items in the i-th cluster. Once the values of SES in subsequent iterations do not show significant changes (changes are less than the required value) the procedure terminate.

3.2 Fuzzy Decision Trees

The proposed forecasting method uses the so-called Fuzzy Iterative Dichotomizer 3 (FID3) which is a generalization of the classical ID3 algorithm. The FID3 algorithm extends the ID3 algorithm so that it can be applied to a set of data with fuzzy attributes. The generalization relies on that FID3 algorithm computes the gain using membership functions [19] instead of crisp values.

Assume that D is a set of items with attributes $A_1, A_2, \ldots, A_p$ and each item is assigned to one class $C_k \in \{C_1, C_2, \ldots, C_n\}$. Additionally assume that each attribute A_i can take l_i fuzzy values $\tilde{F}_{i1}, \tilde{F}_{i2}, \ldots, \tilde{F}_{il_i}$. Let then D^{C_k} be a fuzzy subset in D whose class is C_k and let $|D|$ be the sum of the membership values in a fuzzy set of data D. Then, the algorithm for generating a fuzzy decision tree is the following [20]:

1. Generate the root node that contains all data, i.e., a fuzzy set of all data with the membership value 1.
2. If a node t with a fuzzy set of data D satisfies the following conditions:
   - the proportion of a data set of a class C_k is greater than or equal to a threshold θ_r, that is $|D^{C_k}|/|D| \geq \theta_r$;
   - the number of data set is less than a threshold θ_n that is $|D| \leq \theta_n$;
   - there are no attributes for more classification, then the node t is a leaf and assigned by the class name.
3. If the node t does not satisfy the above conditions, it is not a leaf and the test node is generated as follows:
   (a) For each A_i (i = 1, 2,…, p), calculate the information gains $G(A_i, D)$ (see below) and select the test attribute A_{max} with the maximal gain.
   (b) Divide D into fuzzy subsets D_1, D_2,…, D_l according to A_{max}, where the membership value of the data in D_j is the product of the membership value in D and the value of $\tilde{F}_{\max,j}$ of the value of A_{max}, in D.
   (c) Generate new nodes t_1, t_2, …, t_l for fuzzy subsets D_1, D_2, …, D_l, and label the fuzzy sets $\tilde{F}_{\max,j}$ to edges that connect between the nodes t_j and t.
   (d) Replace D by D_j (j = 1,2,…, l) and repeat recursively starting from 2.

The information gain G(A_i, D) for the attribute A_i is defined by

$$G(A_i, D) = I(D) - E(A_i, D), \tag{1}$$

where:

$$I(D) = -\sum_{k=1}^{n} \left(\left|D^{C_k}\right|/D\right) \log\left(\left|D^{C_k}\right|/D\right) \tag{2}$$

$$E(A_i, D) = \left(\left| D_{F_{ij}} \right| / \sum_{j=1}^{m} \left(D_{F_{ij}} \right) I\left(D_{F_{ij}} \right) \right) \quad (3)$$

As for assigning the class name to the leaf node the following methods are proposed:

1. The node is assigned by the class name that has the greatest membership value, that is, other than the selected data are ignored.
2. If the condition (a) in Step 2 in the algorithm holds, do the same as the method (1). If not, the node is considered to be empty, i.e., the data are ignored.
3. The node is assigned by all class names with their membership values, that is, all data are taken into account.

4 The Proposed Long-Term Forecasting Method

The overview of forecasting methods presented in Sect. 2 shows that the structure and level of apparent consumption of steel products or GDP steel intensity depends on selected macroeconomic parameters characterizing the economy of a country. The value of GDP per capita and the sectorial composition of the GDP are most often used as exogenous variables. The sectorial composition is characterized by the contribution of industry and construction in GDP. Additionally, in the case of industry, the share of sectors determining the level of steel consumption (the so-called steel intensity industries) in the industry is significant for the total GDP. Such sectors include manufacture of finished metal products excluding machinery and equipment, manufacture of electrical equipment, manufacture of machinery and equipment not classified elsewhere, manufacture of motor vehicles and trailers, excluding motorcycles and the manufacture of other transport equipment.

The proposed forecasting method (see Fig. 1) is used to predict the following (endogenous) model variables:

- GDP steel intensity (S_t),
- share of various ranges of products in consumption
  - share of long products (U_d),
  - share of flat products (U_p),
  - share of pipes and hollow sections (U_r),
  - relation of the consumption of organic coated sheets to the consumption of metallurgic products altogether (U_o).

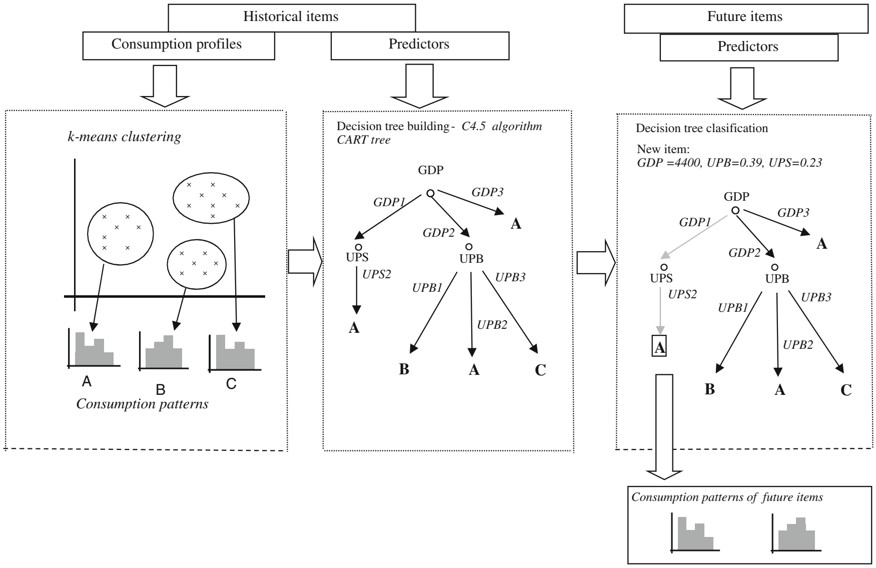

Fig. 1 The proposed forecasting procedure

The exogenous variables in the model (predictors) are:

- value of GDP per capita (GDP),
- share of industrial production and construction in GDP (UPB),
- share of steel intensity industries in making up gross value of industry (UPS).

Endogenous variables create a consumption profile for each country in a particular year. Besides the consumption profile, the historical data include a summary of values of predictors (exogenous variables) for each country and year. Historical data covering consumption profiles and values of predictors are collected from different countries and different periods. For certain countries, only values of predictors were available in the forecasted period. Consumption profiles are forecasted on the basis of these values.

The overall forecasting procedure proposed in this paper is the following. First, consumption patterns are created using the k-means method. A consumption pattern is understood as a GDP steel intensity and a structure of consumption expressed by the share of individual ranges of steel products in a total consumption. Such patterns are defined using data on consumption profiles in different countries over many years. The centroids of clusters identified using the k-means method form the consumption patterns. After defining clusters and their centroids, a fuzzy decision tree is built using historical data. The obtained fuzzy decision tree distinguishes easily interpretable links between predictors and pre-defined consumption patterns. This provides the possibility for forecasting the level and structure of steel products consumption based on the forecasted value of GDP and GDP composition.

5 Industry Application

Historical data used to generate the fuzzy decision tree was derived from the following countries: Austria, Belgium, Denmark, Finland, France, Spain, Holland, Japan, Lithuania, Norway, Czech Republic, Russia, Slovakia, Slovenia, Sweden, United Kingdom, Hungary, Italy and the USA. Data from Japan, the USA and Western European countries are collected over the period 1960–2010, and data for the remaining countries came from the period 1993–2010. It was not possible to gather data on the structure of consumption for all counties in all years of the periods indicated above. Taking into account the gaps in the data, the total number of collected items was 730. The data was divided randomly into two sets: a set of 657 items, which served to build a classifier and a set of 73 items, which were used to test the classifier and compare it with another method. Data on GDP for each country was expressed in dollars according to prices from 2007.

The values of GDP steel intensity and demand structure defined by consumption patterns are used to determine the forecast in the proposed method. These patterns are defined on the basis of historical data describing the consumption profiles of various countries and in various periods. During the time between the period from which comes the data characterizing the consumption profiles and the period for which the forecast is made, there are production technology changes and structural changes of the products in sectors utilizing steel products. Changes also occur in the parameters of steel products in terms of their durability characteristics, technological characteristics. These changes affect the reduction of sectorial indicators of steel intensity. This in turn results in a reduction of GDP steel intensity not directly associated with changes in the GDP and its sectorial composition. Therefore, steel intensity was adjusted in the profiles of the individual countries in various years.

In the first step, 9 classes of patterns of the consumption were obtained from the clustering of steel intensity and demand structure (see Table 1).

To build FID3 the fuzzification was carried out for all independent variables. The fuzzification relied on the division of each attribute on equinumerous bins. The membership to the classes is described by fuzzy numbers. In this case, the

Table 1 Centers of consumption patterns

Cluster	G1	G2	G3	G4	G5	G6	G7	G8	G9
No of items	**118**	**61**	**73**	**56**	**85**	**105**	**76**	**90**	66
S_t (kg/K USD)	15.70	35.83	27.63	47.69	23.37	24.41	23.20	11.63	35.66
U_d	0.39	0.42	0.44	0.44	0.45	0.44	0.47	0.38	0.50
U_p	0.50	0.48	0.48	0.45	0.46	0.47	0.43	0.53	0.40
U_r	0.10	0.11	0.08	0.12	0.09	0.09	0.10	0.09	0.10
U_o	0.20	0.13	0.15	0.12	0.09	0.19	0.16	0.24	0.17

fuzzification was performed using the following algorithm. The attribute A_i is a sequence of J + 1 observation $A_i = \{x_j\}_{j=0,1,\ldots,J}$ such that $x_{j+1} \geq x_j$. Observations are divided into l subsets, where l = 5 + 3.3log(J + 1) (here l = 11). For each subset, a trapezoidal fuzzy number (a, b, c, d) was determined, where values *a*, *b*, *c*, *d* are designated in the following way:

$k = 1$	$\tilde{F}_k = (g_k, g_k, g_{k+q}, (1+m)g_{k+1})$
$1 < k < l+1$	$\tilde{F}_k = ((1-m)g_k, (1-m), g_{k+1}, g_{k+1}(1+m))$
$k = l+1$	$\tilde{F}_k = ((1-m)g_k, g_k, g_{k+1}, g_{k+1})$

where $g_k = x_{j_k}, j_k = (k-1)\left\lfloor \frac{J+1}{l} \right\rfloor, k = 1,\ldots,l+1$ and $m \in [0,1]$.

To improve readability of result, each fuzzy set in attribute A_i was labelled. Each label consists of name of attribute and value of k e.g. GDP1 means fuzzy set of GDP where k = 1. The following values of m was accepted for attributes: $m_{GDP} = 0.05$, $m_{UPB} = 0.01$, $m_{UPS} = 0.04$, A decision tree is built using the FID3 algorithm described in Sect. 3.2, where the stopping criterions are based on the following thresholds: $\theta_r = 3\,\%$, $\theta_n = 90\,\%$.

The final decision tree consists of 216 leaves and 100 nodes. The GDP was important attribute, and UPS attribute is the most rarely tested. This means that it is the least important in determining the level and structure of steel products consumption. Sectorial composition of the GDP is more relevant in this case.

Inference in an ordinary decision tree is executed by starting from the root node and repeating to test the attribute at the node and branch to an edge by its value until reaching at a leaf node, a class attached to the leaf being as the result. The difference between the ordinary and fuzzy tree relies on this that the given case is not credited only to one branch, but to many with some degree of the membership.

On the basis exemplary case (see Table 2) will be elaborated the forecast. For the simplicity of presentation, below are shown only non-zero values of the membership function.

To elaborate the forecast three operations must be executed. First for every branch one ought to designate the indicator G—this is the value of membership function with which the attribute of the case for which we elaborate the forecast satisfies the rule by which the branch is based. On Fig. 2 is indicated the fragment of the tree, where values of indicators F for individual branches are non-zero. Values of indicators F are found on the grey background close to the branch.

In the second step, the value of the membership function of the analyzed case to individual clusters is designated. Suitable calculations are shown below.

Table 2 The exemplary-case for prediction

GDP		UPB	UPS	
GDP7	*GDP8*	*UPB2*	*UPS7*	*UPS8*
0.22	0.78	1.00	0.54	0.46

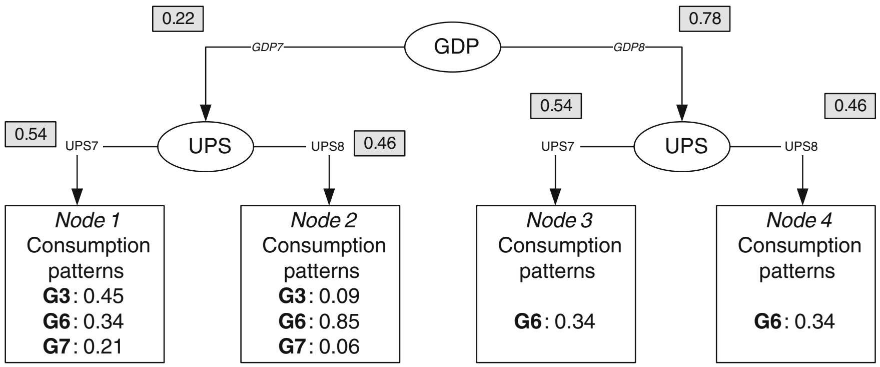

Fig. 2 Part of the tree for exemplary case

Table 3 The forecast for the analyzed case

	G3	G6	G7	Forecast
St, kg/K USD	27.62	24.41	23.20	**24.55**
Ud	0.44	0.44	0.47	**0.44**
Up	0.48	0.46	0.43	**0.46**
Ur	0.08	0.09	0.10	**0.09**
Uo	0.15	0.20	0.16	**0.19**

$$0.22 \cdot \left(0.54 \cdot \begin{bmatrix} 0.45 \\ 0.35 \\ 0.20 \end{bmatrix} + 0.46 \cdot \begin{bmatrix} 0.09 \\ 0.85 \\ 0.06 \end{bmatrix}\right) + 0.78 \cdot \left(0.54 \cdot \begin{bmatrix} 0.00 \\ 0.34 \\ 0.00 \end{bmatrix} + 0.46 \cdot \begin{bmatrix} 0.00 \\ 0.34 \\ 0.00 \end{bmatrix}\right)$$
$$= \begin{array}{cl} \begin{bmatrix} 0.09 \\ 0.90 \\ 0.04 \end{bmatrix} & \begin{matrix} G3 \\ G6 \\ G7 \end{matrix} \end{array}$$

In the third step, on the basis of the computed value of the membership function the forecast for the analyzed case is designated (Table 3).

6 A Comparison of the Results with a Classical Method

It is difficult to compare the proposed method with conventional econometric models, since a vector characterizing the level and structure of consumption is forecasted. The vector components must meet certain conditions. The sum of the forecasted shares of the three main product groups (long products, flat products, pipes) must be 1. In order to compare the effectiveness of the proposed fuzzy method, additional classifiers were also tested based on a traditional (crisp) method which uses Gini coefficient to build the tree.

Table 4 Comparison of the *mean absolute percentage error* between the 2 tested models

	Mean absolute percentage
The proposed method (fuzzy decision tree)	0.028
The *CART* crisp method	0.076

A comparison of the quality of the forecasts was made using 67 selected items. The quality of the forecasts was rated by calculating MAPE. The results of the tests are presented in Table 4.

The data in Table 3 indicate that the proposed fuzzy algorithm provides more accurate forecasts of the level and structure of consumption. The method provides smaller value of the MAPE. In this case, the value of the MAPE constitutes 36.8 % values of the error in the crisp method using the Gini coefficient to select attributes based on which the training set is divided.

7 Final Remarks

In the last decade, data mining methods have become very popular tools for supporting decision-making processes. The forecasting method presented in this paper proved to be effective for the analyzed example. The presented concept provided good results. This allows to recommend the proposed method to be used for long-term forecasting of demand for selected products traded on the industrial market. The method can be classified into the group of analog methods. In such methods a forecast is formulated on the basis of comparator. Most comparators are defined by experts. The proposed method allows for objectifying the selection of comparators by making the selection dependent on the values of chosen predictors. The method makes it possible to forecast the level and structure of demand. The applied means of determining the centroids of clusters in the k-means method allows for correctly forecasting the structure of consumption.

The proposed method was compared with other methods for building decision trees. The conducted test showed that the smaller mean absolute percentage error was obtained using fuzzy ID3 algorithm for building decision trees in comparison with the crisp method.

References

1. Rebiasz, B., Gawel, B., Skalna, I.: Hybrid framework for investment project portfolio selection. In: Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on, pp. 1117–1122 (2014)
2. Malenbaum, W.: World demand for raw materials in 1985 and 2000. McGrow-Hill, New York (1975)

3. Evans, M.: Modeling steel demand in the UK. Ironmaking Steelmaking **1**, 19–23 (1996)
4. Ghosh, S.: Steel consumption and economic growth: evidence from India. Resour. Policy **1**, 7–11 (2006)
5. Labson, B.S.: Changing patterns of trade in the world iron ore and steel market: an econometric analysis. J. Policy Model. **3**, 237–251 (1997)
6. Rebiasz, B., Garbarz, B., Szulc, W.: The influence of dynamics and structure of Polish economic development on home consumption of steel products. Metallurgist-Metall. News **9**, 454–458 (2004)
7. Roberts, M.C.: Predicting metal consumption: the case of US steel. Resour. Policy **1**, 56–73 (1990)
8. Roberts, M.C.: Metal use and the world economy. Resour. Policy **3**, 183–196 (1996)
9. Rebiasz, B.: Polish steel consumption 1974–2008. Resour. Policy **1**, 37–49 (2006)
10. Hougardy, H.P.: Zukünftige Stahlentwicklung. Stahl Eisen **3**, 85–89 (1999)
11. Altman, E.I., Macro, G., Varetto, F.: Corporate distress diagnosis: comparison using linear discriminant analysis and neural networks. J Bank Financ **18**, 505–529 (1994)
12. Andres, J., Landajo, M., Lorca, P.: Forecasting business profitability by using classification technique: a comparative analysis based on Spanish case. Eur. J. Oper. Res. **2**, 518–542 (2005)
13. Chen, Z.: Data Mining and Uncertain Reasoning—An Integrated Approach. Wiley, New York (2001)
14. Rastogi, R., Shim, K: A decision tree classifier that integrate building and pruning. Proceedings of the 24th International Conference on Very Large Databases (VLDB'98), pp. 405-415, (1998)
15. Tsujimo, K., Nishida, S.: Implementation and refinement of decision tree using neural network for hybrid knowledge acquisition. Artif. Intell. Eng. **9**, 265–275 (1995)
16. Thomassey, S., Fiordaliso, A.: A hybrid forecasting system based on clustering and decision trees. Decis. Support Syst. **42**, 408–421 (2006)
17. Hansen, N.K., Salamon, P.: Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 993–1003 (1990)
18. Zhou, Z.H., Jiang, Y.: NeC4.5: neural ensemble based C4.5. IEEE Trans. Knowl. Data Eng. **16**, 770–773 (2004)
19. Umano M.: Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems. Fuzzy Systems, 1994. In: Proceedings of the Third IEEE Conference on IEEE World Congress on Computational Intelligence, vol. 3, pp. 2113–2118 (1994)
20. Crompton, P.: Future trends in Japanese steel consumption. Resour. Policy **2**, 103–114 (2000)