Analysis of drilling conditions by a catalog mining method based on Fuzzy c-means algorithm

Claire-Amelie JANOT*, Hirokazu HIGUCHI*, Kotaro YAMADA*, Akihito ASAKURA*, Toshiki HIROGAKI* and Eiichi AOYAMA*

* Department of Mechanical Engineering, Doshisha University
1-3, Tataramiyakodani, Kyotanabe, Kyoto 610-0321, Japan
E-mail: thirogak@mail.doshisha.ac.jp

Received: 17 February 2020; Revised: 24 June 2020; Accepted: 21 July 2020

Abstract
To support the engineer in the selection of drills and their machining parameters, we propose a novel catalog mining system based on data mining techniques applied on drill catalogs by using twice the Fuzzy c-means method along with the Maximum Information Coefficient (MIC). We first perform the clustering algorithm Fuzzy c-means that returns the membership degree of every point to each cluster on the parameters defining a series of tools. We then use the maximum information coefficient (or MIC, index measuring the correlation between two parameters) to find the drill or material properties that have the most influence on the drilling conditions in each of the clusters in order to realize a second clustering that includes the drilling conditions. The Davies-Bouldin index (DBI), which evaluates the dispersion inside the clusters, is used to assess the result of the second clustering and find its optimal parameters. Finally, a multi linear regression is used to find the equations predicting the drilling conditions in each sub-cluster. The mean squared error indicator is used to validate the result of the prediction. A new flexible index based on the membership degree value computed by the Fuzzy c-Means algorithm is proposed to filter the points and clarify the borders of the clusters in order to optimize the data used in the regression.

Keywords: Data mining, Fuzzy c-Means, Maximum information coefficient, Davies-Bouldin index, Tool catalog, Machining conditions, Optimal parameters, Drill, Computer aided manufacturing

1. Introduction
Nowadays, the trajectories, rotations and moving speeds of the tools are controlled by a numerical control (NC) program implemented in the machine tools. The trajectories in the NC program are automatically generated by a computer aided manufacturing (CAM) software. Still, the selection of the tools and their rotations and moving conditions is still left to the programmer who must input these parameters into the CAM software. Thus, many unexperienced engineers refer to tool catalogs to select the tools and their advised parameters.

Tools catalogs contain very large knowledge on tools and drilling parameters and can then be used to choose the conditions to use for a determined tool. Experimentered engineers of the tool makers perform numerous cutting tests on various kinds of tools in a wide variety of conditions to determine their optimal settings of use and indicate them in their catalogs. They can then advise their users on the values to fix for each parameter depending on the desired hole and on the material used. A catalog can contain thousands of entries, with a broad range of options for each tool. Thus, it is difficult for the unexperienced engineer to extract the desired information from the tremendous quantity of knowledge presented in the catalog. New techniques need to be used to help dealing with such big databases.
It is now common for large databases such as tool catalogs to be processed using data mining techniques. Data mining is an analytical method used to gain useful insights from a mass of data. Figure 1 shows an example of how to find valuable information in the database based on a data mining process: from a large data base, some features of the data are extracted and processed in order to conserve only the useful and correct data. From the analysis of this first data, feedbacks can be used to clean or modify the original data, or to complete the database. The output is usually given after several feedbacks to ensure the pertinence and accuracy of the data.

Fig. 1 The data mining process.

Data mining techniques are widely used in a large range of research fields, such as marketing, medicine, biology or social sciences. Many manufacturing questions now necessitate the use of large databases as sources of knowledge to improve processes. Thus, data mining techniques, such as clustering, classification, prediction, description, characterization or evolution analysis, are increasingly used to extract valuable information. Choudhary et al. (2009) and Shahbaz et al. (2010) reviewed the different domains where data mining has become essential for manufacturing.

Data mining is now helping to resolve many issues in manufacturing domains. By analyzing the information from the data collected for every realization with data mining methods, the performance and quality of the whole process has been consequently enhanced. When used in CAM systems, data mining can increase the automation and consequently limit the need for human intervention in the generation of NC programs, as shown by Huang et al. (2015) who proposed a system based on the reuse of similar machining processes to facilitate CAM programming.

Tool catalogs are also very rich databases that contain machining knowledge accumulated by tool manufacturers over the years. Their sizes make them difficult to be directly handled and treated correctly by the tool users, inducing a loss in the transmission of knowledge. Data mining techniques can help properly extract the information and to realize better machining operations. Therefore, we proposed a novel catalog mining method based on the k-Means algorithm to support engineers in the selection of tools and machining conditions for end-mill processes (Hirogaki, et al., 2011). The evaluation index MIC (maximal information coefficient) was also found effective as hierarchical clustering to mine tool catalog databases of radius end-mill tools (Sakuma, et al., 2019). However, a problem emerged that it is difficult to perform a data mining that would fit different kinds of machining tools (such as drilling, boring and tapping, etc.) and combine numerical and categorical data. The determination of the clusters number is also a challenge for the operator, as it is limited to discrete values with the k-Means method that returns binary results for the belonging of a point to a cluster.

In the present report, we propose an improved data mining method for tool catalog databases based on the clustering method Fuzzy c-Means algorithm, which gives more options for the choice of cluster number. The novel catalog mining system combines twice the Fuzzy c-Means method and the MIC index, to combine numerical and categorical data and resolve the issue of the number of clusters to consider. To complete the knowledge in machining processes from the previous studies, a new type of tool catalogs is used, with solid drill catalogs. As a result, this application of Fuzzy c-Means is found to be effective to give a prediction of drilling conditions, with a better filtering of the noisy data (incomplete or inaccurate) and selection of the number of clusters, compared to the traditional k-Means method. Practical formulas of the spindle speed and the feed rate can be derived from the drill catalog data with the novel system.

2. Data mining methods

2.1 Representative-based clustering algorithm

Representative-based algorithms are the simplest clustering algorithms, as they only rely on visible notion of distance between the data points to form the clusters and associate a representative to them. The representatives of the clusters are points either directly taken from the database or computed by a function of the data points; the geometrical center of a
cluster (i.e., the mean of the coordinates of all the points) is often used. Determining the representatives of the clusters is
the main point of this method: once the representatives are fixed, data points are assigned to the closest cluster by a
distance-based function. As explained in Aggarwal (2015), the final goal of a representative-based algorithm is to
minimize the sum of the distances between each data point and the representative of the cluster it is assigned to. The two
main representative-based methods are the \(k\)-Means algorithm and the Fuzzy \(c\)-Means algorithm. This research focuses
on the Fuzzy \(c\)-Means method.

2.2 Fuzzy \(c\)-Means

The Fuzzy \(c\)-Means (FCM) method is a representative-based clustering process proposed by Bezdeck (1981) to
recognize patterns in a large database. It uses the distance between the points to form \(c\) clusters in the database: the points
are grouped with the closest representative which is computed by a function of all the data points. The Euclidian distance
presented in Eq. (1) is the most common distance used.

\[
D(x, y) = \|x - y\|^2 = \sum_{i=1}^{d}(x_i - y_i)^2
\]

(1)

where \(x = \{x_1, x_2, \ldots, x_d\}\) and \(y = \{y_1, y_2, \ldots, y_d\}\) are two data points in a \(d\)-dimensional space.

FCM retains more information than usual clustering algorithm, such as \(k\)-Means, as it introduces the fuzziness of the
belonging of a point to a cluster: instead of having a binary result for the belonging of every point in the clusters, FCM
returns the membership degree matrix \(U = \{u_{ki}\}_{1 \leq k \leq c, 1 \leq i \leq n}\) which contains all the membership degree values \(u_{ki}\) of a point
\(i\) to a cluster \(k\). This membership degree value is defined by Eq. (2). It can be seen in this equation that this value will
tend to \(1/c\) when a point is far from all the clusters.

\[
u_{ki} = \left(\frac{\sum_{l=1}^{c} D(x_l, v_k)^{-m}}{\sum_{l=1}^{c} D(x_l, v_l)^{-m}}\right)^{-\frac{1}{m-1}}
\]

(2)

with \(\sum_{k=1}^{c} u_{ki} = 1\), \(u_{ki} \in [0, 1]\), \(0 \leq \sum_{i=1}^{n} u_{ki} \leq n\)

Where \(m\) is the parameter determining the amount of fuzziness in the result, \(x_i = \{x_{i1}, \ldots, x_{id}\}\) the \(i\)th \(d\)-dimensional
point of the database and \(D(x,y) = \|x-y\|^2\) is the Euclidian distance between two points. The representative
\(v_k = \{v_{k1}, \ldots, v_{kd}\}\) of the \(k\)th cluster is given by Eq. (3). The final goal of Fuzzy C-Means is to minimize the objective
function presented by Eq. (4).

\[
v_{k,i} = \frac{\sum_{i=1}^{n} u_{ki}^{m} x_{i,j}}{\sum_{i=1}^{n} u_{ki}^{m}}
\]

(3) \hspace{2cm} \(J_m = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ki}^{m} D(x_l, v_k)\)

(4)

A point \(x_i\) can be associated to the cluster corresponding to its maximal value of \(u_{ki}\) defined by Eq. (5), that will be
referred as \(u_{\max}\) in this study. A point with a \(u_{\max}\) value equal to 1 is a cluster center: this value allows to evaluate how
close a point is to a cluster center.

\[
u_{\max}(x_i) = u_{\max,i} = \max_k (u_{ki})
\]

(5)

For the present study, the \textit{fcm} function of Matlab was used with default parameters. This function realizes the FCM
clustering by using two to five entered parameters and returns three objects: \textit{fcm}(\textit{data}, \textit{c}, \textit{m}, \textit{T, e}) = \{\textit{centers}, \textit{U}, \textit{objFunc}\},
with \textit{data} a \(n\) by \(d\) matrix containing the coordinates of the \(N\) \(d\)-dimension points of the database, \(c\) the number of clusters,
m the level fuzziness \((m = 2\) by default), \(T\) the maximum number of iterations \((T = 100\) by default), \(\varepsilon\) the minimum improvement in objective function between two consecutive iterations \((\varepsilon = 10^{-5}\) by default), \(\text{centers}\) a \(c\) by \(d\) matrix containing the coordinates of the final cluster centers, \(U\) a \(n\) by \(c\) matrix giving the membership degree values to each of the \(c\) clusters of each of the \(n\) points, and \(\text{objFunc}\) a column vector containing the value of the objective function at each iteration of the algorithm.

The FCM method is now used in various domains. Its capacities of treating incomplete or noisy data and approximating the clustering result make it particularly used for the treatment of images. Many modifications to the original version of Bezdeck et al. (1984) have been now proposed to reduce its computing time. Zhao et al. (2014) showed FCM and several of its variations to be efficient for the segmentation of heavily noise-corrupted images. FCM also has applications in physical systems observation, diagnosis and prediction. Wang et al. (2015) successfully implemented a diagnosis system based on the original FCM algorithm to automatically identify failures of gas turbines through the measurement of exhaust gas temperature.

Gosh and Dubey (2013) compared the efficiency of the FCM algorithm presented by Bezdeck et al. (1984) with the most common clustering method, \(k\)-Means. The \(k\)-Means method provides a binary result: it assigns a particular cluster to each data point. The Matlab functions \textit{fcm} and \textit{kmeans} were used for this study to ensure the reliability of the comparison of the results. The two methods converged to similar results. It was shown that \(k\)-Means was more efficient to obtain a precise result for most of the databases used, as it had better time complexity (i.e., the time taken to compute the result), which means better convergence. However, FCM was more competent to give approximate results in a shorter time, as well as dealing with incomplete or corrupted data.

Figure 2 presents the different results obtained by applying \(k\)-Means and FCM on a similar data of 60250 entries, the combination of the drill catalogs from two Japanese tool manufacturers, companies A and B, detailed in the next section. Here, the cut diameter \(D_c\) [mm] and the feed rate \(f\) [mm/rev] are shown, considered as typical parameters. Figure 2 (a) shows the assignation of the data points to each of the clusters according to the maximum membership degree value \(u_{\text{max}}\) displayed in Fig. 2 (b). Figure 2 (c) shows the three clusters computed by \(k\)-means, and Fig. 2 (d) highlights in red the points that are assigned to different clusters when comparing the two methods.

The points that change cluster in the two methods are situated in the dark blue zone of Fig. 2 (b): they are points with a low membership degree value in the FCM method. In the middle of two clusters, they are considered as boundary, outlier or noise data, they cannot be used to define a cluster properly, it is coherent that their assignations change when using different clustering methods. Those outlier data points affect the processing of the clustering result, by reducing the accuracy of a regression analysis in a cluster for example. With the \(k\)-Means method, the noise data is hard to identify. However, with FCM, the low membership degree level points can be thought as a supplementary cluster, suggesting that the actual number of clusters in FCM is \(c + 1\) (four in this case). By filtering appropriately with \(u_{\text{max}}\), the unreliable data of the last supplementary cluster can be suppressed. This property of FCM can help constructing a more precise data mining process.
2.3 Data normalization

The result of a representative-based algorithm strongly depends on the distances between points. It is then greatly affected by the ranges of the different parameters of the database. To avoid the range effect, the data can be normalized to ensure the equality of the ranges of all the parameters. The normalization chosen for the study is the min-max normalization to map all the points in a [0, 1] range. It uses $x_{\text{max}}$ and $x_{\text{min}}$, the maximal and minimal values of the parameter, respectively. The normalized value $x'$ of a data value $x$ is defined by Eq.(6).

$$x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

Figure 3 shows the effect of the normalization of the data on the result of its clustering. The data used is a random generation of 100 points. The $x$ values have a range of [0, 100] while the $y$ values have a range of [0, 1]. With no normalization, Fig. 3(a), only the values of $x$ have an effect on the clustering: the clusters and their centers only depend on the level of $x$. With the normalization, Fig. 3(b), both of the parameters have the same influence on the clustering: the points are distributed in the clusters according to the values of $(x, y)$. 

(a) Clustering of the raw data.  (b) Clustering of the normalized data.

Fig. 3 Effect of the normalization on the clustering of a random database.
2.4 Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is an internal validation criterion for metric clustering (Davies and Bouldin, 1979). It uses metric properties of the clusters to evaluate the final clustering scheme. This criterion can be used to find the optimal number of clusters for a database, by comparing the DBI of the clustering results for various $k$ or $c$.

To compute the DBI, intra and inter-cluster measures are observed. The intra-cluster distance $S_i (1 \leq i \leq N)$ determines the dispersion in each of the $N$ clusters as shown in Eq. (7), while the inter-cluster distance between two clusters $i$ and $j$ is measured by $M_{i,j}$, Eq. (8), which reflects the separation between the two clusters.

$$S_i = \left( \frac{1}{T_i} \sum_{j=1}^{T_i} \|X_{i,j} - A_i\|^p \right)^{\frac{1}{p}}$$

$$M_{i,j} = \|A_i - A_j\|^p = \left( \sum_{k=1}^{n} |a_{k,i} - a_{k,j}|^p \right)^{\frac{1}{p}}$$

Where $X_{i,j}$ is a n-dimensional point in the cluster $i$, $A_i = \{a_{i,1}, \ldots, a_{i,n}\}$ is the centroid of the cluster $i$, $T_i$ the number of points in the cluster $i$ and $\| \cdot \|_p$ the Euclidean distance for $p = 2$.

The final measure $R_{i,j}$ evaluates the performance of the clustering results between two clusters: the dispersion intra-cluster should be as small as possible and the distances inter-clusters as big as possible. $R_{i,j}$ is calculated the ratio of the sum of the $S_i$ of the two clusters and $M_{i,j}$:

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

The DBI is finally defined by Eq. (10). It consists of the mean value of the biggest $R_{i,j}$ for each cluster, which corresponds as the cluster $j$ being the closest to the cluster $i$, to evaluate the worst performances of each cluster. The DBI should then be as small as possible to ensure the best clustering result.

$$\text{DBI} = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq 1} R_{i,j}$$

2.5 Maximum information coefficient

The maximum information coefficient (MIC) is a statistic tool to measure the dependence for the relationship of two variables (Reshef et al., 2011). It can detect associations of different types and is not limited to specific correlations. Kinney et al. (2014) proved the MIC to be a great for a better understanding of large datasets in any disciplines. Sakuma et al. (2019) have shown a data mining method based on MIC efficient for end mill catalog databases. The MIC is based on the principle that if the data is plotted on a scatterplot, it exists a grid that partitions the data by enclosing the relationship. By coloring the cases containing points, the resulting contrast can be studied to find the optimal grid and assess the presence of a correlation between the parameters. Figure 4 illustrates the difference of contrast between two data sets when looking at the number and position of the colored cases, with one set that seems to follow a parabolic relation, and one that is randomly distributed in the scatterplot and shows no apparent correlation.

To compute the MIC, all the grids that partition the $N$ data points with a resolution of $n_x \times n_y \leq N^{0.6}$ are studied to find the maximal mutual information $I_G$. The mutual information, estimated for a certain grid $G$ with by Eq. (11), is a measure that quantifies how much the value of one parameter can tell about the value of the other parameter. The final MIC is then a statistic that uses the maximum value of $I_G$ over all the grids and normalize it by taking the logarithm of the minimum value of the grid dimensions, as shown by Eq. (12).
Fig. 4 Contrast appearing in a scatterplot divided by a grid.

\[
I_G = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (11)
\]

\[
MIC = \frac{\max(I_G)}{\log_2(\min(n_X, n_Y))} \quad (12)
\]

Where \(n_X\) and \(n_Y\) are the number of bins into which \(X\) and \(Y\) are partitioned, \(x\) the \(x^{th}\) column of the partition of \(X\) (i.e., the probability of \(x\)), \(y\) the \(y^{th}\) column of the partition of \(Y\), \(p(x)\) the fraction of data points in \(x\) and \(p(x,y)\) the fraction of points in the \(x\)-\(y\) case (i.e., the joint probability of \(x\) and \(y\)). Figure 5 presents the effect of the grid on the value of the mutual information of a given data set. Grids A and B have the same number of rows and columns, but they are differently split on the scatterplot. The probabilities of each row and column as well as the joint probabilities of each cases are affected by the new repartition. For the data presented here, a grid that divides equally the ranges of \(x\) and \(y\) (grid A) does not give the optimal result of mutual information. Grid B shows a better result, with unequal ranges of rows.

The MIC is computed in this study by using the function \texttt{mine} from \texttt{minepy}, the Matlab form of the MINE application developed by Reshef and Reshef (2018). This function computes all the MINE statistics between \(X\) and \(Y\), two row vectors of size \(n\): \texttt{mine(X, Y, alpha, c, est)} = \texttt{[minestats, M]}, with \(alpha\) the coefficient to fix the maximal dimension \(n_x \times n_y\) of the grids that will be used to compute the MIC \((n_x \times n_y \leq n^{\alpha}\)), \(alpha = 0.6\) by default), \(c\) determines how many clumps there will be in the partition of the data (by default \(c = 15\)), \(est\) is the computation method (\(est = \text{‘mic-approx’}\) by default), \texttt{minestats} is a structure containing the 6 MINE statistics that have been computed, and \(M\) is the square matrix.
that contains the maximal value of the MIC for each of the grid dimensions. For this study, only the MIC value was examined. It was obtained from the minestats structure, minestats.mic returning the desired information.

2.6 Mean squared error
The mean squared error (hereinafter, MSE) is an indicator that measures the average of the squared errors of a prediction. It gives an indication about the quality of the prediction. Always positive, it does not have any objective values: for a given data, the best predictor will show the lowest MSE. For the \( n \) points of the original vector \( X = [x_1, ..., x_n] \), the MSE of the predicted vector \( Y = [y_1, ..., y_n] \) is given as follows:

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2
\]  (13)

3. Data-mining of drill catalogs
3.1 Drill catalog database
Tool catalogs are commonly used to select a tool and the machining conditions to apply. Edited regularly by tool manufacturers, they present all the properties of their current collection in details: the dimensions of each tool, its coating, material and type of lubrication. For each tool, or series of tools, the optimal machining conditions are then given according to the material machined, usually depending on its nature and hardness.

3.1.1 Solid drill catalogs
The catalogs used in the present report are solid drill catalogs. In those catalogs, the conditions are given separately for each series of drills. They depend on the cut diameter of the drill (i.e., the external diameter of the teeth that defines the final diameter of the drilled hole), and on the maximal height-by-diameter ratio that can be drilled with the tool, which corresponds to the ratio of the effective length of the drill to the cut diameter. For one given tool, multiple conditions can then be chosen according to the material drilled.

3.1.2 Catalog database
For this report, we have used the catalogs of solid drills from two major Japanese tool manufacturers. Those two companies will be referred as Company A and B. The data used is extracted from the drill catalogs 2017 and 2018, respectively. As there is a broad range of options for the optimal conditions of a single tool depending on the material drilled, one entry of the database corresponds to one of the possible sets of conditions for a tool. For A Co., up to 33 different drilling conditions can be fixed for a drill, while 13 conditions can be given for B Co. The number of entries in the database (i.e., its volume) is then far superior to the number of different tools in the catalog.

Table 1 summarizes all the drill properties given in the catalogs, as well as the drilled material properties, the recommended drilling conditions, and general information, including both quantitative and qualitative variables. The geometrical parameters of the drill are explained on Fig.6 and the drilling conditions on Fig.7. A computed value is added to the list of geometrical parameters, the length-to-diameter ratio, used in the catalogs to determine the drilling conditions.

Fig. 6 Geometrical parameters of a drill.

Fig. 7 Drilling conditions.
Table 1  Ranges of the quantitative parameters and description of the qualitative parameters given in the drill catalogs.

| Type                  | Variable                      | Company A       | Company B       |
|-----------------------|-------------------------------|-----------------|-----------------|
| Drill properties      | Cut diameter $D_c$ [mm]       | 0.5 – 12        | 0.5 – 25        |
|                       | Shank diameter $D_s$ [mm]     | 0.5 – 12        | 2 – 25          |
|                       | Flute length $l$ [mm]         | 3 – 95          | 2.6 – 391.6     |
|                       | Effective length $L_e$ [mm]   | 2.1 – 80        | 0.6 – 361.6     |
|                       | Length-to-diameter ratio $l_e/D_c$ [-] | 2 – 10        | 1 – 40          |
|                       | Shank length $L_s$ [mm]       | 17 – 51         | 3 – 75          |
|                       | Total length $L$ [mm]         | 20 – 142        | 47.1 – 449.6    |
|                       | Point angle $\alpha$ [deg]    | 124 – 150       | 135 – 180       |
|                       | Number of tooth $Z$ [-]       | 2 or 3          | 2               |
|                       | Lubrication                   | int. or ext.    | int. or ext.    |
|                       | Coating                       | 2 types         | 4 types         |
| Material properties   | Type of material              | 10 types        | 6 types         |
|                       | Hardness [HRC]                | 9 – 50          | 10 – 93         |
| Drilling conditions   | Spindle rotation speed $S$ [rpm] | 200 – 48500    | 300 – 25400     |
|                       | Feed rate $f$ [mm/rev]        | 0.003 – 0.53    | 0.006 – 0.6     |
|                       | Feed $F$ [mm/min]             | 10 – 3485       | 35 – 3050       |
|                       | Cutting speed $V$ [m/min]      | 5 – 245         | 7 – 180         |
| Number of distinct tools|                              | 364             | 5561            |
| Number of tool series |                              | 5               | 10              |
| Data volume           |                              | 6272            | 53978           |

The four drilling conditions given in Table 1 and presented in Fig. 7 are the four main conditions for drilling operations. The spindle rotation speed $S$ is the rotational frequency of the spindle of the machine, the cutting speed $V$ is the speed at the periphery of the tool, the feed rate $f$ is the advance of the tool in the direction of rotational axis given for one revolution of the tool, and the feed $F$ is the speed of the tool in the rotational axis direction. The relations between the four conditions are given by Eq. (14) and (15). Those relations are used to compute some of the values of the database: $S$, $f$ and $V$ are fully given in the catalog of A Co. and used to compute $F$, while only $S$ and $f$ are completely given in the catalog of B Co., the values of $V$ and $F$ being given for some tools and computed for the rest. This study will then focus on $S$ and $f$.

\[
V = \frac{D_c \cdot \pi \cdot S}{1000} \quad (14)
\]

\[
F = f \cdot S \quad (15)
\]

3.2 Tool catalog mining

3.2.1 Preparation of the data

3.2.1.1 Normalization of the variables

All the present study has been realized on Matlab. It uses either functions already implemented, applications added to the basic modules or directly written programs.

As shown previously in Table 1, all the quantitative parameters present very different ranges of value. To ensure the good clustering of all these parameters, they have been first normalized by using the min-max normalization presented in Eq.(6), for every database. All the ranges have then been set to [0, 1]. The qualitative variables also presented different distinct values, but they cannot be normalized the same way as the numerical parameters.

3.2.1.2 Handling of the qualitative variables

FCM being a distance-based algorithm, the clustering can only be realized on quantitative (i.e., numerical) variables. Thus, categorical variables such as the type of material drilled, the coating, the type of lubrication, or the material of the drill cannot be used directly in the clustering. However, this information is crucial for any machining processes, as many conditions will directly influence the result of the process. To include this information in the clustering, it was necessary to create numerical parameters corresponding to the categorical values. Assigning arbitrary values would then create hierarchy between the non-hierarchical variables: for
example, when considering the material drilled, if the values 1, 2 and 3 where attributed to the aluminum alloy, titanium alloy and copper alloy, respectively, it would then imply that the aluminum alloy would be closer to the titanium alloy than to the copper alloy, creating a notion of distance between the different materials. To ensure the conservation of the properties of the type of data, sets of equidistant points were created. In a n-dimension space, n + 1 points can be determined such as the distance between any points x and y of the set is equal to a constant value D (i.e., ∀(x, y), ||x - y||2 = D). The distance chosen for this study is 1 to keep the same value range as the normalized variables. For every categorical parameter, their corresponding equidistant points have then been added to the database and will be used for all the clustering in the present report.

3.2.2 Clustering of the parameters defining the series

The drill catalogs give the drilling conditions separately for each series of tool. The first of the catalog mining was to find the parameters differing by series. All the drills diameters and lengths, as well as the material properties evolve the same way for each series. Only the length-to-diameter ratio, the point angle, the number of teeth, as well as the coating and the type of lubrication seem to differ from one series to another when considering the parameters in the database. However, it can be seen in Fig. 8 that the general geometry of the tools also varies. Two new variables were then introduced to the database: the ratio of the diameters D_s/D_l, and the ratio of the lengths l_e/L. The description of the 6 types of coating is given in Table 2. Coatings 3 and 5 have the same components, however they are considered separately: realized by two different companies, the exact composition as well as the coating process might differ, and thus the mechanical properties cannot be considered equal. In Fig. 8, the two modes of lubrication can also be seen as well as two different coatings can be seen: a tool from series No. 2 has external lubrication, with coating number 1 (DP1), while a tool from series No. 13 has internal lubrication, which can be seen with the two holes in the head of the tool, and no coating.

| Coating No. | Coating Name | Company | Description          | Number of tools |
|------------|--------------|---------|----------------------|-----------------|
| 1          | DP1          | B       | Laminated PVD coating| 1984            |
| 2          | DP3          | B       | PVD coating          | 370             |
| 3          | VP           | B       | (Al,Ti)N coating     | 2691            |
| 4          | none         | B       | No coating           | 411             |
| 5          | AITin        | A       | (Al,Ti)N coating     | 432             |
| 6          | TiB₂         | A       | TiB₂ coating         | 75              |

(a) Series number 2. (b) Series number 13. Fig. 8 Drills of two different series from the catalog of B Co.

The Fuzzy c-means algorithm was then performed with a 11-dimension database: point angle \( \alpha \), Z, \( l_e/D_s \), \( D_s/D_l \), l_e/L, lubrication and coating (categorical parameter of 6 distinct values: use of 6 equidistant points of dimension 5). Three to six clusters were formed. For this clustering, each tool appears one time in the data-base, as it has no relation to the type of material drilled. The result of the clustering is presented in Table 3, where the values of the defining parameters are given for each cluster. For most of the cases, knowing the coating or the type of lubrication is enough to assign the tool to a certain cluster. As the number c of clusters increases, it can be seen from Fig. 9 that at each iteration only one cluster seems to be divided each time, when considering the global division of the data base. For instance, the cluster 2 when c=3 gets divided into cl. 2 and cl. 3 at c=4, and then cl. 3 at c=4 is divided into cl.3 and cl. 4 at c=5 based on the coating parameter. One parameter can be easily identified as the guide of the division of the clusters, focusing on the number Z of teeth, the coating and the lubrication type. Especially, the coating seems to have a large influence on the clustering result. This tendency is in good agreement with the results obtained for end-mill tools (Sakuma et al, 2019). However, the lubrication type is found to have less impact on end-mill processes than for drilling operations where it is essential, due to the fact that drills are completely enclosed in the material while end mills have an open side. The number of teeth Z is considered as an essential parameter. However, it has few influence in the partition of the data in this clustering.
is due to the very inequal proportion of tools with 2 or 3 teeth: in the two catalogs, 58900 tools with $Z=2$ and 1350 tools with $Z=3$. Fuzzy $c$-Means algorithm is highly influenced by the number of data points that have a certain value or are in a certain category. The relationships between the parameters at $c=4$ are shown in Fig. 10, where all the parameters except the lubrication are plotted. We look at this typical clustering as its clusters are relatively distinct. Figure 10 (a) presents the position of the different clusters according to the diameters and lengths ratios of the tools. It can be seen that the ratios show different relations when considering each cluster, even when their value ranges are similar. Figure 10 (b) illustrates the coating types, the number of teeth and the point angles in each cluster, according to Table 3.

Table 3 Result of the clustering with FCM for the defining parameters (quantitative and qualitative).

| Cluster | Coating | $a$[deg] | $Z$ [-] | Lubrication | $l_e/D_e$ [-] | $D_c/D_s$ [-] | $l_e/L_e$ [-] |
|---------|---------|----------|-------|-------------|--------------|-------------|-------------|
| 1, $c = 3$ | 3; 4; 5; 6 | 124; 140; 145; 150 | 2; 3 | extern | 1 – 7 | 0.5 – 1 | 0.040 – 0.476 |
| 2, $c = 3$ | 2; 3; 4; 5 | 135; 140; 145 | 2 | intern | 1 – 30 | 0.167 – 1 | 0.013 – 0.774 |
| 3, $c = 3$ | 1; 6 | 124; 140; 180 | 2; 3 | extern | 2 – 40 | 0.5 – 1 | 0.109 – 0.924 |
| 1, $c = 4$ | 3; 4; 5; 6 | 124; 140; 145; 150 | 2; 3 | extern | 1 – 6 | 0.5 – 1 | 0.040 – 0.476 |
| 2, $c = 4$ | 3 | 135; 140; 145 | 2 | intern | 1 – 30 | 0.167 – 1 | 0.013 – 0.774 |
| 3, $c = 4$ | 2; 4; 5 | 135; 140 | 2 | intern | 3 – 10 | 0.5 – 1 | 0.136 – 0.602 |
| 4, $c = 4$ | 1; 6 | 124; 140; 180 | 2; 3 | extern | 2 – 40 | 0.5 – 1 | 0.109 – 0.924 |

Fig. 9 Division of the clusters with the increase of the number of clusters.

(a) Diameters and lengths ratios. (b) Discrete parameters.

Fig. 10 Results of the clustering for the defining parameters and 4 clusters.
3.2.3 MIC: influence of the drill and material properties in each cluster

In order to find the properties having the largest influence on the drilling conditions in each cluster, the MIC of all the pairs condition-parameter were computed, for the raw values (i.e., without normalization). Figure 11 shows the evolution of the MIC for each fully given condition with every numerical parameter. The parameters having only a few discrete values were not used, as their MIC cannot be properly computed. To have better coherence with the catalogs and the way the conditions are given in them, only the limit values of diameters and length-to-diameter ratios that appear explicitly in the used catalogs are considered.

The feed rate \(f\) presents higher values of MIC for most of the parameters. For the two conditions, the smallest MIC values are obtained for the length-to-diameter ratio \(l_e/L\) and for the hardness of the drilled material HRC. This might be explained by the fact that those two parameters only discrete limited values, which often results in misleading MIC result.

For cluster 2, the ratio of lengths \(l_e/L\) also shows a low level of MIC. Cluster 2 actually has only eight distinct values of \(l_e/L\), resulting as the same problem as previously. Considering the other parameters, each cluster has a different order of dependence levels. For the feed amount, both cluster 1 and 3 show their biggest influence of the diameters \(D_c\) and \(D_s\), while cluster 4 has the greatest level for the shank length \(L_s\), and cluster 2 for every lengths and diameters except \(L_s\). For the rotation speed the four clusters tend to have more dependence with \(D_c\) and \(D_s\), this dependency is very clear for cluster 3, but less marked for the three other clusters where similar levels can be obtained for the different lengths.

![Fig. 11 MIC computed on the four clusters, for the two fully given drilling conditions.](image)

3.2.4 FCM: clustering of the clusters with the drilling conditions

The Fuzzy C-Means algorithm was then performed on the limit values database, which was used for the MIC computation. The MIC values obtained previously were used to define the parameters to include in the clustering. The parameters were ranked according to the mean of their MIC values for \(f\) and \(S\), for each of the cluster previously formed.

In total, 320 clusterings were made (for every of the four clusters, for each of the two drilling conditions, clusterings with 2 to 11-dimension and 3 to 6 clusters). The final order of the variables is presented in Tables 4 and 5, for each of the four clusters. A clustering has at least two dimensions, the drilling condition and the parameter with the highest MIC value, the other dimensions are added in the decreasing order of the MIC levels.

| Dimension | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|-----------|-----------|-----------|-----------|-----------|
| 1         | \(S\) [rpm] | \(S\) [rpm] | \(S\) [rpm] | \(S\) [rpm] |
| 2         | \(D_c\) [mm] | \(D_c\) [mm] | \(D_c\) [mm] | \(D_c\) [mm] |
| 3         | \(l_e\) [mm] | \(D_c\) [mm] | \(HRC\) | \(l_e\) [mm] |
| 4         | \(l_e/L\) [-] | \(D_c\) [mm] | \(HRC\) | \(l_e/L\) [-] |
| 5         | \(L_s\) [mm] | \(L_e\) [mm] | \(l_e/L\) [-] | \(l_e\) [mm] |
| 6         | \(L_s\) [mm] | \(D_e/D_s\) [-] | \(L_e\) [mm] | \(D_e/D_s\) [-] |
| 7         | \(l_e\) [mm] | \(D_e/D_s\) [-] | \(l_e/D_c\) [-] | \(l_e/L\) [-] |
| 8         | \(D_e/D_s\) [-] | \(HRC\) | \(L_e\) [mm] | \(L_e\) [mm] |
| 9         | \(l_e/D_c\) [-] | \(HRC\) | \(l_e/L\) [-] | \(HRC\) |
| 10        | \(l_e/D_c\) [-] | \(l_e/D_c\) [-] | \(l_e/D_c\) [-] | \(l_e/D_c\) [-] |

Table 4 Variables used in the clustering according to their mean value of MIC for the rotation speed.
Table 5 Variables used in the clustering according to their mean value of MIC for the feed rate.

| Dimension | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|-----------|-----------|-----------|-----------|-----------|
| 1         | $f$ [mm/rev] | $f$ [mm/rev] | $f$ [mm/rev] | $f$ [mm/rev] |
| 2         | $D_c$ [mm] | $D_c$ [mm] | $D_c$ [mm] | $L_s$ [mm] |
| 3         | $D_s$ [mm] | $D_s$ [mm] | $D_s$ [mm] | $L$ [mm] |
| 4         | $l_e/L$ [-] | $l_e$ [mm] | $l_e$ [mm] | $l_e$ [mm] |
| 5         | $L$ [mm] | $L$ [mm] | $D_s/D_c$ [-] | $D_c/D_s$ [-] |
| 6         | $l$ [mm] | $l$ [mm] | $l_e$ [mm] | $D_s$ [mm] |
| 7         | $l_e/D_c$ [-] | $l_e/D_s$ [-] | HRC | HRC |
| 8         | $D_s/D_c$ [-] | $D_s/D_c$ [-] | HRC | $l_e/D_c$ [-] |
| 9         | HRC | HRC | $l_e/L$ [-] | $l_e/L$ [-] |

3.2.5 DBI: Assessment of the clustering result

The performances of each clustering were then assessed by using the DBI index. In order to find the optimal clustering that would help in the decision of the drilling conditions for a given tool, the DBI index was computed on every clustering by considering 11 variables used in the clustering. The DBI can then assess the separation between the clusters and the dispersion inside them. Figure 12 and 13 show the evolution of the DBI of the different clustering with the number of clusters, for the four clusters and the two drilling conditions. When considering a single cluster, the evolution of the DBI shows similar tendencies for the two drilling conditions, apart from cluster 4. For cluster 2 and 3 the lowest and smoothest DBI levels are observed at $c = 3$. In general, the smallest DBI values are obtained for high dimension clusterings, and for small numbers of clusters, which is coherent with the fact that the DBI considers all of the eleven variables. The higher the number of clusters, the more variations appear in the DBI evolution: more diversity appear in the clusters when forming more groups.

Fig. 12 DBI computed for all the different clustering with the rotation speed.
For the rotation speed the introduction of the hardness of the drilled material, which is the only variable not linked to the drill geometry, leads to a decrease of the DBI in many cases: for cluster 1, the HRC is introduced in the clustering as the 10th dimension, for cluster 2 as the 9th dimension, for cluster 3 as the 4th dimension, and for cluster 4 as the 10th dimension. The decrease is clearly marked with $c = 5$ and 6 for cluster 1, with $c = 4, 5$ and 6 for cluster 2, with $c = 4, 5$ and 6 for cluster 3, and with $c = 6$ for cluster 4. For the feed rate, the hardness is introduced at the 11th position for cluster 1, at the 9th for cluster 2, at the 10th for cluster 3 and at 9th for cluster 4. The decrease is barely visible for cluster 1 and 3, but clearly marked for $c = 5$ and 6 for cluster 2, and for $c = 6$ in cluster 4.

### 3.2.6 MIC: influence of the material and drill properties on the drilling conditions

For each cluster and drilling condition, the two clusterings presenting the lowest values of DBI for every number of clusters were considered for the catalog mining system. Sub-groups of points were realized containing the points of each of the selected clusters. The sub-groups can be considered as the two best clusterings given a number of clusters, where the only the dimension of the clustering differs. For two different number of clusters, the optimal dimension may differ. The same process as presented in section 3.2.4 was then applied on each of the sub-groups: the MIC between the drilling condition they correspond to and each property was computed. The four most influent variables for the conditions in each sub-groups were then extracted.

### 3.2.7 Refining of the cluster borders: filtering with the flexible membership degree $u_{\text{mean}}$

Each of the sub-groups corresponding to each cluster presents different trends that define the drilling conditions. To create more clearly defined groups, the points were filtered according to an index, $u_{\text{mean}}$, based on the membership degree $u_{\text{max}}$. By setting a minimum level of $u_{\text{mean}}$ as shown in Fig. 2 (b), the points on the borders of the clusters can be suppressed. Those points tend to have mix tendencies as they are adjacent to two or more clusters as mentioned in 2.2 section.

The membership degree value $u_{\text{max}}$, given for each point to every cluster, is the specificity of the FCM method for clustering. This index provides information about the distance of a point to the centers of the clusters. For a clustering...
with \( c \) clusters, a value of \( 1/c \) of \( u_{\text{max}} \) means that the point is exactly at the same distance of the \( c \) centers or located at infinity. Considering this property of \( u_{\text{max}} \), it can be understood that the FCM method gives the clusters a spherical definition: \( u_{\text{max}} \) decreases in every direction.

This study first proposed to consider the points by their \( u_{\text{max}} \) value, creating sub-groups that only contained points that a membership degree superior to a fixed level. However, due to the spherical definition of the clusters, points that had extreme values for even one of the clustering parameters were all suppressed as they were too far from the centers. Almeida et al. (2006) showed that FCM had troubles dealing with those extreme values. The trends in some clusters were then lost. To limit this effect the new flexible index \( u_{\text{mean}} \) was proposed as defined in Eq. (16), where \( \max_j(u_{ij}) \), ..., \( \max_c(u_{ij}) \) are the ordered \( u_{ij} \) values of \( x_i \) for a \( c \) clusters clustering, with \( \max_i(u_{ij}) \) being the highest value. This index considers the difference of the membership value of a point \( x_i \) to the different clusters. A value of 1 corresponds to a cluster center and 0 is a point exactly in the middle in the case of two clusters. To compare the effect of the filtering with the two indexes \( u_{\text{max}} \) and \( u_{\text{mean}} \), a random data was filtered with the same level of index. The result is presented in Fig. 14 (the limit value of 0.7 was taken as an example, as it displays clearly the difference between the indexes). It can be seen that the points with high value of data 1 are kept in the cluster 3 with \( u_{\text{mean}} \), when they were suppressed by \( u_{\text{max}} \). This new flexible index allows to keep all the significant points while suppressing the points in between two clusters.

\[
 u_{\text{mean}}(x_i) = \frac{(c-2) \cdot \max_1(u_{ij}) - \sum_{j=2}^{c-1} \max_j(u_{ij})}{(c-2) \cdot \max_1(u_{ij})}
\]  

(16)

![Fig. 14 Difference of filtering with \( u_{\text{max}} \) and \( u_{\text{mean}} \).](image)

3.2.8 Regression in the sub groups

To extract equations from the sub-groups to define the drilling conditions, the multiple linear regression of Matlab was used. For the predicted value \( y \) and the predictors matrix \( X \), the function \( \text{regress}(y, X) = b \) returns the vector \( b \) containing the estimated coefficients. For all the clusters from the clusterings previously chosen, the \( \text{regress} \) function was applied using the drilling conditions as \( y \), and a matrix containing second order combinations of drill and material quantitative parameters as \( X \). A second order surface response is needed to obtain a sufficient prediction accuracy for machining conditions (Sakuma et al., 2019).

Different limit values of \( u_{\text{mean}} \) were used, from 0 to 0.95, with a step of 0.05, for each cluster of the second clustering. As all the possible combinations were tested, 77760 sub-groups were created for each condition. For every of these sub-groups, the MIC was first computed, and the four variables having the highest MIC value was used in the parameters matrix. For each sub-group, the linear regression resulted in an equation linking the four parameters with the drilling condition the clustering depended on, following the second order model of Eq. (17): for the four most influent variables \( x_1, x_2, x_3 \) and \( x_4 \), the optimal coefficients \( c_{ij} \) are determined by the \( \text{regress} \) function. This equation was then applied to the whole sub-cluster (i.e., the sub-group and the points suppressed with the \( u_{\text{mean}} \) filtering) to obtain the list of the predicted condition values. The optimal sub-groups were decided by measuring the mean squared error of the condition given in
the database, with the prediction computed with the equation from the filtered group applied on the whole cluster. The whole predicted condition corresponds then to the combination of each of the best sub-clusters.

\[
cond_{predicted} = c_{0,0} + \sum_{i=1}^{4} c_{i,0}x_i + \sum_{i=1}^{4} \sum_{j=1}^{4} c_{i,j}x_i x_j
\] (17)

4. Results and discussion based on the developed method

The objective of this research is to analyze the drilling conditions in regards with their relation to the variables given in a tool catalog. With the Fuzzy c-means clustering and the use of the DBI and the MIC indexes, smaller groups of points with distinct tendencies were extracted from the database, as shown in the previous parts. Figure 15 showcases the main steps of the developed catalog mining method, which uses two combinations of FMC and MIC. The evaluation of the predictions made with the MSE are used as feedbacks on the quality of the two sets of clusterings realized.

![Main phases of the catalog mining method developed in this research.](image)

In the case of the first clustering with four clusters, with the second clustering realized for the feed rate \(f\), the second clusterings presented in Table 6 were considered as the best ones by the DBI for each of the four clusters. The MSE values of the prediction realized for each of them with the values chosen by the MIC are also given in the table: for a given cluster, the best prediction is obtained for the lowest level of MSE, which can only be compared when considering the same original data. For cluster 1, the clustering with 4 clusters and 8 dimensions appears to have the best predictors. For cluster 2, 6 clusters and 9 dimensions showed better results. For cluster 3, 4 clusters gave also the best result with 7 dimensions. For cluster 4, 5 clusters with 6 dimensions appear to be the most efficient. In each of these clusters, a different equation was obtained to define the rotation speed with the multi linear regression. Here, Eq. (18) shows as example the equation obtained for the second cluster out of five for cluster 3. The MIC had given for most influent parameters the hardness HRC, the cut diameter \(D_c\), the shank diameter \(D_s\), and the shank length \(L_s\) (Fig. 11). For this sub-cluster, the hardness HRC and the shank length \(L_s\) have the biggest coefficients and the highest frequency in Eq. (18). HRC is known to be an essential mechanical parameter for the workpiece and \(L_s\) one of the most important parameters of the tool for data mining methods, as the same tendency was seen in some clusters for endmill catalogs, where the total length \(L\) is the most important parameter (Hirogaki, et al., 2011).
$f_{predicted} = 1.43 \times 10^{-2} \times \text{HRC} + 4.50 \times 10^{-3} \times L_s + 1.21 \times 10^{-5} \times \text{HRC}^2 + 4.90 \times 10^{-5}$

$\times \text{HRC} \times D_c - 2.79 \times 10^{-4} \times \text{HRC} \times L_s + 4.80 \times 10^{-5} \times D_c \times L_s$

$+ 1.03 \times 10^{-5} \times L_s^2 \quad (18)$

Table 6  Best clusterings and their prediction evaluation for the feed rate $f$, for a first clustering of four clusters.

| Cluster | Number of clusters | Dimension of the clustering | MSE [-] | Cluster | Number of clusters | Dimension of the clustering | MSE [-] |
|---------|--------------------|-----------------------------|---------|---------|--------------------|-----------------------------|---------|
| Cluster 1 1227 points | 3 | 10 | 29.1 x10^{-4} | 3 | 3 | 23.2 x10^{-4} |
| | 3 | 11 | 28.9 x10^{-4} | 3 | 8 | 25.9 x10^{-4} |
| | 4 | 8 | 26.4 x10^{-4} | 4 | 4 | 26.4 x10^{-4} |
| | 4 | 9 | 26.8 x10^{-4} | 4 | 8 | 25.2 x10^{-4} |
| | 5 | 7 | 37.6 x10^{-4} | 5 | 7 | 25.5 x10^{-4} |
| | 5 | 8 | 28.7 x10^{-4} | 5 | 9 | 24.9 x10^{-4} |
| | 6 | 8 | 27.8 x10^{-4} | 6 | 9 | 23.1 x10^{-4} |
| | 6 | 11 | 27.4 x10^{-4} | 6 | 11 | 27.0 x10^{-4} |
| Cluster 2 1310 points | 3 | 3 | 23.2 x10^{-4} |
| | 3 | 8 | 25.9 x10^{-4} |
| | 4 | 4 | 26.4 x10^{-4} |
| | 4 | 8 | 25.2 x10^{-4} |
| | 5 | 7 | 25.5 x10^{-4} |
| | 5 | 9 | 24.9 x10^{-4} |
| | 6 | 9 | 23.1 x10^{-4} |
| | 6 | 11 | 27.0 x10^{-4} |
| Cluster 3 953 points | 3 | 8 | 17.5 x10^{-4} | 3 | 8 | 40.0 x10^{-4} |
| | 3 | 11 | 17.9 x10^{-4} | 3 | 9 | 36.0 x10^{-4} |
| | 4 | 7 | 16.2 x10^{-4} | 4 | 4 | 35.3 x10^{-4} |
| | 4 | 11 | 17.8 x10^{-4} | 4 | 11 | 35.0 x10^{-4} |
| | 5 | 6 | 18.0 x10^{-4} | 5 | 6 | 26.3 x10^{-4} |
| | 5 | 7 | 19.0 x10^{-4} | 5 | 11 | 34.5 x10^{-4} |
| | 6 | 8 | 16.9 x10^{-4} | 6 | 6 | 27.6 x10^{-4} |
| | 6 | 11 | 17.7 x10^{-4} | 6 | 9 | 33.3 x10^{-4} |
| Cluster 4 386 points | 3 | 3 | 23.2 x10^{-4} |
| | 3 | 8 | 40.0 x10^{-4} |
| | 4 | 4 | 35.3 x10^{-4} |
| | 4 | 11 | 35.0 x10^{-4} |
| | 5 | 6 | 26.3 x10^{-4} |
| | 5 | 11 | 34.5 x10^{-4} |
| | 6 | 6 | 27.6 x10^{-4} |
| | 6 | 9 | 33.3 x10^{-4} |

The predicted values obtained with the regression for the two drilling conditions are shown in Fig. 16. Their correlation coefficient $r^2$ is also displayed. Most of the points are close to the objective line except some points with high feed rate or spindle speed values, which show a lower predicted value. Those points correspond to the conditions for aluminum or plastic materials. In general, the HRC is considered to be an unsuitable unit to estimate this kind of soft hardness and thus the suggested conditions given in the catalog would be higher values. The proposed data mining method can then be thought to be effective for predicting proper drilling conditions, as well as estimating the data quality and the limits of the database used, especially considering the unbalanced amount of information between the different types of materials. To provide accurate drilling conditions for soft materials such as plastic and aluminum, more information on the proper hardness of the drilled material is needed. This information is actually insufficient in the catalog data base, especially compared to the data given for steels. Conditions for plastics or aluminum were given for few tools only in the catalogs, inducing few entries in the database used and thus smaller weigh in the clustering process. Those points were then included in the sub-clusters with a high proportion of harder materials. The equations are then not in adequacy with plastics and aluminum conditions, but not optimal for the harder materials neither, due to the influence of the softer ones. The representative-based clustering algorithm Fuzzy c-Means was enhanced with the measure coefficient MIC, and was performed twice on two catalogs of solid drills. Catalogs give the optimal drilling conditions according to different parameters: firstly, according to the series of the tools, then according to the type of material drilled, and finally to the diameter of the hole and its relative depth. To follow these steps, the clustering was first realized on the variables identified as the defining parameters for the series. In each of the clusters formed, the MIC, an index measuring the correlation between two parameters, was used to determine the parameters having the largest influence on each of the drilling conditions. A second clustering was then conducted for each of the drilling condition, with varying dimensions according to the influence of the parameters. The best clustering results were then identified with a validation criterion.
for clustering results, the DBI index. Finally, the use of a multilinear regression in each of the clusters of the best clustering allowed to create a model to predict the drilling conditions from the parameters determined by the MIC index. As a result, it can be seen that the developed data mining method is effective not only to predict proper drilling conditions but also to clear the insufficient information in the tool catalogs. To ensure better results of the data mining system, all the materials should be equally represented in the database, with parameters that can be applied correctly to define each of them. Moreover, the validity of the conditions predicted in term of the drilling process depends directly on the data given by the tool manufacturers, as the method developed here uses the raw data of the catalogs.

Fig. 16 Predicted values of the conditions obtained with the developed catalog mining system for four clusters.

Table 7 Correlation in each cluster for $S$ and $f$.

| Condition         | Cluster Number | Coefficient $r^2$ | Condition         | Cluster Number | Coefficient $r^2$ |
|-------------------|----------------|-------------------|-------------------|----------------|-------------------|
| Spindle speed $S$ | 1              | 0.90              | Feed rate $f$     | 1              | 0.81              |
| [rpm]             | 2              | 0.74              |                   | 2              | 0.92              |
|                   | 3              | 0.91              |                   | 3              | 0.90              |
|                   |                |                   |                   | 4              | 0.88              |
|                   |                |                   |                   |                |                   |

5. Conclusion

The following results were obtained with the tool catalog mining method based on Fuzzy c-means algorithm to expand the knowledge of drilling process, method combining twice the clustering algorithm with an evaluation coefficient:

1. The quantitative parameters given in the catalogs have very different value ranges and need to be normalized: the spindle rotation speed $S$ shows values from 200 to 48500 rpm, while the feed rate $f$ is in a $0.003 – 0.6$ mm/rev range. For the qualitative variables, that are essential parameters for drilling processes, by creating associated equidistant points, they can be integrated in the clustering. The clustering of the parameters that define the tool series follows a logical division when the number of clusters increases. It was also shown that with less than three defining parameters, a tool can be assigned a tool to a cluster. The lubrication type along with the coating are enough for most of the cases.

2. Concerning the indexes used in the method, the MIC index is sensitive to the number of discrete values in the database, and tends to consider them not influent. The lengths and diameters of the drills are the parameters with the largest influence on spindle rotation speed and feed rate, however each cluster presents different orders for the influence of its parameters. The lowest DBI values are obtained for high dimension clusterings. The DBI also shows a marked decrease when the hardness of the material is added in the clustering for most of the cases.
(3) After having defined the optimal clusterings with FCM, filtering the database with the maximal membership degree value $u_{\text{max}}$ only will systematically suppress the points with extreme values. However, the use of the new flexible index $u_{\text{mean}}$ makes it feasible to keep the extreme values that give the specificity of a cluster, while suppressing the noise data that can be found in the tool catalogs.

(4) For each sub-cluster, an equation predicting the drilling conditions can be obtained in the properly filtered clusters. These predictions show good general tendencies, with for each cluster, sub-clusters showing the best predictions from different dimension clusterings. Moreover, the predictions show that more information on the workpiece is needed in the catalogs to predict proper drilling conditions for soft materials, such as plastic and aluminum. Moreover, the use of the Fuzzy $c$-means algorithm makes it feasible for an operator to consider more options for the choice of the cluster number in the two clusterings than when using conventional $k$-means method, where only binary results are returned for the belonging of a point to a cluster. As a result, the proposed catalog mining method is found effective to understand the construction of the tool catalog database and to estimate information in it.

References

Aggarwal, C.C., Data mining: the textbook (2015), Cham, Switzerland: Springer.
Almeida, R.J., & Sousa, J.M.C., Comparison of fuzzy clustering algorithms for classification, Conference Proceedings of the 2006 International Symposium on Evolving Fuzzy Systems, (2006), pp.112-117.
Bezdek, J.C., Pattern recognition with fuzzy objective function algorithms (1981), New York: Plenum Press. Bezdek, J.C., Ehrlich, R., & Full, W., FCM: the Fuzzy c-Means clustering algorithm, Computers & Geosciences, Vol.10, No.2-3 (1984), pp.191-203.
Choudhary, A.K., Harding, J.A., & Tiwari, M.K., Data mining in manufacturing: a review based on the kind of knowledge, Journal of Intelligent Manufacturing, Vol.20 (2009), pp.501-521.
Davies, D.L., & Bouldin, D.W., A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.1, No.4 (1979), pp.224-227.
Gosh, S., & Dubey, S.K. Comparative analysis of K-Means and Fuzzy C-Means algorithm. International Journal of Advanced Computer Science and Applications, Vol.4, No.4(2013), pp.34-39.
Hirogaki, T., Aoyama, E., Ogawa, K., Kodama, H. and Kitamura T., Decision methodology of end-milling conditions using data-mining, Precision Engineering, 85 (2011), pp. 197-203.
Huang, R., Zhang, S., Bai, X., Xu, C., & Huang, B., An effective numerical control machining process reuse approach by merging feature similarity assessment and data mining for computer-aided manufacturing models, Journal of Engineering Manufacture, Vol.229, No.7 (2015), pp.1229-1242.
Kinney, J.B., & Gurinder, A.S., Equitability, mutual information, and the maximal information coefficient. PNAS, Vol.111, No.9(2014), pp.3354-3359.
Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Tumbergh, P.J., Lander, E.S., Mitzenmacher, M., & Sabeti, P.C., Detecting Novel Associations in Large Data Sets, SCIENCE, Vol.334 (2011), pp.1518-1524.
Sakuma, T., Hirogaki, T., Aoyama, E., Kubo, K., Kodama, H., Catalog mining using MIC Maximal Information Coefficient of radius end mill Tool, Journal of the Japan Society for Precision Engineering, 85, 3 (2019), pp. 260-266 (in Japanese).
Shahbaz, M., Masood, S.A., Shaheen, M., & Khan, A., Data mining methodology in perspective of manufacturing databases. Journal of American Science, Vol 6, No.11(2010), pp.999-1012.
Wang, Z.T., Zhao, N.B., Wang, W.Y., Tang, R., & Li, S.Y. A fault diagnosis approach for gas turbine exhaust gas temperature based on Fuzzy c-Means clustering and support vector machine. Mathematical Problems in Engineering, 2015(2015), 11 pages.