Knowledge Discovery in Steel Industry Measurements

Ulla ELSILÄ and Juha RÖNING
University of Oulu, Department of Electrical Engineering
Computer Engineering Laboratory, PO BOX 4500, FIN-90401 Oulu, Finland
ulla.elsila@ee.oulu.fi, juha.roning@ee.oulu.fi

Abstract. In today’s world, enormous amounts of data are gathered from many kinds of processes and it is difficult to distinguish useful knowledge from the resulting extensive databases. This is also the case in steel manufacturing. The recorded databases in rolling process contain hundreds of features and new methods are needed to reveal the novel and useful information. In this study, the data from a hot strip rolling process was analyzed in order to identify the conditions in which retentions occur. At the beginning, basic statistical analysis and linear correlation study were used to prepare the data and to make the search for the knowledge easier. With these two tools, the dataset was reduced from its original size of over 200 features into 17%. After this, self-organizing maps, parallel coordinates display and k-means clustering were used to find out the probabilities of the common retention types. As a result, the method presented here revealed the rolling conditions for the four common retention types. And what is most important, the features presenting these conditions can be measured before the finishing mill and, thus, these rules can be used to reduce the number of retentions.

1. Introduction

Nowadays, steel products are expected to be more and more polymorphic and both better and uniform in quality. This has led to a situation where tighter and more precise process control is needed. The processes are also largely automated and, during them, measurements are taken from a variety of different places with very high frequencies. While many features related to the process itself and to product quality are already measured, the correlations between these features are still unknown and they can therefore not be used for process enhancement. Traditionally, improvements in, for example, a rolling process have been pursued with mathematical as well as physical models. The advantage that can be gained using these models has already been exploited to its near maximum. Thus, what are needed are methods that can help to discover previously hidden but useful knowledge from large databases. Methods of this type are called knowledge discovery methods.

In our study, the target of application was the hot strip rolling process at Rautaruukki Steel. The process in question is mainly automated and vast numbers of measurements are taken from it with high frequencies. The product variety is also large and some products are only rarely manufactured. The goal of the study was to discover new knowledge with which the retentions occurring during rolling could be foreseen. Herein, the word retention is used to refer to a situation during hot strip rolling in which a product of rolling (i.e. coil) is not automatically accepted for selling, but has to be retained for re-checking and post-action.

In order to find correlations between hundreds of measured features, different types of data mining methods were used. These were statistical basic analysis, linear correlation [1 pages 56-67, 125-129], self-organizing maps (SOM) [2], parallel coordinates [3] and k-means clustering. Knowledge obtained from experts also played an important role during the study. Statistical basic analysis was used to find out the effects of individual features on retention distribution.
Linear correlation was used in order to reduce the redundancy in the data, because, otherwise, the results from SOM would have been defective. Finally, dependencies between retention codes and groups of features were sought with SOM, parallel coordinates and k-means clustering. The results of our study can be used to enhance the competitiveness, productivity and product quality of the factory and the combined method presented in here can be used to get better results from other knowledge discovery tasks.

2. Knowledge Discovery

Up to the present, the processes of industry have become mostly automated. To some extent, this is allowing new and more precise measurements to be taken from different stages of the industrial processes. Along with better measuring conditions, the actual process controlling is being improved. However, the problem of handling large databases has arisen, because although information is easy to gather, it is slow and expensive to analyze. New methods to solve this problem are constantly under development, and the main goal of these so-called knowledge discovery methods is to reveal previously hidden and useful knowledge.

In scientific literature, the descriptions of the process of knowledge discovery are usually quite alike. For example, although Mannila [4] and Pyle [5] differ in the importance they attribute to the different parts of the process, several similarities can be found in their approaches. However, a feature for which the descriptions vary frequently is the level of automation. Mannila [4] defines the process as an interactive and not a fully automated system. Matheus et al. [6], on the contrary, argue that the greatest challenge for a knowledge discovery process is to automatically handle large quantities of data, find important and meaningful models and represent them in a user-suitable way. Yet, even they admit that the realization of a fully automated system is too far from reach.

The knowledge discovery process of our study is presented in Figure 1. At the beginning, different methods of knowledge discovery along with the database of Rautaruukki Steel are studied. These two components are studied simultaneously because the data to be analyzed will affect the selection of methods and vice versa. During the data mining stage, the acquired dataset is, at first, prepared using expert knowledge, basic statistical analysis and linear correlation and, then, evaluated by SOM, parallel coordinates and k-means clustering. The last stage includes the presentation of the results to the rolling experts and the creation of an application.

![Figure 1. The main stages of our study.](image-url)
3. Data Preparation

A significant part of data preparation consists of the selection of suitable data. If the dataset is too large, it will take unnecessary time and too much capacity to compute, but, on the other hand, if the dataset is too small, one may obtain a result that is an overfitted model or no results at all. Usually, large datasets include redundancy that can, for continuous features, be located by using correlation analysis [1 page 126]. In most cases, datasets are partly imperfect and erroneous, but these deficiencies can be detected by basic statistical values (mean, standard deviation and value range) and diagrams (histogram, pie chart and time series).

There are many different techniques to select the features that will be studied [7, 8]. Pudil and Novovičová [9] emphasize the idea that there is no unique or optimal way to approach the problem of data subset selection and, in their article [9], they give instructions on how to select a suitable method. Cheeseman and Stutz [10] warn about the dangers of undocumented and irreversible data preparation. If the realization of discarding or the transformation of the data is done hastily, one may lose valuable information that is needed in order to obtain good results.

The process of data preparation used in this study is presented in Figure 2, where the numbers beside the arrows indicate the order of the process. The used data preparation tools were expert knowledge, basic statistical analysis and linear correlation. At the beginning of the process, the knowledge from the experts played an important role because the researchers were not familiar with the environment of a hot strip rolling process.

Genetic algorithm techniques [7, 8] could not be used in this study, because our knowledge about the hot strip rolling was limited. The heuristic algorithm, presented by Yang and Honavar [8], had to be abandoned, because, as pointed out by the experts, the correlations between features of a rolling process are too numerous.

In this study, three different datasets were used. The first dataset included almost all features from the rolling process and the measurements taken over a period of one week. This dataset was used to familiarize the researchers with the process. The second dataset, which was used to produce the conditional probabilities, included far less features than the first, but this time the measurements covered a time period of three months. The third dataset was used to evaluate the conditional probabilities and it included measurements recorded during three months, but not within the same period than the measurements of the second dataset.

3.1. Basic Statistical Analysis

At first, the studied dataset was selected by using the expert knowledge. After this, each feature was studied statistically in order to determine the types of features that were included in the dataset. Like Hair et al. [11 page 37] point out, the starting point for understanding the nature of any feature is the description of its distribution. They also remark that one can usually gain enough knowledge about a feature by merely drawing its histogram. In this study, before drawing the histograms, the values of the features were compared to the allowed value ranges provided by the experts. The values outside these ranges were marked as missing values. The dataset itself included some sample vectors, which lacked a few of the feature values. The mean and the standard deviation were calculated for each feature, and the following three diagrams were drawn:

![Figure 2. The process used in data preparation.](image-url)
1. a pie chart illustrating the percentages of the most general values in the entire dataset,
2. a histogram including all the samples and
3. a histogram including the retained samples.

The second histogram was drawn because the purpose of the study was to find the features that affected the number of retentions. This was a means to discover the features that had irregularities in their retention distributions. Although all this information can be quickly calculated without the need to draw diagrams, visualization is used in order to facilitate the apprehension of the data. After this analysis, the results were presented to the experts and, according to their conclusions, many features were discarded while some were added to the second dataset. A new analysis was carried out on all the selected features because the second dataset was fetched from a different time span than the first one.

The following example presents one of the significant features. The corresponding pie chart is shown in Figure 3. The feature in question included three values for which the incidence was over 1%. There were 6903 samples with value 20 (incidence 22.6%), 17204 samples with value 25 (56.3%) and 6173 samples with value 35 (20.2%). The combined incidence for the remaining values was 0.9%. The distribution of sample values is illustrated in the left most histogram in Figure 4. For the three values mentioned above, the proportions of retentions were calculated. For value 20, there were 1733 samples with retention code, corresponding to 25.1% of all the samples with the same value (100%×1733/6903 ≈ 25.1%). For values 25 and 35, the proportions of retentions were 12% and 29.3% respectively. The right most histogram in Figure 4 shows the distribution of the retentions. Although the number of retentions was almost identical for all the three values, the proportions were remarkably higher for values 20 and 35 than for value 25. Thus, this feature had surely a significant impact on the occurrence of the retentions.

**Value 20:**
Included 6903 samples (22.6% of all samples) from which 1733 retained (25.1%).

**Value 25:**
Included 17204 samples (56.3% of all samples) from which 2073 retained (12.0%).

**Value 35:**
Included 6173 samples (20.2% of all samples) from which 1807 retained (29.3%).

Figure 3. A pie chart of a significant feature.

Figure 4. Two histograms of a significant feature.
3.2. Linear Correlation

After the statistical analysis, the second dataset was studied by using linear correlation in order to reduce the redundancy it contained. The necessity of redundancy reduction depends on the tool that is used to explore the data after its preparation. For example, results acquired with SOM will be erroneous if redundancy is not excluded. Correlation can only be calculated for continuous features and, therefore, all the bilateral dependencies could not be studied with this method. It should be remembered that although independent random features are linearly uncorrelated, linearly uncorrelated features are not necessarily independent.

When using correlation values to reduce the number of features, an appropriate limit for exclusion has to be selected. Unfortunately, there are no statistical methods to calculate it when the number of samples is large. Thus, it has to be chosen separately for each study. It is recommendable to use visual inspection of scatter plots while making the decision. In this study, the selected limit for exclusion was 0.9.

4. Data Evaluation

After the data preparation stage, begins data evaluation. In this study, the tools used for this purpose were SOM, parallel coordinates and k-means clustering. Previously, SOM and parallel coordinates tools have been used separately [3, 12] but not simultaneously as was done in this study. The experiments done by Vesanto and Alhoniemi [13] indicated that clustering the SOM instead of directly clustering the data is computationally effective approach. The clustering results using SOM as an intermediate step were also comparable with the results obtained directly from the data. The order in which to use SOM and parallel coordinates does not matter in results point of view. Instead, what matters is that by using either one of them alone the resulting rules would have been too difficult to comprehend. With adding the parallel coordinates tool the results of knowledge discovery are better than when using the SOM and k-means clustering without it.

4.1. Self-Organizing Map (SOM)

One problem with clustering methods is that the interpretation of derived clusters might be difficult. Most methods favor clusters with certain shapes and, furthermore, these clusters might be multidimensional and, thus, additional methods are needed to represent them. The selection of number of the clusters can also influence the results decisively.

Self-organizing maps (SOMs) were developed by Kohonen [2], and a very thorough comparison between SOM and other clustering methods has been proposed by Kaski [12]. In short, SOM is an unguided learning method in which an n-dimensional input space is grouped into a regular 2-dimensional line of knots. In other words, SOM is a projection of the multidimensional density function into the 2-dimensional space. In a way, the resulting map is very general, considering that beforehand assumptions about the shapes and the number of the clusters do not need to be made. The horizontal and vertical axes of the map should, however, not be interpreted generally, because SOM may move the adjustment of the samples in a nonlinear way. In other words, given the tendency of SOM to preserve the local structures in a dataset, the interpretation should also be done locally. In addition to the cluster map, SOM produces pictures of component planes that show the distributions of the component values corresponding to the map. This operation makes the visual interpretation of the clustering both easier and faster.

4.2. Parallel Coordinates and K-Means Clustering

Parallel coordinates is a method of transforming the search of relations among multivariate datasets into a 2-dimensional pattern recognition problem and the real strength of it is in modeling relations between variables [3]. Plainly put, in parallel coordinates, the horizontal axis represents the variables and the vertical axis their values.
At the final stage of deriving the rules, the resulting SOMs were clustered by using a k-means clustering algorithm. The best possible result of clustering was selected according to Davies-Bouldin index [14]. Now, when the SOMs were clearly divided into different clusters, the conditional probabilities could be derived for each of them.

5. The Hot Strip Rolling Process

The used data was gathered from the hot strip mill of Rautaruukki Steel, illustrated in Figure 5. Before the rolling process, the slabs are made and both their dimensions and several chemical parameters are measured and stored in the database. The slab-specific rolling plans are also stored in the database before the rolling. At the beginning of the rolling, the slabs are heated in furnaces. At Rautaruukki, there is one walking beam and three pusher type furnaces. When the slab is discharged from the furnace onto the straight rolling line, the timestamp and the duration of heating are stored in the database. After this, the slab is taken several times through the roughing mill that transforms it into a transfer bar. Next, the transfer bar goes through the finishing mill. Immediately after the passage in the finishing mill, the thickness, profile, flatness, temperature and width of the strip are measured. With these parameters, several deviation parameters between planned and real values are then calculated and stored in the database. Finally, the strip goes through the cooling system, and once again, before the coiler, the temperature is measured.

Since the actual rolling process after the heating is fast, with a mere duration of a few minutes, the process is very sensitive to various errors. The maintenance of the right temperature during the process is a particularly challenging task. At any moment during the rolling, the automation system or the roller is able to store a slab-specific retention code in the database. Each product with this type of code has to be checked manually and sometimes reshaped before it can be approved for selling. In this study, these products are collectively called retained samples.

5.1. Datasets

The first dataset included 238 features with 4293 samples. There were 1086 retained samples, i.e. 25.3% of all samples. Because of the large number of features, the size of the dataset was reduced according to advice received from the rolling experts. This permitted the selection of the 69 most interesting features for a basic statistical analysis. The results from this analysis were then presented to the rolling experts, who decided which of the features were to be included in the second dataset.

![Figure 5. The hot strip mill of Rautaruukki Steel.](image-url)
The second dataset included 41 features, of which 34 were present in the first dataset and seven were not. All of these features were statistically analyzed because the data was now gathered during a different time span than earlier. The dataset also included eight additional features, which were selected to aid in its handling. Thus, the total size was 49 features with 28738 samples. There were 9010 retained samples, i.e. 31.4% of all samples. The methods of linear correlation study, SOM and k-means clustering were applied to this dataset, which resulted in conditional probabilities for retention codes.

Later, a third dataset was fetched in order to evaluate the rules derived from the second dataset. The third dataset included thirteen features with 33524 samples. There were 8743 retained samples, i.e. 26.1% of all samples.

6. Results

6.1. Basic Statistical Analysis

At the beginning of the study, 69 features of the first dataset were analyzed with basic statistical analysis. All of the 238 fetched features were not analyzed due to various reasons, one of which was the fact that some of the features did not include any measurements. Next, these results and comments were presented to the rolling experts, who decided on the contents of the second dataset.

The analysis of the second dataset included 41 features. Thirty-one of these were found to be significant in a way that was presented in Chapter 3.1. However, the ten remaining features were not excluded from the dataset at this point, because these were suspected to have a more complicated impact on the occurrence of retentions.

The final analysis was done to the third dataset with thirteen features. These features were all found to be significant both in the second and the third dataset. In order to visualize the basic differences between these two datasets, parameter comparisons were made between the thirteen variables in both datasets. These comparisons included the mean, standard deviation and value range for both the whole dataset and the set that only included the samples with retention code. The hot strip rolling process, from which the features were measured, was changed between the two time spans of the measurements. A transfer bar coil was added into the line after the roughing mill and it turned out to have an effect on some of the features used in this study. All in all, the changes concerning the whole dataset are similar to the changes concerning the samples with retention code.

6.2. Linear Correlation and SOM

The use of SOM requires an exclusion of the redundancy from the dataset. Thus, the linear correlation analysis was done to the 29 continuous features in the second dataset, leading to the exclusion of eight features. The resulting dataset included 21 continuous and 20 class features and, hence, the size of the dataset was 17% of the original size.

When applying SOM, the features have to be scaled from 0 to 1 or from −1 to 1. In this study, the former scale was selected. It is easy to normalize continuous features, but in the case of class features, the task becomes slightly more time-consuming. In this study, for every class feature, new components were created so that every component represented only one class. For example, if a feature included classes A and B, the first new component had value one if the sample belonged to a class A and value zero otherwise. Correspondingly, the second new component had value one if the sample belonged to a class B and zero otherwise. Thus, a feature with two class values yielded two components, three class values yielded three components, and so forth.

The scaling of the features used in this study produced a total of 63 components. The resulting SOM revealed numerous clusters and, on the basis of a visual inspection of the corresponding component planes, 27 components were selected for future analysis because they seemed to have an effect on exception code distribution in the said clusters.

Figure 6 illustrates the SOM composed after reducing the component count, revealing even larger and clearer clusters than the map of all components. In a SOM, the clusters are shown as light areas separated by the darker areas. In Figure 6, the areas separated with a solid line included
extensively more samples with retention code than the areas marked with dotted line. The visual inspection of the component planes did not reveal any need to reduce the number of the components, but, in order to obtain predictive rules for retention code appearance, the three components measured after finishing mill had to be excluded from further analysis.

6.3. Parallel Coordinates

At this point, in order to find out the explanations for different retention types, six smaller datasets were selected according to the most common retention codes. The idea behind this was that one kind of a set-up situation might never produce a retention of a specific type and, on the other hand, might always produce a retention of another type. Hence, each of these sets included all the samples without the retention code and samples that had a certain type of retention code.

These six datasets were then visualized by using the parallel coordinates method. The purpose of drawing a figure of parallel coordinates is that it allows the viewer to see the distributions of every component simultaneously. Figure 7 presents the parallel coordinates for the first set. The components are aligned on the horizontal axis and their values are on the vertical axis. A dot stands for a successful sample and a circle for a sample with retention code.

In Figure 7, for the large values of the components 5, 6, 7, 8, 23 and 24 there are no samples with retention codes but only successful samples. The same situation also occurs for the small values of the components 16 and 24. What is more, the components 9 and 12 seem to have some differences between the distributions of the successful samples and the samples with retention codes. Hence, the components 5, 6, 7, 8, 9, 12, 16, 23 and 24 were selected for the next stage of the study. The remaining sets were also studied in a similar manner.

6.4. K-Means Clustering

After reducing the components by visually inspecting the parallel coordinates figures, the final SOM groupings were done for the six smaller datasets. The resulting maps were then clustered by using a k-means clustering algorithm. Tens of program runs were done in order to find the best possible result of clustering according to a Davies-Bouldin index [14].

Figure 6. The SOM of 27 components.
Figure 7. Parallel coordinates for the first smaller dataset.

After selecting the suitable clustering, the conditional probabilities were derived for each cluster. In the case of the first smaller dataset, the clusters separated efficiently the parts of the map that only had successful samples from the parts that had multiple samples with retention code. For example, the cluster numbered as 5 included 3484 successful samples and none of the samples with retention code. On the other hand, the cluster numbered as 7 included 862 of the successful samples and 265 samples with retention code. Hence, for cluster 5 the probability of having retention code is zero, and for cluster 7, the corresponding probability is 0.235 (= 265 / (265 + 862)). Thus, there is a large difference in predicting the possible retention code for a sample that resembles more the ones in cluster 5 than those in cluster 7. After studying the component values corresponding the clustering, the conditional probabilities were derived for each cluster.

At the end of the study, the conditional probabilities were tested with new data. Again, six smaller datasets were constructed in the same manner as earlier, before the parallel coordinates were drawn. At this point, the first issue to consider was the way in which the new samples were to be linked to the existing clusters. The best way to do this was to apply the following procedure:

For each sample
Step 1. Find the clusters with suitable component limits, i.e. the value of any component is allowed to exceed the upper or lower limit in conditional probability rule only with the margin of 0.15. For example, if the upper limit for a component is 0.60, the sample value for the corresponding component must not exceed 0.75 or, if the lower limit for a component is 0.40, the sample value for the corresponding component must not exceed 0.25.

Step 2. If one or more suitable clusters are found, select the cluster with most similar mean vector. If suitable clusters are not found, select from all of the clusters the one with most similar mean vector.

In this context, the mean vector of a cluster includes arithmetical mean value for each component of that cluster. With this method, the sample count outside the component limits for different datasets was in the worst case 4.54% and in the best case 0.02%. When comparing the proportional numbers of retention codes in clusters, four of the six smaller datasets had a similar distribution. Hence, the conditional probabilities for these four sets were found to be reliable. The two other datasets had some changes in their distributions of proportional numbers of retentions and they have to be presented to the rolling experts before utilizing the results.

In conclusion, the method presented here revealed the rolling conditions for the four common retention types. This information can be utilized to prevent the retentions and the capability of reducing the occurrence of retentions, in its turn, can increase competitiveness, productivity and product quality of the steel factory.
7. Conclusions

The purpose of the study was to discover new knowledge on the occurrence of retentions in a hot strip rolling process and to find out how this occurrence could be foreseen before the finishing mill. The reason behind is obvious: if a product has to be re-checked after rolling, the productivity of the factory is decreased while the costs are increased in many ways. In the past, the main object to this type of research has been the lack of knowledge on how to utilize the enormous databases resulting from the measurements taken from the rolling process.

As a result of our study, a new method was developed permitting to analyze large datasets. At first, this method helped to reduce the size of the original dataset into 17% and, later on, by using the combination of SOM, parallel coordinates and k-means clustering, it helped to find the conditions for the four common types of retentions. Without use of SOM and parallel coordinates together, the resulting rules would have been all too complicated. These results will help the rolling experts to reduce the number of retentions and, thus, to enhance the process. In the future, this method can be used to study other databases that include vast numbers of features and samples.

References